DEEPFAKE SPEECH DETECTION THROUGH EMOTION RECOGNITION: A SEMANTIC APPROACH

Emanuele Conti[*] *Davide Salvi*[*] *Clara Borrelli*[*] *Brian Hosler*[†] *Paolo Bestagini*[*]
Fabio Antonacci[*] *Augusto Sarti*[*] *Matthew C. Stamm*[†] *Stefano Tubaro*[*]

[*]Dipartimento di Elettronica, Informazione e Bioingegneria - Politecnico di Milano - Milan, Italy
[†]Department of Electrical and Computer Engineering - Drexel University - Philadelphia, PA, USA

ABSTRACT

In recent years, audio and video deepfake technology has advanced relentlessly, severely impacting people's reputation and reliability. Several factors have facilitated the growing deepfake threat. On the one hand, the hyper-connected society of social and mass media enables the spread of multimedia content worldwide in real-time, facilitating the dissemination of counterfeit material. On the other hand, neural network-based techniques have made deepfakes easier to produce and difficult to detect, showing that the analysis of low-level features is no longer sufficient for the task. This situation makes it crucial to design systems that allow detecting deepfakes at both video and audio levels. In this paper, we propose a new audio spoofing detection system leveraging emotional features. The rationale behind the proposed method is that audio deepfake techniques cannot correctly synthesize natural emotional behavior. Therefore, we feed our deepfake detector with high-level features obtained from a state-of-the-art Speech Emotion Recognition (SER) system. As the used descriptors capture semantic audio information, the proposed system proves robust in cross-dataset scenarios outperforming the considered baseline on multiple datasets.

Index Terms— deepfake, audio forensics, deep learning

1. INTRODUCTION

With the term deepfake we refer to a category of videos that have been edited to alter the identity of the depicted person through facial or speech manipulation. This is done by swapping the face, restyling the voice, or modifying what the person is saying [1] , often through the use of deep learning techniques. Thanks to the constant development of new technologies and the massive evolution of neural networks, deepfake generation is nowadays an effortless operation and it is becoming increasingly difficult to distinguish the manipulated material from the original one.

While this opens the door to new challenging and stimulating scenarios, it can also lead to unpleasant situations, primarily when the generated content does not have the approval of the involved people. Deepfakes have been used with malicious intents in several cases, as in the spreading of fake news [2], and fraud cases [3]. This led to some ethical considerations regarding the use of artificial intelligence [4].

Concerning audio modifications, the rapid progress of Text-To-Speech (TTS) synthesis and Voice Conversion (VC) techniques has increased the possibility of impersonating one person's voice. For this reason, it has become of paramount importance to develop techniques capable of determining whether a given multimedia content is authentic or counterfeited [5]. In the past few years, the research community has moved in this direction and has already proposed numerous approaches that analyze both video [6, 7] and audio [8] material. In the audio field, [9] feeds linear filter banks into a Resnet to generate embeddings used as input of a neural network classifier, and in [10] long-term features are used to discriminate fake and real audio tracks. Recently [11] detected for audio deepfakes based on long-term and short-term predictor features, while [12] exploits the traces left by time scaling to discriminate fake audio signals.

This manuscript proposes a method to identify whether a given speech signal is authentic or deepfake exploiting sentiment analysis. We focus on the audio component since generating a well-counterfeited speech is crucial to produce accurate and convincing synthetically generated videos. To perform this classification, we leverage semantically meaningful audio embeddings, as the use of semantic features has already proved successful for both audio and video deepfake analysis [13, 14]. As suggested in [13], we assume that deepfake generators can synthesize low-level voice characteristics but fail to recreate more complex aspects, such as emotions.

