Unmasking Deception: Deepfake and Lie Detection Techniques in Audio Forensics

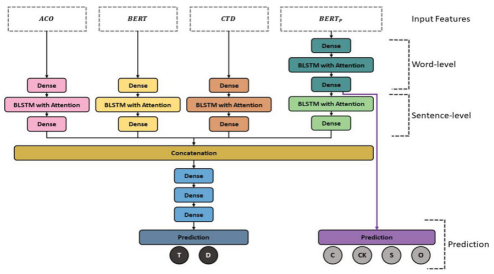
Shyam Sathvik [1] Neermita Bhattacharya [2]

Indian Institute of Technology, Jodhpur, Jodhpur 342037, India
{b22ee036, b22cs092}@iitj.ac.in

Abstract

Audio forensics has proven to be crucial lately due to the increasing challenges imposed by deception and synthetic media manipulation. Lie detection, one of the cornerstones of audio forensics, uses advanced techniques in speech processing to identify deceiving behavior due to psycho-neural changes of speech. In this regard, state-of-the-art methods used include Mel-Frequency Cepstral Coefficients (MFCCs), LSTMs, and CNNs, which have demonstrated promising results in making a distinction between truthful and deceitful speech. These approaches offer non-invasive, scalable, and automated solutions. However, challenges such as speech quality dependency and dataset limitations as well as cultural biases remain, and research in this area needs to further enhance the robustness and generalizability.

Audio forensics also deals with deepfake detection: Identifying synthetic speech that is produced from text-to-speech (TTS) and voice conversion (VC) systems. Advanced models like ARawNet2 and AASIST incorporate spectral, phase-based, and deep learning features to enhance generalization against unseen spoofing attacks. These systems are important for protecting automatic speaker verification systems and audio evidence quality in forensic and legal applications. However, there are some challenges that persist, such as computational complexity and dataset diversity. Lie detection and deepfake detection go hand in hand in audio forensics to serve as a means of countering deception and manipulation of synthetic media in a fast-growing digital world. Code files and analyses are available at github.com/Neermita18 and github.com/boku13.



(a) A Lie Detection Model

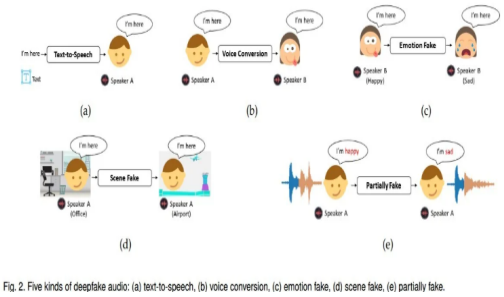


Fig. 2. Five kinds of deepfake audio: (a) text-to-speech, (b) voice conversion, (c) emotion fake, (d) scene fake, (e) partially fake.

(b) Types of Deepfake Audio

Figure 1: Lie and Deepfake detection

Contents

1 Introduction

1.1 Audio Forensics in the Real World

1.1.1 Importance of Lie Detection

Lying is very common in our day-to-day lives. But in the context of audio forensics, lie detection is a critical area of research with significant implications for law enforcement, legal proceedings, and security applications. The ability to discern deceptive speech from truthful speech can provide valuable insights during criminal investigations, interrogations, and courtroom testimonies. In forensic contexts, audio recordings of interviews, or witness statements are often analyzed to determine the veracity of the speaker's claims. Detecting deception in these recordings can help investigators identify inconsistencies, uncover hidden truths, and build stronger cases. Moreover, in high-stakes scenarios such as counter-terrorism or border security, lie detection systems can be used to screen individuals for potential threats, even in situations where traditional polygraph tests are impractical or inadmissible. In our everyday life, prank calls to 911/100, digital arrests, and spam calls can be avoided too.

One of the key advantages of audio-based lie detection is its non-invasive nature. Unlike polygraph tests, which require physical sensors to measure physiological responses (e.g., heart rate, blood pressure), audio-based systems don't need them, reducing the likelihood that the subject will alter their behavior to manipulate the results. This is particularly important in real-world applications where subjects may be aware of being monitored (and they take sedatives, antiperspirants, etc). Additionally, speech characteristics such as pitch, tone, and speech rate are generally more difficult to control voluntarily than physiological responses, making audio-based deception detection potentially more reliable in certain contexts.

The development of advanced machine learning and deep learning techniques has further enhanced the potential of audio-based lie detection. By leveraging large datasets of real-life recordings, such as the RLDD (Real-Life Trial Data for Deception Detection) and RODeCAR (Romanian Deva Criminal Investigation Audio Recordings) databases, researchers can train models to identify subtle patterns and cues associated with deceptive speech. These datasets, which are annotated based on actual trial outcomes or criminal investigation results, provide a more realistic and challenging environment for developing and testing lie detection systems compared to datasets based on simulated behavior.

However, the task of detecting deception from audio is inherently complex. Deceptive speech is influenced by a wide range of factors, including the speaker's emotional state, cognitive load, and cultural background, as well as the context in which the deception occurs. Hence, we require a combination of automatically extracted features from spectrograms and algorithmically extracted acoustic and prosodic features. By fusing these different types of features, the system can capture both low-level and high-level characteristics of speech that may be indicative of deception.

In summary, lie detection in audio forensics is an important tool for enhancing the accuracy and reliability of criminal investigations and legal proceedings. The development of robust and accurate deception detection systems has the potential to significantly improve the ability of law enforcement and security agencies to identify deceptive behavior, uncover hidden truths, and make more informed decisions.

1.1.2 Importance of Deepfake Detection

With the rapid advancement of artificial intelligence and deep learning, deepfake technologies have become increasingly sophisticated, enabling the generation of highly realistic synthetic media. Deepfakes are often used for malicious purposes, such as misinformation, identity theft, and fraud, making their detection a critical research area. The threat of deepfakes extends across multiple domains, including security, journalism, social media, and audio authentication systems.

Research communities across the world, particularly in English- and Chinese-speaking regions, have made significant contributions to deepfake detection. The English-speaking research community has focused extensively on datasets like ASVspoof, WaveFake, and In-the-Wild to develop generalizable detection methods. Meanwhile, the Chinese research community has introduced challenges such as ADD (Audio Deepfake Detection) to address language-specific issues and advance the robustness of detection techniques. These efforts contribute to enhancing the ability to detect manipulated content across different linguistic and cultural contexts.

