Toward comprehensive short utterances manipulations detection in videos

Abderrazzaq Moufidi, David Rousseau, Pejman Rasti

Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-20284-x

Toward comprehensive short utterances manipulations detection in videos

Abderrazzaq Moufidi[1,2] · **David Rousseau**[1] · **Pejman Rasti**[1]

Received: 8 April 2024 / Revised: 21 August 2024 / Accepted: 18 September 2024

Abstract
In a landscape increasingly populated by convincing yet deceptive multimedia content generated through generative adversarial networks, there exists a significant challenge for both human interpretation and machine learning algorithms. This study introduces a shallow learning technique specifically tailored for analyzing visual and auditory components in videos, targeting the lower face region. Our method is optimized for ultra-short video segments (200-600 ms) and employs wavelet scattering transforms for audio and discrete cosine transforms for video. Unlike many approaches, our method excels at these short durations and scales efficiently to longer segments. Experimental results demonstrate high accuracy, achieving 96.83% for 600 ms audio segments and 99.87% for whole video sequences on the FakeAVCeleb and DeepfakeTIMIT datasets. This approach is computationally efficient, making it suitable for real-world applications with constrained resources. The paper also explores the unique challenges of detecting deepfakes in ultra-short sequences and proposes a targeted evaluation strategy for these conditions.

Keywords DeepFake · Biometrics · Multimodality · Late fusion · Presentation attacks · Adversarial attacks

1 Introduction

In an era marked by rapid advancements in artificial intelligence and Generative Adversarial Networks (GANs), the manipulation of multimedia content has become increasingly sophisticated and accessible [1, 2]. This growing accessibility opens new avenues for creative and practical applications but also introduces a significant challenge: multimedia manipulations that are becoming progressively more difficult to detect. These manipulations pose serious risks, as they can mislead both human perception and automated systems, thereby

✉ Pejman Rasti
pejman.rasti@univ-angers.fr

1 Laboratoire Angevin de Recherche en Ingénierie des Systèmes (LARIS), UMR INRAe-IRHS, Université d'Angers, 62 Avenue Notre Dame du Lac, Angers 49000, France

2 Centre d'Études et de Recherche pour l'Aide à la Décision (CERADE), ESAIP, 18 Rue du 8 Mai 1945, Saint-Barthélemy-d'Anjou 49124, France

compromising the integrity of information channels [3]. Current detection methods often rely on resource-intensive deep learning models, which limits their practical use in real-world environments where computational resources are constrained [4–7].

The latest developments in deepfake technology have greatly influenced multimedia manipulation, leading to the creation of various methods capable of generating highly realistic synthetic media [8]. Notably, lip-syncing algorithms, which alter lip movements in videos to align with specific audio tracks, have gained significant attention [8, 9]. These techniques create a convincing illusion that the person in the video is speaking the provided audio. Previous research has indicated that detecting manipulations in the lower part of the face, particularly the lips and mouth, is more challenging compared to other facial features like the eyes, especially when using deep neural networks such as XceptionNet [10, 11]. This is a critical aspect of our study, as many deepfake techniques target the lower facial region to create deceptive effects [1]. The significance of the lip area is further amplified in scenarios where the upper face is obscured, making other facial features less reliable for verifying authenticity. By focusing on lip area manipulation detection, our research tackles a vital component of deepfake techniques and seeks to offer a robust solution for ensuring the authenticity of digital media.

Voice conversion represents a recent advancement in deepfake technology, enabling the transformation of one person's voice to closely mimic the vocal characteristics of another [2, 12]. This technology can be combined with visual manipulations to create even more convincing synthetic media [2, 13]. Additionally, text-to-speech synthesis has advanced to the point where synthetic voices are becoming nearly indistinguishable from human voices, facilitating the creation of highly realistic audio content from text inputs [2, 12, 14]. While these developments are impressive, they highlight the critical need for effective detection mechanisms to ensure the authenticity of multimedia content in this new era of synthetic media.

To tackle the detection of manipulated short utterances in both audio and video, our research centers on the sub-word level-a domain that has not been widely explored in existing deepfake detection methods. This focus is essential for handling short sequences, where the limited amount of information can make pattern recognition particularly challenging. By examining sub-word elements, our study seeks to capture subtle details that are crucial for accurately identifying deepfakes, thereby addressing a significant gap in current research. This innovative approach provides a more precise tool for analyzing brief and nuanced utterances, contributing to the advancement of multimedia manipulation detection.

In the context of these technological advancements, we introduce a novel shallow learning-based method specifically designed for the detection of deepfakes in ultra-short video sequences, ranging from 200 milliseconds (ms) to 600 ms.

The primary contributions of this paper are as follows:

- Novel Detection Method: Introduced a specialized shallow learning technique for detecting deepfake content by analyzing the visual and auditory components of multimedia, specifically targeting the lower region of the face in videos.
- Focus on Ultra-short Segments: Optimized the detection method for ultra-short video segments ranging from 200 ms to 600 ms, addressing a significant challenge in existing methods.
- Multi-scale Audio Analysis: Employed a multi-scale analysis for audio features using the wavelet scattering transform (WST), which effectively captures essential frequency characteristics of audio signals.

