Are you Really Alone? Detecting the use of Speech Separation Techniques on Audio Recordings

Davide Salvi, Mirco Pezzoli, Sara Mandelli, Paolo Bestagini, Stefano Tubaro
Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB), Politecnico di Milano
Piazza Leonardo Da Vinci 32, 20133 Milano, Italy
{name.surname}@polimi.it

2023 IEEE International Workshop on Information Forensics and Security (WIFS) | 979-8-3503-2491-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/WIFS58808.2023.10374717

***Abstract*—The pervasive influence of digital media has brought about new challenges in verifying the authenticity and integrity of audio recordings. The ease of editing and altering audio has raised concerns regarding the potential malicious use of speech separation techniques, where multiple speakers' voices can be extracted from a mixed recording. In light of these emerging threats, the need for robust forensic detectors that can identify the presence of speech separation forgeries becomes increasingly crucial. In this paper, we propose a novel forensic detector designed to discern between original single-speaker speech recordings and those obtained using speech separation techniques applied to audio recordings containing multiple speakers. Leveraging the power of Convolutional Neural Networks (CNNs), we explore the efficacy of different Short-Time Fourier Transform (STFT) representations in tackling the task. While many conventional approaches in the literature employ the audio spectrogram (i.e., the STFT magnitude) as input for CNNs, our study explores the use of the STFT real and imaginary parts, as well as the STFT magnitude and phase. In doing so, we ensure the preservation of all essential information embedded within the speech signal. Results show that the proposed signal representation improves over the sole use of the spectrogram. Moreover, the proposed approach is able to generalize to datasets and speech separation techniques never seen in training. Finally, our proposed detector shows promising results on preliminary experiments performed on synthetically generated audio tracks.**

***Index Terms*—Forensics, Audio, Speech, Speech Separation**

I. Introduction

The widespread impact of digital media has revolutionized the creation and dissemination of audio content. Nonetheless, this technological advancement has also opened avenues for potential audio manipulation and tampering, raising concerns regarding the integrity and authenticity of online audio recordings. Among the several existing audio manipulation methods, speech separation techniques have emerged, enabling the extraction of individual speakers' voices from mixed audio tracks [1]–[3]. These introduced new challenges in audio forensics and the detection of manipulated speech recordings.

Instances of malicious use of speech separation techniques underscore the critical need for effective forensic detectors to combat audio manipulation. For instance, speech separation techniques can be utilized to fabricate convincing voice impersonations. By isolating and manipulating individual voices, perpetrators can create audio recordings of individuals saying things they never said. This deceptive practice can be employed to frame innocent individuals or spread false information, leading to reputational damage or even legal ramifications.

Another area of concern is the potential manipulation of audio evidence in legal proceedings. By employing speech separation, an individual could selectively edit or remove specific speech segments from a mixed audio recording, altering the context and distorting the original meaning of the conversation. Such tampering could potentially undermine the reliability and trustworthiness of audio evidence, jeopardizing the fairness and integrity of legal proceedings.

In the last few years, the multimedia forensic community has focused on developing techniques to expose a wide range of audio manipulations. Multiple approaches have been proposed to detect synthetic speech and audio deepfakes, encompassing diverse techniques [4]. Specific methods concentrate on identifying artifacts left by the speech generators [5]–[7], others emphasize the examination of higher-level aspects [8], [9], and additional strategies focus on audio signal verification by means of environmental traces left on the recording [10], [11]. An entire branch of approaches is dedicated to the analysis of traces left by the acquisition device [12], [13]. Additionally, extensive research has been conducted on audio splicing detection and localization [14]–[16]. However, to the best of our knowledge, little to no work has been conducted to detect the use of speech separation techniques.

In this paper we present a novel forensic detector that aims to accurately discern between original single-speaker speech recordings and those that have undergone speech separation techniques. The proposed detector leverages the capabilities of Convolutional Neural Networks (CNNs) and explores the effectiveness of various time-frequency representations of the speech signal.

