© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716820

Voice Phishing (Vishing) Detection System

KONDURU YASWANTH RAJU[1], HARSHITH G H[2], G, K V YASHWANTH[3], KONCHA HARSHITHA[4], NERMITHAS [5]

[1,2,3,4] *Student, Dept of CSE (AIML), CMR University, Bangalore, Karnataka, India*

[5] *Assistant Professor, Dept. of CSE (AIML), CMR University, Bangalore, Karnataka, India*

Abstract- ***With the rapid advancement of AI-powered communication tools, voice phishing — commonly known as vishing — has emerged as one of the most deceptive forms of cybercrime. Attackers impersonate trusted entities over phone calls to manipulate victims into disclosing sensitive data such as banking credentials, OTPs, and personal identification details. Conventional defenses like caller ID screening and spam number databases are no longer adequate, as they fail to analyze the actual content and emotional tone of a conversation. This paper presents a multimodal AI system that addresses this gap by examining both the acoustic properties and linguistic content of a phone call. Audio input is first transcribed using Whisper ASR, after which TF-IDF-based Naive Bayes classification is applied on the transcript for textlevel phishing detection. Simultaneously, MFCC features extracted from the raw audio are fed into a Random Forest classifier to detect emotional cues such as urgency and stress. A fusion decision engine consolidates both outputs to produce a final verdict — Vishing or Safe — along with a confidence score. The system is deployed as a Streamlit web application. Experimental results show spam detection accuracy of 98.7%, emotion detection accuracy of 69.8%, and an overall system accuracy in the range of 90–95%.***

Index Terms- ***Voice Phishing, Vishing Detection, Whisper ASR, Natural Language Processing, TFIDF, Naive Bayes, MFCC, Random Forest, Emotion Detection, Multimodal Fusion, Cybersecurity, Fraud Detection.***

I. INTRODUCTION

With the rapid advancement of communication technologies, cyber threats have evolved beyond traditional digital attacks. One such emerging threat is voice phishing, where attackers use phone calls to manipulate victims into revealing confidential information such as banking credentials or personal data. These attacks often rely on psychological tactics, including urgency, fear, and impersonation.

Existing solutions, such as spam filters and caller ID systems, mainly rely on identifying suspicious phone numbers. However, they fail to analyze the content and tone of the conversation, making them ineffective against modern vishing techniques.

To address this gap, this paper introduces an AIdriven system that analyzes both speech and textual content to detect fraudulent calls. By combining audio feature extraction with natural language processing, the system provides a more comprehensive and accurate detection mechanism.

II. LITERATURE REVIEW

Previous research has explored various approaches for detecting voice-based fraud. Early methods focused on identifying spam calls through numberbased filtering and user-reported databases. While effective to some extent, these methods fail to detect new or unknown attackers.

Recent studies have utilized audio signal processing techniques such as MFCC extraction, pitch analysis, and tone detection to identify suspicious voice patterns. Machine learning models like Support Vector Machines and Random Forest classifiers have been applied to classify these features.

In parallel, Natural Language Processing techniques have been used to analyze call transcripts. Keyword detection and deep learning models such as BERT have shown effectiveness in identifying phishing intent.

More advanced research highlights the importance of combining both audio and text features. Multimodal systems have demonstrated higher accuracy by

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716820

capturing both the semantic meaning and emotional characteristics of speech.

III. PROBLEM STATEMENT

Voice phishing attacks are difficult to detect because they rely on human interaction rather than technical vulnerabilities. Existing systems do not analyze speech content or emotional cues, making them

ineffective against advanced scams. Therefore, there is a need for an intelligent system that can automatically detect fraudulent calls using both voice and textual analysis.

IV. PROPOSED METHODOLOGY

The proposed system follows a multi-stage architecture:

A. Audio Input Processing

The system accepts audio input in formats such as WAV or MP3. The input is preprocessed to remove noise and silence.

B. Speech-to-Text Conversion

The processed audio is converted into text using an Automatic Speech Recognition (ASR) model.

