Proceedings of the 3rd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/2025.22707

Construction and Optimization of Deception Detection Models Based on Audio Features

Xiaohao Kuang

Beijing University of Posts and Telecommunications, Beijing, China
kuangxiaohao@bupt.edu.cn

Abstract: This project aims to construct and optimize a deception model based on audio feartures, aiming to distinguish between truthful and deceptive narratives. In the era of rapidly spreading digital information, the detection of deceptive audio content has become increasingly critical to combat misinformation. This research collected a dataset of 100 audio clips, each 60 seconds in duration, recorded by speakers from various countries in their native and non-native languages. Key audio features, including pitch, power, jitter, shimmer, and zero-crossing rate (ZCR), were extracted from the audio clips using the librosa library, a widely used tool for audio signal processing. To classify the audio data, several machine learning models were implemented, such as Histogram-based Gradient Boosting Decision Trees (HGBDT), Random Forest, Support Vector Classifier (SVC), and Logistic Regression. These models were chosen for their complementary strengths in handling different types of data and classification tasks. To further improve the model's performance, an ensemble approach was adopted using the VotingClassifier, which combines the predictions of the individual models. The experimental results demonstrasted that the proposed model achieved an accuracy of 75%, successfully identifying subtle differences between truthful and deceptive speech patterns. This research highlights the effectiveness of audio-based deception detection in multilingual contexts and lays the groundwork for future studies.

Keywords: Audio analysis, Deception detection, Machine Learning, Multilingual settings, Ensemble learning

1. Introduction

In today's highly developed internet era, the rapid and widespread dissemination of deceptive audio content has made accurately distinguishing the authenticity of audio narratives increasingly important. Deception is a psychological process whereby an individual consciously tries to convince another person of a false beliefs [1].This project aims to predict whether a narrated story is true or not. Specifically, it explores how audio features such as pitch, speech rhythm, and voice quality can be leveraged to detect deception across different languages and cultural contexts. It also investigates the impact of native versus non-native language use on deception detection accuracy, as well as the challenges posed by varying speech patterns and emotional states in multilingual settings.

The dataset includes stories in multiple languages. Specifically, a machine learning model that takes as an input an audio recording of 60 seconds of duration is constructed to predict whether the story being narrated is true story or deceptive. The methodological approach involves preprocessing the dataset to extract pitch-related features such as pitch values, voiced/unvoiced flags, and voiced

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.22707

probability. These features, along with mean and standard deviation of pitch, total power, ratio of voiced regions, jitter, shimmer, and zero-crossing rate (ZCR), are organized into feature vectors and paired with labels to form the training dataset. The performance is evaluated using accuracy scores. This research aims to provide a new technical solution to prevent the spread of false information and advance machine learning in multilingual settings, ultimately contributing to more credible virtual communication platforms.

2. Audio feature-based deception detection model construction

2.1. Dataset construction

The original dataset comprises 100 audio clips, each ranging from 2 to 3 minutes in duration. These audio clips were recorded by individuals from various countries, narrating stories either in their native language or in English. The stories include both true and false narratives. For each audio clip, the corresponding information such as nationality, language, and the truthfulness of the story is known. The primary objective is to utilize machine learning techniques to distinguish between true and false stories.

Existing forensic studies primarily focus on how people perceive the credibility of statements made by native and non-native speakers. Numerous studies have consistently shown that observers are more likely to consider statements made by native speakers as truthful compared to those made by non-native speakers [2]. However, the model constructed in this study does not determine the veracity of a story based on whether it was narrated in the speaker's native language. Given that 2 to 3 minute audio clips are too long for the intended analysis, only the first 60 seconds of each clip are extracted. It is known that specific schemes can extract features in as little as 2 to 3 seconds, indicating that 60 seconds is sufficient for audio analysis and feature extraction [3].

The complete original dataset, including audio clips and corresponding metadata, is first downloaded. Subsequently, the audio is extracted, and the first 60 seconds of each clip are trimmed, maintaining the original order and storing them in a separate file. The remaining metadata, including the language used, whether it is the speaker's native language, and the truthfulness of the story, are kept in the original file in the same order. This results in a standardized dataset comprising 100 audio clips, each 60 seconds long, along with their associated language, native language status, and truthfulness information.

2.2. Audio feature extraction

Lying typically takes longer than truthful speech [4] and induces changes in autonomic nervous responses [5]. This difference in speech patterns and physiological responses can be captured through various audio features, which form the basis of the machine learning model for deception detection.

