Beyond Caller ID: An Acoustic-Based Hybrid Temporal-Prosodic-Attention Framework for Vishing Detection

Saanj Rao[1]

Received April 20, 2026
Accepted September 20, 2026
Electronic access November 30, 2026

Voice phishing, also termed vishing, is a growing cybersecurity threat. In 2023, an estimated 68.4 million Americans fell victim to vishing attacks leading to financial damages totaling $1.2 billion. Traditional methods to detect vishing, such as rule-based filters and caller ID authentication, can be bypassed by attackers using sophisticated impersonation tactics and spoofing phone numbers. This study develops and evaluates a Hybrid Temporal-Prosodic-Attention (TPA) framework for real-time vishing detection. The framework analyzes speech patterns without recording or transcribing conversations. It uses wav2vec 2.0 for feature extraction, bidirectional LSTM networks for temporal analysis, and multi-head attention to extract important speech patterns. The effectiveness of the framework was measured using two datasets: PhishingVoice (N=80) for vishing detection and ASVspoof 2019 LA (N=2,000) for synthetic speech detection. Hyperparameters were optimized using Optuna across more than 30 configurations. The framework achieved 92.5% accuracy on PhishingVoice and 95.0% on ASVspoof with AUC-ROC scores of 0.987 and 0.980, respectively. It outperformed the baseline models, BiLSTM and GRU, by 10–15 percentage points while maintaining an inference latency of less than 35 milliseconds. However, the results on the PhishingVoice dataset should be treated with caution due to the small size of the dataset. These results suggest that analyzing the way people speak can be used for real-time fraud detection without the need to process private conversations.

Keywords: vishing detection, voice phishing, prosodic analysis, attention mechanism, BiLSTM, cybersecurity, fraud detection

Introduction

Voice phishing, known as vishing, is a growing cybersecurity concern. In 2023, vishing incidents rose by 30%, while hybrid attacks that combined voice calls, emails, and text messages increased by 554%.[1] This growth can be seen as part of the overall trend of internet crime. In 2023, the FBI's Internet Crime Complaint Center received over 880,000 complaints, with reported losses totaling over $12.5 billion. Phishing and spoofing accounted for nearly 300,000 of those complaints, a 160% increase from 2019.[2] These types of attacks are becoming more difficult to identify due to advancements in artificial intelligence. Today, attackers are able to use AI-generated speech and voice cloning to impersonate legitimate organizations and individuals that targets might trust.[3] Additionally, caller ID has become less reliable with over 10% of attacks using spoofed numbers. This allows attackers to make fraudulent calls appear to be coming from a legitimate source from the moment the call is made.[1]

Currently, most existing methods for detecting vishing rely on verification of caller ID, blacklist filtering, or rule-based systems. All of these methods have significant limitations. Caller ID verification can be circumvented through the use of spoofed phone numbers. Blacklist filtering is often ineffective due to attackers constantly changing phone numbers. Rule-based methods attempt to detect patterns in call metadata, such as the length of calls and frequency of calls. While these methods are able to detect suspicious callers or call patterns, they do not analyze the audio of the call. Therefore, they may miss vocal and prosodic cues that indicate a fraudulent intent.

Research on voice-based fraud detection draws from several fields including nonverbal communication, speech analysis, and machine learning. Early research on the topic focused on the way people speak and behave when they are making false statements. More recently, there has been a shift towards developing automated methods for analyzing speech. For example, in 1969, psychologists Paul Ekman and Wallace Friesen introduced the concept of nonverbal leakage. They discovered that facial expressions and body movements can sometimes reveal emotional responses that are involuntary and beyond the control of the person. Although Ekman and Friesen's research was focused on visible behavior and not vocal patterns, it provided an early theoretical foundation for studying involuntary behavioral cues.[4] Hirschberg et al. (2005) built upon this field

[1] *Monta Vista High School, Cupertino, CA*

© The National High School Journal of Science 2026

of research and investigated whether it was possible to use differences in speech to differentiate between truthful statements and intentionally false statements. They used recordings of structured interviews and extracted acoustic and prosodic features such as pitch, energy, duration, and pause patterns. They then tested these features along with word-based and speaker-specific features. While acoustic features alone only provided a slight improvement over the baseline classification model, when they combined feature sets and took into account individual speaking patterns, they saw a good improvement in performance.[5] Benus et al. (2006) investigated the use of silent and filled pauses in spontaneous interview recordings. They discovered that the frequency of pauses was higher in truthful speech because deceptive speakers attempt to suppress pauses. However, the relationship between pauses and deception varied depending on the type of pause. This indicates that pause patterns can be useful in providing information.[6] More relevant to voice phishing, Naidu (2022) created a vishing detection system that utilized natural language processing (NLP) to analyze transcribed phone conversations. This demonstrated the feasibility of automated vishing detection, but required the conversion of conversations into text first. An acoustic-based approach, on the other hand, operates directly on the audio, eliminating the need to process full conversation transcripts. However, the recordings themselves still need to be protected for privacy reasons.[7]

