Journal of Cognitive Human-Computer Interaction

♦ ISSN: 2771-1463 Vol: 07, No. 02, 2024 ♦ PP. 05–09

Article DOI: https://doi.org/10.54216/JCHCI.070202

FULL LENGTH ARTICLE

Extraction of Signal Features in Voice Signals to Train Machine Learning-Based Classifier Algorithms for Emotion Detection

Simran Somani[1] **Bhagyashree Shah**[1] **Bhisaji C. Surve**[2]

[1] Student of MBA TECH (IT), MPSTME, NMIMS University, Mumbai, India

[2] Asst. Professor, Dept. of IT, MPSTME, NMIMS University, Mumbai, India

Emails: simran.somani@nmims.in · bhagyashree.shah@nmims.in · bhisaji.surve@nmims.edu

Received: October 28, 2023 Revised: January 24, 2024 Accepted: March 21, 2024 ⋆ Corresponding author

ABSTRACT

This research aims to detect human emotions using speech signals through the development and implementation of methodologies, namely frequency-domain synthesis. To achieve improved results, various machine-learning and deep-learning models were applied, and their resulting model performance was analysed. The research findings revealed that each model exhibited different accuracy rates for different emotions, while the weighted accuracy is best for the deep-learning-based model. This study provides valuable insights into the feasibility and effectiveness of utilizing different methodologies and models for emotion detection through voice-signal synthesis. The audio signals are synthesized for Mel-Frequency Cepstrum Coefficients (MFCC), Chroma, and MEL characteristics, which are then used as features to train various machine-learning-based classifiers. Python libraries such as Librosa, Sklearn, Pyaudio, Numpy, and sound files are used to analyse voice modulations and identify emotions.

Keywords: MFCC ▪ Emotion Detection ▪ Machine Learning ▪ Neural network

1. INTRODUCTION

Emotion is a natural element of every living being, including plants as some scientific experiments demonstrate it. In the case of human beings, it is vital because humans are social animals and have various means to express emotions. The various means by which humans can express emotions are voice in speech, facial expression, or text used in written communication. Emotion detection can be a very useful use case for technology-based tool development. In verbal talks, audio signals can be processed to detect emotions in speech; for facial detection, video signals can be processed; and in textual matter, natural language processing can be employed.

It is a very challenging task to detect exact emotional states from audio signals due to various reasons. Hence, identification of the best representative features of the respective emotions to be used for machine learning is a vital task. There are various applications of speech emotion detection in human interactions, such as customer service, teaching-learning, medical analysis, voice synthesis, personal interviews, and forensics.

Another challenge is the presence of different languages, accents, phrases, speaking styles, and speakers because these factors directly affect most extracted attributes, such as pitch and energy. Additionally, because each emotion correlates with a different component of speech signals, it is possible to have many instances of a given emotion in a single speech signal. As a result, establishing the limits between different emotional components is difficult. Studies have considered classifying emotions across multiple languages, but audio-based speech synthesis can be language-independent.

2. LITERATURE SURVEY

A method for calculating MFCC coefficients, STE (Short-Term Energy), and audio pitch to identify irritation, pleasure, and depression using audio samples was performed by Girija Deshmukh et al. [1] using the entire RAVDESS dataset and a multi-class SVM classifier.

Peng Shi [2] developed and demonstrated that, compared with SVM and ANN algorithms, DBN (Deep Belief Networks) performed well and achieved about 5% accuracy improvement. J. Uma Maheswari and A. Akila [3] used GLCM (Gray Level Co-occurrence Matrix) and MFCC as audio features, and employed KNN and PRNN on the extracted features.

The audio properties such as sound, format, and phoneme, detected through MFCC, can be used as signal features for training and emotion prediction [4]. Varol et al. [5] suggest that various machine-learning procedures can be employed on numerous datasets with varied tests to yield better success rates.

3. RESEARCH METHODOLOGY

The research journey is executed through a structured process that begins with dataset selection and feature extraction, continues with model training, and ends with prediction and evaluation.

3.1 Research Framework

The proposed step-by-step process of machine-learning-based emotion detection is shown in Figure 1. The trained classifier can later be deployed in an online application for emotion detection from speech signals in various contextual applications.

Figure 1. Framework.

Figure 2. Application of the trained model for new audio-signal emotion detection.