- High-frequency Video Analysis; Developed a video feature extraction method based on high-frequency spatial analysis using discrete cosine transforms (DCT), focusing on the lip region for enhanced detection accuracy.
- Versatility: Designed the method to be versatile, applicable in both unimodal and multimodal settings, leveraging visual and auditory cues for a comprehensive evaluation.
- Performance and Efficiency: Demonstrated that the proposed method not only excels in accuracy for ultra-short segments but also scales efficiently to longer video lengths, making it suitable for real-world applications with constrained computational resources.

The remainder of this paper is structured as follows: Section 2 reviews existing literature in audio and visual deepfake detection, emphasizing the limitations and computational challenges of current approaches. Section 3 describes our hand-crafted methods for audio and visual deepfake detection. Section 4 details the experimental design and datasets used, presents the evaluation of our proposed method, and provides a comprehensive analysis of the results, followed by the conclusion and a discussion of future research directions.

2 Related works

Previous studies in deepfake detection have charted a range of technical methodologies for analyzing audio and visual components, employing strategies that extend from detailed low-level hand-crafted features to advanced high-level neural network architectures [2, 6].

In the realm of combating audio deepfakes, the use of Mel-Frequency Cepstral Coefficients (MFCC) is widespread, analyzed through 2D neural network architectures such as VGG16 and EfficientNet [4, 5, 15]. These methods are promising, yet they demand significant computational resources and have not been examined when dealing with ultra-short audio samples. We propose to explore shallow learning methods which require less resources in this article.

Pianese et al. [16] employed a distinct approach for audio deepfake detection by harnessing the Person of Interest (POI) concept, echoing the core ideas of speaker verification systems [17]. Their strategy focuses on assessing the similarity between the voice under scrutiny and a pre-existing reference collection of the claimed identity, employing two unique non-supervised methods: centroid-based and maximum-similarity testing [16]. The primary challenge of this method is its reliance on a comprehensive reference set for each identity analyzed. Our methodology aims to address this limitation by proposing an alternative approach that minimizes the need for such extensive reference collections, thereby enhancing the practicality and scalability of audio deepfake detection.

Visual-based deep fake detection methods have seen a diverse range of strategies. Some leverage 3D networks for in-depth sequence analysis [6, 18]. Zhou et al. [18] proposed a system that exploits the intrinsic synchronization between audio and visual elements, particularly focusing on the lips' movement and corresponding audio at the word level. Employing a multimodal neural network, they experimented with three types of fusion mechanisms based on attention mechanisms [18]. However, it's noteworthy that [18] employed 3D networks for video feature extraction and attention mechanisms, processes that are known for their high resource demands.

On the other side, alternative approaches to deepfake video detection primarily employ image-based methods, placing a significant focus on facial features as discussed in references

[4, 6]. In other words, these methods involve the independent analysis of each frame within the video input by the network, culminating in a conclusive decision through hard or soft voting [4, 6]. While these techniques demand fewer resources compared to their 3D counterparts, they lack the incorporation of temporal information essential for thoroughly examining videos.

In light of these limitations, our methodologies designed for audio-visual detection rely on shallow learning and only the lips part of the face, demanding fewer resources and demonstrating a comprehensible and interpretable nature. Our visual deepfake detection is based on sequence level, in other words, we exploit the temporal anomalies with less computational cost.

3 Methodological framework

3.1 Proposed methodology pipeline

The global view of the proposed method is depicted in Fig. 1 and will be detailed in this section. After some standard pre-processing steps, we apply a dual-phase method for deepfake detection. Initially, we extract features separately from the video's facial imagery and the corresponding audio. We then proceed to independently classify the authenticity of both video and audio. Finally, we bring together these independent classifications at a decision level, harnessing the temporal information inherent in both modalities.

3.1.1 Pre-processing: lips part selection

As illustrated in the pre-processing module of our pipeline in Fig. 1, our methodology focuses on the precise isolation of the lip region from the talking face within the video frame. We employ Google's Mediapipe tool [19], an existing algorithm specifically designed for lip detection and known for its proficiency in identifying facial landmarks. Mediapipe attains a lip detection precision of 94.4%. The majority of errors arises from faces in extreme poses, where the technique struggles to locate the lip position.

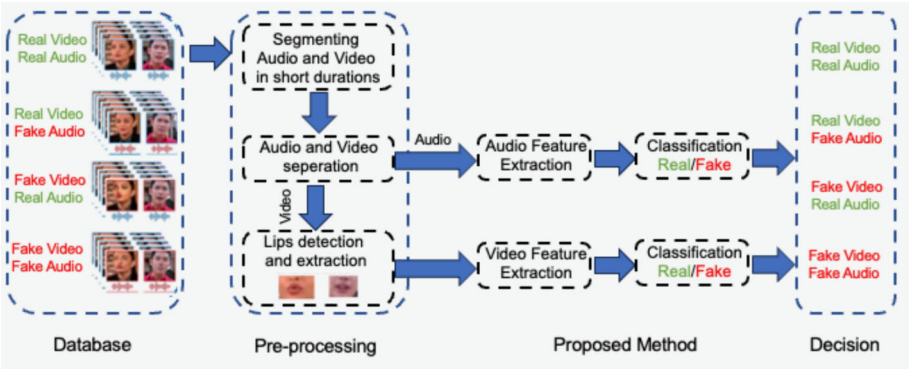

Fig. 1 Proposed pipeline for short utterances manipulations detection in videos

3.2 Methodology

We detail our method, which is based on new hand-crafted techniques for extracting features from both audio and video. This is shown in the proposed method module of our pipeline in Fig. 1.