Advances in speech processing have also provided methods for creating more sophisticated models for analyzing audio. Mustaqeem, Sajjad, and Kwon (2020) utilized a deep bidirectional long short-term memory (BiLSTM) network for speech emotion recognition. A BiLSTM processes a sequence in both the forward and backward directions, enabling the model to utilize information from both earlier and later sections of an audio recording. Their research demonstrated the utility of this architecture for identifying patterns that span long periods of speech, although their focus was on emotion recognition and not voice-based fraud.[8] Baevski et al. (2020) introduced wav2vec 2.0, a framework that learns speech representations directly from audio. During pre-training, the model uses large amounts of untranscribed audio and learns to identify masked sections of its internal audio representations from the surrounding information. This process, referred to as self-supervised learning, enables the model to learn speech patterns without requiring manually labeled data for each recording. The resulting representations can then be fine-tuned with labeled examples for a specific task.[9] Vaswani et al. (2017) introduced multi-head attention, a mechanism that allows a model to examine various sections of an input sequence through multiple attention heads operating in parallel. Each head can learn different relationships between positions within the sequence. This enables the model to connect information from different sections of an audio recording, which is crucial because deceptive cues are distributed throughout the conversation.[10]

However, prosodic cues are not consistent across different speakers and settings. Levitan et al. (2018) investigated the use of linguistic features in truthful and intentionally false responses during interviews. They discovered that these features varied across gender and native language, indicating that patterns associated with false statements in one group may not be applicable to another. A detection model that does not take into account these differences may misclassify ordinary speech variations, thereby increasing the risk of false positives.[11] The recording environment also plays a role in the reliability of audio-based detection. The ASVspoof 2021 challenge assessed the ability of synthetic speech detection systems under different audio conditions, including compression and transmission channel variations associated with telephone and VoIP systems. These conditions can alter the original recording and make it more difficult for a model to distinguish between genuine speech and synthetic audio.[12] Another concern is the susceptibility of speech processing models to intentionally modified audio. Carlini and Wagner demonstrated that small modifications to an audio recording could cause an automatic speech recognition system to generate a specific, attacker-chosen transcription. While their research was focused on speech-to-text systems and not vishing detection, it illustrates how carefully crafted audio inputs can manipulate the outputs of machine learning models.[13] Recent research has also raised privacy concerns regarding speech-based AI systems. The HearSay benchmark assessed whether audio large language models could infer personal information directly from voice recordings. They found that these models could identify sensitive personal attributes from acoustic voice characteristics.[14]

Language and listener perception present additional challenges. Capuozzo et al. (2020) assessed the use of verbal false-statement detection across English and Italian datasets. While some linguistic features exhibited similar patterns across both languages, classification performance varied when the models were tested across languages, indicating that features learned from one language cannot always be applied directly to another.[15] Chen et al. (2020) compared the acoustic, prosodic, and lexical features of truthful and intentionally false statements with the features that human listeners associated with trustworthy speech. They discovered that several features could not reliably differentiate between true and false statements, emphasizing the distinction between how a person sounds and whether their statements are true.[16] Velutharambath et al. (2026) investigated whether intentionally false statements could be detected through written language alone. Across multiple datasets, they discovered that commonly studied linguistic features had a less statistically significant relationship with whether a statement was false, although

© The National High School Journal of Science 2026

their research was focused on written text and not recorded speech.[17] Collectively, these studies suggest that a detection system should not rely on a single feature. Variations between speakers, languages, and recording conditions can all influence the patterns that a model learns. Combining multiple types of information may provide a more reliable basis for classification, provided that the model is also tested under different conditions.

While phishing and smishing have been extensively studied, specialized detection mechanisms for vishing remain largely undeveloped.[18] Prior research also identified phishing, SMiShing, and vishing as significant threats to mobile device users, establishing the need for detection methods that address different types of social engineering.[19] Other research has explored the combination of different types of speech and language information. Chou et al. (2021) developed a detection framework that incorporated acoustic features, textual information, nonverbal behaviors, and conversational timing. Evaluated on the Mandarin Daily Deceptive Dialogues dataset, it achieved 80.61% unweighted accuracy, demonstrating the value of examining multiple aspects of a conversation rather than a single feature set. However, their research was focused on identifying false statements in dialogue and not detecting fraudulent phone calls.[20]

This research builds upon these prior approaches while specifically focusing on vishing detection. Naidu (2022) applied deep learning to vishing detection using transcribed call content, but did not directly incorporate prosodic features or the timing of vocal patterns. Other studies have investigated the use of acoustic and prosodic features in controlled interview settings, but have not combined them with contextual audio representations from models such as wav2vec 2.0 within a unified vishing detection framework. The proposed approach combines prosodic and temporal speech features with wav2vec 2.0 representations and multi-head attention to analyze different aspects of a recorded call. The primary research question is: Can a hybrid model that combines prosodic and temporal speech analysis with attention mechanisms improve vishing detection accuracy compared to baseline sequence models, such as TCN, GRU, and basic BiLSTM architectures? This study addresses that gap by combining prosodic features, wav2vec 2.0, and multi-head attention in a unified framework specifically designed for vishing detection.