In the transformation phase, this research extract these features from raw audio data and convert them into a format suitable for model input. Specifically, the getPitch_librosa and getXy functions are utilized to extract features such as pitch, power, jitter, shimmer, and zero-crossing rate (ZCR). The following inputs and outputs are employed in this phase:

Inputs:

files: A list of audio file paths.

labels_file: A DataFrame containing label information.

scale_audio: A boolean indicating whether to normalize the audio.

power_avg_by_fs: A boolean indicating whether to average power by second during calculation.

Outputs:

X: A feature matrix of shape (number of samples, number of features).

y: A label vector of shape (number of samples,).

DOI: 10.54254/2755-2721/2025.22707

Feature Extraction Steps:

1.Read the audio file x and sample rate fs.

2.Extract pitch features f0 and voiced flags voiced_flag.

3.Compute the total power of the audio power.

4.Calculate the mean pitch_mean and standard deviation pitch_std of the pitch.

5.Calculate the ratio of voiced regions voiced_fr.

6.Compute jitter and shimmer.

7.Calculate the zero-crossing rate zcr.

8.Consolidate the extracted features into a feature vector xi.

9.Add the feature vector xi and corresponding label yi to the feature list X and label list y.

Functional Importance of Key Features:

Pitch feature f0 and voiced_flag: These features reflect the basic tone of the speech and whether a segment is voiced, which can be indicative of the speaker's emotional state and stress levels. In addition to their role in detecting emotional cues, these features are particularly impactful in improving the performance of Automatic Speech Recognition (ASR) systems, especially for tonal languages where pitch variations are crucial for conveying meaning [6]. By leveraging f0 and voiced_flag, the model can better capture the linguistic and paralinguistic characteristics of speech, enhancing its ability to distinguish between true and deceptive narratives in multilingual settings. This dual utility—both in emotional analysis and linguistic accuracy—makes these features indispensable for advancing the robustness and reliability of audio-based deception detection systems.Power feature power: This helps understand the overall energy distribution of the audio, which can be influenced by the speaker's effort and physiological state when lying.

Mean and standard deviation of the pitch (pitch_mean and pitch_std): These provide statistical information about pitch variation, which can be higher in deceptive speech due to increased stress or cognitive load.

Ratio of voiced regions voiced_fr: This helps distinguish between voiced and unvoiced segments, which is crucial for speech recognition and can provide insights into the speaker's voice production.

Jitter and shimmer: These features reflect the stability and quality of the speech signal, aiding in emotion analysis and authenticity judgment. Jitter measures the frequency perturbation, while shimmer measures the amplitude perturbation. Since they characterise some aspects concerning particular voices, it is expected to find differences in the values of jitter and shimmer among speakers, making them valuable features for distinguishing subtle vocal patterns, such as those associated with deceptive speech [7].

Zero-crossing rate (ZCR): This feature is crucial for detecting short-term changes in the speech signal, which are often more frequent in deceptive speech as speakers may attempt to modulate their voice to appear more convincing. While more complex features could be employed, the computational efficiency and reliable performance of ZCR make it a practical and effective choice for large datasets. This aligns well with the requirements of audio-based deception detection systems, where efficient and scalable feature extraction is essential for analyzing subtle vocal patterns across multilingual narratives [8].

By extracting and analyzing these features, our model aims to capture the subtle differences in speech patterns and physiological responses that distinguish truthful from deceptive speech, ultimately improving the accuracy of deception detection in multilingual settings.

2.2.1. Selection of appropriate audio feature extraction tools and methods

For feature extraction, the librosa library, a widely used tool in audio signal processing was empolyed. Librosa provides efficient functions for extracting pitch, power, and other spectral features. Additionally, custom functions were utilized to calculate jitter, shimmer, and zero-crossing rate

DOI: 10.54254/2755-2721/2025.22707

(ZCR). These tools were selected for their robustness, ease of use, and ability to handle multilingual audio data.

2.2.2. Extraction of acoustic, prosodic, and speech quality features

Three categories of features were extracted:

The extracted features can be categorized into three main groups: acoustic features, prosodic features, and speech quality features. Acoustic features, such as pitch, power, and zero-crossing rate (ZCR), capture the fundamental physical properties of the audio signal. Prosodic features, including pitch mean, pitch standard deviation, and voiced region ratio, reflect the rhythm and intonation of speech, providing insights into the speaker's vocal patterns. Pitch features, particularly those related to pitch level, have been shown to be highly effective in distinguishing subtle vocal variations, as evidenced by studies on prosodic behavior in speaker recognition [9]. When combined with duration and energy characteristics, these features offer valuable information about habitual speech patterns, making them particularly relevant for applications such as audio-based deception detection, where identifying voluntary but consistent vocal behaviors is crucial. Additionally, speech quality features, such as jitter and shimmer, measure the stability and quality of the speech signal, further enhancing the model's ability to detect subtle differences in voice production. Together, these features provide a comprehensive representation of the audio signal, enabling the model to effectively distinguish between truthful and deceptive speech.