Conventional audio forensic methods employed in speech analysis typically rely on the audio spectrogram as input for CNNs [17]. However, the spectrogram represents only a partial view of the information contained within the audio recording. This limitation arises from the fact that the spectrogram is derived from the Short Time Fourier Transform (STFT), which is a complex-valued representation. In contrast, the spectrogram discards the complex component of the STFT, capturing only the magnitude information.

In our research, we address this limitation by exploring a novel approach that preserves the entire information embedded within the audio signal. We investigate the use of both the real

and imaginary parts of the STFT, as well as its magnitude and phase. By mapping the complex STFT representation into two real signals, we retain the comprehensive information present in the original audio recording. By employing this comprehensive representation of the STFT, our proposed technique offers a more holistic view of the speech signals under analysis. This approach allows the CNN classifier to capture fine-grained details and subtle variations present in the audio, potentially enhancing the performance of the forensic detector in distinguishing between original single-speaker speech recordings and those generated through speech separation techniques.

II. Background on Speech Separation

In today's technological landscape, speech analysis has become increasingly prevalent and crucial in numerous scenarios, particularly in the realm of voice assistants. As these AI-powered devices continue to gain popularity, their ability to accurately understand speech becomes crucial. However, real-life audio recordings are rarely noise-free and frequently involve mixtures of speakers, making it challenging to extract meaningful information from isolated individual voices.

In this context, speech separation techniques are invaluable. These methods can effectively separate and enhance speech signals from complex mixtures, enabling improved speech analysis and enhancing the overall performance in speech-related applications [1], [2].

Speech Separation Problem. Speech separation refers to the process of extracting individual speech sources from a given mixture of audio signals, characterized by the presence of multiple, possibly overlapping, speakers [1].

With reference to Fig. 1, let us consider an audio track $\mathbf{x}_{\text{mix}}$ containing the recording of N speakers speaking simultaneously. The speech signal belonging to the n-th speaker is $\mathbf{x}^n_{\text{speech}}$. In a noiseless environment, the mixture audio recording can be defined as

$$\mathbf{x}_{\text{mix}} = \sum_{n=0}^{N-1} \mathbf{x}^n_{\text{speech}}. \tag{1}$$

The goal of a speech separation system is to estimate each individual speech track (or one specific reference speaker track) starting from the mixture. Formally, let us define as $\hat{\mathbf{x}}^n_{\text{speech}}$ the estimate of $\mathbf{x}^n_{\text{speech}}$. The output of a speaker separation system is

$$\hat{\mathbf{x}}^0_{\text{speech}},\ \hat{\mathbf{x}}^1_{\text{speech}},\ \ldots,\ \hat{\mathbf{x}}^{N-1}_{\text{speech}} = \mathcal{S}(\mathbf{x}_{\text{mix}}), \tag{2}$$

where $\mathcal{S}$ represent the speech separation operator. In case a reference target speaker is selected (e.g., $n = 0$), the speech separation system may return just one speaker's estimate (e.g., $\hat{\mathbf{x}}^0_{\text{speech}}$), rather than all of them.

Speech Separation Methods. Several speech separation methods have been proposed in the literature. Depending on the number of microphones used during the recording stage, it is possible to split speech separation methods into two main classes: i) array-based (i.e., multi-microphone) techniques; ii) monaural (i.e., single microphone) techniques.

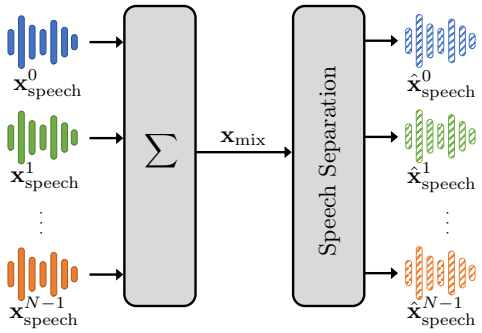


Fig. 1: Representation of the speech separation problem. N speech signals $\mathbf{x}^n_{\text{speech}}$ are simultaneously recorded in a single mix $\mathbf{x}_{\text{mix}}$. A speech separation techniques should be able to demix the speech signals obtaining each speech estimate $\hat{\mathbf{x}}^n_{\text{speech}}$.