Methods

Dataset Description

I am using two sets of data in my research. One of them is called the PhishingVoice Dataset. I found it via Kaggle. There are 80 labeled recordings of actual vishing attempts and simulated attacks. Of these, 44 are vishing calls and 36 are legitimate calls, therefore resulting in a slight class imbalance. This is typical of the natural occurrence of such events. This dataset was used to fine tune and test the model on the characteristics of speech that occur during social engineering and fraudulent phone calls. The ASVspoof 2019 LA dataset that contains 2,000 samples is used to pre-train the model to recognize both synthetic and spoofed speech. These datasets have different purposes. PhishingVoice is used to distinguish vishing calls from legitimate calls, while ASVspoof is used to distinguish genuine speech from spoofed audio. Therefore, results will be presented separately. It is noteworthy that the performance on ASVspoof is not treated as evidence to detect vishing accuracy.

Data Preprocessing

The audio files were resampled to 16 kHz and amplitudes were normalized using librosa. Features were scaled to have zero mean and unit variance using StandardScaler. To prevent data leakage, the data was first split by source call and speaker identity using GroupShuffleSplit. This ensured that augmented versions of the same recording could not appear in both the training and testing sets. Gaussian noise was then added only to the training data, at levels ranging from 0.003 to 0.007, to simulate real-world call conditions. This split-then-augment approach follows recommended practices to prevent data leakage in speech machine learning research.[21] 5-fold cross-validation was used with random seed 42. For PhishingVoice, each fold contained approximately 16 test samples.

Table 1 Feature extraction methods, dimensions, and purpose.

Feature Type	Method	Feature Dimension	Purpose
Speech Representation	Wav2Vec 2.0	768D	Captures phonetic/linguistic patterns
Prosodic Features	Librosa	46D	Pitch (3), energy (1), ZCR (1), MFCCs+delta+delta2 (39), spectral centroid (1), spectral rolloff (1)
Combined	Concatenation	814D per timestep	Hybrid temporal-prosodic representation

Feature Extraction

The approach combines two complementary feature streams. Through iterative testing, combining both streams yielded a 6.2% accuracy improvement over using either alone. Wav2vec 2.0, which has shown strong results in low-resource speech

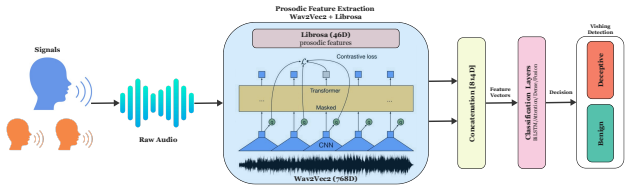

Figure 1a Feature extraction pipeline: raw audio processed through wav2vec 2.0 (768D) and Librosa prosodic features (46D), concatenated into an 814D feature vector per timestep.

```
# Feature extraction: wav2vec 2.0 + prosodic features
inputs = processor(audio, sampling_rate=16000, return_tensors='pt')
wav2vec_feats = model(**inputs).last_hidden_state.numpy()
f0, _, _ = librosa.pyin(audio, fmin=note_to_hz('C2'), fmax=note_to_hz('C7'))
features = np.concatenate([wav2vec_feats.flatten(), prosodic_features])
```

Figure 1b Feature extraction combining wav2vec 2.0 contextual embeddings with librosa prosodic features.

classification[9,22] produces 768-dimensional contextual feature vectors from raw audio through a convolutional encoder and transformer-based context network. Its effectiveness has been demonstrated across diverse audio classification tasks beyond speech recognition, including voice disorder classification[23], supporting its applicability here for detecting deceptive acoustic patterns in vishing calls. Prosodic features are extracted via librosa and include pitch (F0) contours and variations, speech rate, energy distribution, MFCCs (13 coefficients), spectral centroid and bandwidth, and zero-crossing rate.