3.2 Study of Dataset

The RAVDESS dataset was selected because it contains song and speech files classified into eight diverse emotions at two intensity levels, such as sad, angry, calm, happy, fearful, disgust, neutral, and surprised. The dataset provides speech recordings that are useful for studying emotional states under controlled conditions.

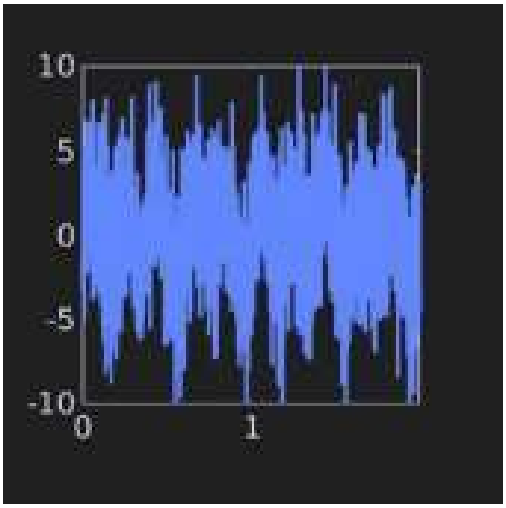

Figure 3. Sample waveform of an audio signal.

3.3 Audio Signal Processing

The audio signal is loaded and analysed through Python libraries. The signal may be represented in the time domain, where variations over time are shown, or transformed into the frequency domain to display the percentage of the signal included within each frequency band.

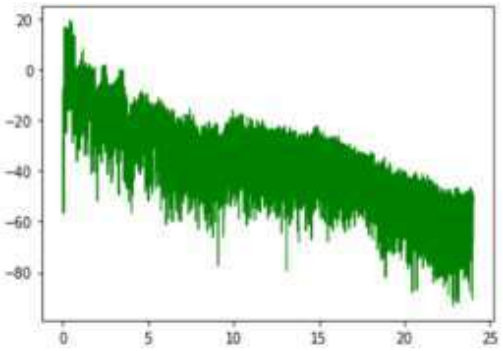

Figure 4. Audio-signal waveform representation.

Some of the principal audio features are as follows:

- **Mel scale:** examines how people perceive frequency as a scale of pitches perceived to be equally spaced.
- **Pitch:** represents the highness or lowness of sound; a higher pitch indicates a higher frequency depending on context.
- **Frequency:** measures wave cycles per second and is the rate at which sound vibrates.
- **Chroma:** projects the spectrum onto 12 bins, each standing for one of the twelve semitones.
- **Fourier transform:** moves signals between time and frequency domains.

3.4 Feature Extraction

Physical features and perceptual features are the two categories of audio features. Physical properties, such as energy, spectrum, cepstral coefficients, and fundamental frequency, are mathematical measurements calculated directly from the sound wave. Perceptual qualities describe how humans perceive sounds, including loudness, brightness, pitch, timbre, and rhythm. This work employs the Zero-Crossing Rate, Mel spectrogram, Chroma, delta and delta-delta MFCCs, and Mel Frequency Cepstral Coefficients. Using Librosa, a 525-length feature vector can be generated for each speech segment.

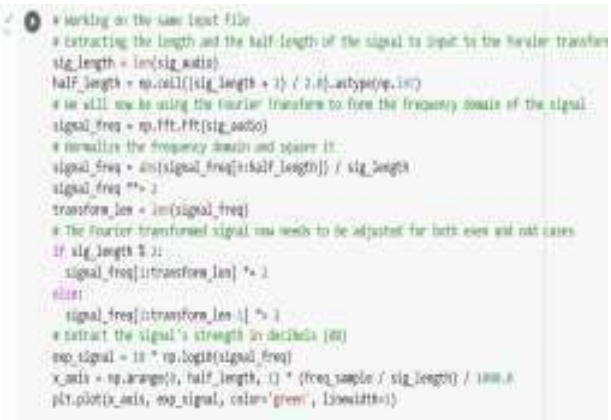

Figure 5. Feature-extraction implementation excerpt.

3.4.1 Spectrograms

Spectrograms show signal strength at various frequencies over time. The Mel-spectrogram is extracted from the power spectrum using Mel-spaced filter banks. By being more discriminative at lower frequencies and less discriminative at higher frequencies, the Mel scale seeks to emulate the nonlinear human-ear perception of sound.