Speech Emotion Recognition (SER) is a field that has been increasingly investigated in recent years and refers to the problem of automatically recognizing the emotion perceived by the talking person from the analysis of its speech (i.e., audio recording). Many different networks have been designed for this purpose, both considering audio (speech) only [15, 16], and multi-modal approaches [17, 18]. In this work, we propose a novel transfer-learning method, using the semantic features extracted from a SER network as input of a deepfake classifier. The method is focused on the detection of TTS and mixture TTS/VC deepfakes, while does not take into account pure VC algorithms. This is because we exploit speech semantic information to detect anomalies, and pure VC fakes do include such content, being generated from a real voice and then altered with style transfer techniques. We performed a large-scale experiment on 123 effective hours of speech recordings from different datasets, in both clean and noisy setups. Results show promising performance, reaching a balanced accuracy close to 94 % in clean conditions and above 83 % in the case of deepfake speech corrupted by noise during tests on the ASVspoof [19] evaluation set.

This work was supported by the PREMIER project, funded by the Italian Ministry of Education, University, and Research within the PRIN 2017 program. This material is based on research sponsored by DARPA and Air Force Research Laboratory (AFRL) under agreement number FA8750-20-2-1004. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA and Air Force Research Laboratory (AFRL) or the U.S. Government.

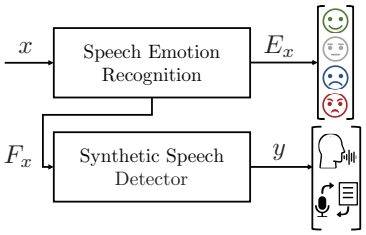

Fig. 1. Architecture of the proposed system.

2. PROPOSED SYSTEM

In this paper we propose a method to detect if a speech recording belongs to a real person or has been synthetically generated through some deepfake technique. Given a speech audio signal x under analysis, the goal is to estimate its class

$$y \in \{\text{REAL}, \text{DF}\}, \tag{1}$$

where REAL indicates that the speech signal is authentic, and DF corresponds to deepfake audio tracks.

Figure 1 shows the pipeline of the proposed method. The process is composed of two blocks. The first block is a SER system that exploits the architecture recently proposed in [15]. Starting from an input speech signal x, it estimates the expressed emotion E_x and extracts a set of features F_x. The second block is the Synthetic Speech Detector (SSD) system that associates a class y to the input features F_x. In the following, we provide details about each block.

Speech Emotion Recognition. The first part of the proposed pipeline extracts a set of features F_x able to express the emotional content of the speech audio signal x under analysis. Motivated by the state-of-the-art performances provided by deep-learning methods, we decided to explore data-driven neural networks for the purpose rather than to use hand-crafted feature extraction. The considered emotional features are computed making use of the 3D-Convolutional Recurrent Neural Network (CRNN) proposed in [15]. We refer to the complete paper for further details. The authors address the problem of speech emotion recognition as a classification problem using a categorical approach, i.e., N possible emotion classes are considered [20]. Therefore, given a speech utterance x, the output of the network is $E_x \in \{e_1, e_2, ..., e_N\}$, where e_i is the i-th emotion class (e.g., happy, sad, angry, etc.). As reported in [15], the input signal x must be pre-processed to be fed to the following neural network. We do so by computing the spectrum of x through an Short Time Fourier Transform (STFT) in the mel-frequency domain and applying a logarithmic transform to the STFT magnitude. This returns a log-mel spectrogram defined as

$$\mathbf{S}_{\text{mel}} \in \mathbb{R}^{M \times K}, \tag{2}$$

where M is the number of windows and K is the number of mel bins. Then, we compute the first and second discrete derivatives of $\mathbf{S}_{\text{mel}}$ along its second dimension (frequency axis), obtaining $\Delta\mathbf{S}_{\text{mel}}$ and $\Delta\Delta\mathbf{S}_{\text{mel}}$. By stacking the log-mel spectrogram and its derivatives along a third dimension, we obtain the final 3D matrix $\mathbf{X}$ defined as

$$\mathbf{X} = [\mathbf{S}_{\text{mel}}, \Delta\mathbf{S}_{\text{mel}}, \Delta\Delta\mathbf{S}_{\text{mel}}] \in \mathbb{R}^{M \times K \times 3}. \tag{3}$$