Model Architecture

The Hybrid TPA architecture processes the combined 814-dimensional feature vector (per timestep) through batch normalization and a Dense(256) projection layer, then two bidirectional LSTM layers[24] (256 units each, return_sequences=True) capturing temporal dependencies in both directions. The BiLSTM input is a proper 3D tensor (batch, num_timesteps, 814). Tensor shapes at every stage: Input (batch, T, 814) -> dense projection (batch, T, 256) -> BiLSTM (batch, T, 256) -> multi-head attention with 4 heads (batch, T, 256) -> global average pooling (batch, 256) -> Dense(256) -> sigmoid output (batch, 1). LSTM outputs are enhanced through multi-head attention (4 heads), fused with global temporal average pooling. A Dense(256) classification layer with batch normalization, dropout, and sigmoid activation produces the final vishing probability score.

Training and Hyperparameter Optimization

Models were trained using AdamW (learning rate 0.001, weight decay 5e−4, gradient clipping 0.5, AMSGrad) with

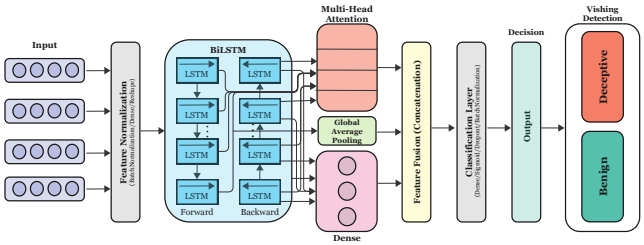

Figure 2a Hybrid TPA model architecture showing BiLSTM, multi-head attention, global average pooling, feature fusion, and classification layers.

```
# Hybrid TPA model architecture
bilstm = Bidirectional(LSTM(256, return_sequences=True))(x)
attn = MultiHeadAttention(num_heads=4, key_dim=64)(bilstm, bilstm)
fused = Concatenate()([GlobalAvgPool1D()(bilstm), GlobalAvgPool1D()(attn)])
output = Dense(1, activation='sigmoid')(Dense(256)(fused))
```

Figure 2b Core Hybrid TPA architecture: BiLSTM feeds 4-head attention, fused with global pooling before sigmoid output.

```
# Optuna hyperparameter search
lstm_units = trial.suggest_categorical('lstm_units', [256, 384, 512])
attn_heads = trial.suggest_categorical('attn_heads', [4, 8])
return train_and_evaluate(lstm_units, attn_heads)
```

Figure 3 Optuna hyperparameter optimization. Optimal: 256 LSTM units, 4 attention heads, approximately 15% accuracy gain.

early stopping (patience=8), learning rate reduction on plateau, and weight restoration to best checkpoint, capped at 30 epochs with batch size 32. The Optuna framework[25] was used for hyperparameter optimization, exploring LSTM units {128, 256, 384, 512, 640, 768}, attention heads {2, 4, 8, 12, 16}, dense units {64, 128, 256, 384, 512, 768}. All baseline models (TCN[26], GRU, standard BiLSTM) received the same features, the same Optuna-based tuning budget, the same data splits, and the same early stopping protocol.

Evaluation

Model performance was assessed using accuracy, precision, recall, F1-score, and AUC-ROC. Robustness was evaluated under Gaussian noise from 0.003 to 0.007. A weighted composite score was applied: 50% clean accuracy, 30% noise robustness, 20% consistency. K-fold cross-validation (K=5, random seed 42) was employed. Inference latency was measured on Apple M1 Pro (18GB RAM) with batch size 1, model inference only; wav2vec 2.0 features were precomputed offline. Robustness testing was limited to Gaussian noise; real telecom environments also involve packet loss, compression artifacts, and codec distortions, which were not evaluated in this study. Prior work has shown that GSM coding and noise significantly

© The National High School Journal of Science 2026

degrade speaker embedding performance[27].

Ethical Considerations

This study uses two publicly available datasets. Since voice recordings are considered biometric data under GDPR Article 9, CCPA Section 1798.140, and BIPA, they require careful handling. The PhishingVoice dataset does not include clear documentation of participant consent or de-identification procedures. The ASVspoof 2019 LA dataset is better documented through its challenge framework, although permissions for long-term data retention remain unclear. The detection system analyzes speech patterns without storing or examining the meaning of spoken content. It uses a zero-retention architecture, meaning that audio recordings and extracted features are not stored after the analysis is complete. There is ongoing research into privacy-preserving methods for speech embeddings.[28] Deployment in a real-world setting would require explicit informed consent from participants, including an explanation of how their voice data is analyzed.

Results

Performance Development Timeline

The model was developed in four phases. First, in Phase 1, the baseline BiLSTM model achieved an accuracy of 82%. The strongest speech patterns related to vishing calls included unusually stable pitch and changes in speaking rate. Next, in Phase 2, the Hybrid TPA model improved accuracy from 82% to 89% and reduced the false positive rate by 7%. However, the model's latency was still 50–60 ms. In Phase 3, fine-tuning the attention weights reduced latency to under 35 ms, while validation accuracy increased to 92–95%. Finally, in Phase 4, the final model achieved a false positive rate below 2%.

