ISSN 2392-0300 „Civitas et Lex" 2024 / 1(41)
DOI 10.31648/cetl.9393

NAUKI O BEZPIECZEŃSTWIE / SECURITY STUDIES

Karol Jędrasiak

AUDIO STREAM ANALYSIS FOR DEEP FAKE THREAT IDENTIFICATION*

The rapid advancements in deep learning technologies, particularly in text-to-speech (TTS)[1] and voice conversion (VC)[2], have revolutionized the generation of human-like natural speech. These technologies have found widespread applications, enhancing user experiences in various domains such as car navigation systems, e-book readers, and intelligent robotics. However, the flip side of these innovations is their potential misuse in the form of deepfake technologies[3]. Originally recognized for their ability to seamlessly swap faces in videos, deepfakes have evolved to include sophisticated audio manipulations, creating challenges in distinguishing between authentic and falsified ecordings. This progression has profound implications for social security and political stability.

One notable incident illustrating the potential danger of audio deepfakes involved an employee who was tricked into transferring a significant amount of money, amounting to USD 243.000, due to a fraudulent voice mimicking his superior's[4]. This episode starkly highlights the emerging risks associated with deepfake technologies, where artificial intelligence (AI) can be exploited for malicious purposes, such as financial fraud and identity theft.

In response to these growing threats, the field of audio deepfake detection has emerged as a critical area of research[5]. This field aims to differentiate genuine utterances from falsified ones using machine learning techniques. Predominantly,

Karol Jędrasiak – Akademia WSB w Dąbrowie Górniczej, ORCID: https://orcid.org/0000-0002-2254-1030, e-mail: kjedrasiak@wsb.edu.pl

***Acknowledgment**: The work was cofinanced as part of the implementation of the project Interdisciplinary research projects of WSB University academics.

[1] T. Dutoit, *High-quality text-to-speech synthesis: An overview*, "Journal Of Electrical And Electronics Engineering Australia" 1997, no 17(1), pp. 25–36.

[2] A.F. Machado, M.G.D. Queiroz, *Voice conversion: A critical survey*, "In Proceedings" 2010.

[3] P. Swathi, S. Sk, *Deepfake creation and detection: A survey*, "2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)" 2021, pp. 584–588.

[4] Internet source: <https://cyware.com/news/fraudsters-make-away-with-243000-by-impersonating-company-ceo-in-new-voice-phishing-attack-c8dc188d>, accessed: 06.11.2023.

[5] M.S. Rana, M.N. Nobi, B. Murali, A.H. Sung, *Deepfake detection: A systematic literature review*, "IEEE access" 2022, no. 10, pp. 25494–25513.

the research has been divided into two approaches: the pipeline method, which combines frontend feature extraction with backend classification, and the more recent end-to-end methods that optimize these processes directly on raw audio data. Despite significant advancements, the research remains fragmented and predominantly focused on protecting automatic speaker verification (ASV) systems[6].

The urgency in addressing these threats is underscored by the ease of access to deepfake generation technologies and their rapidly improving sophistication. These technologies have the potential to be used for benign purposes, such as in the entertainment industry, but also for more nefarious activities, such as spreading misinformation, influencing political narratives, or compromising security systems.

Problem statement

This article addresses the critical need of distinguishing fake from real recordings by introducing new method of audio deepfake detection. I focus on identifying common discriminative audio features relevant to deepfake detection and the computational methodologies for developing effective, generalized automatic systems. I started by exploring microfeatures like Voicing Onset Time (VOT) and coarticulation[7], examining their potential in distinguishing between authentic and synthesized speech. Finally, I delved into the potential of vocal emotion analysis (VEA) and sentiment analysis in enhancing deepfake detection. Through extensive experimentation on various datasets, I aimed to provide a balanced audio deepfake detection technology. The work seeks to contribute to the field of digital communication security in an era increasingly dominated by synthetic audio. By presenting a detailed and comprehensive analysis of audio deepfake detection techniques, it was aimed to pave the way for innovative and effective strategies to mitigate the risks posed by deepfakes, thereby ensuring the integrity and security of digital communications globally.

Related work

The domain of audio deepfake detection[8] has experienced significant growth, driven by advancements in deepfake technologies, competitions, datasets, evaluation metrics, and detection methods. Audio deepfakes, essentially audio

[6] A.E. Rosenberg, *Automatic speaker verification: A review*, "Proceedings of the IEEE" 1976, no. 64(4), pp. 475–487.

[7] A.S. Abramson, D.H. Whalen, *Voice Onset Time (VOT) at 50: Theoretical and practical issues in measuring voicing distinctions*, "Journal of phonetics" 2017, no. 63, pp. 75–86.

[8] Z. Almutairi, H. Elgibreen, *A review of modern audio deepfake detection methods: challenges and future directions*, "Algorithms" 2022, no. 15(5), p. 155.

recordings in which key attributes have been artificially manipulated using AI technologies, present a unique challenge due to their retention of perceived naturalness. Studies in this field have primarily focused on five types of deepfake audio: text-to-speech (TTS), voice conversion (VC), emotion fake, scene fake, and partially fake.