3.2.1 Audio feature extraction

Our approach aims to refrain from making assumptions regarding the specific frequency domain impacted by deep fake alterations in the audio signal. As a result, we have chosen to employ a multi-scale analysis for the development of our audio features. The literature presents various multi-scale decomposition methodologies, such as Mel Frequency Cepstrum Coefficients (MFCC) and Mel-Spectrograms, along with the Wavelet Scattering Transform (WST). For the purposes of this study, we have selectively employed the WST. The selection was based on WST's proven efficacy in capturing essential frequency characteristics of audio signals, which are crucial for identifying the subtle alterations introduced by deepfake techniques.

Consider an audio signal $x \in \mathbb{R}^{t_x}$ sampled at $16kHz$ and maximally normalized, where $t_x \in \mathbb{N}^*$. We partition this signal into its positive $x_p \in \mathbb{R}^{t_x}$ and negative $x_n \in \mathbb{R}^{t_x}$ components as follows:

$$\begin{cases} x_p = ReLU(x) \\ x_n = ReLU(-x) \\ x = x_p - x_n \end{cases} . \tag{1}$$

Let us introduce Ψ, denoting the Wavelet Scattering Transform (WST) designed for one-dimensional signals as presented in existing works [20, 21]. The fundamental idea behind WST is the iterative application of the wavelet transform [22] coupled with a modulus operation serving as a non-linear function and subsequently averaging the result through a Gaussian filter. This transformation technique is subject to multiple hyper-parameters, including the invariance scale (window length), the transform depth, and the quality factors which determine the number of wavelets per octave. This WST is suited for our purpose as it gives a frequency characterization of an audio.

A preliminary assessment of disparities between an authentic audio and its cloned using the audio deepfake generation method described in [12]. The two scattergrams (c-d) in the Fig. 2 illustrate variations in the WST between the upper and lower samples of the two audio sources, while the scattergrams (a-b) of the two signals without decomposition have small dissimilarities. The two plots in (c-d) shows that the fake audio is characterized by predominantly negative values with a positive value at $0Hz$, whereas the authentic one primarily consists of positive values and exhibits null values at $0Hz$.

In line with the existing research on audio processing [20], we have set the window length at $32ms$ and chosen four layers, which are adequate for capturing the majority of the signal's energy [23]. With this setup and a sampling frequency of $16kHz$, we derive 153 WST coefficients. Subsequently, we compute the WST Ψ of our signal x and both negative x_n and positive x_p components. This results in the output matrices $\mathbf{X}_n, \mathbf{X}_p, \mathbf{X} \in \mathbb{R}^{C \times T_x}$, where ($C \in \mathbb{N}^*$) denotes the number of WST coefficients, ranging from zero order up to the fourth order, here $C = 153$.

The core concept of our method hinges on using transformed matrices to calculate the lower bounds of the Pearson correlation coefficient, focusing on the interactions between $\mathbf{X}_p$, $\mathbf{X}_n$, $\mathbf{X}$, and $\mathbf{X}_p - \mathbf{X}_n$. This approach is integral to our analysis, as it leverages only the

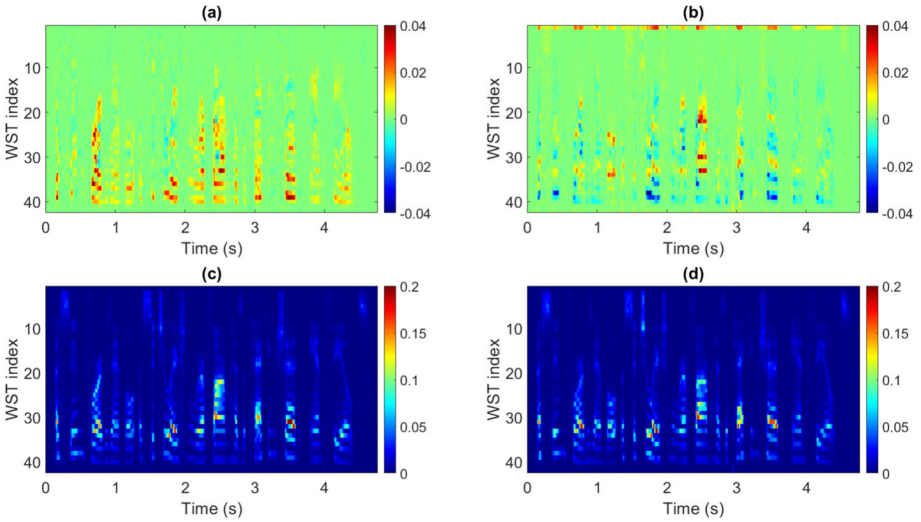

Fig. 2 (a) The difference between the WST of positive and negative sample of Real audio. (b) The difference between the WST of positive and negative sample of the fake audio represented in (a). (c) The WST of the same real signal. (d) The WST of the fake signal. Only the zero-th and the first order WST are presented and the fake signal was generated using the method described in [12]

inherent elements of the audio signal, thus eliminating the need for external benchmarks. Mathematically, the framework of our proposed method can be described as follows:

$$\begin{aligned}\mathbf{S}(c) = \min \Big(& \rho(\mathbf{X_n}(c), \mathbf{X_p}(c) - \mathbf{X_n}(c)), \\ & \rho(\mathbf{X}(c), \mathbf{X_p}(c) - \mathbf{X_n}(c)), \\ & \rho(\mathbf{X_p}(c), \mathbf{X_p}(c) - \mathbf{X_n}(c)) \Big),\end{aligned} \tag{2}$$

where $c = 1, .., C$ is the channel index of the WST coefficients, $\mathbf{S} = (\mathbf{S}(1), ..., \mathbf{S}(C)) \in \mathbb{R}^C$ is the features vector to detect a fake audio and ρ is the Pearson coefficient across temporal axis.