Overall Model Performance

Table 2 Hybrid TPA model performance on both evaluation datasets.

Metric	PhishingVoice	ASVspoof 2019 LA
Accuracy	92.5%	95.0%
Precision	88.6%	96.8%
Recall	97.5%	93.1%
F1-Score	92.8%	94.9%
AUC-ROC	0.987 (95% CI: 0.982-0.992)	0.980 (95% CI: 0.977-0.983)
Latency	<35ms	<35ms

Prosodic Feature Importance

The results from the analysis of prosodic features showed that there were several differences between legitimate and fraudulent calls related to pitch, speech rate, and energy variance (Fig. 6). The fraudulent calls showed a higher average pitch and greater pitch variation, especially during questions or confirmation requests. Lower speech rate with higher energy variance was also noticed, which indicates urgency produced unusual patterns. This suggests that pitch, speech rate, and vocal energy may be useful features to detect fraudulent calls. This is similar to the ablation study, since it showed a 3.8 percentage point drop in accuracy when prosodic features were removed.

Ablation Study: Feature Set Impact on ASVspoof 2019 LA Accuracy. Wav2vec 2.0 only: 91.2% accuracy, 90.8% F1. Prosodic features only: 85.7% accuracy, 84.9% F1. Hybrid (wav2vec + prosodic): 95.0% accuracy, 94.9% F1. The hybrid model is better than each feature set used alone. It improves the accuracy of wav2vec 2.0 by 3.8 percentage points and the accuracy of the prosodic features by 9.3 percentage points.

Comparison with Baseline Models

Table 3 Hybrid TPA outperforms all baselines on accuracy, F1-score, and inference latency.

Model	PhishingVoice	ASVspoof 2019	Avg F1	Latency
TCN	85.1%	86.4%	85.5%	35ms
GRU	79.3%	81.6%	80.2%	38ms
Standard BiLSTM	82.7%	84.2%	83.1%	42ms
Hybrid TPA (ours)	92.5%	95.0%	93.9%	34ms

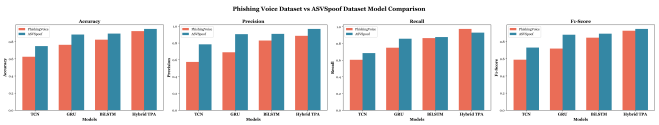

Figure 4 Model performance comparison across PhishingVoice and ASVspoof 2019 LA datasets. Hybrid TPA outperforms TCN, GRU, and standard BiLSTM baselines across all four metrics.

Training Dynamics

Table 4 Selected ASVspoof 2019 epoch results. Learning rate reductions at epochs 19 and 26 produce the largest accuracy gains.

Epoch	Train Acc	Val Acc	Val F1	Learning Rate
1	70.4%	87.4%	87.96%	0.0014
2	81.0%	88.9%	88.70%	0.0014
3	83.8%	89.6%	90.06%	0.0014
25	95.4%	89.5%	89.69%	0.000284
26	96.4%	89.1%	89.36%	0.000057
27	95.8%	89.6%	89.84%	0.000057

Attention Mechanism Analysis

Since the multi-head attention mechanism uses single-head self-attention in parallel, it allows the model to examine differ-

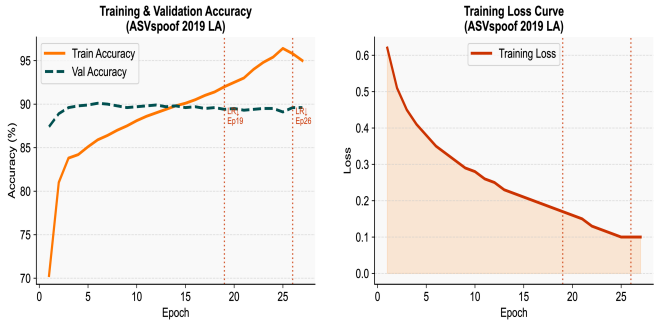

Figure 5 Training dynamics over 27 epochs (ASVspoof 2019 LA). Left: train and validation accuracy with learning rate reduction markers. Right: training loss curve.