- **Text-to-Speech (TTS)**. TTS[9], also known as speech synthesis, aims to create intelligible and natural speech from any given text using machine learning models. Recent developments in deep neural networks have enabled TTS systems to generate increasingly realistic and human-like speech. TTS systems typically encompass text analysis and speech waveform generation modules, with two major methods in speech waveform generation being concatenative and statistical parametric TTS. The latter often involves an acoustic model and a vocoder. Recent advancements include end-to-end models like Variational Inference with adversarial learning for end-to-end Text-to Speech (VITS) and FastDiff-TTS, which produce high-quality audio.
- **Voice Conversion (VC)**. VC[10] focuses on digitally cloning a person's voice to alter the timbre and prosody of a speaker's speech to match that of another while keeping the content unchanged. VC systems process natural utterances from a given speaker, with three main technological approaches: statistical parametric, frequency warping, and unit-selection. Recent years have seen the proposal of end-to-end VC models aimed at mimicking a person's voice characteristics.
- **Emotion Fake**. The technique known as Emotion Fake intricately alters the emotional undercurrents of spoken language[11]. This method intricately tweaks the perceived emotional expression within a speech sample without disturbing the consistency of the speaker's identity or the verbal content itself. For example, an originally joyful message can be transformed to convey sadness, fundamentally shifting the listener's perception while maintaining the message's verbal integrity. The intricacies of vocal emotion can be complex, involving nuances such as tone, pitch, and rhythm, which this method manipulates. Techniques for achieving such emotional manipulation vary, including those that rely on parallel datasets where aligned pairs of different emotional utterances by the same speaker are available and those that operate on non-parallel datasets, which do not require direct pairing and thus offer a broader application potential.
- **Scene Fake**. Scene Fake[12] is a sophisticated audio manipulation technique that transforms the environmental context, or "scene", in which the original

[9] M.R. Hasanabadi, *An overview of text-to-speech systems and media applications*, "arXiv preprint arXiv:2310.14301" 2023.

[10] K.B. Bhangale, M. Kothandaraman, *Survey of deep learning paradigms for speech processing*, "Wireless Personal Communications" 2022, no. 125(2), pp. 1913–1949.

[11] A. Mittal, M. Dua, *Automatic speaker verification systems and spoof detection techniques: review and analysis*, "International Journal of Speech Technology" 2021, vol. 25, pp. 105–134.

[12] J. Yi, C. Wang, J. Tao, Z. Tian, C. Fan, H. Ma, R. Fu, *Scenefake: An initial dataset and benchmarks for scene fake audio detection*, "ArXiv" 2022, vol. abs/2211.06073.

speech was recorded. It utilizes advanced speech enhancement technologies to superimpose a different acoustic environment onto the original audio. This can involve adding background noise characteristics of different settings or altering the reverberation to match a distinct location, such as an open-air market or a crowded train station. The primary objective is to shift the listener's perceived setting of the speech, which can have profound implications for the audio's authenticity and integrity. Such alterations can lead to changes in the semantic interpretation of the speech, as the context in which words are spoken often influences their meaning.

- **Partially Fake**. This subset of deepfake audio manipulation is characterized by its targeted approach, focusing on the alteration of select portions of speech rather than the entirety of the audio[13]. The strategy involves splicing certain words or phrases within an utterance with either authentic or artificially synthesized audio segments that maintain the original speaker's vocal characteristics. This form of tampering can be particularly deceptive as it preserves the overall sound and timbre of the speaker's voice, making it difficult to discern the modifications. The resulting audio appears seamless and can mislead listeners or automated systems into misinterpreting the speaker's true intentions or statements. Partial fakes may be employed in scenarios where only specific segments of speech need to be falsified to change the meaning or outcome of the audio recording.

In terms of detection methods, feature extraction is a critical module in pipeline detectors. The goal is to capture audio fake artifacts from speech signals to learn discriminative features. Previous studies[14] have categorized these features into four types: short-term spectral, long-term spectral, prosodic, and deep features. Short- and long-term spectral features, largely reliant on digital signal processing algorithms, describe the acoustic correlates of voice timbre and long-range speech signal information, respectively. However, short-term spectral features have limitations in capturing the temporal characteristics of speech. In contrast, prosodic features, spanning over longer segments like phones and syllables, offer a broader perspective. Traditionally, many of these features were handcrafted, leading to biases due to the limitations of human-designed representations. To address these gaps, deep features extracted via deep neural network-based models have been increasingly employed.

This section sets the foundation for my study, which aims to build upon these existing methodologies and contribute novel insights to the field of audio deepfake detection. The characteristics and relationships of different features are listed in Fig. 1: RFCC[15], ModSpec[16], Global M[17], SDC[18],

[13] J. Yi, Y. Bai, J. Tao, H. Ma, Z. Tian, C. Wang, T. Wang, R. Fu, *Half-truth: A partially fake audio detection dataset*, "Proc. Of Interspeech" 2021.

[14] J. Yi, C. Wang, J. Tao, X. Zhang, C.Y. Zhang, Y. Zhao, *Audio Deepfake Detection: A Survey*, "arXiv preprint arXiv:2308.14970" 2023.

[15] M. Sahidullah, T. Kinnunen, C. Hanilci, *A comparison of features for synthetic speech detection*, "Proc. of INTER SPEECH" 2015.

[16] Ibidem.

[17] Ibidem.

[18] Ibidem.

FDLP[19], CQTMGD[20], LEAF[21], LMS[22], GD[23], MGD[24], BPD[25], RLMS[26], IF[27], CEP[28], LFCC[29], IMFCC[30], MFCC[31], MFPC[32], MWPC[33], LPCC[34], MGDCC[35], Pitch Pattern[36], CosPhase[37], RPS[38], LBP[39], CQTgram[40], CQCC[41],

[19] Ibidem.