Deepfake manipulation techniques can be broadly classified into several categories: (1) Text-to-Speech (TTS), which generates artificial speech from text input; (2) Voice Conversion (VC), which modifies the identity of a speaker's voice while maintaining the linguistic content; (3) Emotion Fake, where the emotional tone of speech is altered; (4) Scene Fake, in which the acoustic environment is modified; and (5) Partially Fake, where only specific segments of an audio recording are manipulated. These manipulations pose severe risks, as they can be used to deceive individuals and automated systems in authentication and forensic applications.

Several benchmark datasets and competitions have been instrumental in pushing the boundaries of deepfake detection. The ASVspoof challenges have played a pioneering role in developing countermeasures against spoofing attacks in speaker verification systems. The WaveFake dataset focuses on evaluating state-of-the-art TTS systems and their vulnerabilities. The ADD challenge, initiated in China, has further expanded research into low-quality fake detection, partially fake detection, and adversarial attacks. Despite these efforts, existing detection models still struggle with generalization, particularly in real-world and unseen attack scenarios.

Detecting deepfakes presents several challenges, including the need for large-scale, diverse datasets that reflect real-world variations, the difficulty of distinguishing between highly realistic fakes and genuine content, and the computational demands of state-of-the-art detection models. Furthermore, interpretability remains a major concern, as most deep learning models function as black boxes, making it difficult to understand their decision-making processes. Addressing these challenges requires continued interdisciplinary research, collaboration between academia and industry, and the development of robust, explainable, and generalizable detection models.

In summary, deepfake detection is an essential field of study due to its implications for security, trust, and authenticity in digital communications. With ongoing contributions from researchers worldwide, particularly in English- and Chinese-speaking communities, the field is making significant strides. However, overcoming generalization issues, improving dataset diversity, and enhancing model interpretability remain crucial areas for future research.

2 State-of-the-Art (SOTA) methods

2.1 SOTA methods for Lie Detection

To analyze the SOTA models in the domain of lie detection, we focused on reviewing papers [1], [2], [5] and [6].

2.1.1 Paper [1]: Lie detection using speech processing techniques

IOPScience
The primary motivation behind this research was to develop a non-invasive and efficient method for detecting deception using speech signals. Traditional lie detection methods, such as polygraph tests and behavioral analysis, require trained experts and are time-consuming. Moreover, they are invasive (require blood pressure cuffs, rubber tubes, etc.) and people have found ways to cheat the system by taking sedatives. The authors aim to leverage speech processing techniques to identify lies by analyzing psycho-neural changes reflected in speech. This approach is advantageous as it does not require direct contact with the subject and can be automated using machine learning algorithms.

Methodology:

Speech signals are extracted from video clips containing both truthful and deceptive statements. They are cleaned using the STFT-based Noise Reduction process which involves computing the STFT of the noisy signal, applying a thresholding function, and then computing the inverse STFT to obtain the cleaned signal. Mel Frequency Cepstral Coefficients (MFCCs) are extracted from the pre-processed speech signals. The process includes pre-emphasis, frame blocking, windowing, DFT spectrum calculation, Mel spectrum calculation, log transformation, and Discrete Cosine Transform (DCT) to obtain the MFCCs. The mean of the MFCCs across all frames of a speech utterance is calculated to represent the utterance. Next, PCA is used to reduce the dimensionality of the feature vectors, to reduce complexity. The extracted features are fed into an SVM classifier to distinguish between truthful and deceptive speech. The SVM uses Gaussian and polynomial kernels to find the optimal hyperplane for classification.

Dataset Utilized:

The dataset used in this study is called "Real-life trial data," created by Rada Mihalcea, Veronica Perez-Rosas, and colleagues. It consists of 121 video clips from public court trials, including 61 deceptive and 60 truthful statements. The dataset features 21 unique female and 35 unique male speakers, aged between 16 and 60 years. The average duration of the video clips is approximately 28 seconds.

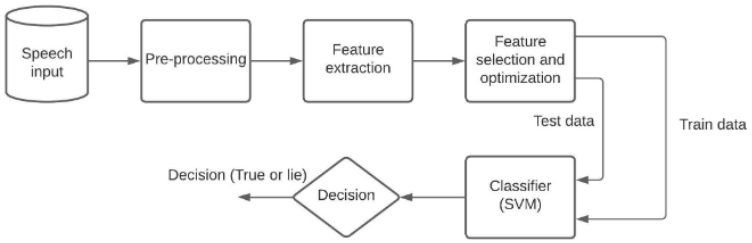


Figure 2: Workflow of Paper [1]

Advantages:

Non-invasive: The method only requires speech signals, making it less intrusive compared to physiological lie detection methods like polygraphs.
Automation: The use of machine learning algorithms allows for automated lie detection, reducing the need for human experts.
Efficiency: The method achieves a relatively high accuracy of 81% using polynomial kernel SVM, making it a promising approach for real-world applications.

Scalability: The method can be easily scaled to analyze large datasets, as it relies on computational techniques rather than manual analysis.

Disadvantages:

Dependence on Speech Quality: The accuracy of the method is highly dependent on the quality of the speech signal. Background noise or poor audio quality can significantly affect performance.
Limited Dataset: The dataset used is relatively small (121 samples), which may limit the generalizability of the results.
Complexity of Feature Extraction: The process of extracting MFCCs and other features is computationally intensive and requires significant preprocessing.
Cultural and Linguistic Bias: The method may not generalize well across different languages and cultures, as speech patterns and deception cues can vary.

Performance:

Polynomial Kernel (Degree 3): Achieved an accuracy of 81.48% with a standard deviation of 4.80.
Gaussian Kernel ($\sigma = 0.02$): Achieved an accuracy of 78.70% with a standard deviation of 6.54.
Polynomial Kernel (Degree 2): Achieved an accuracy of 78.70% with a standard deviation of 1.81.
Gaussian Kernel ($\sigma = 0.01$): Achieved an accuracy of 75.00% with a standard deviation of 8.32.

2.1.2 Paper [2]: Automatic deception detection using multiple speech and language communicative descriptors in dialogs

Cambridge University Press
The main motivation behind this research was to improve automatic deception detection by integrating a comprehensive set of speech and language features. Humans are generally poor at detecting deception, with an average accuracy of only 54%. The authors aimed to leverage prior domain knowledge in deceptive behavior understanding to develop a more effective and robust deception detection framework. The study focused on Mandarin Chinese speakers, as deception behaviors can vary significantly across cultures, and there is a lack of research on deception detection in Eastern cultures.

