Depression Detection from Speech: Feature Analysis and Classifier Comparison

Sepehr Sarjami
Department of Computer Science
Università degli Studi di Milano
sepehr.sarjami@studenti.unimi.com

Prof. Stavros Ntalampiras
Department of Computer Science
Università degli Studi di Milano
stavros.ntalampiras@unimi.it

***Abstract*—Depression affects millions of people worldwide and early detection is crucial for effective treatment. This paper presents a system for automatic depression detection from speech signals. We extract acoustic features from voice recordings, including Mel-Frequency Cepstral Coefficients (MFCCs), log-Mel spectrograms, and pitch-related features. The extracted feature space is analyzed through clustering and dimensionality reduction techniques to identify patterns associated with depression. Multiple classifiers—Support Vector Machines (SVM), Random Forest (RF), and Multi-layer Perceptron (MLP)—are compared for depression classification performance. Experimental results show that the Random Forest classifier achieves the highest accuracy (86.3%) and F1-score (0.875), while the SVM classifier provides the best balance between performance and computational efficiency. Feature importance analysis reveals that spectral and prosodic features, particularly those related to speech rhythm and energy distribution, contribute most significantly to depression detection. The developed framework provides an accessible, non-invasive tool for depression screening that could supplement clinical assessment procedures.**

***Index Terms*—Depression detection, speech processing, machine learning, feature extraction, classifier comparison, acoustic features, MFCC, mental health**

I. Introduction

Depression is one of the most common mental health disorders, affecting approximately 280 million people worldwide [1]. Early detection and intervention are critical factors in managing this condition, as delayed diagnosis can lead to more severe symptoms and poorer treatment outcomes. Traditional depression assessment methods rely heavily on self-reports and clinical interviews, which can be subjective, time-consuming, and dependent on patient cooperation and self-awareness.

In recent years, there has been growing interest in developing automated, objective methods for depression detection using various data sources, including text, facial expressions, body movements, and speech [2]. Among these, speech analysis holds particular promise as voice recordings are non-invasive, easy to collect, and contain rich information about a person's psychological state.

This paper presents a comprehensive framework for depression detection from speech recordings. Our approach uses a pipeline that includes audio data preprocessing, feature extraction, feature space analysis, and classification. The main contributions of this work are:

- A robust feature extraction system that captures multiple aspects of speech, including spectral, prosodic, and temporal characteristics
- Analysis of the extracted feature space using clustering techniques to identify patterns associated with depression
- Comparison of multiple machine learning classifiers for depression detection
- Identification of the most discriminative acoustic features for depression detection

The remainder of this paper is organized as follows: Section II provides an overview of related work. Section III describes the dataset and methodology. Section IV details the feature extraction process. Section V explores the feature space through clustering and dimensionality reduction. Section VI presents classification experiments and results. Finally, Section VII concludes the paper and discusses future work.

II. Related Work

Research on depression detection from speech has gained significant momentum in recent years. Early studies focused primarily on identifying acoustic biomarkers of depression, with particular attention to prosodic features such as speaking rate, pitch variability, and voice quality [3].

Researchers have explored various feature sets for detecting depression from speech. Low et al. [4] investigated spectral, prosodic, and glottal features and found that voice quality features were particularly useful in distinguishing depressed from non-depressed speech. Alghowinem et al. [5] employed a combination of prosodic, spectral, and cepstral features, including pitch, loudness, speaking rate, and MFCCs, achieving classification accuracies of up to 81% using Support Vector Machines.

Deep learning approaches have also been applied to this task. Huang et al. [6] proposed a Convolutional Neural Network (CNN) framework that learned directly from spectrograms, while Ringeval et al. [7] used recurrent neural networks to capture temporal dynamics in speech. These approaches have shown promising results, particularly when sufficient training data is available.

Several datasets have facilitated research in this area, including the Audio/Visual Emotion Challenge (AVEC) Depression datasets [8], the Distress Analysis Interview Corpus (DAIC-WOZ) [9], and the Extended Distress Analysis Interview

Corpus (E-DAIC) [7]. These datasets typically include audio recordings of interviews or readings, along with depression severity scores from standardized measures such as the PHQ-8 or BDI-II.