[20] Ibidem.

[21] N. Zeghidour, O. Teboul, F. Quitry, M. Tagliasacchi, *Leaf: A learnable frontend for audio classification*, "ICLR" 2021.

[22] X. Xiao, X. Tian, S. Du, H. Xu, H. Li, *Spoofing speech detection using high dimensional magnitude and phase features: the ntu approach for asvspoof 2015 challenge*, "Interspeech" 2015.

[23] Ibidem.

[24] Ibidem.

[25] Ibidem.

[26] X. Tian, Z. Wu, X. Xiong, E.S. Chng, H. Li, *Spoofing detection from a feature representation perspective*, "2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)" 2016.

[27] Ibidem.

[28] L. Rabiner, B.H. Juang, *Fundamentals of speech recognition*, "Fundamentals of speech recognition" 1999.

[29] M. Todisco, H. Delgado, K.A. Lee, M. Sahidullah, N.W.D. Evans, T.H. Kinnunen, J. Yamagishi, *Integrated presentation attack detection and automatic speaker verification: Common features and gaussian back-end fusion*, "Interspeech" 2018.

[30] Ibidem.

[31] L. Chen, W. Guo, L. Dai, *Speaker verification against synthetic speech*, "7th International Symposium on Chinese Spoken Language Processing" 2010, pp. 309–312.

[32] S. Novoselov, A. Kozlov, G. Lavrentyeva, K. Simonchik, V. Shchemelinin, *Stc anti-spoofing systems for the asvspoof 2015 challenge*, "2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)" 2016.

[33] Ibidem.

[34] S. Chakroborty, A. Roy, G. Saha, *Improved closed set text-independent speaker identification by combining mfcc with evidence from flipped filter banks*, "World Academy of Science, Engineering and Technology, International Journal of Electrical, Computer, Energetic, Electronic and Communication Engineering" 2008, vol. 2, pp. 2554–2561.

[35] Z. Wu, X. Xiong, E.S. Chng, H. Li, *Synthetic speech detection using temporal modulation feature*, "IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)" 2013.

[36] Z. Wu, P.L. De Leon, C. Demiroglu, A. Khodabakhsh, S. King, Z.H. Ling, D. Saito, B. Stewart, T. Toda, M. Wester, J. Yamagishi, *Anti-spoofing for text-independent speaker verification: An initial database, comparison of countermeasures, and human performance*, "IEEE/ACM Transactions on Audio, Speech, and Language Processing" 2016, vol. 24, no. 4, pp. 768–783.

[37] E.S.C. Zhizheng Wu, H. Li, *Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition*, "Interspeech" 2012.

[38] J. Sanchez, I. Saratxaga, I. Hernaez, E. Navas, D. Erro, T. Raitio, *Toward a universal synthetic speech spoofing detection using phase information*, "IEEE Transactions on Information Forensics & Security" 2010, vol. 10, no. 4, pp. 810–820.

[39] F. Alegre, R. Vipperla, A. Amehraye, N.W.D. Evans, *A new speaker verification spoofing countermeasure based on local binary patterns*, "Interspeech" 2013.

[40] X. Cheng, M. Xu, T.F. Zheng, *Replay detection using cqt-based modified group delay feature and resnewt network in asvspoof 2019*, "2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)" 2019.

[41] M. Todisco, H. Delgado, N. Evans, *A new feature for automatic speaker verification antispoofing: Constant q cepstral coefficients*, "Processings of Odyssey 2016" 2016.