Methodology:

The proposed method involves a multi-task learning architecture with two main tasks:
Implicature Classification: This task models non-verbal and pragmatic behaviors at the word level to classify four types of implicatures (complications, common knowledge details, self-handicapping strategies, and others).
Deception Detection: This task detects deceptive or truthful statements using acoustic, textual, and conversational temporal dynamics features.

Feature extraction was done by extracting Acoustic-Prosodic Features (ACO) using the openSMILE toolbox, including fundamental frequency, intensity, loudness, MFCCs, and more. Conversational Temporal Dynamics (CTD) features such as silence-duration ratio, utterance-duration ratio, backchannel times, etc., inspired by conversational analysis literature were extracted. Textual Embeddings were extracted using BERT (Bidirectional Encoder Representations from Transformers) for both sentence-level and word-level representations. Non-verbal and pragmatic behaviors are also annotated and included in the word-level embeddings.

The Model Architecture is defined by using Word-Level Modeling which captures the relationship between acoustic-pragmatic behaviors and implicatures using a BLSTM-DNN (Bidirectional Long Short-Term Memory with Deep Neural Network) and a Sentence-Level Modeling which encodes each answering turn into a sentence-level representation and trains additional BLSTM-DNN models for deception detection using different feature sets. Deception Modeling finally concatenates the outputs of all models and fine-tunes the embeddings with three additional dense layers for late fusion.
Dataset Utilized:

The study uses the Daily Deceptive Dialogues corpus of Mandarin (DDDM), which contains 27.2 hours

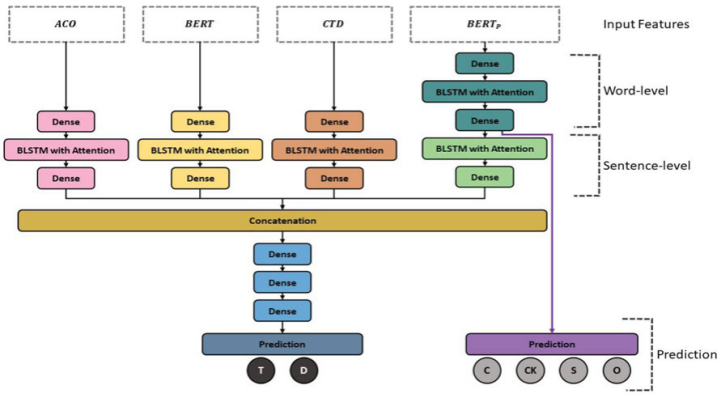


Figure 3: Workflow of Paper [2]

of audio recordings from 96 unique Mandarin native speakers. The dataset includes 7126 sentence-like utterances, which are converted into 2764 question-answering (QA) pair turns. The speakers are asked to discuss three daily life topics, with the goal of deceiving the interlocutors in their answers to one or two of the questions.

Advantages:

Comprehensive Feature Integration: The method integrates a wide range of features, including acoustic, textual, conversational dynamics, and implicatures, providing a holistic approach to deception detection.
Multi-Task Learning: The framework simultaneously models implicature classification and deception detection, leveraging shared information between the tasks.
State-of-the-Art Performance: The proposed model achieves an unweighted average recall (UAR) of 80.61%, outperforming previous methods on the DDDM dataset.
Cultural Relevance: The study focuses on Mandarin Chinese speakers, addressing the gap in deception detection research for Eastern cultures.

Disadvantages:

Complexity: The method involves multiple stages of feature extraction, modeling, and fusion, which can be computationally intensive and time-consuming.
Dependence on Annotation Quality: The accuracy of the model relies heavily on the quality of the annotated features, such as non-verbal behaviors and implicatures, which require manual labeling.
Limited Generalizability: The model is trained and evaluated on a specific dataset (DDDM), and its performance may not generalize well to other languages or cultural contexts.
Imbalanced Data: The dataset has an imbalanced distribution of implicature classes, which may affect the model's performance in classifying rare classes.

Performance:

Overall Unweighted Average Recall (UAR): 80.61%
Deception UAR: 80.34%
Truth UAR: 80.87%
Weighted F1 Score: 79.95%
Macro-Precision: 81.37%

2.1.3 Paper [5]: Novel Lie Speech Classification by using Voice Stress

SciTePress
The motivation behind this research was to develop an efficient and reliable method for lie detection using voice stress analysis. Traditional methods like polygraphs are time-consuming, require physical presence, and lack conclusive proof of accuracy. Voice stress analysis (VSA) offers a non-invasive alternative by

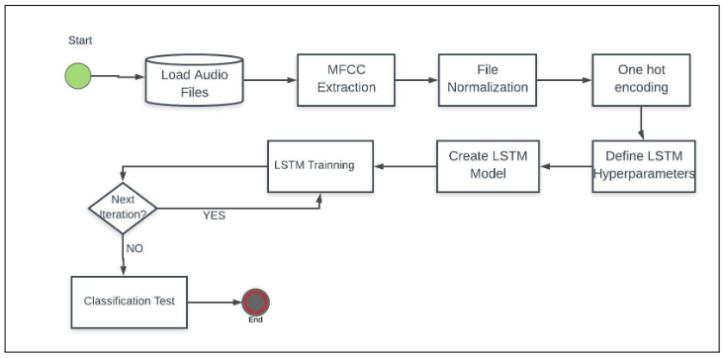


Figure 4: Workflow of Paper [5]

detecting stress in a person's voice, which can indicate deception. The authors aim to implement a neural network to classify speech as truthful or deceptive based on voice stress, leveraging the capabilities of Long Short-Term Memory (LSTM) networks to capture temporal patterns in speech.

Methodology:

Audio recordings were collected from 10 male subjects in Brazilian Portuguese. Each subject was interviewed twice—once lying and once telling the truth—resulting in 220 audio files (110 truthful and 110 deceptive). Then, the audio files were edited using Audacity to remove silence and segment the answers. The dataset was split into 180 files for training and 40 for testing. Mel-Frequency Cepstral Coefficients (MFCC) were extracted from the audio files using the Librosa library. The number of MFCCs extracted varied (13, 20, or 40) to test different scenarios. The variable-length MFCC sequences were normalized using padding to ensure uniform input size for the neural network. The data was labeled using one-hot encoding. An LSTM neural network was implemented with multiple hidden layers and cells. The network was trained using the Adam optimizer and cross-entropy loss function. The model was trained on the extracted MFCC features to classify speech as truthful or deceptive. Different hyperparameters (e.g., number of layers, cells, MFCCs, batch size, learning rate) were tested to optimize the model's performance. The best-performing model achieved an accuracy of 72.5%.