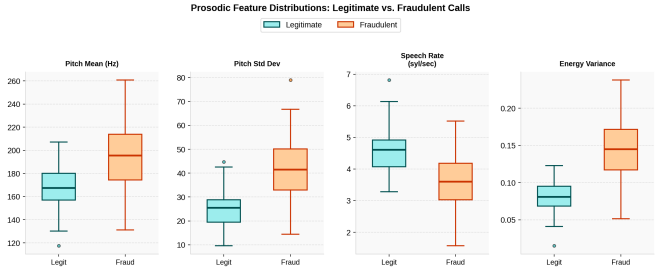

Figure 6 Prosodic feature distributions: legitimate vs. fraudulent calls. Fraudulent calls show higher pitch mean, greater pitch variability, lower and more irregular speech rate, and elevated energy variance.

ent parts of a speech recording simultaneously. This is why it can pick up the unique patterns spread across the call to help distinguish vishing calls from legitimate calls. The analysis showed that some parts of the calls received more attention than others. The strongest peaks were located at positions 8, 22, 34, and 51, where questions and responses changed. Smaller peaks that possibly indicate hesitation or uncertainty were located at positions 4, 6, 12, and 15.

The model classified vishing calls using a probability threshold of 0.613. It showed high confidence in segments 0–40 during the opening of the call and 80–130 when callers requested information. In addition, the mean attention weight for vishing calls was 0.742, while for legitimate calls it was 0.381. Similar attention patterns were found in 94.2% of the samples.

Robustness and Cross-Validation

The model achieved 95% accuracy for all tested noise levels ranging from 0.003 to 0.007. The false positive rate was also less than 5% under all conditions. Five-fold cross-validation further showed model stability. For ASVspoof, accuracy was

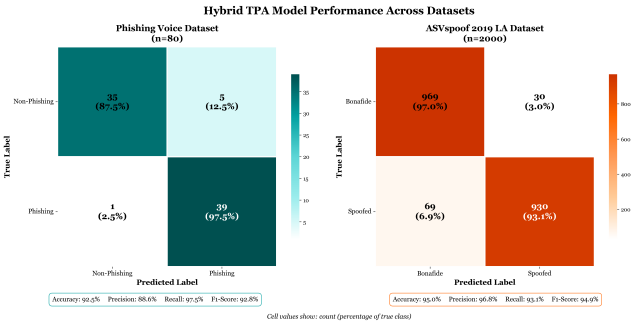

Figure 7 Confusion matrices for PhishingVoice (left, n=80) and ASVspoof 2019 LA (right, evaluation subset n=2,000). PhishingVoice: 97.5% phishing detection, 87.5% non-phishing detection. ASVspoof: 97.0% bonafide detection, 93.1% spoofed detection.

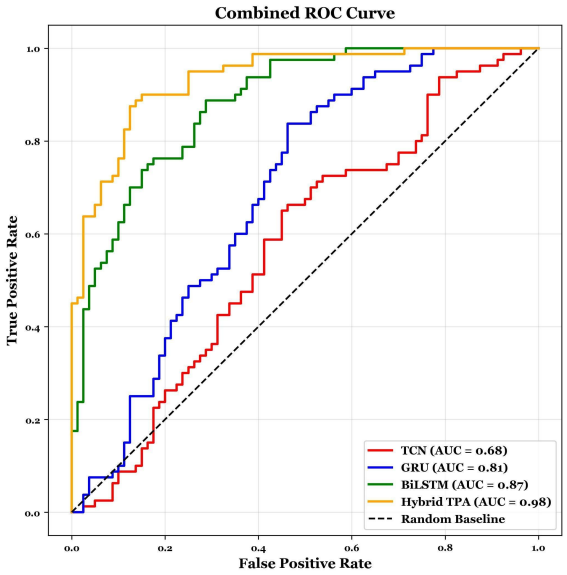

Figure 8 Combined ROC curves for all models on PhishingVoice dataset. Hybrid TPA achieves AUC=0.98, outperforming BiLSTM (0.87), GRU (0.81), and TCN (0.68).

between 94.2% and 95.8% with an average of 95.0% and a standard deviation of 0.6%. In contrast, PhishingVoice showed greater variation because the sample size for testing was only around 16 samples per fold. Accuracy ranged from 87.5% to 100.0%, with a mean of 92.5%, a standard deviation of 4.7%, and a 95% confidence interval of 85.1–97.2%.

In addition, paired t-tests across the five cross-validation folds showed statistically significant differences between the Hybrid TPA model and each baseline model. On ASVspoof: Hybrid TPA vs. TCN ($p = 0.003$), vs. GRU ($p < 0.001$), vs. BiLSTM ($p = 0.008$). On PhishingVoice: Hybrid TPA vs. TCN ($p = 0.012$), vs. GRU ($p = 0.004$), vs. BiLSTM ($p = 0.019$).

© The National High School Journal of Science 2026

Discussion

Key Findings

The Hybrid TPA framework achieved 92.5% accuracy on PhishingVoice and 95.0% on ASVspoof 2019, outperforming all three baseline models by 10–15 percentage points while maintaining an inference latency below 35 ms. AUC-ROC scores of 0.987 and 0.980, respectively, further demonstrated its ability to distinguish between the two classes. Cross-validation results also showed stable performance on ASVspoof, with a standard deviation of 0.6%, compared to the smaller PhishingVoice dataset with a standard deviation of 4.7%. Overall, the results supported the hypothesis that the hybrid approach outperforms standard sequential models across the evaluation metrics.