We employ the Pearson Correlation Coefficient to elucidate the interrelationships among the distinct components of the audio signal. Analyzing the correlation patterns between the positive and negative aspects of the WST coefficients allows us to detect inconsistencies characteristic of deepfake manipulations. Such irregularities are indicative of alterations, given that genuine audio signals typically demonstrate a stable and consistent correlation pattern, which is often disrupted in manipulated audio.

3.2.2 Video feature extraction

Following the exposition of our audio deepfake detection methodology, we now introduce the algorithm we designed to discern the authenticity of visual sequences. This method hinges on both the spatial and temporal attributes of a video, focusing primarily on detecting any anomalies or inconsistencies in the motion or appearance of a speaking subject in a video with a zero head pose. Given that deepfake generation algorithms exhibit difficulties in accurately

replicating the high-frequency characteristics inside the mouth area such as the teethes [24, 25], our method emphasizes the high-frequency components of the video signal.

Let $V \in \mathbb{R}^{t_v \times 3 \times N_x \times N_y}$ be a video sequence depicting the lip movements of a subject speaking without head pose. The video frames are converted to gray-scale for analysis. For each temporal instance $t = 1, ..., t_v$, we take the fourth-order spatial derivative of the frame $V(t) \in \mathbb{R}^{N_x \times N_y}$ with respect to both x and y axes to yield $\frac{\partial^4 V}{\partial x^2 \partial y^2} \in \mathbb{R}^{t_v \times 3 \times N_x \times N_y}$. This operation effectively shifts the energy of the signal toward the high-frequency domain.

Subsequently, we apply the two-dimensional discrete cosine transform (DCT2), denoted as Φ, to $\frac{\partial^4 V}{\partial x^2 \partial y^2}$, resulting in the frequency representation of the rate of intensity change across frames. Drawing inspiration from existing work [26], which leverages temporal variations in Pearson correlation coefficient to identify talking regions, we compute this coefficient for successive frames. Mathematically, the process can be described as follows

$$\begin{cases} \Phi(\frac{\partial^4 V}{\partial x^2 \partial y^2})(t) = (v_{i,j}(t))_{1 \le i \le N_x, 1 \le j \le N_x} \in \mathbb{R}^{N_x \times N_y} \\ \rho(t) = pearson(\Phi(\frac{\partial^4 V}{\partial x^2 \partial y^2})(t), \Phi(\frac{\partial^4 V}{\partial x^2 \partial y^2})(t-1)) \in \mathbb{R}, \ t = 2, ..., t_v \end{cases}, \tag{3}$$

where $\Phi\left(\frac{\partial^4 V}{\partial x^2 \partial y^2}\right)(t)$ represents the high-frequency components obtained from the DCT2 transform Φ of the frame $V(t)$, and $\rho(t)$ denotes the Pearson correlation coefficient between two successive frames $V(t)$ and $V(t+1)$. To perform an analysis over a specific duration or the entire length of the video, we construct a scatter plot using the temporal mean μ and the standard deviation σ of the ρ values.

4 Experimental analysis

In this section, we detail the experimental procedures used to validate our method against existing state-of-the-art (SOTA) solutions [5, 6, 16, 27] for both audio and visual deepfake detection. Our validation process includes experiments with two reputable datasets which would be explained in the following.

4.1 Datasets

We introduce benchmark and reference datasets that are widely accepted and commonly employed in the field of deepfake detection to assess the effectiveness of our proposed methods. Note that this may not be an exhaustive list.

4.1.1 FakeAVCeleb

The FakeAVCeleb dataset serves as a comprehensive and developed resource for audio-visual deepfake detection [28]. Originating from the frequently cited VoxCeleb2 multimodal corpus [29], FakeAVCeleb stands out for its frame rate of 25 fps and an average video duration of approximately 7.8 seconds. Employing an array of SOTA deepfake generation methods such as lip-syncing and face-swapping technologies [30–33], the authors have fabricated videos that pose a realistic and formidable challenge in discerning their authenticity.

To bolster its applicability, the dataset has been curated to offer ethnic and gender diversity, thus paving the way for equitable and representative evaluation. Structurally, it comprises four distinct categories of audio-visual data:

- A collection of 500 videos with real-audios and real-videos ($R_v R_a$),
- Another set of 500 videos where the audio has been cloned or generated from a speech text while the visual content remains authentic ($R_v F_a$),
- A larger set of 9, 700 videos featuring real audio coupled with manipulated visual content ($F_v R_a$),
- Finally, a comprehensive set of 10, 800 videos in which both the visual and audio components are synthetic ($F_v F_a$).