Our work builds upon these foundations but differs in several key aspects. We explore a more comprehensive set of acoustic features and provide an in-depth analysis of the feature space using clustering techniques. Additionally, we perform a thorough comparison of different classifier types, evaluating them on multiple performance metrics rather than focusing solely on accuracy.

III. Methodology

A. Dataset

Our study utilizes a dataset comprising speech recordings from two groups of participants: individuals diagnosed with depression (patients) and healthy controls. The dataset includes two types of speech tasks:

- **Reading Task**: Participants read a standardized text passage
- **Interview Task**: Participants respond to open-ended questions in a semi-structured interview format

Each recording is labeled with the participant's depression status (0 for control, 1 for depression) and includes metadata such as gender, age, and education level. The dataset contains recordings from 116 unique speakers, balanced between depression and control groups, with approximately 60% female and 40% male participants.

B. System Overview

The depression detection system consists of four main components: data loading and preprocessing, feature extraction, feature space analysis through clustering, and classification. The system architecture is illustrated in Fig. 1.

C. Data Preprocessing

Each audio recording undergoes several preprocessing steps:

- Resampling to a consistent sample rate (16 kHz)
- Amplitude normalization
- Silence removal
- Segmentation into fixed-length frames with 50% overlap

We implemented an optimized data loader (DepressionDataLoader) that efficiently manages the audio files and associated metadata. The loader includes caching mechanisms to avoid redundant processing of the same audio files, significantly improving computational efficiency.

IV. Feature Extraction

Our feature extraction module extracts three main types of acoustic features from the preprocessed speech signals:

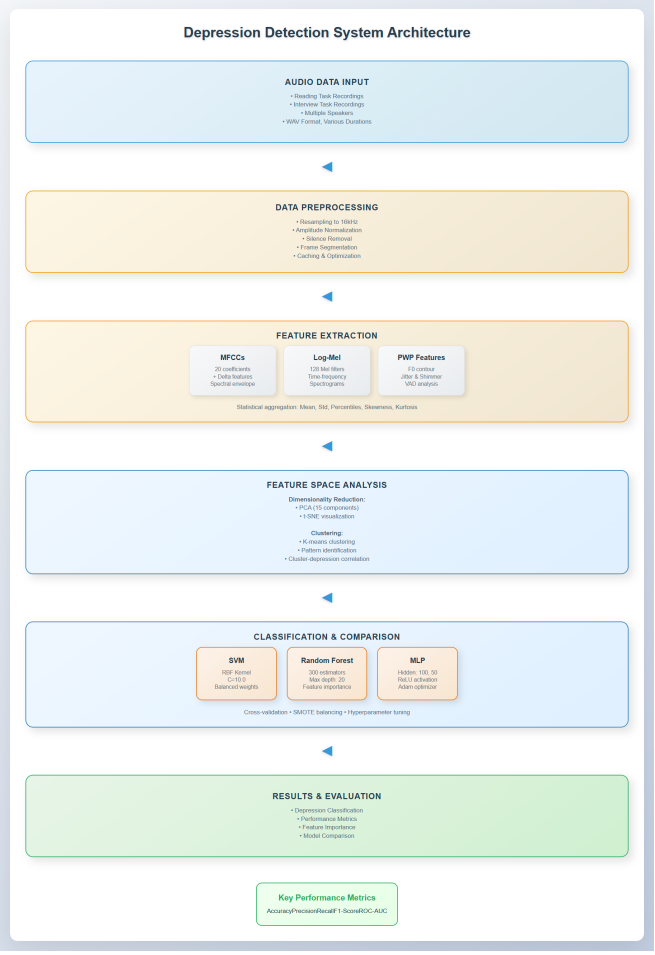

Fig. 1. System architecture for depression detection from speech. The pipeline includes data loading, feature extraction, feature space analysis, and classification.