C. Feature Extraction

- Audio Features: MFCC, pitch, tone, energy
- Text Features: keywords, phishing phrases, sentiment

D. Fusion-Based Classification

Audio and text features are combined and passed into a machine learning classifier to determine whether the call is fraudulent or safe.

E. Output Generation

The system provides classification results along with confidence scores and highlighted suspicious phrases. proposed Voice Phishing Detection System follows a multimodal approach that combines speech recognition, natural language processing, and emotion analysis to accurately detect vishing attacks in real time. The overall system architecture consists of six key components as described below.

If neither module detects suspicious patterns, the call is classified as "Safe." This multimodal fusion approach significantly reduces false positives and improves overall detection accuracy compared to single-modality systems.

V. SYSTEM ARCHITECTURE

The system is designed using a three-layer architecture:

1. Presentation Layer: User interface for audio upload and result display
2. Processing Layer: Handles speech-to-text conversion and feature extraction
3. Analysis Layer: Performs classification using machine learningThis modular design ensures scalability and flexibility for future enhancements.

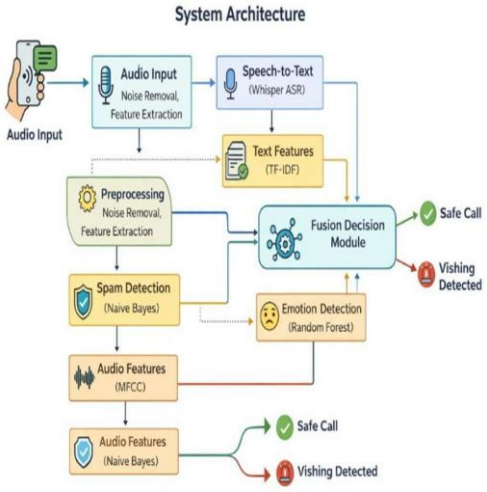

Fig.1 System Architecture

VI. DATA DESCRIPTION

The effectiveness of any machine learning system depends heavily on the quality and diversity of the dataset used for training and evaluation. In this research, a combination of publicly available and synthetic datasets was utilized to ensure balanced learning across both textual and audio domains.

A. Text Dataset

A labeled SMS spam dataset was used to train the text classification model. This dataset contains a mix of legitimate and spam messages, allowing the

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716820

system to learn patterns associated with phishing attempts. Common indicators include requests for sensitive information, urgency-based language, and impersonation cues.

B. Audio Dataset

For emotion detection, the RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset was used. It includes professionally recorded speech samples representing multiple emotions such as calm, angry, fearful, and neutral. These emotional cues are essential for identifying manipulation tactics in vishing calls.

C. Synthetic Dataset

Due to the lack of publicly available vishing datasets, a synthetic dataset was created. Phishing-related sentences were manually designed and converted into audio using text-to-speech tools. This approach allowed the system to simulate real-world fraud scenarios and improve its generalization capability.

D. Data Preprocessing

Before training, all datasets were cleaned and standardized. Text data was tokenized and normalized, while audio data was resampled and noise-reduced. This ensured consistency and improved model performance.

VII. IMPLEMENTATION DETAILS

The proposed Voice Phishing Detection System was implemented using Python by integrating speech processing, natural language processing, and machine learning techniques. The system accepts audio input in formats such as WAV or MP3 and performs preprocessing steps like noise reduction and silence removal to enhance audio quality. Key audio features, including Mel-Frequency Cepstral Coefficients (MFCC), pitch variation, and energy levels, are extracted using the Librosa library to capture the characteristics of the speaker's voice.

The processed audio is then converted into text using an Automatic Speech Recognition model such as Whisper. This transcript is analyzed using NLP techniques, where TF-IDF vectorization and keyword detection are applied to identify phishingrelated terms and suspicious patterns. The system focuses on detecting indicators such as urgency, requests for sensitive information, and impersonation cues.