The FakeAVCeleb dataset introduces three distinct classes of manipulated content, alongside a dedicated category for wholly authentic videos. For testing our method, videos from this dataset have been chunked into segments ranging from 200 ms to 600 ms in duration.

4.1.2 DeepfakeTIMIT

The DeepfakeTIMIT dataset emerges as an essential benchmark in the realm of deepfake detection, featuring an average video length of approximately 4.25 seconds. Originating from the VidTIMIT database [34], it employs GAN-based face-swapping techniques to generate manipulated content [35, 36]. This corpus is bifurcated into two main categories: the first contains 320 clips with manipulated visuals yet authentic audio ('Fake Video - Real Audio'), while the second consists of 430 clips preserving both the original video and audio ('Real Video - Real Audio').

Our investigation is primarily aimed at the low-quality segment of the DeepfakeTIMIT collection. This specific tier provides invaluable insights into the robustness of deepfake detection methods when operating under suboptimal conditions. It thereby furnishes a more nuanced understanding of algorithmic performance constraints, especially considering that diminished visual and auditory quality exacerbate the challenges of distinguishing authentic content from fabricated instances. For the purposes of our study, we implemented a preprocessing step that involved the removal of silent segments from the video clips in the dataset. This ensured that the subsequent segmentation into durations ranging from 200 ms to 600 ms resulted in utterances containing at least a word, thereby guaranteeing meaningful audio-visual data for analysis.

4.2 Experimental strategy

4.2.1 The first experiment

The first experiment assesses the capability of our model in detecting deepfakes within short utterances. For this purpose, we segmented the datasets into frames of 200 to 600 milliseconds, with a 50% overlap between consecutive frames. This approach ensures that the dataset is both robust and continuous, enabling a thorough analysis of the model's performance.

The data was split into training, validation, and test sets using a 60-20-20% ratio, which is a standard practice to ensure a balanced evaluation of the model. We utilized a Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel to classify the features extracted as described in (2). The RBF kernel was chosen for its effectiveness in handling non-linear patterns within the data. During the training and validation phases, the regularization coefficient was carefully tuned to 0.01, optimizing the model's performance by controlling the balance between model complexity and classification accuracy.

Table 1 Performance of our proposed method on the whole audio length modality of FakeAVCeleb compared to the SOTA methods (Real: 500 videos from $R_v R_a$, Fake: 500 videos from $F_v F_a$). Bold entries indicate the best performance

Model	Accuracy
MFCC + XceptionNet [5]	76.6%
Mel-Spectrograms + DST-Net [4]	97.51%
MFCC + DST-Net [4]	88.5%
X-vectors + SVM [7, 16, 37]	**99.98%**
ECAPA-TDNN + SVM [16]	99.97%
Our hand-crafted method	99.83%

4.2.2 The second experiment

The second experiment evaluates the effectiveness of our model in detecting deepfakes within the visual channel of videos, both in short utterances and across entire video frames. Similar to the approach used for audio signals, we segmented the video datasets into short frames ranging from 200 to 600 milliseconds. This segmentation allows us to focus on brief visual segments, which are particularly challenging in deepfake detection.

We applied the same cross-validation strategy as in the first experiment to ensure consistent and rigorous evaluation of the model's performance in the visual domain. By maintaining this uniform approach, we were able to accurately assess the model's ability to detect visual manipulations in both short and extended video sequences.

4.3 Exprimental results

4.3.1 Experiment 1

Our experimental framework, as detailed in (2), is designed to evaluate the effectiveness of our hand-crafted method for detecting deepfake audio. We rigorously tested our approach using the FakeAVCeleb dataset [28], which predominantly consists of synthesized audio samples that pose significant challenges for detection algorithms. The results, shown in Tables 1 and 2, indicate that our method not only surpasses SOTA deep learning techniques in accurately identifying deepfakes within short audio utterances but also maintains a high level of performance when applied to the full duration of the audio.

4.3.2 Experiment 2

We present the empirical evaluation of our visual deepfake detection approach, formulated as per (3). The focus of this investigation is restricted to the lip region of the subject, contingent

Table 2 Accuracy of our method in (2) on short utterances and full audio modality in FakeAVCeleb, compared to SOTA methods (FC denotes Fully Connected layers followed by a softmax function). Bold entries indicate the best performance

	X-vectors + SVM [7, 16, 37]	ECAPA-TDNN + SVM [16]	Our Method
Whole video	**99.98%**	99.97%	99.83%
600 ms	96.35%	89.03%	**96.83%**
200 ms	85.62%	75.13%	**87.02%**

upon the maintenance of a zero-degree head pose throughout the recording session. Consequently, this evaluation is exclusively conducted on the DeepfakeTIMIT dataset.

Our aim is to track the temporal behavior of the Pearson coefficient defined in (3), then we decide the authenticity of the visual sequence. We considered 20 fake videos from DeepfakeTIMIT and 20 real videos from VidTIMIT. We then split every video into chunks of $600ms$ and an overlap of 50%, we obtain 279 fake and 279 real videos. To imprint the temporal dynamics of our parameter in (3), we choose the mean and the standard deviation.