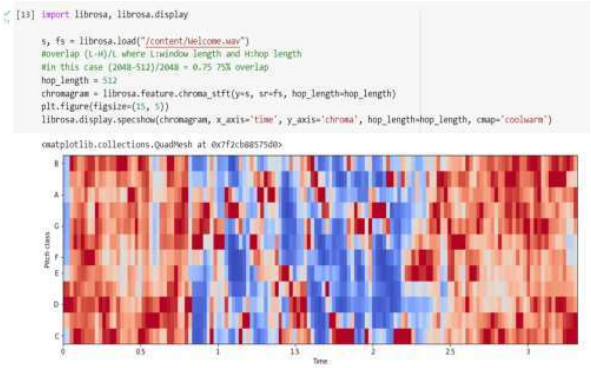

Figure 6. Chromagram.

The Mel-frequency cepstrum is useful for recognising audio and modelling the subjective pitch and frequency content of the audio stream. The triangular band-pass filter bank filters FFT power coefficients before computing MFCCs. Differential and acceleration coefficients are also called delta and delta-delta MFCCs; they help understand the dynamics of the power spectrum through time.

The mapping from actual frequency to the Mel scale is:

$$\text{mel}(f) = 1127 \ln\left(1 + \frac{f}{700}\right), \tag{1}$$

where ln is the logarithm with base e.

$$mel(f) = 1127 \ln(1 + \frac{f}{700})$$

Figure 7. Formula for mapping frequency to the Mel scale.

MFCC is a major technique used to extract audio-signal features mapped to emotions. In this technique, raw audio signals are converted into a compressed illustration that represents temporal and frequency-domain information. By applying inverse discrete Fourier transforms, MFCC extracts the first 12 coefficients of the signal and also captures the energy of the signal sample, resulting in 13 basic features. First-order and second-order derivatives add 26 more features that help describe transitions.

3.5 Data Preparation

The researcher first formulates data for training. Data must be statistical, the most common example being real values. Categorical data, such as sex traits with values “male” and “female”, and emotion elements such as “happy”, “sad”, “angry”, and “neutral”, are converted to a real-valued representation by one-hot encoding. The total dataset is split into an 80:20 ratio, where 80% is used for training and 20% for testing.

4. MACHINE LEARNING MODELS

The researchers used machine-learning classifiers that take numerical data extracted from the RAVDESS dataset as training features. Broadly, there are two categories of machine-learning algorithms: non-neural-network-based models and neural-network-based models.

In this research, three conventional classifiers were first tested: SVM, Decision Tree, and Random Forest. The researchers then worked with neural-network-based deep-learning algorithms, namely MLP (Multilayer Perceptron) and CNN (Convolution Neural Network). Training and testing are vital processes in machine learning. For training, 80% of the dataset is used, while 20% is assigned for testing to evaluate model performance.

4.1 Prediction and Model Evaluation

Once a model has been trained, it can be used to make predictions. Predictions are made on test data to estimate performance on unseen data. Accuracy and other merit ratios are checked using the testing dataset. Model performance for four emotions is evaluated using a confusion matrix and related ratios.

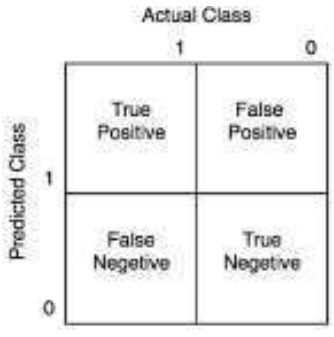

Figure 8. Confusion matrix.

Precision measures how many positive predictions are correct:

$$\text{Precision} = \frac{TP}{TP+FP}. \quad (2)$$

Recall, or sensitivity, measures how many positive cases are correctly predicted:

$$\text{Recall} = \frac{TP}{TP+FN}. \quad (3)$$

The F1-score combines precision and recall using the harmonic mean:

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (4)$$

$$2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Figure 9. F1-score formula.

All merit ratios for model evaluation should be higher, and values close to 1 are better.

4.2 Non-Deep-Learning-Based Classifiers

Support Vector Machine, Decision Tree, and Random Forest models were implemented for the extracted features. Their performance was evaluated using classification reports including precision, recall, F1-score, and support.

0.625

	precision	recall	f1-score	support
angry	0.72	0.70	0.71	47
happy	0.56	0.70	0.62	40
neutral	0.43	0.40	0.42	25
sad	0.69	0.61	0.65	56
accuracy			0.62	168
macro avg	0.60	0.60	0.60	168
weighted avg	0.63	0.62	0.62	168

Figure 10. Classification report for one of the evaluated models.