The use of microphone arrays opens the doors to spatial audio processing methods. It is possible to apply beamforming or spatial filtering to attenuate interference from other directions, thus isolating a target speaker of interest from the surrounding auditory environment [3], [18]. Unfortunately, these techniques can be hardly used in scenarios in which the speech signal under analysis is acquired with an unknown setup, possibly using a single microphone.

Monaural methods focus on analyzing a single mono audio recording, which is by far the most common practical scenario. Seminal work is based on the use of pure signal processing to extract each speech component according to some hand-crafted model. This is the case of the work in [19], where the authors propose an iterative technique to isolate a speaker based on harmonicity and temporal continuity. However, with the rise of deep learning, monaural speech separation has been cast in terms of classification problem, and data-driven approaches have become the de-facto standard [1].

In this work we consider different speech separation techniques belonging to the deep-learning monaural class. The SepFormer [2] is a transformer-based neural network for speech separation that learns short- and long-term dependencies with a multi-scale approach. In addition to the features of the previous model, the RE-SepFormer [20] uses non-overlapping chunks in the latent space reducing by half the number of chunks to process.

III. Speech Separation Detection

This section introduces the formal definition of the considered problem and the details of the proposed detection method.

Problem Formulation. In this paper, we consider the task of speech separation detection, i.e., we aim at determining whether a speech track containing a single speaker's voice was actually recorded like that or extracted from a mixture track containing signals from multiple speakers.

Formally, given a generic speech track $\mathbf{x}_{\text{speech}}$, we can assign it to the class $y \in \{0, 1\}$, where 0 means the speech was originally acquired from a single speaker and 1 means that

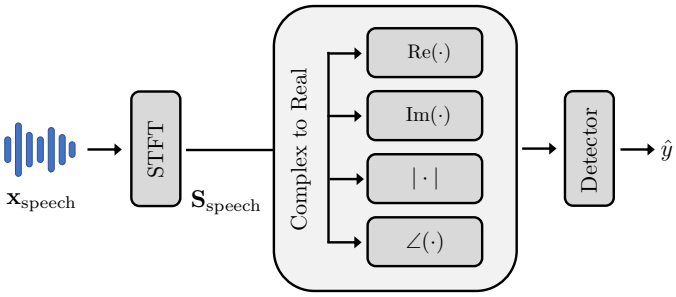


Fig. 2: Sketch of the proposed methodology. The detection system attributes a score $\hat{y}$ to the track $\mathbf{x}_{\text{speech}}$ under analysis.

$\mathbf{x}_{\text{speech}}$ is the result of a speaker separation technique $\mathcal{S}$ applied to a mixture signal $\mathbf{x}_{\text{mix}}$.

Our goal is to develop a speech separation detector $\mathcal{D}$ that estimates the class of the speech signal $\mathbf{x}_{\text{speech}}$, returning a real score $\hat{y} \in [0, 1]$ which indicates the likelihood that $\mathbf{x}_{\text{speech}}$ has been extracted from a mixture $\mathbf{x}_{\text{mix}}$.

Proposed Methodology. The speech separation detector $\mathcal{D}$ we propose is a CNN fed with pre-processed audio tracks. We explore different types of pre-processing techniques for the input tracks, analyzing multiple signal components in their time-frequency representations. Fig. 2 shows an overview of the proposed pipeline, which can be divided into three main steps:

- *STFT computation.* Given an audio track $\mathbf{x}_{\text{speech}}$, we compute its STFT over T time frames and F frequency bins and represent it as a 2D matrix with size $F \times T$. Formally,

$$\mathbf{S}_{\text{speech}} = \text{STFT}(\mathbf{x}_{\text{speech}}),\ \mathbf{S}_{\text{speech}} \in \mathbb{C}^{F \times T}. \quad (3)$$