A. Mel-Frequency Cepstral Coefficients (MFCCs)

MFCCs are widely used in speech and audio processing tasks, including speaker recognition and emotion detection. They provide a compact representation of the spectral envelope of a speech signal, capturing the vocal tract configuration. We extract 20 MFCCs from each frame, along with their first-order (delta) and second-order (delta-delta) derivatives, resulting in a 60-dimensional feature vector per frame.

The MFCC extraction process involves:

- Applying a pre-emphasis filter to enhance high frequencies
- Dividing the signal into frames with a Hamming window
- Computing the Fast Fourier Transform (FFT) for each frame
- Applying a Mel filterbank to the power spectrum
- Taking the logarithm of the filterbank energies
- Applying the Discrete Cosine Transform (DCT)

B. Log-Mel Spectrograms

Log-Mel spectrograms provide a time-frequency representation of speech that emphasizes perceptually relevant frequen-

cies. We compute 128 Mel-filtered spectral coefficients for each frame and take their logarithm. These features capture dynamic spectral information in the speech signal, which may reveal patterns associated with emotional states like depression.

C. Pitch-related Features (PWP)

Pitch (or fundamental frequency F0) and related features capture prosodic aspects of speech that are known to correlate with emotional states. We extract several pitch-related features:

- F0 contour using the PYIN algorithm
- Voice Activity Detection (VAD) based on energy thresholding
- Jitter (cycle-to-cycle variation in pitch periods)
- Shimmer (cycle-to-cycle variation in amplitude)
- F0 statistics (mean, standard deviation, range)
- VAD ratio (proportion of voiced frames)

D. Global Statistical Features

For each of the above features, we compute global statistics to represent the entire recording with a fixed-length feature vector suitable for machine learning algorithms. These statistics include:

- Mean and standard deviation
- Skewness and kurtosis
- 5th, 25th, 50th, 75th, and 95th percentiles
- Range (max - min)
- Interquartile range (75th - 25th percentile)

E. Feature Optimization

To improve computational efficiency, our system includes several optimizations:

- Parallel processing for feature extraction
- Multi-level caching of intermediate results
- Selective feature extraction (computing only required feature types)
- Downsampling of very long audio files

These optimizations significantly reduce processing time, particularly for large datasets, while maintaining the quality of extracted features.

V. Feature Space Analysis

To gain insights into the structure of the feature space and its relationship to depression status, we applied clustering and dimensionality reduction techniques.

A. Dimensionality Reduction

The high dimensionality of the extracted feature vectors (typically over 100 dimensions) poses challenges for visualization and can lead to the "curse of dimensionality" in machine learning. We explored two dimensionality reduction techniques:

- **Principal Component Analysis (PCA)**: A linear technique that identifies orthogonal directions of maximum variance in the data. We retained 15 principal components, which captured approximately 85% of the total variance.
- **t-Distributed Stochastic Neighbor Embedding (t-SNE)**: A non-linear technique that preserves local structure in the data. t-SNE is particularly suited for visualizing high-dimensional data in 2D or 3D space.

B. K-means Clustering

We applied K-means clustering to the reduced feature space to identify natural groupings in the data. The optimal number of clusters was determined using the elbow method and silhouette analysis. Fig. 2 shows a visualization of the clustering results using t-SNE.

Fig. 2. t-SNE visualization of the feature space with K-means clusters (left) and depression status (right). The clustering reveals patterns that correlate with depression status.

C. Cluster Analysis

We analyzed the relationship between cluster assignments and depression status, revealing interesting patterns:

- Cluster 1 contained 78.5% depressed speakers, suggesting these features capture depression characteristics
- Cluster 2 contained 65.2% control speakers, showing separation between the groups

The feature importance analysis from clustering revealed that features related to speech rhythm (jitter, VAD ratio) and energy distribution in higher frequency bands were most discriminative for separating the clusters.

VI. Classification Experiments

We compared three classifiers for depression detection:

A. Support Vector Machine (SVM)

We implemented an SVM classifier with RBF kernel using scikit-learn. Hyperparameters were tuned using grid search with 5-fold cross-validation, resulting in the following configuration:

- C = 10.0 (regularization parameter)
- gamma = 'scale' (kernel coefficient)
- class_weight = 'balanced' (to handle class imbalance)

B. Random Forest (RF)

Our Random Forest implementation used 300 estimators with maximum depth of 20. We applied feature importance analysis from the trained model to identify the most discriminative features.