Dataset Utilized:

The dataset used in this study was created by the authors through interviews with 10 male subjects. Each subject was interviewed twice—once lying and once telling the truth—resulting in 220 audio files (110 truthful and 110 deceptive). The interviews were conducted in Brazilian Portuguese, and the audio files were preprocessed to remove silence and segment the answers. The dataset was split into 180 files for training and 40 for testing.

Advantages:

Non-Invasive: The method relies on voice analysis, which does not require physical contact with the subject.
Automation: The use of neural networks allows for automated lie detection, reducing the need for human intervention.
Scalability: The method can be applied to larger datasets and potentially extended to other languages and accents.
State-of-the-Art Performance: The model achieved an accuracy of 72.5%, which is comparable to or better than similar works in the literature.

Disadvantages:

Limited Dataset: The dataset is relatively small (220 audio files) and limited to male subjects speaking Brazilian Portuguese, which may affect generalizability.
Dependence on MFCC Features: The model's performance heavily relies on MFCC features, which may

not capture all aspects of voice stress.
Computational Complexity: Training LSTM networks can be computationally intensive, especially with larger datasets.
Overfitting Risk: The model showed signs of overfitting when the number of iterations exceeded 200, leading to poor performance on new data.

Performance:

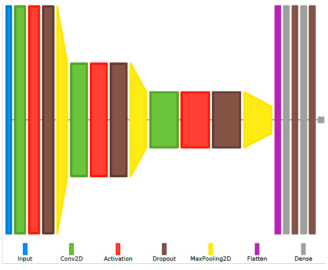
Model Accuracy: 72.5%
False Positives and Negatives: 6 false positives, 5 false negatives

2.1.4 Paper [6]: Using voice activity detection and deep neural networks with hybrid speech feature extraction for deceptive speech detection

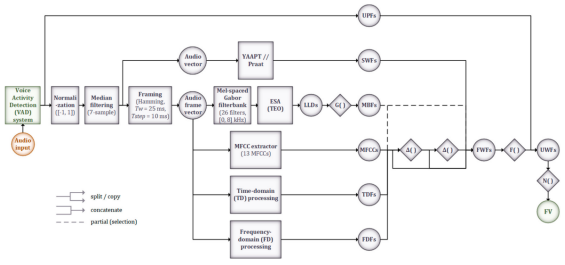
MDPI
The motivation behind this research was to improve the accuracy and robustness of Voice Activity Detection (VAD) and Deceptive Speech Detection (DSD) systems using advanced deep learning techniques. So this paper focussed on these two tasks. VAD is important for identifying speech segments in audio signals, which is a very basic step in various speech processing applications. The authors aimed to enhance VAD performance by exploring different deep neural network (DNN) architectures and postprocessing techniques. Additionally, they extend the application of the VAD system to the more task of DSD, which is particularly relevant in forensic and law enforcement scenarios where detecting deceptive speech can provide valuable insights during interviews or interrogations. The authors also aim to address the limitations of existing DSD datasets, which often rely on simulated behavior, by using more realistic datasets like RLDD and RODeCAR.

Methodology:

The authors proposed a VAD system based on deep neural networks (DNNs), including multilayer perceptrons (MLPs), recurrent neural networks (RNNs), and convolutional neural networks (CNNs). They investigatd the performance of these networks using different feature extraction methods, such as algorithmically extracted features (e.g., energy, zero-crossing rate, MFCCs) and automatically extracted features from raw time-domain samples or frequency-domain representations (e.g., DFT, spectrograms). The CNN-based VAD system, showed the best performance. For the DSD task, the authors proposed a hybrid CNN-MLP network that uses a fusion of automatically and algorithmically extracted speech features. The system first uses the VAD system to detect and split the input audio into utterances. Then, it extracts a comprehensive set of features, including acoustic, prosodic, spectral, and cepstral features, and applies high-level statistical functions to these features. The hybrid CNN-MLP network combines the automatically extracted feature maps from the CNN with the best-performing subset of algorithmically extracted features, selected using a Kolmogorov-Smirnov (KS) feature ranking algorithm. The final system is trained and tested on the RLDD and RODeCAR datasets, which contain real-life trial and criminal investigation recordings, respectively.



(a) CNN-MLP Model



(b) Preprocessing for DSD

Figure 5: Workflow of Paper [6]

Dataset Utilized:

CENSREC-1-C and TIMIT were used for VAD. RLDD (Real-Life Trial Data for Deception Detection) was used containing 121 audiovisual recordings from real-life trials, annotated as truthful or deceptive based on trial outcomes. RODeCAR (Romanian Deva Criminal Investigation Audio Recordings) was also utilized; a dataset developed by the authors, containing audio recordings from criminal investigations, annotated as truthful or deceptive based on case outcomes.

Advantages:

Improved VAD Performance: The CNN-based VAD system achieves high accuracy (up to 99.13% on CENSREC-1-C and 97.60% on TIMIT) by leveraging deep learning and postprocessing techniques.
Realistic DSD Datasets: The use of RLDD and RODeCAR datasets, which contain real-life recordings, provides a more realistic and challenging environment for DSD compared to simulated datasets.
Hybrid Feature Extraction: The combination of automatically and algorithmically extracted features in the hybrid CNN-MLP network improves DSD performance by capturing both low-level and high-level speech characteristics.
Utterance-Level DSD: The proposed system focuses on detecting deception at the utterance level, which is more relevant for forensic and law enforcement applications than speaker-level or recording-level approaches.

Disadvantages:

Computational Complexity: The hybrid CNN-MLP network, while effective, is computationally expensive due to the large number of parameters (over 20 million) and the need for zero-padding to handle variable-length utterances.
Limited Generalization: The performance of the VAD system drops in low signal-to-noise ratio (SNR) conditions, which limits its applicability in the real world because real world is noisy.
Imbalanced Datasets: The RODeCAR dataset is slightly imbalanced in terms of class distribution (53.5% truthful vs. 46.5% deceptive), which may affect the model's performance and require class weighting techniques to mitigate bias.

Performance:

RLDD: 63.7% unweighted accuracy (UA) with the hybrid CNN-MLP network.
RODeCAR: 62.4% unweighted accuracy (UA) with the hybrid CNN-MLP network.

2.2 SOTA methods for Deepfake Detection

To analyze the SOTA models in the domain of deepfake detection, we focused on reviewing papers [3].