  $\mathbf{S}_{\text{speech}}$ is a time-frequency representation of the audio signal (frequency information along rows and time information along columns). Notice that $\mathbf{S}_{\text{speech}}$ is complex by definition.
- *Complex to real conversion.* The complex signal contained in $\mathbf{S}_{\text{speech}}$ is converted to real components. We investigate four possible complex-to-real conversions, extracting the real part, the imaginary part, the magnitude and the wrapped phase of $\mathbf{S}_{\text{speech}}$.
- *Speech Separation Detector.* Once the audio track has been converted into a 2D real signal, it is fed to the CNN-based detector $\mathcal{D}$ that returns a score $\hat{y}$ associated with the likelihood of the input signal $\mathbf{x}_{\text{speech}}$ being the result of a speech separation operation.

The network we consider as speech separation detector $\mathcal{D}$ is a ResNet-based model. This was initially introduced for the synthetic speech detection task in [21], where it was trained on the magnitude representation of input speech signals. Considering the strong resemblance between that problem and the one at hand, we believe this model is well-suited for our specific purpose, and it can also work on different time-frequency representations of the input signals.

The architecture is a residual CNN that creates shortcuts between layers by skipping connections that help stabilize training. The network comprises six residual blocks positioned after the input layer, followed by a set of dense layers that end in a softmax output. For the complete network description, we refer the reader to [21].

IV. Experimental Setup

This section provides information on the evaluation setup we used during the experiments. First, we describe the datasets used for training and testing the systems. Then, we provide details regarding the models used to perform the speech separation. Finally, we specify the parameters considered to train the detectors.

Datasets. Since during all our experiments we wanted to ensure complete control over the considered speech mixtures and the separated speech tracks resulting from them, we generated the signals $\mathbf{x}_{\text{mix}}$ starting from single-speaker state-of-the-art datasets, namely LibriSpeech [22] and VCTK [23]. By mixing the tracks of these datasets, we created three new sets known as Libri2Mix, Libri3Mix, and VCTK2Mix, where the number in the name of each set indicates the amount of speakers considered when generating the mixtures. The generation of this type of data is well-established for speech separation and these three splits have been proposed in [24].

Libri2Mix is the only dataset we consider during training, where the "authentic" tracks are those of LibriSpeech, while the "manipulated" ones are extracted using a speech separator $\mathcal{S}$. All the other sets are considered in test only.

To test the proposed method in more challenging scenarios, we also considered another dataset known as REAL-M. REAL-M [25] is a dataset for speech separation in real-life settings, obtained through crowd-sourcing.

To generate the actual data we use to train and test the proposed detectors, we apply speech separator models to the aforementioned datasets. We consider two distinct separators, namely SepFormer and RE-SepFormer, and for both, we use pre-trained versions from the SpeechBrain Hugging Face repository [26]. In particular, for the datasets derived from LibriSpeech and VCTK, we consider the RE-SepFormer model trained on WSJ0-Mix dataset [27]. For separating the REAL-M dataset, we employ the SepFormer model trained on the WHAMR! dataset [28].

Training Stategy. During the training phase, we trained the detectors to discriminate between authentic speech and signals obtained by speech separation methods. We fed all the considered models assuming an input time window of 3.0 seconds. We used these values because, from preliminary experiments, they turned out to be the best compromise between the shortness of the windows and the performance, which is ideal in a real-world scenario.

All the hyperparameters of the networks have been fine-tuned to maximize their accuracy. These are the sets of parameters used, chosen after verifying the convergence of the models. For all the models, we considered a maximum

TABLE I: Area Under the Curve (AUC) and Balanced Accuracy (BA) values obtained on the coherent test set through single component analysis and fusion strategies. Best result in bold; second best result in italic.