This matrix is then standardised by means of z-score normalization. The processed input is fed to a set of 3D convolutional layers, followed by a linear layer, a BLSTM and an attention layer. Finally, a sequence of dense layers outputs a probability measure of each emotion class, from which we extract the prediction E_x. Adopting a transfer-learning strategy, we extract a feature vector F_x of dimensionality M from an intermediate network layer. Specifically we consider the output of the final attention layer, presented by the authors as the *utterance-level emotional representation*. Formally, we can express the feature extraction block as a function $\mathcal{F}$ such that

$$F_x = \mathcal{F}(x) \in \mathbb{R}^M \tag{4}$$

The selected feature vector does not simply have discriminative power for its original task (i.e., estimating the *quality* of the emotion) but also for synthetic speech detection (i.e., estimating the *intensity/quantity* of the emotions expressed). This is because TTS deepfake algorithms reach excellent results in terms of speech naturalness but still fail in modeling the emotional properties of the human voice correctly. We can therefore exploit this weakness with neural networks' ability to create powerful and flexible embeddings, using F_x as input to a classifier trained for deepfake detection.

Synthetic Speech Detector. In the second part of the proposed pipeline, a binary classifier takes as input the feature vector F_x and estimates the class y to which the input signal x belongs. It is worth noting that we can use any supervised classification method at this stage. However, since this work aims to explore the deepfake discriminatory power of the selected semantic features, we decided to use well-known classical classifiers. Our experiments show that a Random Forest Classifier is capable of discriminating between real and fake audio with high accuracy.

3. EXPERIMENTAL SETUP

Datasets. In this section we present the datasets that have been used to train the SER stage and to train and test the deepfake detection method. The considered datasets include both real and deepfake speech samples for 123 hours of audio recordings. We use multiple datasets to ensure that our proposed technique does not overfit to one dataset or domain, and is appropriate for real-world conditions.

- *ASVspoof 2019* [21] is a speech audio dataset created to develop antispoofing techniques for automatic speaker verification. It contains both real and deepfake speech data. We consider its Logical Access (LA) partition and we select only samples generated using TTS or TTS/VC hybrid methods. Therefore, in our *train* and *dev* partitions we include authentic signals along with speech samples generated with 4 different algorithms (named A01, A02, A03, A04). In *eval* partition, we keep real signals and samples from 10 other algorithms (A07, ..., A16).
- *LibriSpeech* (LS) [22] is an open-source dataset containing about 1000 hours of authentic speech. From this corpus we considered the subset *train-clean-100*.
- *LJSpeech* (LJS) [23] contains audio clips of a single speaker reciting pieces from non-fiction books.
- *Cloud2019* corresponds to the dataset proposed in [8]. It includes tracks from different TTS cloud services: Amazon AWS Polly (PO), Google Cloud Standard (GS), Google Cloud WaveNet (GW), Microsoft Azure (AZ) and IBM Watson (WA).
- *Interactive Emotional Dyadic Motion Capture (IEMOCAP)* (IEM) [24] contains video and speech recordings annotated with the speaker's emotions. Speech tracks are segments of scripted or improvised dialogues performed by actors emphasizing a particular emotion. From the improvised dialogues subset, which is the

one considered in [15], we analyzed all the tracks labeled with the four classes of our SER system (i.e., angry, happy, sad, neutral).

Input pre-processing and transformation. To avoid detecting dataset-specific artifacts, we pre-process all tracks to make them as uniform as possible. We convert all tracks to mono and, if necessary, downsample them to a standard sampling frequency $F_s = 16$ kHz. Then, we filter all speech signals using a Butterworth band-pass digital filter with order 6, considering a lowcut frequency $F_l = 250$ Hz and a highcut frequency $F_h = 3600$ Hz. Finally, we normalize each track using infinity norm. We compute the input of the SER block starting from a time-frequency transform, as detailed in [15]. Each track of the datasets is reduced to have a common length $L_{cut} = 3$ s, using zero-padding if necessary. Then, we compute the STFT of x using a Hamming windows of length $L_w = 0.025$ s and a hop-size $L_h = 0.01$ s. Only the magnitude of the STFT is considered. The spectrum is then processed using a bank of mel-spaced filters and further scaled using the natural logarithm function. In our implementation we consider $M = 300$ windows and $K = 40$ mel bins.