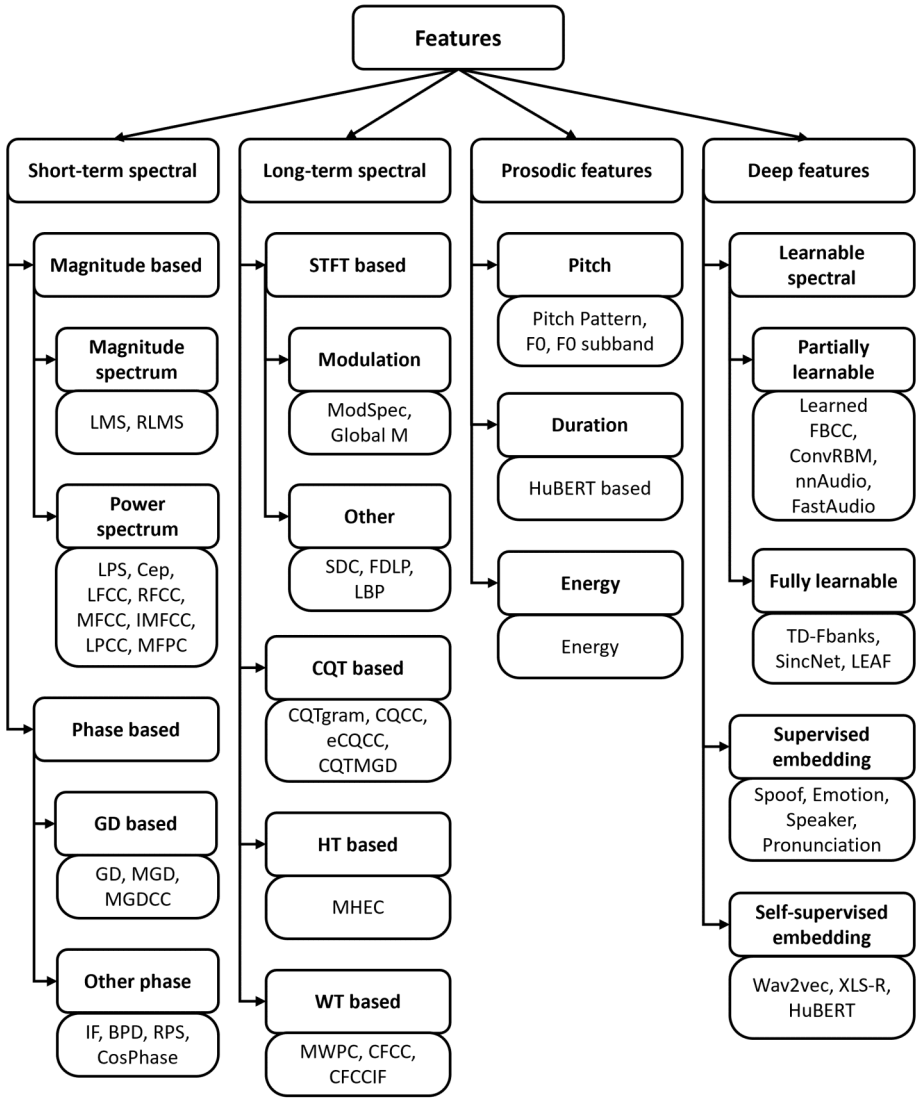

Fig. 1. The commonly employed features in prior research can generally be categorized into four groups: short-term spectral features, long-term spectral features, prosodic features, and deep features

eCQCC[42], CFCC[43], CFCCIF[44], F0[45], HuBERT[46], Energy[47], Pronounciation[48], Learned FBCC[49], ConvRBM[50], nnAudio[51], FastAudio[52], TD-Fbanks[53], SincNet[54], Spoof[55], Emotion[56], Speaker[57], Wav2vec[58], XLS-R[59], LPS[60].

[42] R.K. Das, J. Yang, H. Li, *Assessing the scope of generalized countermeasures for anti-spoofing*, "IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2020" 2020.

[43] T.B. Patel, H. Patil, *Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech*, "Conference of International Speech Communication Association" 2015.

[44] Ibidem.

[45] M. Pal, D. Paul, G. Saha, *Synthetic speech detection using fundamental frequency variation and spectral features*, "Computer Speech & Language" 2018, vol. 48, pp. 31–50.

[46] C. Wang, J. Yi, J. Tao, C. Zhang, S. Zhang, X. Chen, *Detection of cross-dataset fake audio based on prosodic and pronunciation features*, "Interspeech" 2023.

[47] Ibidem.

[48] Ibidem.

[49] Y. Hong, Z.H. Tan, Z. Ma, J. Guo, *Dnn filter bank cepstral coefficients for spoofing detection*, "IEEE Access" 2017, vol. 5, no. 99, pp. 4779–4787.

[50] H.B. Sailor, D.M. Agrawal, H.A. Patil, *Unsupervised filterbank learning using convolutional restricted boltzmann machine for environmental sound classification*, "Interspeech" 2017.

[51] K.W. Cheuk, H. Anderson, K. Agres, D. Herremans, *nnaudio: An on-the-fly gpu audio to spectrogram conversion toolbox using 1d convolutional neural networks*, "IEEE Access" 2020, vol. PP, no. 99, pp. 1–1.

[52] Q. Fu, Z. Teng, J. White, M.G. Powell, D.C. Schmidt, *Fastaudio: A learnable audio front-end for spoof speech detection*, "ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)" 2021, pp. 3693–3697.

[53] N. Zeghidour, N. Usunier, I. Kokkinos, T. Schatz, G. Synnaeve, E. Dupoux, *Learning filterbanks from raw speech for phone recognition*, "IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)" 2018, pp. 5509–5513.

[54] M. Ravanelli, Y. Bengio, *Speaker recognition from raw waveform with sincnet*, "IEEE Spoken Language Technology Workshop (SLT)" 2018, pp. 1021–1028.

[55] N. Chen, Y. Qian, H. Dinkel, B. Chen, K. Yu, *Robust deep feature for spoofing detection – the sjtu system for asvspoof 2015 challenge*, "Interspeech" 2015.

[56] E. Conti, D. Salvi, C. Borrelli, B. Hosler, P. Bestagini, F. Antonacci, A. Sarti, M.C. Stamm, S. Tubaro, *Deepfake speech detection through emotion recognition: A semantic approach*, "IEEE International Conference on Acoustics, Speech and Signal Processing", ICASSP 2022, Virtual and Singapore, 23–27 May 2022, pp. 8962–8966.

[57] J.Y. Pan, S. Nie, H. Zhang, S. He, K. Zhang, S. Liang, X. Zhang, J. Tao, *Speaker recognition-assisted robust audio deepfake detection*, "Interspeech" 2022.

[58] Y. Xie, Z. Zhang, Y. Yang, *Siamese network with wav2vec feature for spoofing speech detection*, "Interspeech" 2021.

[59] J.M. Martín-Doñas, A. Álvarez, *The vicomtech audio´ deepfake detection system based on wav2vec2 for the 2022 add challenge*, 2022, pp. 9241–9245.

[60] Y. Zhang, W. Wang, P. Zhang, *The effect of silence and dual band fusion in anti-spoofing system*, "Interspeech" 2021.