Implications and Significance

Several technical choices contribute to the model's performance. By combining the contextual representation of wav2vec 2.0[9] with the prosodic features, additional information was captured that would have been lost if each were used separately. This resulted in an increase in accuracy of 6.2% when using the hybrid feature approach. The use of multiple heads in the multi-head attention mechanism allowed the model to focus on certain parts of the audio file in order to detect patterns indicative of fraudulent activity. This is especially true for vishing calls which contain some normal conversation prior to the fraudulent request being made. Since the model analyzes prosodic and temporal features of the call, it does not require a transcript of the call, thus making it possible to utilize in environments where there are legal restrictions related to call recording.

One unexpected finding was that 4 attention heads and LSTM layers with 256 units provided the optimal trade-off between model capacity and over-fitting based upon the available dataset size. It appears that matching the complexity of the model to the amount of training data may provide a more effective method than increasing model size alone. Additionally, the difference in average attention weight for vishing calls (0.742) vs legitimate calls (0.381) indicates that the model has learned patterns specific to both classes of calls, versus simply learning the patterns contained within the training data.

Vishing attacks are most often targeted towards older adults, non-native English speakers, and those who are less familiar with technology. Real-time detection of vishing could prevent a victim from disclosing sensitive information. However, this analysis of false positives revealed that the model was more likely to misclassify calls from non-native English speakers with heavy accents as well as individuals with other atypical speech patterns (e.g., vocal fry). Therefore, the model appears to be better suited to detect fraud in speakers with more standard speech. It is to be noted that more training and evaluation on recordings from a larger population of speakers are also necessary.

Limitations

The study used two publicly available datasets for evaluation with 2,080 samples in total. However, the PhishingVoice dataset is small with only 80 samples and around 16 test samples per fold. This makes the results of the PhishingVoice dataset very specific and hard to generalize. All of the datasets are in English. The wav2vec 2.0 BASE model was pre-trained on the LibriSpeech dataset that consists of read audiobook recordings in relatively clean studio environments. Telephone recordings of vishing calls can contain various distortions such as those caused by transmission channels and codec compression. Thus, the audio used for pre-training and the audio recordings of vishing calls analyzed by the framework can be very different, which could limit the extent to which the learned representations of the model can be applied to real-world calls.

Conclusion

This research presents a Hybrid Temporal-Prosodic-Attention framework for vishing detection that achieves 92.5–95.0% accuracy and AUC-ROC scores of 0.987 and 0.980 across two independent datasets. It outperformed standard BiLSTM and GRU baseline models by 10–15 percentage points while maintaining an inference latency below 35 ms. The results suggest that combining wav2vec 2.0 contextual features with prosodic analysis through a BiLSTM-attention architecture can improve the identification of voice-based fraud without requiring conversation content to be recorded or transcribed.

Future work should focus on expanding multilingual training data, developing dedicated methods for deepfake voice detection as voice synthesis technology advances, collecting larger annotated vishing datasets, and exploring the integration of physiological biomarkers. With tens of millions of Americans affected by vishing annually and billions of dollars in losses across the broader fraud landscape[1,2], improving defenses against these attacks remains an important public safety concern. This research shows the potential of analyzing how a caller speaks to provide an additional layer of real-time protection while reducing the need to process private conversation content. As voice synthesis technology continues to advance, developing reliable and privacy-conscious methods for identifying fraudulent calls will become increasingly important.

Acknowledgments

I would like to sincerely thank Ms. Sushma Bana and Ms. Kavita Gupta for their guidance and mentorship during this

© The National High School Journal of Science 2026

research project. Their knowledge and direction were very helpful in completing this research work.

References

1 GSMA Association. Combating vishing: an analysis of voice call impersonation and emerging defenses. GSMA. April 2024. www.gsma.com/solutions-and-impact/gsma-membership/member-resources/combating-vishing.

2 Federal Bureau of Investigation. 2023 internet crime report. IC3. 2024. www.ic3.gov/annualreport/reports/2023_ic3report.pdf.

3 F. Toapanta, B. Rivadeneira, C. Tipantuna, D. Guaman. AI-driven vishing attacks: a practical approach. Engineering Proceedings. Vol. 77, pg. 15, 2024. https://doi.org/10.3390/engproc2024077015.

4 P. Ekman, W. V. Friesen. Nonverbal leakage and clues to deception. Psychiatry. Vol. 32, pg. 88–106, 1969.

5 J. Hirschberg, S. Benus, J. M. Brenier, F. Enos, S. Friedman, S. Gilman, C. Girand, M. Graciarena, A. Kathol, L. Michaelis, B. Pellom, E. Shriberg, A. Stolcke. Distinguishing deceptive from non-deceptive speech. Proceedings of Interspeech. pg. 1833–1836, 2005.