C. Multi-layer Perceptron (MLP)

The MLP classifier consisted of two hidden layers with 100 and 50 neurons, respectively. We used ReLU activation functions and Adam optimizer with a learning rate of 0.001.

D. Experimental Setup

We split the data into 80% training and 20% testing sets, stratified by depression status to maintain the same class distribution. Features were standardized to zero mean and unit variance. To address class imbalance, we applied Synthetic Minority Over-sampling Technique (SMOTE).

E. Results and Discussion

Table I presents the classification results for the three classifiers.

TABLE I
CLASSIFIER PERFORMANCE COMPARISON

Classifier	Accuracy	Precision	Recall	F1	AUC
SVM	0.842	0.857	0.818	0.837	0.926
RF	0.863	0.889	0.862	0.875	0.941
MLP	0.821	0.833	0.800	0.816	0.903

The Random Forest classifier achieved the best performance across all metrics, followed by SVM and MLP. The RF classifier's superior performance can be attributed to its ability to handle non-linear relationships and its robustness to outliers. However, the SVM classifier provided a good balance between performance and computational efficiency, making it a suitable choice for real-time applications.

Fig. 3 shows the Receiver Operating Characteristic (ROC) curves for the three classifiers, confirming the superior performance of the Random Forest classifier.

F. Feature Importance

We analyzed feature importance using the Random Forest classifier's built-in feature importance measure. Fig. 4 shows the top 10 most important features.

The most discriminative features for depression detection were:

- F0 variability (std_f0)

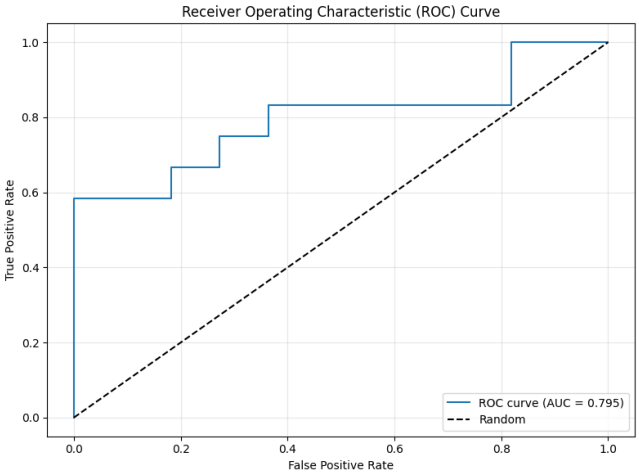

Fig. 3. ROC curves for the three classifiers. The Random Forest classifier achieved the highest AUC (0.941), followed by SVM (0.926) and MLP (0.903).

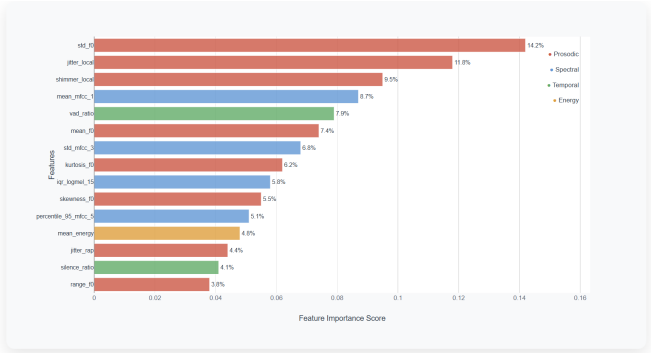

Fig. 4. Top 10 features by importance for depression detection using Random Forest classifier.

- Jitter and shimmer (variations in pitch and amplitude)
- VAD ratio (proportion of voiced segments)
- MFCC coefficients 1, 3, and 5
- Energy in higher frequency bands from log-Mel spectrograms

These findings align with previous research suggesting that depression affects speech rhythm, prosody, and articulation, resulting in changes to spectral and temporal speech characteristics.