Figure 3 provides a visual demonstration of our method's effectiveness. The figure presents a scatter plot that includes various measurements, such as the average and variation (standard deviation) of the Pearson coefficient. This visual arrangement effectively separates real videos from fake ones. The top-left part of the plot illustrates the application of the DCT coefficient to unaltered video frames. In contrast, the top-right part shows the results of applying the first spatial derivative to the video frames. Notably, the bottom-left section of the plot underscores our method's strong capability in distinguishing real from fake videos. Based on this last observation, the authenticity of the video can be done by setting a threshold on the inverse coefficient of variation C_v^{-1} plotted at the bottom right of the Fig. 3. For various values of the threshold C_v^{-1} ranging from -1 to 10, we plot the receiver operating characteristic (ROC) curve (not shown) and the optimal threshold to distinguish genuine and fake videos giving a maximum detection accuracy 98.39% was found at 1.

Setting the threshold at 1 allowed us to achieve a remarkable detection accuracy of 99.87% across the entire VidTIMIT and DeepfakeTIMIT video datasets. This result strongly validates the effectiveness of our method in reliably distinguishing between authentic and manipulated content. To further highlight the capability of our approach, we conducted a performance comparison with SOTA methods, as detailed in Tables 3 and 4.

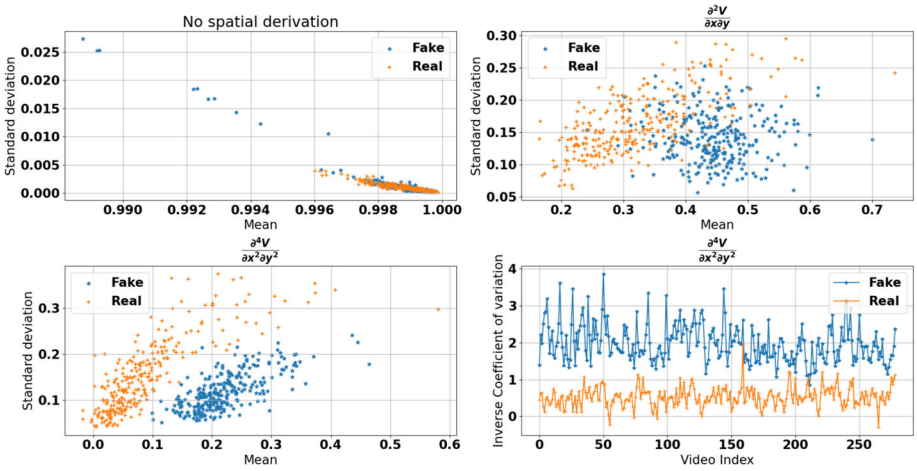

Fig. 3 Analysis of low-quality 20 fake and genuine videos split into $600ms$ with an overlap of 50% using our method. Top-left: Mean and standard deviation without spatial derivation. Top-right: After second-level spatial derivation. Bottom-left: After fourth-level spatial derivation. Bottom-right: Temporal inverse coefficient of variation of videos from DeepfakeTIMIT and VidTIMIT

Table 3 Performance of the SOTA and our proposed method on whole length visual sequence from low-quality videos of DeepfakeTIMIT. Bold entries indicate the best performance

Model	Accuracy
XceptionNet (Image level detection) [28]	65.98%
I3D (Sequence level detection) [6]	96.38%
Our hand-crafted method	**99.87%**

4.4 Comparison and discussion

The methodology we have formulated for the detection of fake audio signals exhibits numerous distinct advantages, with its performance and generalizability being particularly noteworthy.

Unlike some methods that depend on a predefined set of reference speakers, as discussed in [16], our approach distinguishes itself by removing the need for these comparisons and trainable parameters. This distinction is clearly demonstrated in Fig. 4, which shows the Pearson coefficients across WST channels for both authentic and fake audios produced using Text-To-Speech (TTS) techniques [12].

Moreover, our technique demonstrates a robust capacity to handle short audio utterances, maintaining satisfactory performance which incrementally improves with the length of the audio sample. This trend is particularly advantageous for scenarios commonly faced in the real world, where audio clips are often brief.

On the visual modality, the robustness of our visual-level deepfake detection method is underscored by its efficacy under a range of challenging conditions. Specifically, the technique exhibits high performance even when subjected to low-quality videos. What sets our method apart is its unique focus on the lip region for detection-a region notoriously difficult for traditional deepfake detection methods to analyze. This focus doesn't merely serve to fill a gap in existing methodologies; it provides our system with a marked advantage over SOTA deep learning-based approaches. Moreover, this is achieved with very limited number of hyperparameters, which significantly reduces the computational overhead and simplifies the implementation.

The crux of our method is its strategic utilization of high-frequency spatial energy patterns. By pushing spatial energy towards these higher frequencies, our system is able to significantly amplify the contrast between real and fake visual sequences, a fact that is empirically supported by the results. This approach not only serves to enhance detection capabilities but also fortifies our system's adaptability. This adaptability is further substantiated by the system's performance metrics under scenarios involving short utterances, ranging in length from $200ms$ to $600ms$. Even under these non-ideal conditions, the system was able to maintain a reasonable performance level.

However, it's important to note that the scope of the current study did not extend to evaluating the technique's robustness against videos with varying luminosity or noise levels.