New method of audio stream analysis for deep fake threat identification

This paper introduces a novel method for detecting whether a speech recording is authentic or synthetically generated using deepfake techniques. The proposed approach consists of two primary components: the Vocal Emotion Analysis (VEA) and the Deepfake Audio Identifier (DAI).

1. **Vocal Emotion Analysis** (VEA). The first component of the pipeline focuses on extracting a set of features, F_x, that represent the emotional content of the speech audio signal, x. Leveraging the advancements in deep-learning, the method utilize data-driven neural networks instead of traditional hand-crafted feature extraction methods. Specifically, I employ a 3D Convolutional Recurrent Neural Network (CRNN), as proposed in recent literature. This network classifies speech emotion into N possible categories, such as happiness, sadness, anger, etc.

The process begins with preprocessing the input signal, x, to create a log-mel spectrogram, $S_{mel} \in \mathrm{R}^{MxK}$, via a Short Time Fourier Transform (STFT) in the mel-frequency domain, complemented by a logarithmic transformation. Next step is to compute the first and second discrete derivatives of S_{mel} along the frequency axis, yielding ΔS_{mel} and $\Delta\Delta S_{mel}$. These components are stacked together to form a 3D matrix X, which undergoes z-score normalization. This matrix then passes through a series of 3D convolutional layers, a linear layer, a Bidirectional Long Short-Term Memory (BLSTM), and an attention layer. A sequence of dense layers subsequently outputs a probability measure for each emotion class, from which the prediction E_x should be extracted. Using a transfer-learning approach, next step is to extract a feature vector F_x from the output of the final attention layer, which provides an utterance-level emotional representation.

2. **Deepfake Audio Identifier** (DAI). The second component of the pipeline is a binary classifier that takes the feature vector F_x as input and estimates the class y to which the input signal x belongs. This classifier is designed to distinguish between real and fake audio. Notably, any supervised classification method could be employed at this stage. However, given presented in the article focus on the deepfake discriminatory power of the selected semantic features, well-established classical classifiers were used. Conducted experiments demonstrated that a Random Forest Classifier effectively discriminates between real and fake audio with high accuracy.

The innovative aspect of the proposed method lies in its exploitation of the emotional content of speech as a discriminative factor. TTS deepfake algorithms, while achieving remarkable results in terms of speech naturalness, often fall short in accurately modeling the emotional properties of the human voice. This gap presents an opportunity for the adopted approach, where neural networks' ability to create potent and adaptable embeddings is utilized. The feature vector F_x, derived from the emotional analysis of the speech, serves as a powerful input to the classifier trained specifically for deepfake detection.

Results

Achieved study's results are derived from extensive testing using a variety of datasets, each contributing to a comprehensive understanding of the effectiveness of the proposed method for deepfake speech detection. Multiple datasets were employed, including ASVspoof 2019, Cloud2019, and Interactive Emotional Dyadic Motion Capture (IEMOCAP), LJS and own voice dataset, totaling 133 hours of audio recordings. These datasets encompass both real and deepfake speech samples, crucial for training the Vocal Emotion Analysis (VEA) stage and testing the deepfake detection method.

- **Pre-processing and Input Transformation**. Uniformity across datasets was achieved through rigorous pre-processing. This involved mono conversion, downsampling to a standard frequency, Butterworth band-pass filtering, and normalization. Next step is to transform these pre-processed tracks using time-frequency transforms, creating a common length and computing STFTs to obtain log-mel spectrograms.
- **Training Parameters**. The training of the test implementation of the proposed method was two-fold. The first stage involved training the feature extractor for VEA using the IEMOCAP dataset, focusing on four emotional classes and employing the Adam optimizer. The second stage, training for Deepfake Audio Identifier (DAI), used the features extracted from each dataset. Next step was to create the balanced train set and performed a grid search to select optimal hyperparameters for the Random Forest classifier.
- **Evaluation of Deepfake Audio Identifier (DAI)**. Achieved best-performing RF classifier used information gain as the quality criterion and had 280 learners. Three baseline systems using Receiver Operating Characteristic (ROC) curves were compared with this. The method significantly outperformed these, achieving an Area Under Curve (AUC) of 0.96, indicating superior discrimination capability against classic CNN-based methods. This confirmed that using VEA-trained architectures as feature extractors for DAI tasks enhances deepfake detection accuracy.
- **Performance Across Datasets and Conditions**. The balanced detection accuracy of the binary classifier was evaluated across various datasets and conditions (tab. 1, 2). Excellent performance for pristine signal samples and most deepfake generation algorithms was found, with a few exceptions like T15 or A1 from CDataset. The performance generally degraded with increased noise levels, indicating a tendency of the classifier to label noisy samples as authentic, raising false negatives. However, when trained with noise-augmented data, the proposed system showed resilience to noise in the test set.