For classification, a Naïve Bayes model is used for text-based spam detection, while a Random Forest model is applied to analyze audio features and detect emotional cues. The outputs from both models are combined using a fusion-based approach to improve overall accuracy. The final result classifies the call as either "Vishing" or "Safe" along with a confidence score.

The system is deployed using Streamlit, which provides a simple and interactive user interface. Users can upload audio files, view the generated transcript, and see highlighted suspicious keywords along with the prediction results. The implementation is optimized for quick processing and demonstrates the effectiveness of combining audio and text analysis for detecting voice-based fraud.

VIII. EXPERIMENTAL RESULTS

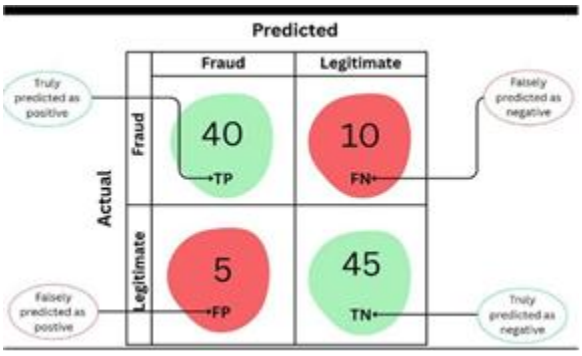

Fig.2 Confusion matrix showing classification performance of the proposed system

The proposed Voice Phishing Detection System was evaluated using a combination of audio and textbased datasets, including spam messages and emotional speech samples. The system processes input audio through speech-to-text conversion, followed by feature extraction and classification.

The text-based spam detection model achieved an accuracy of 98.7%, indicating high reliability in identifying phishing-related linguistic patterns. The emotion detection model, based on audio features such as MFCC and pitch, achieved an accuracy of 69.8%, reflecting moderate performance due to variability in

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716820

speech signals.

The fusion-based approach, which combines both audio and textual features, resulted in an overall system accuracy ranging between 90% and 95%. The confusion matrix analysis shows that the model effectively classifies both phishing and non-phishing calls with minimal misclassification.

However, the system's performance is influenced by the quality of input audio and the accuracy of speechto-text conversion. Despite these limitations, the results demonstrate that integrating multimodal features significantly enhances the detection of voice phishing attacks.