2.2.1 Paper [3]: Advanced RawNet2 with Attention-based Channel Masking for Synthetic Speech Detection

Interspeech
The motivation behind this research was to improve the generalization ability of synthetic speech detection systems, particularly against unseen spoofing attacks. Automatic speaker verification (ASV) systems are vulnerable to various spoofing attacks, including text-to-speech (TTS) and voice conversion (VC) algorithms. The authors aimed to enhance the robustness of the RawNet2 model by introducing an attention-based channel masking (ACM) block, which simulates the human auditory system's ability to focus on discriminative features in speech. This approach is designed to improve the model's ability to detect synthetic speech across different datasets and unseen attack algorithms.

Methodology:

The authors propose an advanced version of RawNet2 (ARawNet2) by incorporating an attention-based channel masking (ACM) block. The ACM block consists of three key components:
1) **Squeeze-and-Excitation (SE) Block**: This component recalibrates the channel-wise correlations of high-level global acoustic feature maps, enhancing the model's ability to focus on important features.
2) **Channel Masking**: This component randomly masks partial features during training to enhance the model's robustness against unseen attacks.
3) **Global-Local Feature Aggregation**: This component combines global and local feature maps to fully exploit their complementarity, improving the model's ability to capture both broad and fine-grained features.
The ARawNet2 model replaces the 1-dimensional convolution layers in the original RawNet2 with 2-dimensional convolution layers and removes the feature map scaling (FMS) blocks. The model is trained using raw waveforms as input, and the ACM block is applied during training to improve the model's generalization ability. The model is evaluated on both the ASVspoof 2019 and ASVspoof 2021 datasets, which include a variety of unseen spoofing attacks.

Dataset Utilized:

The study uses the ASVspoof 2019 and ASVspoof 2021 datasets. The ASVspoof 2019 dataset includes logical access (LA) tasks with 6 known spoofing algorithms in the training set and 13 unseen algorithms in the evaluation set. The ASVspoof 2021 dataset includes both LA and deepfake (DF) tasks, with the LA task containing the same spoofing algorithms as ASVspoof 2019 but degraded by unknown transmission channels. The DF task includes over 100 undisclosed spoofing algorithms, making it a more challenging dataset for evaluating the model's generalization ability.

Advantages:

Improved Generalization: The ACM block enhances the model's ability to detect synthetic speech across different datasets and unseen attack algorithms.
Robustness: The channel masking component improves the model's robustness by simulating the human auditory system's ability to focus on discriminative features.
State-of-the-Art Performance: The ARawNet2 model achieves significant improvements over the baseline RawNet2, with relative EER reductions of 12.00% and 14.97% on the ASVspoof 2021 LA and DF tasks, respectively.
End-to-End Learning: The model directly processes raw waveforms, eliminating the need for handcrafted features and allowing for end-to-end learning.

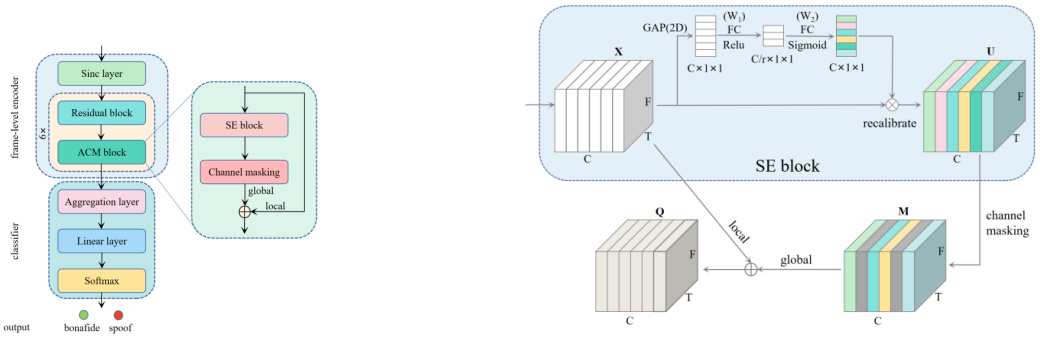
Disadvantages:

Computational Complexity: The addition of the ACM block and 2-dimensional convolution layers increases the computational complexity of the model.
Dependence on Dataset Quality: The model's performance is highly dependent on the quality and

diversity of the training data, particularly in handling unseen spoofing attacks.
Limited to Audio Data: The model is designed specifically for synthetic speech detection and may not generalize well to other types of deepfake content, such as video or image-based deepfakes.

Performance:

ASVspoof 2019 LA Task: The ARawNet2 achieves an EER of 4.61%.
ASVspoof 2021 LA Task: The ARawNet2 achieves an EER of 8.36%, a 12.00% relative reduction over the baseline RawNet2.
ASVspoof 2021 DF Task: The ARawNet2 achieves an EER of 19.03%, a 14.97% relative reduction over the baseline RawNet2.



(a) ARawNet2 Architecture

(b) attention-based channel masking (ACM) block

Figure 6: Paper [3]

2.2.2 Paper [7]: AASIST: Audio Anti-Spoofing Using Integrated Spectro-Temporal Graph Attention Networks

IEEE Xplore
The motivation behind this research was to develop a more efficient and unified approach to detecting audio spoofing attacks. Previous systems typically relied on computationally intensive ensemble methods, where each subsystem was specialized for detecting specific types of spoofing artifacts. The authors recognized that spoofing artifacts could be present in both spectral and temporal domains and sought to create a single, efficient system capable of detecting various spoofing attacks without requiring score-level ensembles. This would make the system more practical for real-world applications while potentially improving detection accuracy.

Methodology:

The authors propose AASIST, a novel end-to-end spoofing detection system based on graph neural networks. The system consists of several key components:
1) **RawNet2-based Encoder**: Processes raw waveform inputs directly, extracting high-level representations through six residual blocks with pre-activation. The encoder interprets the sinc-convolution layer output as a 2D image with a single channel.
2) **Heterogeneous Stacking Graph Attention Layer (HS-GAL)**: Implements a heterogeneous attention mechanism to model spectral and temporal graphs simultaneously. Introduces a stack node to accumulate heterogeneous information. Uses three different projection vectors for attention weights between nodes in the spectral domain, nodes between spectral and temporal domains, and nodes in the temporal domain.
3) **Max Graph Operation (MGO)**: Employs two parallel branches, each containing two HS-GAL layers. Applies an element-wise maximum operation to combine branch outputs. Includes graph pooling layers after each HS-GAL. Shares stack nodes between sequential HS-GALs in each branch.
4) **Modified Readout Scheme**: Performs node-wise maximum and average operations. Concatenates average and maximum nodes with the stack node. Feeds into the final output layer with two nodes.
5) The authors also developed **AASIST-L**, a lightweight variant with only 85K parameters, optimized