Table 1. Results (balanced accuracy) of the evaluation of the proposed system for different datasets and TTS algorithms using clean and augmented training sets. Selected results part 1

SNR [dB]	**Aug.**	**LJS**	**IEM**	**OWN**	**T10**	**T11**	**T12**
∞	No	0.939	0.938	0.969	0.985	0.902	0.893
25	No	0.944	0.947	0.969	0.880	0.877	0.859
20	No	0.962	0.945	0.997	0.703	0,796	0.781
15	No	0.984	0.941	0.999	0.439	0.581	0.539
10	No	0.984	0.930	0.995	0.227	0.361	0.331
∞	Yes	0.858	0.825	0.869	0.991	0.962	0.970
25	Yes	0.859	0.828	0.891	0.982	0.949	0.959
20	Yes	0.863	0.831	0.909	0.933	0.937	0.959
15	Yes	0.795	0.825	0.847	0.889	0.921	0.942
10	Yes	0.653	0.811	0.803	0.823	0.904	0.912

Table 2. Results (balanced accuracy) of the evaluation of the proposed system for different datasets and TTS algorithms using clean and augmented training sets. Selected results part 2

SNR [dB]	**Aug.**	**T13**	**T14**	**T15**	**A1**	**A2**	**A3**
∞	No	0.829	0.922	0.765	0.855	0.924	0.859
25	No	0.738	0.879	0.713	0.613	0.848	0.613
20	No	0.635	0.692	0.555	0.339	0,636	0.334
15	No	0.454	0.306	0.221	0.073	0.241	0.094
10	No	0.337	0.091	0.091	0.023	0.044	0.044
∞	Yes	0.916	0.962	0.882	0.927	0.972	0.923
25	Yes	0.863	0.956	0.874	0.801	0.947	0.823
20	Yes	0.824	0.923	0.839	0.687	0.911	0.702
15	Yes	0.815	0.872	0.793	0.623	0.848	0.623
10	Yes	0.847	0.833	0.756	0.655	0.775	0.688

- **Impact of Training Data Augmentation**. Training data augmentation played a significant role in system performance under different Signal-to--Noise Ratios (SNRs) (fig. 2). While the system trained on clean data performed better in noise-free conditions, its performance was significantly lower than the noise-augmented trained system in noisier environments. This trend was evident in the comparative analysis of ROC curves under varying SNR levels, highlighting the benefit of training with augmented data in real-world conditions where noise presence is inevitable (fig. 3).

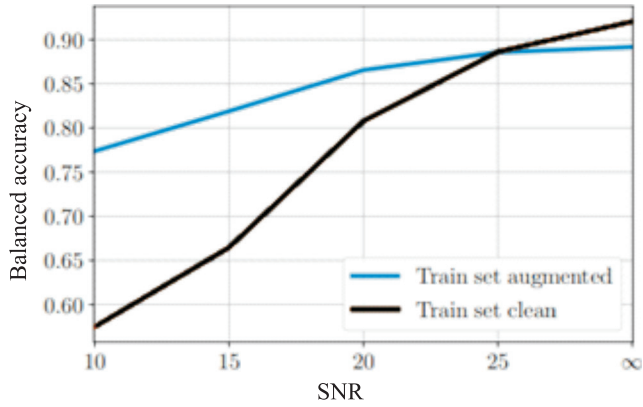

Fig. 2. Balanced accuracy scores for varied Signal-to-Noise Ratios (SNR)

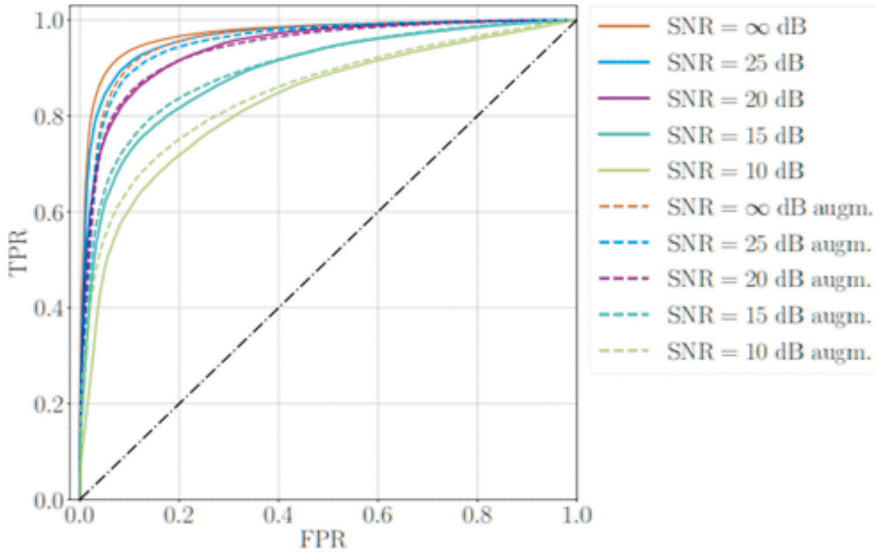

Fig. 3. Receiver Operating Characteristic (ROC) trajectories for the suggested approach, employing both clean and noise-enhanced training datasets across various levels of artificially introduced SNR power

In conclusion, presented results demonstrate that the proposed method for deepfake speech detection, leveraging emotional embeddings and robust training strategies, is highly effective across a range of datasets and conditions. The method's ability to maintain performance in the presence of noise, especially when trained on augmented data, underscores its practical applicability in diverse real-world scenarios.

Conclusions

This paper introduced a pioneering method for audio stream analysis for deep fake threat identification. Most of the deep fakes are generated using synthetic speech, therefore I put effort into leveraging high-level semantic feature extraction with a focus on the emotional voice content of deepfake speech tracks generated by Text-to-Speech (TTS) algorithms. The proposed method's effectiveness is rooted in its two-component system: Vocal Emotion Analysis (VEA) Network and Supervised Classifier for Deepfake Detection. The first component is an VEA network, trained on a dataset annotated for emotions expressed by speakers. This network serves as an emotional feature extractor. By employing a transfer learning approach, we successfully repurposed the network to create an embedding space. This space is significant for both its original purpose – VEA – and targeted application, audio deep fake identification. The second component is a supervised classifier that utilizes the extracted emotional features to discern between real and deepfake speech tracks. This classifier forms the core of the proposed detection system, interpreting emotional nuances to identify synthetic speech.