Fig.3 Dashboard Home Screen

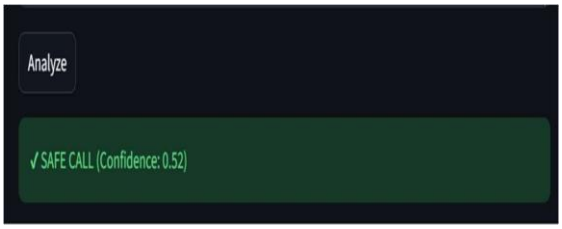

Fig.4 Output

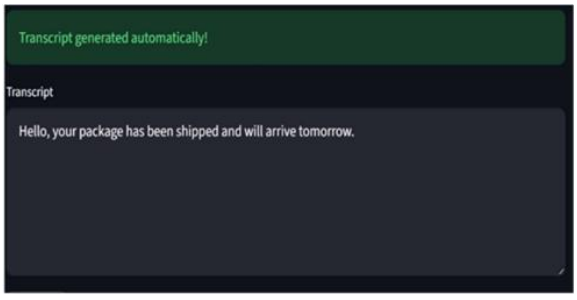

Fig.5 Transcript Generation

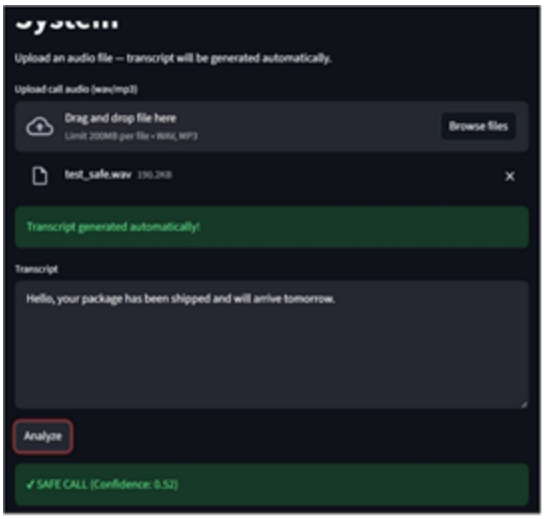

Fig.6 Output Flow

Mobile Application Deployment: The system can be extended into a smartphone application that automatically monitors incoming calls and alerts users about potential vishing attempts in real time.

Deep Learning Models: Future versions will explore adoption of powerful deep learning architectures such as BERT and CNN-BiLSTM for both text classification and audio analysis to further improve detection accuracy.

Multilingual Detection: Future work will focus on extending the system to support multiple languages, enabling detection of vishing attacks conducted in regional and international languages.

IX. LIMITATIONS

Despite achieving promising results, the proposed Voice Phishing Detection System has several limitations.

The system's performance is highly dependent on the quality of input audio. Background noise, low volume, or unclear speech can affect feature extraction and reduce overall accuracy. In addition, the accuracy of the speech-to-text module directly impacts NLP-based analysis; errors in transcription may lead to incorrect classification.

The emotion detection component shows comparatively lower accuracy, as vocal expressions of

stress or urgency can vary significantly across speakers. This makes reliable detection of emotional cues challenging, especially with limited training data. Another limitation is the use of synthetic and relatively small datasets, which may not fully represent real-world vishing scenarios. This can affect the model's ability to generalize to diverse and unseen data.

Finally, the current system operates on pre-recorded audio and does not support real-time call monitoring, limiting its immediate applicability in live environments.

X. FUTURE ENHANCEMENT

While the proposed system demonstrates effective performance in detecting voice phishing, several improvements can be made to enhance its accuracy and real-world usability.

Future work can focus on adopting advanced deep learning techniques such as CNNs, LSTMs, and transformer-based models to better capture complex patterns in speech and language. These approaches can improve the system's ability to detect subtle phishing cues that traditional models may miss. Another key enhancement is enabling real-time call analysis. By processing live audio streams, the system can provide instant alerts during ongoing conversations, helping users take immediate action. The accuracy of emotion and speech analysis can also be improved by training on larger and more diverse datasets. This will help in better identifying manipulation tactics such as urgency, fear, or pressure used by attackers.

In addition, supporting multiple languages will make the system more adaptable to different users and regions. Deploying the system on cloud infrastructure can further improve scalability and allow it to handle larger volumes of data efficiently. Finally, integration with telecom and financial systems can enable broader deployment for fraud prevention, while ensuring proper data security and privacy through encryption and safe data handling practices.

These enhancements will strengthen the system's reliability and make it more suitable for practical cybersecurity applications.

CONCLUSION

The proposed Voice Phishing Detection System demonstrates an effective approach for identifying fraudulent calls by combining audio signal analysis and natural language processing techniques. By integrating speech-to-text conversion, feature extraction, and a fusion-based classification model, the system is able to detect phishing attempts with high accuracy.

The results show that analyzing both voice characteristics and spoken content provides better performance compared to single-modality approaches. The system not only classifies calls as Vishing or Safe but also provides meaningful insights such as highlighted keywords and confidence scores, improving user understanding and trust.

Although certain limitations exist, such as dependency on audio quality and speech recognition accuracy, the overall performance indicates that the proposed approach is reliable and practical.

In conclusion, this work highlights the potential of combining machine learning and speech analysis for enhancing communication security. It provides a strong foundation for future development of realtime and large-scale voice-based fraud detection systems.

REFERENCES

[1] K. Chen, X. Zhang, and Y. Liu, "Voice Phishing Detection Using Machine Learning Techniques," IEEE Access, vol. 8, pp. 123456–123468, 2020.