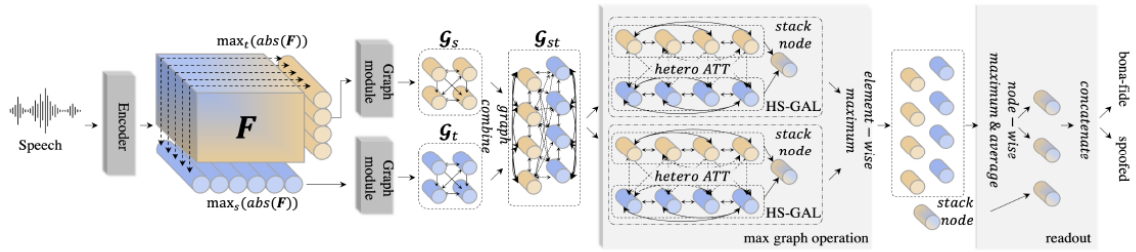


Figure 7: Overall framework of the proposed AASIST. [7]

using a population-based training algorithm.

Dataset Utilized:

The research used the ASVspoof 2019 logical access (LA) dataset, which consists of three subsets: the training and development sets containing attacks from six spoofing algorithms (A01-A06), the evaluation set with attacks from thirteen algorithms (A07-A19), and a dataset specifically focused on voice conversion and text-to-speech-based attacks. All experiments were conducted with three different random seeds to ensure reliability.

Advantages:

State-of-the-art performance with a 20% relative improvement over previous systems. Single unified system instead of ensemble approaches. End-to-end architecture eliminating the need for hand-crafted features. Efficient processing of both spectral and temporal information. Lightweight variant (AASIST-L) maintains competitive performance. Robust performance across different types of spoofing attacks. Effective handling of heterogeneous information through the stack node.

Disadvantages

Complex architecture requiring careful implementation. Performance can vary significantly with different random seeds. Requires raw waveform inputs, which may increase computational load. Limited to audio spoofing detection only. May require specialized hardware for optimal performance. Training process requires multiple runs with different seeds for reliability.

Metrics:

Evaluation Metrics: Minimum tandem detection cost function (min t-DCF) and Equal Error Rate (EER).

Performance:

AASIST (full model) achieved a min t-DCF of 0.0275 (best seed) and an EER of 0.83% (best seed).
AASIST-L (lightweight) obtained a min t-DCF of 0.0309 and an EER of 0.99%.
Individual Attack Performance: Improved performance on 9 out of 13 attack conditions, with up to a 35% relative improvement on specific attacks (e.g., A15).

2.2.3 Features Used in Deepfake Detection Models

The effectiveness of deepfake detection models largely depends on the features they use to differentiate between real and synthetic audio. Recent studies have explored a wide range of features, which can be broadly categorized into spectral, phase-based, prosodic, and deep learning-based features.

Spectral features include short-term and long-term spectral representations. Short-term spectral features, such as Mel-Frequency Cepstral Coefficients (MFCC), Linear Frequency Cepstral Coefficients (LFCC), and Constant-Q Cepstral Coefficients (CQCC), are widely used due to their ability to capture essential

Category		Description	Methods
Hand -crafted spectral features	Magnitude/ power-based spectral coefficients	Short-term: Computing a frequency domain transform on each temporal window of the audio signal to enhance time resolution at lower frequencies.	MFCC [50], IMFCC [28], LFCC [4], RFCC [85], LPCC [66], SSFC [188], SCFC [188], SCMC [188], Gaussian-IMFCC [163], Multi-resolution STFT [137], GTCC [26], WPT [69], STFT [126], PLP [124], DTW [80], Spec-Env [100], Spec-Contrast [100]
		Long-term: Longer temporal window.	CQCC [207], CQBC [127], eCQCC [259], CLBC [258], L-VQT [119], SCC [200], CQMOC [261], CFCC [177], Global M [67]
	Phase-based spectral coefficients	Working effectively as a complement to the magnitude features. However, it may not be helpful for unknown attacks, especially for VC attacks.	MGDCC [245], APGDF [174], Relative phase [221], Quadrature phase [83], MMPS [261], IF [177]
	Bispectrum	Describing the higher-order spectral correlation in the Fourier domain; is useful for the known attacks.	Statistics of bispectral correlations [3]
	Image-like	Propagating in time to show the variations in frequencies and intensities of an audio signal in a 2D feature, which are interpreted as images.	STFT-spec [16], Mel-spec [182], CQT-spec [1], E-Spect [246], C-CQT spectrogram [170], LBP [77], MLTP [89], SDC [101]
DL features	Filter-learning features	Utilizing DL techniques to construct learnable filterbanks or approximate the standard filtering process.	nnAudio [36], DNN-FBCC [270], FastAudio [65], SincNet [273], TD-FBanks [272], LEAF [271]
	Supervised embeddings	Constructing deep embeddings using DL models through supervised training	CNN [244], ResNet [193], X-vector [29], auto-encoder [15], Bi-LSTM [101], U-net [30]
	Pre-trained embeddings	Utilizing SSL models or other DL models pre-trained by external large datasets to extract latent representations of the raw audio waveform.	wav2vec2.0 [227], WavLM [298], HuBERT [123], TDNN [154], HiFi-GAN [56], and ImageNet [144]
Analysis -oriented features	Prosody/ semantic features	Focusing on the prosody and emotion of the speech sounds, which works effectively on TTS-based spoofed audio, not the VC.	Vocal tract estimation [20], shimmer [120], phoneme duration [218], pronunciation [218], prosody [10], emotion [43], VOT [53], coarticulation [53]
	The impact of silence	Contributing effectively to the current anti-spoofing detection models.	Silence portion [169], BTS-Encoder [57]
	Frequency sub-band feature	Focusing on one or more specific portions of the frequency band, rather than the entire frequency range.	F0 [60], 0-4kHz [291], 4-8kHz [168]
	Other possible directons	Including recent attempts on the development of anti-spoofing features.	Varied input length [226], energy loss [52], face embedding [252], dual channel stereo feature [135], Compressed coding metadata [254]

Figure 8: The Categorization of Feature Extraction Methods [4]

speech characteristics. Long-term spectral features, such as modulation spectrum and frequency domain linear prediction, help in capturing temporal dependencies and improving robustness against unseen deepfake attacks.