	Real	Imag	Magnitude	Phase	$\text{Fusion}_{\text{Real-Imag}}$	$\text{Fusion}_{\text{Magn-Phase}}$	Global Fusion
AUC	0.9074	0.8995	0.9373	0.7607	0.9249	*0.9410*	**0.9494**
$\text{BA}_{0.5}$	0.8220	0.8082	0.8630	0.6694	0.8411	*0.8692*	**0.8737**
$\text{BA}_{\text{best } t}$	0.8263	0.8118	0.8648	0.6792	0.8423	*0.8696*	**0.8751**

number of epochs equal to 100 and an early stopping patience of 20, weighted cross-entropy as loss function and Adam optimization, a batch size of 128 and a learning rate value equal to 10^{-5}.

Regarding the computation of the input features, we compute the STFT of the input audio considering a Hamming window of 2048 samples and 25% overlap, which is a typical solution when computing this kind of audio representation [21]. Regarding the real and imaginary representations of the audio signal, we simply consider the real and imaginary parts of the STFT, and normalize them in dB scale. Regarding the magnitude, we compute the log-magnitude representation of the STFT [21] Regarding the phase representation, we extract the phase component of the STFT not considering any unwrapping procedure. During preliminary experiments we considered also the unwrapped and differential versions of the phase but, due to the obtained results, we decided to stick with the unwrapped version.

V. Results

In this section, we first report the evaluation metrics, then we show the achieved results of our experimental campaign. We start investigating results on a test set coherent with the training data, then we test the robustness of the proposed methodology to different scenarios unseen at training stage, evaluating performances on different datasets and different speech separation methods.

Evaluation metrics. For each binary classification experiment, we consider the AUC of the Receiver Operating Characteristic (ROC) curve related to the tackled problem, together with the BA. In both situations, the higher the metrics (ideally approaching 1), the better the performance. As ROC curve we consider the representation of False Positive Rate (FPR) vs. True Positive Rate (TPR) obtained by thresholding $\hat{y}$.

We evaluate the BA computed by thresholding the detector scores with a fixed threshold 0.5 ($\text{BA}_{0.5}$), but also $\text{BA}_{\text{best } t}$, computed at the optimal threshold value t determined to maximize the BA value. We do so to test the robustness of the proposed detector when dealing with unseen scenarios. If $\text{BA}_{\text{best } t}$ presents similar values to $\text{BA}_{0.5}$, our detector proves robust to unknown data and does not need further calibration.

In case of Real-M dataset, given the absence of original single-speaker tracks, we cannot report binary classification metrics. In these scenario, we always measure the achieved rate of correct detection over the class characterizing the tackled data (i.e., class 1).

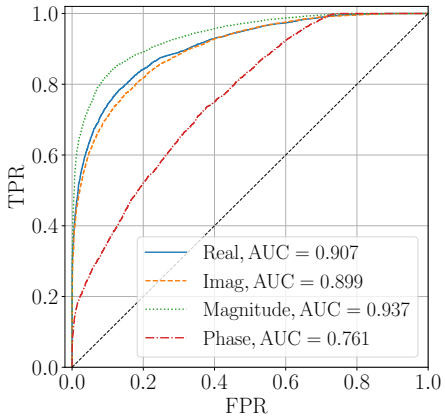


Fig. 3: ROC curve results related to the test set split of Libri2Mix dataset.

Results on Coherent Test Set. In this first scenario, we evaluate the proposed detector in a coherent scenario, i.e., on the test split related to the training dataset.

Fig. 3 shows the achieved results, which are comparable for the real part, the imaginary part and the magnitude of the signal's STFT, though the magnitude brings the most effective detection traces. The wrapped phase proves less discriminative for the tackled problem. As a matter of fact, it has been already shown that phase signal is often less accurate then magnitude for what regards audio forensics tasks [16].

Together with the analysis of single component results, we also perform some preliminary investigations on the fusion of multiple signal components to potentially enhance the detection performance. To this purpose, we propose a straightforward fusion strategy: averaging the scores returned by the single components involved. Table I depicts the overall results.