Table 4 Accuracy comparison of SOTA performance with our proposed approach using visual sequences from low-quality DeepfakeTIMIT videos over various time segments. Bold entries indicate the best performance

	XceptionNet [28]	Multi-view video CNN [27]	Our Method
Whole video	65.98%	98.43%	**99.87%**
600 ms	68.19%	99.12%	**99.39%**
200 ms	61.98%	99.36%	**99.88%**

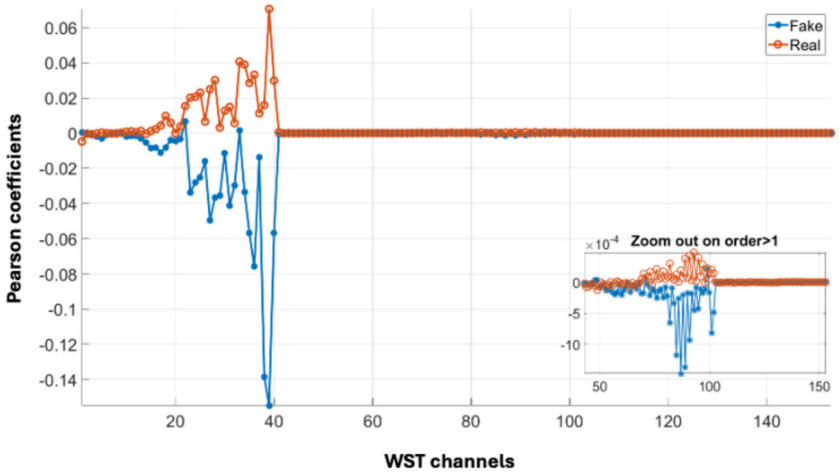

Fig. 4 Pearson coefficient value (y-axis) described in (2) per each channel (x-axis) (orange circles refers to the real audio and the blue star stands to the fake audio)

Given the demonstrated performance of the method, this remains a crucial avenue for future research, to comprehensively assess the system's applicability under a wider array of real-world conditions. As long as the luminosity does not change temporally or spatially in temporal or spatial frequency ranges which are discriminant between real and fake this should not have impact on the performance of our method.

5 Conclusion and future work

In this study, we introduced a full-pipeline approach to detect fake audio and video content, leveraging hand-crafted features for audio and distinctive visual cues, particularly in the lip region. Our method demonstrates high interpretability and computational efficiency, achieving robust performance on the FakeAVCeleb and DeepfakeTIMIT datasets. This unified strategy underscores the synergy between auditory and visual elements, reflecting a comprehensive stance against the rising tide of deepfake technologies. The robustness of our approach is evident as it maintains performance metrics across various conditions without requiring a reference dataset or a complex train-test split, marking a significant advancement over existing deep learning methods.

Future work will focus on enhancing this synergy by integrating audiovisual features at multiple levels of abstraction, potentially leading to a more resilient detection mechanism against sophisticated deepfake manipulations. We also aim to optimize performance for short utterances and expand the model's robustness to accommodate diverse environmental variables such as noise and lighting conditions. These efforts will contribute to ensuring the authenticity and trustworthiness of digital media in an era of rapidly advancing synthetic content technologies.

Supplementary information

Data sharing not applicable to this article as no datasets were generated during the current study. The article references two databases that are publicly accessible.

Author Contributions Conceptualization, A.M.; Methodology, A.M., D.R. and P.R.; Software, A.M.; Validation, D.R. and P.R.; Formal analysis, D.R. and P.R.; Investigation, P.R.; Writing-original draft, A.M., D.R. and P.R.; Writing-review & editing, A.M., D.R. and P.R.; Visualization, A.M.; Supervision, D.R. and P.R. All authors have read and agreed to the published version of the manuscript.

Funding Open access funding provided by Université d'Angers. The Ph.D. grant of Abderrazzaq Moufidi is funded by Angers Loire Metropole (ALM).

Data Availability The article references two databases that are publicly accessible.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Zhang T (2022) Deepfake generation and detection, a survey. Multimed Tools Appl 81(5):6259–6276
2. Huang T-h, Lin J-h, Lee H-y (2021) How far are we from robust voice conversion: A survey. In: 2021 IEEE Spoken Language Technology Workshop (SLT), pp 514–521. IEEE
3. Kandasamy V, Hubálovskỳ Š, Trojovskỳ P (2022) Deep fake detection using a sparse auto encoder with a graph capsule dual graph cnn. PeerJ Comput Sci 8:953
4. Ilyas H, Javed A, Malik KM (2023) Avfakenet: A unified end-to-end dense swin transformer deep learning model for audio-visual deepfakes detection. Appl Soft Comput 136:110124
5. Khalid H, Kim M, Tariq S, Woo SS (2021) Evaluation of an audio-video multimodal deepfake dataset using unimodal and multimodal detectors. In: Proceedings of the 1st workshop on synthetic multimedia-audiovisual Deepfake generation and detection, pp 7–15
6. Zi B, Chang M, Chen J, Ma X, Jiang Y-G (2020) Wilddeepfake: A challenging real-world dataset for deepfake detection. In: Proceedings of the 28th ACM international conference on multimedia, pp 2382–2390
7. Salvi D, Liu H, Mandelli S, Bestagini P, Zhou W, Zhang W, Tubaro S (2023) A robust approach to multimodal deepfake detection. J Imaging 9(6):122
8. Ling J, Tan X, Chen L, Li R, Zhang Y, Zhao S, Song L (2022) StableFace: analyzing and improving motion stability for talking face generation
9. Dagar D, Vishwakarma DK (2022) A literature review and perspectives in deepfakes: generation, detection, and applications. Int J Multimed Inf Retrieval 11(3):219–289
10. Tolosana R, Romero-Tapiador S, Vera-Rodriguez R, Gonzalez-Sosa E, Fierrez J (2022) Deepfakes detection across generations: Analysis of facial regions, fusion, and performance evaluation. Eng Appl Artif Intell 110:104673
11. Thing VL (2023) Deepfake detection with deep learning: Convolutional neural networks versus transformers. arXiv e-prints, 2304
12. Jiang Z, Liu J, Ren Y, He J, Zhang C, Ye Z, Wei P, Wang C, Yin X, Ma Z et al (2023) Mega-tts 2: Zero-shot text-to-speech with arbitrary length speech prompts. arXiv:2307.07218