Phase-based features exploit the inconsistencies introduced by deepfake generation models. These include group delay, modified group delay, and relative phase shift features, which analyze the phase components of speech signals to identify artifacts not captured by magnitude-based features.

Prosodic features focus on aspects such as pitch, intonation, and speaking rate. Since deepfake audio synthesis often struggles to maintain natural prosody, these features help in detecting unnatural variations in synthesized speech.

Deep learning-based features have gained prominence with the advancement of neural networks. Models such as Wav2Vec, HuBERT, and XLS-R extract self-supervised embeddings from raw audio signals, improving generalization across different types of deepfake manipulations. Other deep features are derived using convolutional neural networks (CNNs) and recurrent neural networks (RNNs) trained specifically for deepfake detection tasks.

A combination of these features is often used in state-of-the-art deepfake detection systems to improve accuracy and robustness against evolving attacks.

Publication		Data augmentation	Feature	Classifier	Loss funcion	# Params	ASVspoof			Access-ibility
							19-LA	21-LA	21-DF	
[291]	INTERSPEECH'21	w/o	Mel-Spec on 0-4kHz	SE-ResNet-18	AM-Softmax	1.1M	1.14	-	-	No
[204]	INTERSPEECH'21	channel masking	RawNet2*	GAT	CE	440K	1.06	6.92	-	Yes[1]
[71]	INTERSPEECH'21	channel masking	SincNet	Raw PC-DARTS	MSE	24.4M	1.77	6.43	-	Yes[2]
[87]	SPL'21	mix-up	E2E: CNN→ResNet→MLP		CE	350M	1.64	-	-	Yes[3]
[65]	ICASSP'22	w/o	FastAudio	ECAPA-TDNN	CE	Unknown	1.54	-	-	Yes[4]
[119]	DSP'22	w/o	L-VQT	DenseNet	CE	338K	2.19	-	-	No
[212]	ICPR'22	w/o	RawNet2+(CQT →ECAPA-TDNN)	CNN→MLP	CE	7.19M	1.11	-	-	Yes[5]
[114]	INTERSPEECH'22	w/o	wav2vec2.0-XLSR	MLP	CE	317M	0.31	-	-	No
[59]	INTERSPEECH'22	w/o	wav2vec2.0-960	MLP	CE	Unknown	0.40	-	-	No
[39]	INTERSPEECH'22	frequency masking	CQT-Spec	LCNN	CE	135K	1.35	-	-	No
[227]	ODYSSEY'22	w/o	wav2vec2.0-XLSR	Bi-LSTM →MLP	CE	317M	1.28	6.53	4.75	No
[209]	ODYSSEY'22	RawBoost	wav2vec2.0-XLSR	AASIST	CE	Unknown	-	**0.82**	**2.85**	Yes[6]
[144]	DDAM'22	RawBoost	ImageNet + Jitter + Shimmer	MLP	AM-Softmax	Unknown	0.87	10.06	27.08	No
[217]	DDAM'22	w/o	wav2vec2.0-Large	DARTS	Unknown	Unknown	1.08	-	7.89	No
[92]	ICASSP'22	w/o	RawNet2	GAT	CE	297K	0.83	5.59	-	Yes[7]
[117]	SPL'22	w/o	LFCC	OCT	Focal loss	250K	1.06	-	-	No
[132]	APSIPA'22	adding noise, RIRs	wav2vec2.0	LCNN	CE	Unknown	0.24	-	-	No
[100]	MAD'23	w/o	Mel-spec + Spec-Env + Spec-Contrast	Transformer →CNN	CE	603K	0.95	-	-	No
[218]	INTERSPEECH'23	w/o	Duration + pronunciation + wav2vec2.0-XLSR	LCNN →Bi-LSTM →MLP	CE	Unknown	1.58	-	-	No
[151]	SPL'23	w/o	(LFCC →ResNet) + (CQT-Spec →ResNet)	GRL →MLP	CE	Unknown	0.80	-	-	Yes[8]
[139]	ICASSP'23	w/o	RawNet2	Rawformer	CE	370K	0.59	4.98	4.53	Yes[9]
[158]	ICASSP'23	FIR filter	wav2vec2.0-XLSR	MLP	OC-Softmax	300M	-	3.54	6.18	No
[32]	ICASSP'23	time & frequency masking	LFB-Spec	GCN	CE	Unknown	0.58	-	-	No
[276]	ALGORITHM'23	RawRoost	wav2vec 2.0	Transformer	CE	Unknown	-	1.18	4.72	No
[101]	ICASSP'24	FIR filter, codec, noises, shift	SDC + Bi-LSTM	Auto-encoder →SE-ResNeXT	CE	Unknown	**0.22**	3.50	3.41	No

Figure 9: The Cross-dataset Performance of Single-System State-of-the-art Models. All Models are trained or fine-tuned on ASVspoof2019-LA Training and Development Set, and evaluated on ASVspoof2019-LA Evaluation Set and In-The-Wild Dataset. [4]

2.2.4 Results of Spoof Detection Systems

Recent evaluations of spoof detection systems, particularly in the ASVspoof challenge, demonstrate significant progress. Various deep learning architectures, including ResNet, LCNN, Graph Neural Networks (GNNs), and self-supervised learning models like Wav2Vec and HuBERT, have been employed to improve detection accuracy. The integration of attention mechanisms and differentiable architecture search (DARTS) has further optimized performance across different spoofing conditions.

2.2.5 Evaluation Metrics for Spoof Detection

Evaluating the effectiveness of spoof detection models requires robust metrics that measure accuracy, generalizability, and reliability. The most commonly used evaluation metrics include:

1. **Equal Error Rate (EER)**: The point where false acceptance and false rejection rates are equal, commonly used in ASVspoof challenges.
2. **F1-score**: Balances precision and recall, especially useful in unbalanced datasets.
3. **Accuracy**: Measures the proportion of correct predictions, but can be biased in unbalanced datasets.
4. **Tandem Detection Cost Function (t-DCF)**: Assesses the cost of detection errors in conjunction with speaker verification systems.
5. **Range-based EER**: Specifically designed for partially spoofed detection, measuring misclassified segment durations at fine resolutions.