Dataset augmentation. To test the robustness of our system against possible audio degradation, we create a second version of the dataset using data augmentation techniques. We do so by adding white noise to the speech tracks considering two different approaches. For the train and validation sets, we perform noise injection according to a double-layer probability distribution. The first layer injects white noise randomly between 30 dB and 15 dB of power Signal-to-Noise Ratio (SNR) with probability $p_1 = 0.8$. The second layer randomly injects white noise between 15 dB and 10 dB of power SNR, with probability $p_2 = 0.3$. For the test set, instead, power SNR is fixed in the range SNR $= [25, 20, 15, 10]$ dB. In other words, training data contains a wide variety of noise, whereas test data is obtained in a controlled scenario to enables results analysis.

Training parameters. Our proposed system contains 2 parts which are trained independently. First, the feature extractor is trained to perform Speech Emotion Recognition (SER) following the procedure proposed in [15]. Specifically, we use the IEMOCAP dataset and we consider the classes *angry*, *happy*, *sad*, *neutral*, hence $N = 4$. Since the IEMOCAP dataset is divided in 5 dialogue sessions, we select sessions 1 to 4 for training and session 5 for development and testing. We use Adam optimizer with learning rate $l_r = 10^{-5}$ and categorical cross-entropy as loss function. Our trained feature extractor achieved comparable results to those presented in [15] with a balanced accuracy of 0.6 of four classes. The dimension of the feature vector F_x is $M = 256$.

The second stage of our proposed system was trained to perform Synthetic Speech Detection (SSD), using features extracted from each of the clean and augmented datasets. The composition of training, development and test dataset are presented in Table 1. The train set is balanced, adjusting the learning weights inversely proportional to the class frequencies in the input data. The hyper-parameters for the Random Forest (RF) have been selected using a grid search on the validation set, using Balanced Accuracy (BA) as a metric. The considered parameters are the criterion of split quality and the number of learners. In particular, we tested as quality criterion function both Gini impurity and information gain. Regarding the number of learners, we consider $N_{\text{RF}} = [10, 30, 100, 300]$.

Baseline. We compared the performance of our system to those of other well-established state of the art methods. In particular, we considered two different baselines, VGGish and RawNet2. VGGish [25] is a popular Convolutional Neural Network (CNN) architecture

Table 1. Composition of train, development and test sets for the SSD block

	Real	DF	N. Tracks
Train	ASVspoof2019 *train* LibriSpeech	ASVspoof2019 *train*	46319
Dev	ASVspoof2019 *dev*	ASVspoof2019 *dev*	17412
Test	ASVspoof2019 *eval* IEMOCAP LJSpeech	ASVspoof2019 *eval* Cloud2019	83660

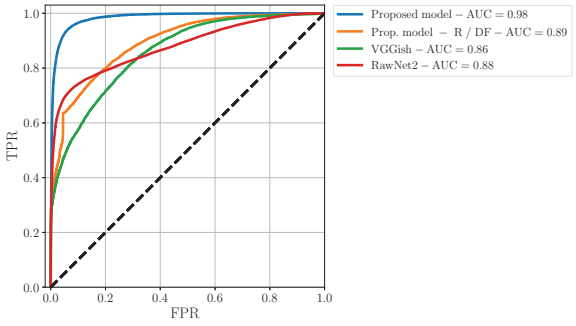

Fig. 2. ROC curves for the proposed method and the considered baselines on clean noiseless datasets.

firstly proposed for audio event classification and trained on an extensive dataset of audio tracks [26]. Due to its generalization capacity, it has often been used as an embedding extractor for other audio analysis tasks. In our case, we add a final classification layer to the standard VGGish embedding extractor architecture and we fine-tune it for the task at hand, i.e., synthetic speech detection, using binary cross-entropy as loss function. The second baseline, RawNet2 [27], is an end-to-end network aimed at audio anti-spoofing detection. We trained it using our considered dataset starting from the pre-trained model provided as a baseline in the ASVspoof2021 challenge [19]. As a third baseline, we compared the results of our transfer-learning approach with those obtained by training the first network of our pipeline on the SSD task directly. This test aims to verify that the use of emotions really benefits our system and is relevant in increasing its performance.