The proposed method was rigorously tested across several datasets, demonstrating its versatility and adaptability. Additionally, to enhance the robustness of the proposed method, data augmentation techniques were incorporated, specifically adding white noise to the used training data. This approach aimed to simulate more realistic and challenging auditory environments, further testing the resilience of the system against various noise levels. The performance results of the system validate the hypothesis that semantic features, particularly those related to emotional content, are highly effective in audio deepfake detection. The method's ability to interpret and analyze emotional nuances in speech presents a novel approach in the landscape of synthetic speech detection. This approach not only advances the field technically but also opens up new avenues for understanding the subtleties of human speech and its replication in AI-generated audio.

In conclusion, the presented research results contributes significantly to the ongoing efforts to detect and combat the challenges posed by deepfake technologies. By focusing on emotional content as a discriminative feature, a unique perspective and a robust solution to identify synthetic speech was presented, a crucial step in safeguarding the authenticity and integrity of digital communication in an era increasingly dominated by sophisticated AI technologies.

BIBLIOGRAPHY

Abramson A.S., Whalen D.H, *Voice Onset Time (VOT)*, "50: Theoretical and practical issues in measuring voicing distinctions", "Journal of phonetics" 2017, no 63, pp. 75–86.

Alegre F., Vipperla R., Amehraye A., Evans N.W.D., *A new speaker verification spoofing countermeasure based on local binary patterns*, "Interspeech" 2013.

Almutairi Z., Elgibreen H., *A review of modern audio deepfake detection methods: challenges and future directions*, "Algorithms" 2022, no. 15(5), p. 155.

Bhangale K.B., Kothandaraman M., *Survey of deep learning paradigms for speech processing*, "Wireless Personal Communications" 2022, no. 125(2), pp. 1913–1949.

Chakroborty S., Roy A., Saha G., *Improved closed set text-independent speaker identification by combining mfcc with evidence from flipped filter banks*, "World Academy of Science, Engineering and Technology, International Journal of Electrical, Computer, Energetic, Electronic and Communication Engineering" 2008, vol. 2, pp. 2554–2561.

Chen L., Guo W., Dai L., *Speaker verification against synthetic speech*, "7th International Symposium on Chinese Spoken Language Processing" 2010, pp. 309–312.

Chen N., Qian Y., Dinkel H., Chen B., Yu K., *Robust deep feature for spoofing detection – the sjtu system for asvspoof 2015 challenge*, "Interspeech" 2015.

Cheng X., Xu M., Zheng T.F., *Replay detection using cqt-based modified group delay feature and resnewt network in asvspoof 2019*, "Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)" 2019.

Cheuk K.W., Anderson H., Agres K., Herremans D., *nnaudio: An on-the-fly gpu audio to spectrogram conversion toolbox using 1d convolutional neural networks*, "IEEE Access" 2020, vol. PP, no. 99, pp. 1–1.

Conti E., Salvi D., Borrelli C., Hosler B., Bestagini P., Antonacci F., Sarti A., Stamm M.C., Tubaro S., *Deepfake speech detection through emotion recognition: A semantic approach*, "IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022", Virtual and Singapore, 23–27 May 2022, pp. 8962–8966.

Das R.K., Yang J., Li H., *Assessing the scope of generalized countermeasures for anti-spoofing*, "IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2020" 2020.

Dutoit T., *High-quality text-to-speech synthesis: An overview*, "Journal Of Electrical And Electronics Engineering Australia" 1997, no. 17(1), pp. 25–36.

Fu Q., Teng Z., White J., Powell M.G., Schmidt D.C., *Fastaudio: A learnable audio front-end for spoof speech detection*, "ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)" 2021, pp. 3693–3697.

Hasanabadi M.R., *An overview of text-to-speech systems and media applications*, "arXiv preprint arXiv:2310.14301" 2023.

Hong Y., Tan Z.H., Ma Z., Guo J., *Dnn filter bank cepstral coefficients for spoofing detection*, "IEEE Access" 2017, vol. 5, no. 99, pp. 4779–4787.

Machado A.F., Queiroz M.G.D, *Voice conversion: A critical survey*, "Proceedings" 2010.

Martín-Doñas J.M., Álvarez A., *The vicomtech audio deepfake detection system based on wav2vec2 for the 2022 add challenge*, 2022, pp. 9241–9245.

Mittal A., Dua M., *Automatic speaker verification systems and spoof detection techniques: review and analysis*, "International Journal of Speech Technology" 2021, vol. 25, pp. 105–134.

Novoselov S., Kozlov A., Lavrentyeva G., Simonchik K., Shchemelinin V., *Stc anti-spoofing systems for the asvspoof 2015 challenge*, "IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)" 2016.

Pal M., Paul D., Saha G., *Synthetic speech detection using fundamental frequency variation and spectral features*, "Computer Speech & Language" 2018, vol. 48, pp. 31–50.

Pan J.Y., Nie S., Zhang H., He S., Zhang K., Liang S., Zhang X., Tao J., *Speaker recognition-assisted robust audio deepfake detection*, "InterSpeech" 2022.

Patel T.B., Patil H., *Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech*, "Conference of International Speech Communication Association" 2015.