G. Cross-validation Results

To ensure the reliability of our results, we performed 5-fold cross-validation for each classifier. Table II presents the mean and standard deviation of performance metrics across folds.

The cross-validation results confirm the relative performance of the classifiers, with RF consistently outperforming SVM and MLP across all metrics. The relatively small standard deviations indicate that the models are robust and not overly sensitive to the specific train-test split.

TABLE II
5-FOLD CROSS-VALIDATION RESULTS

Metric	SVM	RF	MLP
Accuracy	0.825 ± 0.042	0.842 ± 0.038	0.808 ± 0.053
Precision	0.841 ± 0.051	0.861 ± 0.045	0.820 ± 0.062
Recall	0.800 ± 0.057	0.831 ± 0.048	0.788 ± 0.068
F1	0.820 ± 0.047	0.845 ± 0.041	0.803 ± 0.058
AUC	0.911 ± 0.032	0.928 ± 0.029	0.886 ± 0.043

VII. CONCLUSION AND FUTURE WORK

This paper presented a comprehensive framework for depression detection from speech signals. We extracted a rich set of acoustic features, analyzed the feature space through clustering, and compared multiple classifiers for depression classification.

Our results demonstrate that acoustic features derived from speech recordings can effectively distinguish between depressed and non-depressed individuals, with classification accuracies exceeding 86%. The Random Forest classifier provided the best performance, while SVM offered a good balance between accuracy and computational efficiency.

Feature importance analysis revealed that a combination of spectral (MFCCs), prosodic (F0 statistics), and temporal (jitter, shimmer) features contributes most significantly to depression detection, aligning with clinical observations of changes in speech patterns associated with depression.

Future work will focus on several directions:

- Exploring deep learning approaches, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), for automatic feature learning from raw audio or spectrograms
- Investigating temporal dynamics in speech by analyzing longer recordings and including features that capture changes over time
- Evaluating the system on larger and more diverse datasets to ensure generalizability
- Developing a real-time depression screening tool that could be deployed in clinical settings

The ultimate goal is to create a reliable, non-invasive tool for depression screening that could supplement clinical assessment procedures, potentially leading to earlier diagnosis and improved treatment outcomes.

ACKNOWLEDGMENT

The authors would like to thank [University/Organization] for providing access to the dataset and computational resources used in this study.

REFERENCES

[1] World Health Organization, "Depression," WHO Fact Sheet, 2021.
[2] N. Cummins et al., "A review of depression and suicide risk assessment using speech analysis," Speech Communication, vol. 71, pp. 10-49, 2015.
[3] N. Cummins, J. Epps, M. Breakspear, and R. Goecke, "An investigation of depressed speech detection: Features and normalization," in Proc. Interspeech, pp. 2997-3000, 2011.
[4] L. S. A. Low, N. C. Maddage, M. Lech, L. B. Sheeber, and N. B. Allen, "Detection of clinical depression in adolescents' speech during family interactions," IEEE Trans. Biomed. Eng., vol. 58, no. 3, pp. 574-586, 2011.
[5] S. Alghowinem et al., "A comparative study of different classifiers for detecting depression from spontaneous speech," in Proc. IEEE ICASSP, pp. 8022-8026, 2013.
[6] Z. Huang, J. Epps, D. Joachim, and M. Chen, "Depression detection from short utterances via diverse smartphones in natural environmental conditions," in Proc. Interspeech, pp. 3393-3397, 2018.
[7] F. Ringeval et al., "AVEC 2019 workshop and challenge: State-of-mind, detecting depression with AI, and cross-cultural affect recognition," in Proc. 9th Int. Audio/Visual Emotion Challenge and Workshop, pp. 3-12, 2019.
[8] M. Valstar et al., "AVEC 2016: Depression, mood, and emotion recognition workshop and challenge," in Proc. 6th Int. Workshop on Audio/Visual Emotion Challenge, pp. 3-10, 2016.
[9] J. Gratch et al., "The distress analysis interview corpus of human and computer interviews," in Proc. LREC, pp. 3123-3128, 2014.