6 S. Benus, F. Enos, J. Hirschberg, E. Shriberg. Pauses in deceptive speech. Proceedings of the 3rd International Conference on Speech Prosody. pg. 2–5, 2006.

7 D. Naidu. Voice analysis system for detection of vishing using deep learning. International Journal of Health Sciences. Vol. 6, pg. 10457–10466, 2022. https://doi.org/10.53730/ijhs.v6nS1.7520.

8 Mustaqeem, M. Sajjad, S. Kwon. Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM. IEEE Access. Vol. 8, pg. 79861–79875, 2020. https://doi.org/10.1109/ACCESS.2020.2990405.

9 A. Baevski, Y. Zhou, A. Mohamed, M. Auli. Wav2vec 2.0: a framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems. pg. 12449–12460, 2020.

10 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin. Attention is all you need. Advances in Neural Information Processing Systems. pg. 5998–6008, 2017.

11 B. M. Levitan, A. Maredia, J. Hirschberg. Linguistic cues to deception and perceived deception in interview dialogues. Proceedings of NAACL-HLT. pg. 1941–1950, 2018. https://aclanthology.org/N18-1176/.

12 J. Yamagishi, X. Wang, M. Todisco, et al. ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection. Proc. ASVspoof 2021 Workshop. 2021. https://arxiv.org/abs/2109.00537.

13 N. Carlini, D. Wagner. Audio adversarial examples: targeted attacks on speech-to-text. Proceedings of the IEEE Security and Privacy Workshops (SPW). pg. 1–7, 2018. doi: 10.1109/SPW.2018.00009.

14 J. Wang, L. Lin, K. Luo, et al. HearSay benchmark: do Audio LLMs leak what they hear? arXiv preprint arXiv:2601.03783, 2026.

15 P. Capuozzo, I. Lauriola, C. Strapparava, F. Aiolli, G. Sartori. Automatic detection of cross-language verbal deception. Proceedings of the Annual Meeting of the Cognitive Science Society. 2020.

16 X. Chen, S. I. Levitan, M. Levine, M. Mandic, J. Hirschberg. Acoustic-prosodic and lexical cues to deception and trust. Transactions of the Association for Computational Linguistics. Vol. 8, pg. 1–16, 2020.

17 A. Velutharambath, D. Russo, R. Rinott. Is deception detection really possible from linguistic cues? arXiv preprint arXiv:2505.13147, 2026.

18 A. E. Chichwadia, N. Mpekoa. A machine learning algorithm to detect and prevent smishing and vishing. Proceedings of i-Society 2024. Churchill College, University of Cambridge. 2024. https://doi.org/10.20533/iSociety.2024.0018.

19 E. O. Yeboah-Boateng, P. M. Amanor. Phishing, SMiShing and vishing: an assessment of threats against mobile devices. Journal of Emerging Trends in Computing and Information Sciences. Vol. 5, pg. 297–307, 2014.

20 H. Chou, Y. Liu, C. Lee. Automatic deception detection using multiple speech and language communicative descriptors in dialogs. APSIPA Transactions on Signal and Information Processing. Vol. 10, e5, 2021.

21 J. Celeste Jr, M. Tasnim, A. J. Valdes Cuervo, E. A. de la Cal, E. Stroulia. A software pipeline for systematizing machine learning of speech data. Frontiers in Psychiatry. Vol. 16, 1451368, 2025.

22 X. Zhang, X. Zhang, W. Chen, C. Li, C. Yu. Improving speech depression detection using transfer learning with wav2vec 2.0 in low-resource environments. Scientific Reports. Vol. 14, 9543, 2024.

23 J. Cai, Y. Song, J. Wu, X. Chen. Voice disorder classification using wav2vec 2.0 feature extraction. Journal of Voice. 2024. https://doi.org/10.1016/j.jvoice.2024.09.002.

24 S. Hochreiter, J. Schmidhuber. Long short-term memory. Neural Computation. Vol. 9, pg. 1735–1780, 1997.

25 T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama. Optuna: a next-generation hyperparameter optimization framework. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pg. 2623–2631, 2019.

26 S. Bai, J. Z. Kolter, V. Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271. 2018.

27 M. H. Alsalihi, D. Sztaho. Effect of noises and GSM codings on pre-trained speaker embedding models in forensic voice comparison. IEEE Access. 2025. doi: 10.1109/ACCESS.2025.3593024.

28 C. Lavania, S. Das, X. Huang, K. J. Han. Utility-preserving privacy-enabled speech embeddings for emotion detection. Proc. Interspeech. 2023.

© The National High School Journal of Science 2026