2.3. Model training and evaluation

To achieve accurate deception detection, this study employed several widely used machine learning models, including Histogram-based Gradient Boosting Decision Trees (HGBDT), Random Forest Classifier, Support Vector Classifier (SVC), and Logistic Regression. These models were chosen for their complementary strengths.

HGBDT is efficient for large-scale datasets, handles various data types, and performs well with complex nonlinear relationships.

Random Forest prevents overfitting by aggregating multiple decision trees and is suitable for high-dimensional datasets.

SVChandles nonlinear classification problems using kernel functions and is resistant to noise and outliers.

Logistic Regression provides interpretable feature weights and is computationally efficient.

To further enhance model performance, this study adopted an ensemble learning strategy, integrating multiple models using a Voting Classifier. This approach aggregates predictions from different classifiers and determines the final classification result based on majority voting, thereby improving robustness and overall accuracy.

The model's performance was evaluated using accuracy as the primary metric. The ensemble model achieved an accuracy of 75%, demonstrating the effectiveness of an audio-based deception detection approach in distinguishing between truthful and deceptive speech.

3. Model optimization and extension

3.1. Data augmentation techniques

To address the limited dataset size, audio data augmentation techniques such as time-shifting, frequency shifting, and noise addition were applied in this model. These techniques increase the diversity of the training data, improving the model's generalization ability and robustness.The concept of time-shifting, in particular, aligns with broader strategies used in network congestion

DOI: 10.54254/2755-2721/2025.22707

management, where incentives are provided to shift data demand from peak to off-peak periods, optimizing resource utilization. Similarly, in audio data augmentation, time-shifting helps simulate varied temporal patterns, enhancing the model's ability to handle real-world variability in speech data [10].

3.2. Feature selection

Recursive Feature Elimination (RFE) and L1 regularization were employed to identify and retain the most relevant features. This approach reduced the dimensionality of the feature space thereby improving model efficiency and performance.

3.3. Hyperparameter tuning

Grid search and cross-validation were utilized to optimize the hyperparameters of each model. This process ensured that the models were fine-tuned to achieve optimal performance on the given dataset. Given the potential influence of spatial autocorrelation on performance estimation, as highlighted in ecological modeling studies, the choice of resampling methods was carefully considered to avoid overoptimistic results. While spatial and non-spatial partitioning showed minimal differences in hyperparameter sensitivity for some algorithms, spatial tuning was prioritized to maintain consistency with bias-reduced performance estimation, ensuring more reliable and generalizable model outcomes [11].

3.4. Other machine learning models

In addition to the four classifiers, experiments were conducted with other models such as Neural Networks and Gradient Boosting Machines (GBM). These models were evaluated to determine their potential for further improving deception detection accuracy.

4. Results discussion

4.1. Effectiveness of feature extraction

The extracted features, particularly pitch, power, and jitter, were highly effective in capturing the differences between truthful and deceptive speech. The combination of acoustic, prosodic, and speech quality features provided a robust representation of the audio signal.

4.2. Improvement in model performance

The ensemble model achieved an accuracy of 75%, outperforming individual classifiers. Data augmentation and feature selection further improved the model's performance, demonstrating the importance of these techniques in machine learning pipelines.

```
print('The shape of X is', X.shape)
print('The shape of y is', y.shape)
print('The labels vector is', y)
```

```
The shape of X is (100, 7)
The shape of y is (100,)
The labels vector is [False  True False False False False  True False False  True False  True
 False  True False  True  True  True False  True  True False False  True
  True  True False  True  True  True False False False  True  True False
 False False False False False False  True  True False False False  True  True
 False False False  True False  True  True False  True False  True False
  True  True  True  True False  True False  True False False False False
  True  True  True  True False  True  True  True False  True  True  True
  True False  True  True False False  True  True False False False False
  True False  True False]
```

Figure 1: True/False labels for the 100 audio narratives

DOI: 10.54254/2755-2721/2025.22707

```
[False  True  True  True False  True False  True  True False False False
 False False False  True  True False  True False]
[False  True  True False False  True False  True False False  True False
  True False False  True  True False  True  True]
Voting Classifier Accuracy: 0.75
```