4. RESULTS

In this section we present the results relative to the SSD task. The best hyperparameter setup of the RF classifier corresponds to information gain as quality criterion function and a number of learners $N_{RF} = 300$. Figure 2 compares the Receiver Operating Characteristic (ROC) curves of our proposed method, against our 3 baselines. These systems were trained using the clean dataset (without noise injection), and evaluated using the ASVspoof2019 *eval* partition without noise injection. As can be seen from the Figure, our method outperforms both VGGish and RawNet2, reaching a value of AUC $= 0.98$. This first experiment confirms that the proposed approach allows achieving higher discrimination capability if compared to more classic CNN based methods. Figure 2 also shows that training architecture proposed in [15] directly for the task of SSD achieves worse results than training it for SER and then use it as feature extractor for the SSD task. This shows that extracting emotional embeddings from an audio track creates a strong feature set and can

Table 2. Results (balanced accuracy) of the evaluation of the proposed system for different datasets and TTS algorithms using clean and augmented training sets. Real and deepfake dataset names are coherent with definitions in Section 3.

SNR	Train	Real			Deepfake														
[dB]	Augm.	LJS	IEM	BF	A07	A08	A09	A10	A11	A12	A13	A14	A15	A16	PO	AZ	GS	GW	WA
∞		0.941	0.943	0.970	0.948	0.988	1.000	0.900	0.895	0.890	0.831	0.763	0.927	0.898	0.483	0.812	0.993	0.921	0.855
25		0.947	0.944	0.996	0.911	0.883	0.992	0.875	0.861	0.803	0.740	0.710	0.872	0.736	0.421	0.558	0.966	0.851	0.610
20		0.965	0.943	0.999	0.814	0.699	0.917	0.800	0.783	0.539	0.632	0.547	0.687	0.439	0.304	0.264	0.819	0.639	0.331
15		0.982	0.942	0.999	0.565	0.421	0.587	0.576	0.542	0.202	0.446	0.216	0.303	0.129	0.138	0.032	0.361	0.238	0.060
10		0.988	0.934	0.999	0.342	0.224	0.223	0.355	0.334	0.128	0.314	0.084	0.093	0.080	0.093	0.000	0.051	0.038	0.020
∞	✓	0.854	0.828	0.865	0.975	0.994	1.000	0.957	0.973	0.940	0.910	0.877	0.965	0.941	0.603	0.832	0.996	0.966	0.920
25	✓	0.857	0.829	0.894	0.969	0.970	1.000	0.952	0.969	0.922	0.863	0.870	0.956	0.901	0.584	0.768	0.994	0.942	0.803
20	✓	0.861	0.829	0.904	0.947	0.926	0.999	0.939	0.961	0.892	0.824	0.834	0.927	0.837	0.533	0.522	0.978	0.907	0.697
15	✓	0.797	0.823	0.842	0.927	0.884	0.995	0.923	0.946	0.845	0.809	0.783	0.868	0.758	0.497	0.259	0.955	0.845	0.617
10	✓	0.656	0.807	0.800	0.886	0.817	0.984	0.907	0.916	0.843	0.836	0.764	0.829	0.748	0.466	0.268	0.887	0.767	0.676

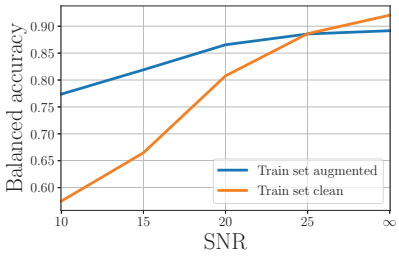

Fig. 3. Balanced accuracy values for arbitrary SNR.