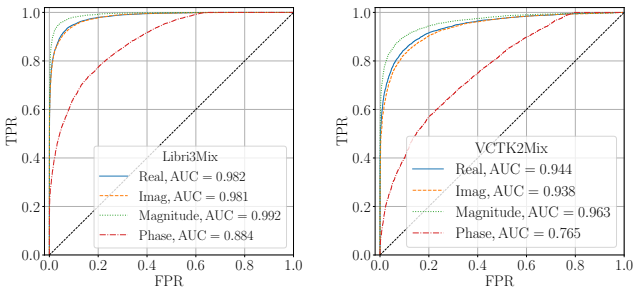
Averaging the scores always provides better performance than single components. The best option is the global fusion, meaning that all the components bring useful information for the fusion. As expected, since the test set is coherent with training data, the detector is well calibrated ($\text{BA}_{0.5}$ and $\text{BA}_{\text{best } t}$ show similar values).

Robustness to Unknown Test Datasets. In this scenario, we evaluate the detector over two state-of-the-art audio mixture datasets which are unknown at training stage, i.e., Libri3Mix and VCTK2Mix. ROC curve results are shown in Fig. 4, while Table II reports all the evaluation metrics.

As shown in the previous experiment, global fusion proves

TABLE II: AUC and BA values obtained through single component analysis and fusion strategies for Libri3Mix and VCTK datasets. Best result in bold; second best result in italic.

Dataset		Real	Imag	Magnitude	Phase	$\text{Fusion}_{\text{Real-Imag}}$	$\text{Fusion}_{\text{Magn-Phase}}$	Global Fusion
Libri3Mix	AUC	0.9818	0.9809	*0.9916*	0.8837	0.9902	0.9904	**0.9954**
	$\text{BA}_{0.5}$	0.8981	0.8958	0.9356	0.7846	0.9271	*0.9468*	**0.9533**
	$\text{BA}_{\text{best } t}$	0.9310	0.9268	0.9558	0.7881	0.9492	*0.9612*	**0.9686**
VCTK2Mix	AUC	0.9440	0.9378	0.9626	0.7651	0.9535	*0.9660*	**0.9688**
	$\text{BA}_{0.5}$	0.8695	0.8446	0.8987	0.6818	0.8689	**0.9022**	*0.8935*
	$\text{BA}_{\text{best } t}$	0.8742	0.8644	0.9037	0.6850	0.8865	*0.9077*	**0.9111**



(a) Libri3Mix. (b) VCTK2Mix.

Fig. 4: ROC curve results related to Libri3Mix and VCTK2Mix datasets.

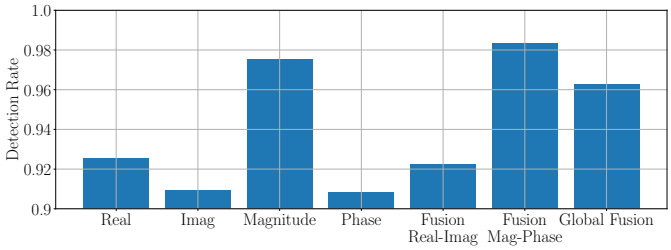


Fig. 5: Rate of correct detection achieved by testing the Real-M dataset.

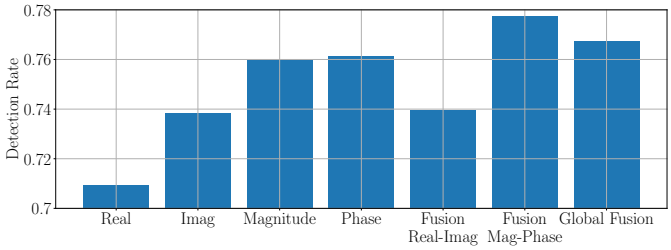


Fig. 6: Rate of correct detection achieved by testing the TIMIT-TTS dataset.

the best strategy overall and the magnitude represents the most informative signal component for the task at end. The phase remains the worst signal component to be analyzed.