13. Masood M, Nawaz M, Malik KM, Javed A, Irtaza A, Malik H (2023) Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. Appl Intell 53(4):3974–4026
14. Seow JW, Lim MK, Phan RC, Liu JK (2022) A comprehensive overview of deepfake: Generation, detection, datasets, and opportunities. Neurocomputing 513:351–371
15. Afchar D, Nozick V, Yamagishi J, Echizen I (2018) Mesonet: a compact facial video forgery detection network. In: 2018 IEEE international Workshop on Information Forensics and Security (WIFS), pp 1–7. IEEE
16. Pianese A, Cozzolino D, Poggi G, Verdoliva L (2022) Deepfake audio detection by speaker verification
17. Desplanques B, Thienpondt J, Demuynck K (2020) Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. arXiv:2005.07143
18. Zhou Y, Lim S-N (2021) Joint audio-visual deepfake detection. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 14800–14809
19. Lugaresi C, Tang J, Nash H, McClanahan C, Uboweja E, Hays M, Zhang F, Chang C-L, Yong M, Lee J, Chang W-T, Hua W, Georg M, Grundmann M (2019) Mediapipe: A framework for perceiving and processing reality. In: Third workshop on computer vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR). https://research.google/pubs/pub48292/
20. Andén J, Mallat S (2014) Deep scattering spectrum. IEEE Trans Signal Process 62(16):4114–4128
21. Mallat S (2012) Group invariant scattering. Commun Pure Appl Math 65(10):1331–1398
22. Stephane M (1999) A wavelet tour of signal processing. Elsevier
23. Rasti P, Ahmad A, Samiei S, Belin E, Rousseau D (2019) Supervised image classification by scattering transform with application to weed detection in culture crops of high density. Remote Sens 11(3):249
24. Garrido P, Valgaerts L, Sarmadi H, Steiner I, Varanasi K, Perez P, Theobalt C (2015) Vdub: Modifying face video of actors for plausible visual alignment to a dubbed audio track. In: Computer Graphics Forum, vol 34, pp 193–204. Wiley Online Library
25. Verdoliva L (2020) Media forensics and deepfakes: an overview. IEEE J Sel Top Signal Process 14(5):910–932
26. Cutler R, Davis L (2000) Look who's talking: Speaker detection using video and audio correlation. In: 2000 IEEE international conference on multimedia and expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No. 00TH8532), vol 3, pp 1589–1592. IEEE
27. Moufidi A, Rousseau D, Rasti P (2023) Attention-based fusion of ultrashort voice utterances and depth videos for multimodal person identification. Sensors 23(13):5890
28. Khalid H, Tariq S, Kim M, Woo SS (2021) Fakeavceleb: A novel audio-video multimodal deepfake dataset. arXiv:2108.05080
29. Chung JS, Nagrani A, Zisserman A (2018) Voxceleb2: Deep speaker recognition. arXiv:1806.05622
30. Korshunova I, Shi W, Dambre J, Theis L (2017) Fast face-swap using convolutional neural networks. In: Proceedings of the IEEE international conference on computer vision, pp 3677–3685
31. Nirkin Y, Keller Y, Hassner T (2019) Fsgan: Subject agnostic face swapping and reenactment. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 7184–7193
32. Prajwal K, Mukhopadhyay R, Namboodiri VP, Jawahar C (2020) A lip sync expert is all you need for speech to lip generation in the wild. In: Proceedings of the 28th ACM international conference on multimedia, pp 484–492
33. Jia Y, Zhang Y, Weiss R, Wang Q, Shen J, Ren F, Nguyen P, Pang R, Lopez Moreno I, Wu Y et al (2018) Transfer learning from speaker verification to multispeaker text-to-speech synthesis. Advances in neural information processing systems, vol 31
34. Sanderson C (2001) Vidtimit audio-video dataset. Zenodo
35. Korshunov P, Marcel S (2018) Deepfakes: a new threat to face recognition? assessment and detection. arXiv:1812.08685
36. Sanderson C, Lovell BC (2009) Multi-region probabilistic histograms for robust and scalable identity inference. In: Advances in biometrics: third international conference, ICB 2009, Alghero, Italy, June 2-5, 2009. Proceedings 3, pp 199–208. Springer
37. Snyder D, Garcia-Romero D, Sell G, Povey D, Khudanpur S (2018) X-vectors: Robust dnn embeddings for speaker recognition. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 5329–5333. IEEE