3 Our Methods

3.1 Lie Detection

3.1.1 Methodology

The workflow mentioned in paper [1] was referred to. We follow the same structured approach combining audio signal processing, feature extraction, and machine learning classification. Additionally, there are annotations from the All_Gestures_Deceptive and Truthful sheet (non-verbal gestures, facial expressions, and prosodic cues). We use these features to train the classifiers too. The samples are read, and speech clips are categorized as truthful or deceptive based on their filenames. Each audio file is loaded, converted to mono (16 kHz sample rate), and normalized to ensure consistent amplitude levels. Noise reduction is then applied using Short-Time Fourier Transform (STFT), where low-magnitude noise components are filtered out before reconstructing the cleaned signal with inverse STFT (ISTFT). Next, feature extraction is performed using Mel-Frequency Cepstral Coefficients (MFCCs), which capture the spectral characteristics of speech. The first 13 MFCCs are computed and averaged over time, forming a fixed-length feature vector. The MFCCs and the additional features were concatenated together. To reduce feature dimensionality, Principal Component Analysis (PCA) is applied, retaining 10 principal components that preserve essential variations while removing noise. The dataset is then split into training (80%) and testing (20%) sets. Two classification models are trained: Support Vector Machines (SVMs) with an RBF kernel and Random Forests with 100 trees. The models learn to differentiate truthful vs. deceptive speech using the reduced feature set. Finally, the trained models make predictions on the test set, and their performance is evaluated using accuracy, precision, recall, and F1-score.

3.1.2 Dataset Description

The dataset we used was the Real-Life Deception Dataset from the paper [3]. It consists of 121 videos having 61 deceptive and 60 truthful trial clips. The average video length is 27.7 seconds and 28.3 seconds for the deceptive and truthful clips, respectively. The data consists of 21 unique female and 35 unique male speakers, with their ages approximately ranging between 16 and 60 years.

Table 1: Data Shape

Type	Shape
Train data	(96, 10)
Train labels	(96, 1)
Test data	(25, 10)
Test labels	(25, 1)

On concatenating the MFCC and annotation features, there were in total 52 features (13 MFCCs + 39 annotation features). Since PCA with n_components=10 was applied, the training and testing samples finally had 10 features.

3.1.3 Results

SVM Classification Accuracy: 68.00%

	precision	recall	f1-score	support
0	0.80	0.57	0.67	14
1	0.60	0.82	0.69	11
accuracy			0.68	25
macro avg	0.70	0.69	0.68	25
weighted avg	0.71	0.68	0.68	25

(a) SVM results on RLDD

Random Forest Classification Accuracy: 76.00%

	precision	recall	f1-score	support
0	0.83	0.71	0.77	14
1	0.69	0.82	0.75	11
accuracy			0.76	25
macro avg	0.76	0.77	0.76	25
weighted avg	0.77	0.76	0.76	25

(b) RF results on RLDD

Figure 10: Results of classifiers on the RLDD dataset

The Support Vector Machine (SVM) reported an accuracy of 68% while the Random Forest (RF) reported an accuracy of 75%.

(a) SVM confusion matrix on RLDD

(b) RF confusion matrix on RLDD

Figure 11: Confusion matrices of classifiers on the RLDD dataset

4 Thoughts and Future Research Directions

4.1 Lie Detection

Stylometry is the quantitative analysis of writing style. Stylometric features are typically used to analyze and classify authorship or identify patterns in writing, and can also be applied to spoken language (like in lie detection) to analyze a person's speech patterns. Coupled with all other features, these may enhance the performance of various models.

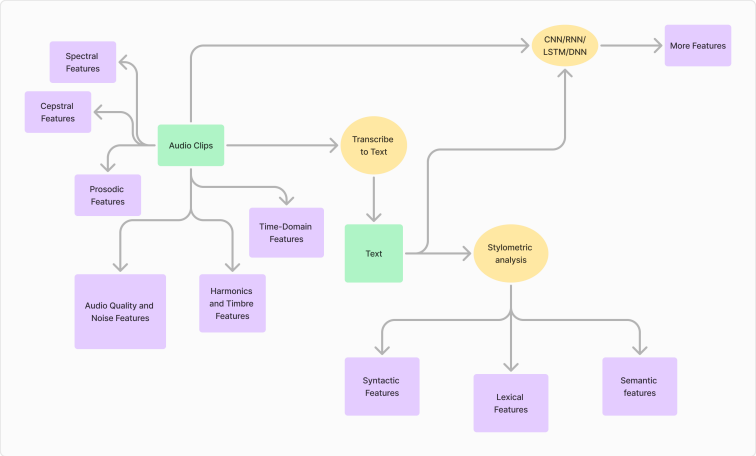


Figure 12: Feature Extraction for Lie Detection

4.2 DeepFake Detection

Future Research Priorities:

- **Enhanced Generalization** should be pursued through multiple approaches, including advanced adversarial training techniques, comprehensive cross-dataset evaluation methodologies, and the development of more diverse and realistic training datasets that better reflect real-world conditions.
- **Model Optimization** efforts need to focus on creating more efficient detection systems that maintain high accuracy while reducing computational requirements, enabling broader deployment across different devices and platforms.

- **Explainable AI Development** must be prioritized to create detection systems that can provide clear, understandable explanations for their decisions, thereby increasing trust and enabling more effective system validation and improvement.

References

[1] EP Fathima Bareeda, BS Shajee Mohan, and KV Ahammed Muneer. Lie detection using speech processing techniques. In *Journal of Physics: Conference Series*, volume 1921, page 012028. IOP Publishing, 2021.

[2] Huang-Cheng Chou, Yi-Wen Liu, and Chi-Chun Lee. Automatic deception detection using multiple speech and language communicative descriptors in dialogs. *APSIPA Transactions on Signal and Information Processing*, 10:e5, 2021.

[3] Jing Li, Yanhua Long, Yijie Li, and Dongxing Xu. Advanced rawnet2 with attention-based channel masking for synthetic speech detection. In *Interspeech 2023*, pages 2788–2792, 2023. doi:10.21437/Interspeech.2023-542.

[4] Menglu Li, Yasaman Ahmadiadli, and Xiao-Ping Zhang. Audio anti-spoofing detection: A survey, 2024. URL https://arxiv.org/abs/2404.13914.

[5] Felipe Mateus Marcolla, Rafael de Santiago, and Rudimar LS Dazzi. Novel lie speech classification by using voice stress. In *ICAART (2)*, pages 742–749, 2020.

[6] Serban Mihalache and Dragos Burileanu. Using voice activity detection and deep neural networks with hybrid speech feature extraction for deceptive speech detection. *Sensors*, 22(3):1228, 2022.

[7] Jee weon Jung, Hee-Soo Heo, Hemlata Tak, Hye jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks, 2021. URL https://arxiv.org/abs/2110.01200.