0.5714285714285714

	precision	recall	f1-score	support
angry	0.71	0.72	0.72	47
happy	0.44	0.45	0.44	40
neutral	0.48	0.48	0.48	25
sad	0.59	0.57	0.58	56
accuracy			0.57	168
macro avg	0.55	0.56	0.56	168
weighted avg	0.57	0.57	0.57	168

Figure 11. Classification report for another evaluated model.

0.6964285714285714

	precision	recall	f1-score	support
angry	0.88	0.74	0.80	47
happy	0.53	0.70	0.60	40
neutral	0.75	0.60	0.67	25
sad	0.71	0.70	0.70	56
accuracy			0.70	168
macro avg	0.72	0.69	0.69	168
weighted avg	0.72	0.70	0.70	168

Figure 12. Classification report for the neural-network-based model.

4.3 Deep-Learning-Based Classifiers

Deep-learning models were also implemented using extracted audio features. In particular, ANN/MLP and CNN-based approaches were explored. These models learn non-linear relations between features and emotional labels and are suitable for complex speech-emotion patterns.

Figure 13. Implementation excerpt for machine-learning model training and evaluation.

5. RESULTS AND DISCUSSION

The comparative performance of all models is summarized in Table 1. The ANN model achieved the highest accuracy among the tested models, followed by Random Forest, SVM, and Decision Tree.

Table 1. Comparison of accuracy for all models.

Classifier Model Type	Accuracy
ANN	80.21%
SVM	62.50%
Random Forest	69.64%
Decision Tree	57.01%

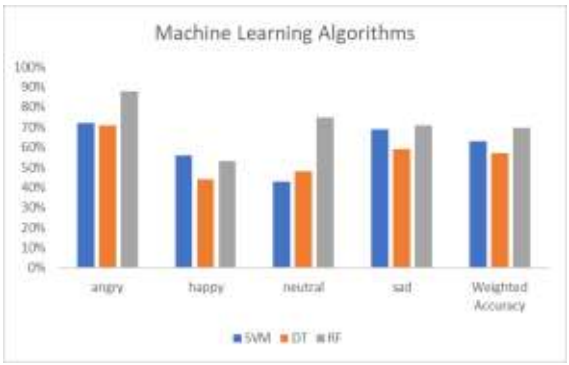

Figure 14. Accuracy comparison of machine-learning algorithms.

The results indicate that neural-network-based models can capture complex relationships in speech-emotion features more effectively than conventional models. The ANN model with appropriate hyperparameter settings provides the best performance and is recommended for real-life deployment.

6. SUMMARY AND CONCLUSIONS

Automatic emotion detection in any conversation at run time can be a very useful application in contexts such as recruitment interviews and customer support. Hence, the development of such a system with practically reasonable accuracy is the objective of this research. Any machine-learning model development begins with an appropriate dataset. The RAVDESS dataset used in this research work is comprehensive and well composed for the development of machine-learning-based models.

The next task is to convert the dataset into vital features that characterize important aspects of emotions, which act as class identifiers for the classifiers. In this research, major efforts were made to extract maximum features that appropriately capture emotion types. Once suitable features are extracted, the next objective is to develop the most accurate model. The researchers implemented different algorithms using Python library functions. Every model was evaluated using various merit parameters and compared to identify the most suitable model.

The results show that ANN with appropriate hyperparameter settings gives the best results and is recommended for real-life applications. Future improvement can be achieved with more varied datasets and by experimenting with additional signal features and model parameters.

REFERENCES

[1] G. Deshmukh, A. Gaonkar, G. Golwalkar, and S. Kulkarni, "Speech based emotion recognition using machine learning," in *IEEE Conference Proceedings*, Mar. 2019.

[2] P. Shi, "Speech emotion recognition based on deep belief network," in *IEEE Conference Proceedings*, Mar. 2018.

[3] J. U. Maheswari and A. Akila, "An enhanced human speech emotion recognition using hybrid of prnn and knn," in *IEEE Conference Proceedings*, Feb. 2019.

[4] S. R. R. Gupta, M. S. Likitha, A. U. Raju, and K. Hasitha, "Speech based human emotion recognition using mfcc," in *IEEE Conference Proceedings*, Mar. 2017.

[5] Y. S. Ü. Sonmez and A. Varol, "New trends in speech emotion recognition," in *IEEE Conference Proceedings*, Jun. 2019.