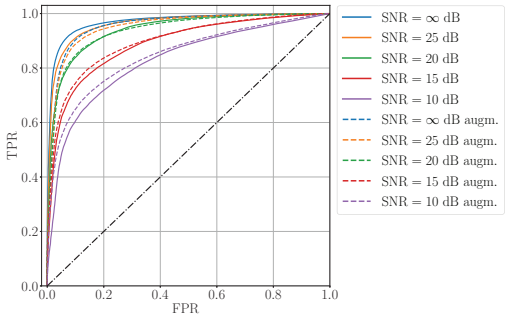

Fig. 4. ROC curves for the proposed method with clean and augmented train sets at different arbitrarily injected power SNR levels.

improve deepfake detection accuracy.

Table 2 shows the balanced detection accuracy of the proposed binary classifier for the two training configurations. Classification accuracies are computed separately for each dataset of real speech tracks and each TTS algorithm used to generate deepfake speech tracks. The top half of Table 2 shows the performance of the system trained on clean data, while the bottom half shows the performance of the system trained with noise-augmented data. In the first row of Table 2, the test set has not been augmented with noise, hence SNR $= \infty$. We can observe that performances are very good for all pristine signal samples and most deepfakes generation algorithms. Algorithm A14 from ASVspoof2019 and PO from Cloud2019 are the only cases where the accuracy is below 0.8. We suspect that this is because algorithm A14 is a mixed TTS/VC system that has been built starting from a very efficient VC system [19]. Hence real emotional qualities are probably still present in the audio tracks, affecting the efficiency of the proposed system. For all the other deepfake systems, the balanced accuracy value is close to or greater than 0.9. We can observe in rows 2 to 4 that, as the noise level increases, the performances of the synthetic speech detector degrades more and more. As the noise increases, the classifier tends to label all samples as authentic, as we notice a remarkable increase in the false-negative rate. This behavior encourages the use of data augmentation strategy on the training set.

When the training set is augmented, the presence of noise in the testing set does not significantly affect the detector performance. To further analyze the effects of training data augmentation, in Figure 3 we report the balanced accuracy values on the entire dataset for different SNRs, training both on clean and augmented dataset. From Table 2, we see that the system trained on clean data achieves higher accuracy on clean data. However, the latter outperforms the former in direct proportion to the decrease in SNR, reaching a difference of almost 20% in the noisiest experiment, i.e., for SNR $=$ 10 dB.

Figure 4 confirms this trend by showing the ROC curves obtained with the proposed method considering clean (solid) and augmented train sets (dashed). In this case, True Positive Rate (TPR) and False Positive Rate (FPR) are computed simply considering all the samples in the test set. We can observe that, when SNR is high, training on clean data is more advantageous than using the augmented training set. As the test set SNR level decreases, the ratio between true positives and false positives generally lowers, but it drops more in the case of the classifier trained on clean data for the one trained on augmented data.

5. CONCLUSIONS

In this paper we have presented a novel method for synthetic speech detection based on high-level semantic feature extraction. We focused on detecting deepfake speech tracks generated with TTS algorithms exploiting their emotional voice content. The system is composed of two main components. The first one is a SER network trained on a speech dataset annotated with the emotion expressed by the speaker and used as emotional feature extractor. By applying a transfer learning approach to this network, we can create an embedding space that is meaningful not only for the original task, i.e., SER, but also for the task at hand, i.e., synthetic speech detection. The second component is a supervised classifier that takes as input the emotional features and predicts if the given speech track is real or deepfake. We tested the proposed system on several different datasets. Moreover, to further increase the robustness of our method, we apply data augmentation with additive white noise. The performances of the proposed system validate the idea of exploiting semantic features for audio deepfake detection.

6. REFERENCES

[1] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner, "Face2face: Real-time face capture and reenactment of RGB videos," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[2] E Howcroft, "How faking videos became easy and why that's so scary," *Bloomberg: New York, NY, USA*, 2018.

[3] BBC News, "Deepfake app causes fraud and privacy fears in china," 2019.