Rabiner L., Juang B.H., *Fundamentals of speech recognition*, "Fundamentals of speech recognition" 1999.

Rana M.S., Nobi M.N., Murali B., Sung A.H., *Deepfake detection: A systematic literature review*, "IEEE access" 2022, no. 10, pp. 25494–25513.

Ravanelli M., Bengio Y., *Speaker recognition from raw waveform with sincnet*, "IEEE Spoken Language Technology Workshop (SLT)" 2018, pp. 1021–1028.

Rosenberg A.E., *Automatic speaker verification: A review*, "Proceedings of the IEEE" 1976, no 64(4), pp. 475–487.

Sahidullah M., Kinnunen T., Hanilci C., *A comparison of features for synthetic speech detection*, "Proc. of INTER SPEECH" 2015.

Sailor H.B., Agrawal D.M., Patil H.A., *Unsupervised filterbank learning using convolutional restricted boltzmann machine for environmental sound classification*, "Interspeech" 2017.

Sanchez J., Saratxaga I., Hernaez I., Navas E., Erro D., Raitio T., *Toward a universal synthetic speech spoofing detection using phase information*, "IEEE Transactions on Information Forensics & Security" 2015, vol. 10, no. 4, pp. 810–820.

Swathi P., Sk S., *Deepfake creation and detection: A survey*, "2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)" 2021, pp. 584–588.

Tian X., Wu Z., Xiong X., Chng E.S., Li H., *Spoofing detection from a feature representation perspective*, "2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)" 2016.

Todisco M., Delgado H., Evans N., *A new feature for automatic speaker verification antispoofing: Constant q cepstral coefficients*, "Processings of Odyssey 2016" 2016.

Todisco M., Delgado H., Lee K.A., Sahidullah M., Evans N.W.D., Kinnunen T.H., Yamagishi J., *Integrated presentation attack detection and automatic speaker verification: Common features and gaussian back-end fusion*, "Interspeech" 2018.

Wang C., Yi J., Tao J., Zhang C., Zhang S., Chen X., *Detection of cross-dataset fake audio based on prosodic and pronunciation features*, "Interspeech" 2023.

Wu Z., De Leon P.L., Demiroglu C., Khodabakhsh A., King S., Ling Z.H., Saito D., Stewart B., Toda T., Wester M., Yamagishi J., *Anti-spoofing for text-independent speaker verification: An initial database, comparison of countermeasures, and human performance*, "IEEE/ACM Transactions on Audio, Speech, and Language Processing" 2016, vol. 24, no. 4, pp. 768–783.

Wu Z., Xiong X., Chng E.S., Li H., *Synthetic speech detection using temporal modulation feature*, "IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)" 2013.

Xiao X., Tian X., Du S., Xu H., Li H., *Spoofing speech detection using high dimensional magnitude and phase features: the ntu approach for asvspoof 2015 challenge*, "Interspeech" 2015.

Xie Y., Zhang Z., Yang Y., *Siamese network with wav2vec feature for spoofing speech detection*, "Interspeech" 2021.

Yi J., Bai Y., Tao J., Ma H., Tian Z., Wang C., Wang T., Fu R., *Half-truth: A partially fake audio detection dataset*, "Proc. Of Interspeech" 2021.

Yi J., Wang C., Tao J., Tian Z., Fan C., Ma H., Fu R., *Scenefake: An initial dataset and benchmarks for scene fake audio detection*, "ArXiv" 2022, vol. abs/2211.06073.

Yi J., Wang C., Tao J., Zhang X., Zhang C.Y., Zhao Y., *Audio Deepfake Detection: A Survey*, "arXiv preprint arXiv:2308.14970" 2023.

Zeghidour N., Teboul O., Quitry F., Tagliasacchi M., Leaf: *A learnable frontend for audio classification*, "ICLR" 2021.

Zeghidour N., Usunier N., Kokkinos I., Schatz T., Synnaeve G., Dupoux E., *Learning filterbanks from raw speech for phone recognition*, "IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)" 2018, pp. 5509–5513.

Zhang Y., Wang W., Zhang P., *The effect of silence and dual band fusion in anti-spoofing system*, "Interspeech" 2021.

Zhizheng Wu E.S.C., Li H., *Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition*, "Interspeech" 2012.

Internet source:

<https://cyware.com/news/fraudsters-make-away-with-243000-by-impersonating-company-ceo-in-new-voice-phishing-attack-c8dc188d>, accessed: 06.11.2023.

AUDIO STREAM ANALYSIS FOR DEEP FAKE THREAT IDENTIFICATION

SUMMARY

The article introduces a new method for identifying deepfake threats in audio, focusing on detecting synthetic speech generated by text-to-speech algorithms. Central to the presented method are two elements: the Vocal Emotion Analysis (VEA) Network and the Supervised Classifier for Deepfake Detection. The VEA Network detects emotional nuances in speech, while the Classifier uses these features to differentiate between real and fake audio. This approach exploits the inability of deepfake algorithms to replicate the emotional complexity of human speech, adding a semantic layer to the detection process. The system's effectiveness has been confirmed through tests on various datasets, including in challenging real-world conditions simulated with data augmentation, such as adding white noise. Results show consistent, high accuracy across different datasets and in noisy environments, particularly when trained with noise-augmented data. This method, leveraging voice's emotions content and advanced machine learning, offers a robust defense against audio manipulation, enhancing the integrity of digital communications amidst the rise of synthetic media.

KEYWORDS: audio modification detection, voice analysis, fake audio detection