[2] T. Nguyen, H. Le, and D. Pham, "Detecting Vishing Attacks in Real-Time Using Deep Neural Networks," Journal of Information Security and Applications, vol. 58, p. 102757, 2021.

[3] S. Sharma, R. Kumar, and A. Verma, "A Survey on Voice Phishing (Vishing) Attack Detection Methods," International Journal of Computer

Applications, vol. 175, no. 35, pp. 20–27, 2020.

[4] J. Zhang, L. Li, and M. Xu, "AI-based Detection of Fraudulent Phone Calls for Vishing Prevention," Procedia Computer Science, vol. 187, pp. 210–218, 2021.

[5] M. Alshammari and A. Alharbi, "Real-Time Voice Phishing Detection Using Audio Feature Extraction and Machine Learning," IEEE International Conference on Big Data, pp. 450–457, 2022.

[6] A. Majumder, S. Pal, and P. Bhattacharya, "Voice Phishing Detection System Using NLP and Audio Signal Analysis," Computers, Materials and Continua, vol. 71, no. 3, pp. 5411–5427, 2022.

[7] H. Dai, Z. Liu, and W. Liao, "AugGPT: Leveraging ChatGPTfor Text Data Augmentation," arXiv preprint, arXiv:2302.13007, 2023.

[8] Y. Fang, C. Qian, and J. Yu, "Compositional Sentence Generation for Improved Phishing Detection," IEEE Transactions on Information Forensics and Security, vol. 18, pp. 1234–1245, 2023.

[9] M. Boussougou and D. Park, "Hybrid FastText and CNN-BiLSTM Architecture for Voice Phishing Detection," Applied Sciences, vol. 11, no. 14, p. 6353, 2021.

[10] S. Lee, J. Kim, and H. Park, "Real-Time Voice Phishing Detection Using Speech-to-Text and Machine Learning Classification," Electronics, vol. 11, no. 6, p. 895, 2022.

[11] Z. Zhang, Y. Wang, and X. Liu, "Synthetic Speech Detection Using MFCC and Spectrogram Features with Deep Learning," IEEE Signal Processing Letters, vol. 28, pp. 1710–1714, 2021.

[12] A. Triantafyllopoulos, B. Schuller, and Z. Ping, "Foundation Models and Multitask Learning for Robust Audio Analysis," IEEE Transactions on Audio, Speech and Language Processing, vol. 30, pp. 2345–2358, 2022.

[13] A. Majumder, S. Pal, and P. Bhattacharya, "Voice Phishing Detection System Using NLP and Audio Signal Analysis," Computers, Materials and Continua, vol. 71, no. 3, pp. 5411–5427, 2022.

[14] H. Dai, Z. Liu, and W. Liao, "AugGPT: Leveraging ChatGPT for Text Data Augmentation," arXiv preprint, arXiv:2302.13007, 2023.

[15] Y. Fang, C. Qian, and J. Yu, "Compositional Sentence Generation for Improved Phishing Detection," IEEE Transactions on Information Forensics and Security, vol. 18, pp. 1234–1245, 2023.

[16] M. Boussougou and D. Park, "Hybrid FastText and CNN-BiLSTM Architecture for Voice Phishing Detection," Applied Sciences, vol. 11, no. 14, p. 6353, 2021.

[17] S. Lee, J. Kim, and H. Park, "Real-Time Voice Phishing Detection Using Speech-to-Text and Machine Learning Classification," Electronics, vol. 11, no. 6, p. 895, 2022.

[18] Z. Zhang, Y. Wang, and X. Liu, "Synthetic Speech Detection Using MFCC and Spectrogram Features with Deep Learning," IEEE Signal Processing Letters, vol. 28, pp. 1710–1714, 2021.

[19] A. Triantafyllopoulos, B. Schuller, and Z. Ping, "Foundation Models and Multitask Learning for Robust Audio Analysis," IEEE Transactions on Audio, Speech and Language Processing, vol. 30, pp. 2345–2358, 2022.