[4] Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, et al., "The malicious use of artificial intelligence: Forecasting, prevention, and mitigation," *arXiv preprint arXiv:1802.07228*, 2018.

[5] Luisa Verdoliva, "Media forensics and deepfakes: an overview," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, pp. 910–932, 2020.

[6] S. Agarwal, H. Farid, O. Fried, and M. Agrawala, "Detecting deep-fake videos from phoneme-viseme mismatches," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020.

[7] N. Bonettini, E. D. Cannas, S. Mandelli, L. Bondi, P. Bestagini, and S. Tubaro, "Video face manipulation detection through ensemble of CNNs," in *International Conference on Pattern Recognition (ICPR)*, 2021.

[8] A. Lieto, D. Moro, F. Devoti, C. Parera, V. Lipari, P. Bestagini, and S. Tubaro, ""Hello? Who Am I Talking to?" A Shallow CNN Approach for Human vs. Bot Speech Classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[9] Tianxiang Chen, Avrosh Kumar, Parav Nagarsheth, Ganesh Sivaraman, and Elie Khoury, "Generalization of audio deepfake detection," in *Odyssey Speaker and Language Recognition Workshop*, 2020.

[10] Madhu R Kamble, Hardik B Sailor, Hemant A Patil, and Haizhou Li, "Advances in anti-spoofing: from the perspective of ASVspoof challenges," *APSIPA Transactions on Signal and Information Processing*, 2020.

[11] Clara Borrelli, Paolo Bestagini, Fabio Antonacci, Augusto Sarti, and Stefano Tubaro, "Synthetic speech detection through short-term and long-term prediction traces," *EURASIP Journal on Information Security*, vol. 2021, pp. 1–14, 2021.

[12] Michele Pilia, Sara Mandelli, Paolo Bestagini, and Stefano Tubaro, "Time scaling detection and estimation in audio recordings," in *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2021.

[13] B. Hosler, D. Salvi, A. Murray, F. Antonacci, P. Bestagini, S. Tubaro, and M. C. Stamm, "Do deepfakes feel emotions? A semantic approach to detecting deepfakes via emotional inconsistencies," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021.

[14] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha, "Emotions Don't Lie: An Audio-Visual Deepfake Detection Method using Affective Cues," in *ACM International Conference on Multimedia*, 2020.

[15] Mingyi Chen, Xuanji He, Jing Yang, and Han Zhang, "3-D convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 25, pp. 1440–1444, 2018.

[16] Ruhul Amin Khalil, Edward Jones, Mohammad Inayatullah Babar, Tariqullah Jan, Mohammad Haseeb Zafar, and Thamer Alhussain, "Speech emotion recognition using deep learning techniques: A review," *IEEE Access*, vol. 7, pp. 117327–117345, 2019.

[17] Didan Deng, Yuqian Zhou, Jimin Pi, and Bertram E Shi, "Multimodal utterance-level affect analysis using visual, audio and text features," *CoRR*, vol. abs/1805.00625, 2018.

[18] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Mazumder, Amir Zadeh, and Louis-Philippe Morency, "Multi-level multiple attentions for contextual multimodal sentiment analysis," in *IEEE International Conference on Data Mining (ICDM)*, 2017.

[19] Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, et al., "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, pp. 101114, 2020.

[20] Giovanna Colombetti, "From affect programs to dynamical discrete emotions," *Philosophical Psychology*, vol. 22, pp. 407–425, 2009.

[21] Massimiliano Todisco, Xin Wang, Ville Vestman, Md Sahidullah, Héctor Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Tomi Kinnunen, and Kong Aik Lee, "ASVspoof 2019: Future horizons in spoofed and fake audio detection," *arXiv preprint arXiv:1904.05441*, 2019.

[22] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2015.

[23] Keith Ito and Linda Johnson, "The LJSpeech dataset," 2017.

[24] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.

[25] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson, "CNN architectures for large-scale audio classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[26] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[27] Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher, "End-to-end anti-spoofing with RawNet2," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021*.