Figure 2: Predicted labels (true/false) and model accuracy

5. Conclusion and future work

This project successfully constructed and optimized a deception detection model based on audio features. The model demonstrated strong performance in distinguishing between truthful and deceptive speech, particularly in multilingual settings. The incorporation of ensemble learning and feature selection techniques further enhanced its accuracy and robustness. However, several limitations should be acknowledged. The dataset size is relatively small, which may affect the model's generalization ability. Additionally, the model relies solely on audio features and does not incorporate multimodal data, such as visual or textual cues, which could provide complementary information for more accurate deception detection. The proposed model exhibits several strengths, including its language-agnostic design, which enables applicability across multilingual settings. Additionally, the use of ensemble learning significantly enhances model stability and accuracy. However, certain limitations must be acknowledged. The dataset size is relatively small, which may restrict the model's generalization ability. Furthermore, the model relies exclusively on audio features and does not incorporate contextual or semantic information, potentially limiting its ability to capture more nuanced aspects of deceptive behavior. These limitations highlight areas for future improvement, such as expanding the dataset and integrating multimodal data to enhance the model's robustness and accuracy.

To enhance deception detection accuracy, multimodal data is first introduced by incorporating visual and textual features. Subsequently, to boost the model's generalization ability, the dataset size should be increased through the collection of more audio samples. To further improve performance, the model architecture will be optimized by experimenting with advanced deep learning architectures such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). Finally, to explore real - world applications, the development of a real - time deception detection system for virtual communication platforms is aimed at.

By addressing these limitations and exploring future directions, this research aims to advance deception detection technology and contribute to more credible and trustworthy virtual communication environments.

References

[1] *DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. Psychological Bulletin, 129(1), 74-112. https://doi.org/10.1037/0033-2909.129.1.74*

[2] *Cheng, K. H. W., & Broadhurst, R. (2005). The detection of deception: The effects of first and second language on lie detection ability. Psychiatry, Psychology and Law, 12(1), 107-118. https://doi.org/10.1375/pplt.2005.12.1.107*

[3] *Kalluri, S. B., Vijayasenan, D., & Ganapathy, S. (2020). Automatic speaker profiling from short duration speech data. Speech Communication, 121, 16-28. https://doi.org/10.1016/j.specom.2020.03.008*

[4] *Suchotzki, K., Verschuere, B., Van Bockstaele, B., Ben-Shakhar, G., & Crombez, G. (2017). Lying takes time: A meta-analysis on reaction time measures of deception. Psychological Bulletin, 143(4), 428-453. https://doi.org/10.1037/bul0000087*

[5] *Furedy, J. J., Davis, C., & Gurevich, M. (1988). Differentiation of deception as a psychological process: A psychophysiological approach. Psychophysiology, 25(6), 683-688. https://doi.org/10.1111/j.1469-8986.1988.tb01908.x*

DOI: 10.54254/2755-2721/2025.22707

[6] *Ghahremani, P., BabaAli, B., Povey, D., Riedhammer, K., Trmal, J., & Khudanpur, S. (2014). A pitch extraction algorithm tuned for automatic speech recognition. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 2494-2498. https://doi.org/10.1109/ICASSP.2014.6854049*

[7] *Farrús, M., Hernando, J., & Ejarque, P. (2007). Jitter and shimmer measurements for speaker recognition. INTERSPEECH 2007, Antwerp, Belgium, 778-781.*

[8] *Gouyon, F., Pachet, F., & Delerue, O. (2000). On the use of zero-crossing rate for an application of classification of percussive sounds. Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-00), Verona, Italy.*

[9] *Shriberg, E., Ferrer, L., Kajarekar, S., Venkataraman, A., & Stolcke, A. (2005). Modeling prosodic feature s equences for speaker recognition. Speech Communication, 46(3-4), 455-472. https://doi.org/10.1016/j.specom. 2005.02.018*

[10] *Sen, S., Joe-Wong, C., Ha, S., & Chiang, M. (2012). Incentivizing time-shifting of data: A survey of time-de pendent pricing for internet access. IEEE Communications Magazine, 50(11), 91-99. https://doi.org/10.1109/ MCOM.2012.6353688*

[11] *Schratz, P., et al. (2019). Hyperparameter tuning and performance assessment of statistical and machine-lea rning algorithms using spatial data. Ecological Modelling, 406, 109-120. https://doi.org/10.1016/j.ecolmodel. 2019.06.002*