Our proposed detector proves valid to test unknown data, without the need of calibration. AUC achieves 0.99 on Libri3Mix and 0.91 on VCTK2Mix. Notice that $\text{BA}_{0.5}$ and $\text{BA}_{\text{best } t}$ present very similar values, meaning for no need of calibration. Performances on VCTK2Mix are slightly lower than on Libri3Mix. We believe that Libri3Mix tracks belonging to class 1 are likely to contain more artifacts as they are the output of a more challenging task, i.e, separating three speakers rather than just two. This can make binary classification easier to perform.

Robustness to Unknown Speech Separators. In this section, we test single-speaker audio tracks that are the result of a speech separation technique different from that used in training. The considered dataset is the Real-M. Fig. 5 shows the achieved rate of correct detection for all signal components and fusions.

The achieved detection rate is always above 0.9, meaning for good robustness to different speech separators. Differently from the previous results, the most effective strategy is the fusion between magnitude and phase components. Real and imaginary parts do not bring useful contributions to the fusion, being slightly less informative than previous experiments.

VI. Preliminary Extension to Synthetic Speech

Due to the widespread use of synthetic speech, this section shows some preliminary experiments in which we test our methodology on audio tracks of single-speaker's voice that have been synthetically generated. The goal is to understand whether synthetic speech is also detected as single-speaker speech, rather than coming from a mixture.

The considered dataset for this experiment is the *Clean* subset of TIMIT-TTS [29]. TIMIT-TTS is a synthetic speech dataset containing signals generated with 12 different Text-to-Speech (TTS) methods considering sentences from the TIMIT corpus.

Fig. 6 depicts the achieved detection rate (i.e., the probability of correctly attributing label 0 to synthetic speech). Synthetic speech seems more challenging to be correctly detected, achieving a maximum detection rate of 0.78. Nonetheless, it is worth mentioning that we never train on synthetic data. Overall, the proposed detector shows promising, though leaving rooms for improvement.

Interestingly, the phase is the most effective among the single components. This behavior is in contrast with what was shown by the other experiments. Given the growing impact and spreading of synthetic speech in everyday life, we will devote more thorough investigations to this topic in the future.

VII. Conclusions

In this paper we considered the problem of detecting if a single-speaker speech track represents the real recording of a single speaker, or it has been obtained by applying speech separation to a mixture. In doing so, we propose an analysis of different audio representations derived from the decomposition into real components of the complex STFT.

Results show that it is worth exploiting the totality of the STFT information rather than resorting just the its magnitude as often done in the literature. Future work will explore the possibility of applying this idea to other audio classification tasks typically tackled by means of STFT-based analysis.

Acknowledgment

This material is based on research sponsored by the Defense Advanced Research Projects Agency (DARPA) and the 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 AFRL or the U.S. Government. This work was supported by the PREMIER project, funded by the Italian Ministry of Education, University, and Research within the PRIN 2017 program. This work was supported by the European Union under the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, partnership on "Telecommunications of the Future" (PE00000001—program "RESTART").

References

[1] D. Wang and J. Chen, "Supervised speech separation based on deep learning: an overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 26, pp. 1702–1726, 2018.

[2] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[3] M. Olivieri, L. Comanducci, M. Pezzoli, D. Balsarri, L. Menescardi, M. Buccoli, S. Pecorino, A. Grosso, F. Antonacci, and A. Sarti, "Real-time multichannel speech separation and enhancement using a beamspace-domain-based lightweight CNN," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[4] L. Cuccovillo, C. Papastergiopoulos, A. Vafeiadis, A. Yaroshchuk, P. Aichroth, K. Votis, and D. Tzovaras, "Open challenges in synthetic speech detection," in *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2022.

[5] C. Borrelli, P. Bestagini, F. Antonacci, A. Sarti, and S. Tubaro, "Synthetic speech detection through short-term and long-term prediction traces," *EURASIP Journal on Information Security*, vol. 2021, pp. 1–14, 2021.

[6] A. Pianese, D. Cozzolino, G. Poggi, and L. Verdoliva, "Deepfake audio detection by speaker verification," in *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2022.

[7] M. Hafizur Rahman, M. Graciarena, D. Castan, C. Cobo-Kroenke, M. McLaren, and A. Lawson, "Detecting synthetic speech manipulation in real audio recordings," in *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2022.

[8] E. Conti, D. Salvi, C. Borrelli, B. Hosler, P. Bestagini, F. Antonacci, A. Sarti, M. C. Stamm, and S. Tubaro, "Deepfake Speech Detection Through Emotion Recognition: a Semantic Approach," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[9] L. Attorresi, D. Salvi, C. Borrelli, P. Bestagini, and S. Tubaro, "Combining Automatic Speaker Verification and Prosody Analysis for Synthetic Speech Detection," in *International Conference on Pattern Recognition (ICPR)*, 2022.

[10] M. Papa, C. Borrelli, P. Bestagini, F. Antonacci, A. Sarti, and S. Tubaro, "A data-driven approach for acoustic parameter similarity estimation of speech recording," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[11] M. Baum, L. Cuccovillo, A. Yaroshchuk, and P. Aichroth, "Environment classification via blind roomprints estimation," in *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2022.

[12] L. Cuccovillo, S. Mann, M. Tagliasacchi, and P. Aichroth, "Audio tampering detection via microphone classification," in *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 2013.

[13] A. Giganti, L. Cuccovillo, P. Bestagini, P. Aichroth, and S. Tubaro, "Speaker-independent microphone identification in noisy conditions," in *European Signal Processing Conference (EUSIPCO)*, 2022.

[14] D. Moussa, G. Hirsch, and C. Riess, "Towards unconstrained audio splicing detection and localization with neural networks," in *International Conference on Pattern Recognition (ICPR)*, 2023.

[15] D. U. Leonzio, L. Cuccovillo, P. Bestagini, M. Marcon, P. Aichroth, and S. Tubaro, "Audio splicing detection and localization based on acquisition device traces," *IEEE Transactions on Information Forensics and Security (TIFS)*, pp. 1–15, 2023.

[16] M. Pilia, S. Mandelli, P. Bestagini, and S. Tubaro, "Time Scaling Detection and Estimation in Audio Recordings," in *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2021.

[17] Z. Xiang, A. K. S. Yadav, S. Tubaro, P. Bestagini, and E. J. Delp, "Extracting efficient spectrograms from MP3 compressed speech signals for synthetic speech detection," in *ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec)*, 2023.

[18] M. Pezzoli, J. J. Carabias-Orti, M. Cobos, F. Antonacci, and A. Sarti, "Ray-space-based multichannel nonnegative matrix factorization for audio source separation," *IEEE Signal Process. Lett.*, vol. 28, pp. 369–373, 2021.

[19] G. Hu and D. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 18, pp. 2067–2079, 2010.

[20] C. Subakan, M. Ravanelli, S. Cornell, F. Lepoutre, and F. Grondin, "Resource-efficient separation transformer," *arXiv preprint arXiv:2206.09507*, 2022.

[21] M. Alzantot, Z. Wang, and M. Srivastava, "Deep Residual Neural Networks for Audio Spoofing Detection," in *Conference of the International Speech Communication Association (INTERSPEECH)*, 2019.

[22] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[23] J. Yamagishi, C. Veaux, K. MacDonald *et al.*, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," 2019.

[24] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "Librimix: An open-source dataset for generalizable speech separation," *arXiv preprint arXiv:2005.11262*, 2020.

[25] C. Subakan, M. Ravanelli, S. Cornell, and F. Grondin, "Real-m: Towards speech separation on real mixtures," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[26] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.

[27] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[28] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. Le Roux, "WHAM!: Extending Speech Separation to Noisy Environments," in *Conference of the International Speech Communication Association (INTERSPEECH)*, 2019.

[29] D. Salvi, B. Hosler, P. Bestagini, M. C. Stamm, and S. Tubaro, "TIMIT-TTS: A text-to-speech dataset for multimodal synthetic media detection," *IEEE Access*, vol. 11, pp. 50 851–50 866, 2023.