Vulnerability of Automatic Identity Recognition to Audio-Visual Deepfakes

Pavel Korshunov Haolin Chen Philip N. Garner Sébastien Marcel
Idiap Research Institute, Martigny, Switzerland.
{pavel.korshunov,haolin.chen,phil.garner,sebastien.marcel}@idiap.ch

Abstract

The task of deepfakes detection is far from being solved by speech or vision researchers. Several publicly available databases of fake synthetic video and speech were built to aid the development of detection methods. However, existing databases typically focus on visual or voice modalities and provide no proof that their deepfakes can in fact impersonate any real person. In this paper, we present the first realistic audio-visual database of deepfakes SWAN-DF, where lips and speech are well synchronized and video have high visual and audio qualities. We took the publicly available SWAN dataset of real videos with different identities to create audio-visual deepfakes using several models from DeepFaceLab and blending techniques for face swapping and HiFiVC, DiffVC, YourTTS, and FreeVC models for voice conversion. From the publicly available speech dataset LibriTTS, we also created a separate database of only audio deepfakes LibriTTS-DF using several latest text to speech methods: YourTTS, Adaspeech, and TorToiSe. We demonstrate the vulnerability of a state of the art speaker recognition system, such as ECAPA-TDNN-based model from SpeechBrain, to the synthetic voices. Similarly, we tested face recognition system based on the MobileFaceNet architecture to several variants of our visual deepfakes. The vulnerability assessment show that by tuning the existing pretrained deepfake models to specific identities, one can successfully spoof the face and speaker recognition systems in more than 90% of the time and achieve a very realistic looking and sounding fake video of a given person.

1. Introduction

The original predictions that deepfakes would pose a significant danger to society[1] are turning out to be correct with more and more reports of money extortion with voice cloning[2], duping politicians using video deepfakes[3], or using deepfake pornography for revenge or harassment online[4].

Many databases with deepfake videos were created to help develop and train deepfake detection methods. One of the first freely available databases was Deepfake-TIMIT [32], followed by the FaceForensics database with deepfakes generated from 1000 Youtube videos [49], and which was later morphed into FaceForensics++ with more types of deepfakes and a separate set of original and deepfake videos provided by Google and Jigsaw [50]. Several independent extensions of FaceForensics++ were also proposed, including the HifiFace [57] and DeeperForensics [21] datasets. Another 5000 videos-large database of deepfakes generated from Youtube videos is Celeb-DF v2 [38]. Facebook [8] also created their own database with more than $100K$ deepfake videos, which was used in Deepfake Detection Challenge 2020 hosted by Kaggle[5]. Mobio-DF[19] is a dataset of $45K$ videos but with an unusually larger set of real videos compared to deepfakes. However, the largest database of deepfake videos to date is the Korean Deepfake (KoDF) dataset [35] with about $175K$ fake videos.

One of the important issues with the most of the existing deepfake databases is that very little is known about the quality of the deepfake videos in terms of their ability to actually impersonate a targeted person. Besides a limited study of face recognition vulnerability [32], the authors of datasets do not provide any justification of whether their deepfakes even look like a person, let alone a specific person. It means that without a verification of how fake those deepfakes are, even a slight distortion of an original video (e.g., an applied color correction) may be considered as a deepfake. This actually happens, as we can observe some videos from Facebook [8] dataset, which are labeled as be-

[1] https://edition.cnn.com/interactive/2019/01/business/pentagons-race-against-deepfakes/

[2] https://www.nbc15.com/2023/04/10/ive-got-your-daughter-mom-warns-terrifying-ai-voice-cloning-scam-that-faked-kidnapping/

[3] https://www.theguardian.com/world/2022/jun/25/european-leaders-deepfake-video-calls-mayor-of-kyiv-vitali-klitschko

[4] https://edition.cnn.com/2023/02/16/tech/nonconsensual-deepfake-porn/index.html

[5] https://www.kaggle.com/c/deepfake-detection-challenge

ing deepfake, to only have a moving patch of Gaussian noise as the only visible difference from the original real version. The lack of clear understanding of what constitutes a deepfake in a dataset leads to an over-fit problem, when detection algorithms, trained on such *so called* deepfakes, may end up detecting distortions that are irrelevant to those manifested in the realistic and dangerous deepfakes.

Another important issue, often overlooked by the authors of deepfake databases and the researchers who develop detection methods, is that the blending techniques used during face swapping process (or reenactment) arguably have as much effect on the accuracy of the detection as the generative adversarial networks (GANs) used to generated fake faces. This effect is well illustrated by the authors of [55], who demonstrate that uncompressed GAN-generated images can be detected with a nearly 100% accuracy, while the accuracy of the same detection methods degrade significantly on deepfake videos, where a similarly GAN-generated image is blended in and blurred into the compressed frame. It shows the possibility that many proposed deepfake detection methods may detect the distortions that come from blending and compression instead of the signatures that come from GANs. Arguably, this is why deepfake detection methods do not generalize well when they are tested on a database that used the same GAN architecture as the training videos but different blending methods [10, 33, 29].

Looking at the voice deepfakes, there are fewer databases that rely on neural networks for text to speech or voice conversion methods. The most notable database is the one used in ASVSpoof challenge in 2019 and 2021 [59], which has a separate subset of audio deepfakes. These deepfakes were generated by several modern text to speech models to train and test detection methods but the methods generate either a voice of a single speaker or a limited pre-defined set of multiple speakers. No identity transfer was done by the methods used to generate these fakes and that is why the authors did not provide identity information for samples in the deepfake subset. WaveFake [12] is the latest dataset of deepfake speech that used public LJSpeech [18] dataset of the single-speaker recordings as the original source to generate the fake samples. Until recently, most of the work in deep learning based methods that generate fake speech focused on producing the realistically sounding voice samples and little was done to preserve identity in that sample. However, recent advances in text to speech and voice conversion methods, allow to create datasets of truly deepfake speech which would preserve or transfer personal identity into the generated fake sample.

In this paper, we present the first high fidelity publicly available dataset of realistic deepfakes SWAN-DF (see examples in Figure 1) where both faces and voices appear and sound like the target person. The SWAN-DF dataset is based on the public SWAN database [45] of real videos recorded in HD on iPhone and iPad Pro (in year 2019). For 30 pairs of people, we swapped faces and voices using several autoencoder-based face swapping models form the well-known open source repo DeepFaceLab[6] [42] and voice conversion (or voice cloning) methods, including zero-shot YourTTS [4] and various models from FreeVC [36]. In addition to the audio-visual deepfake dataset, we also built LibriTTS-DF database (from a well-known LibriTTS [61] database), which contains fake speech samples for 39 speakers generated with either text to speech methods that preserve intended identity, including our own adaptation of Adaspeech TTS model [5] and diffusion-based TorToiSe TTS[7] or YourTTS [4] zero-shot voice conversion approach.

For video deepfakes, we also have put an effort into creating a large variety of different versions of generated videos in terms of models and blending techniques used. We have employed three different models with resolutions 160px, 256px, and 320px, all pretrained on a large variety of faces by the contributors of DeepFaceLab. For each of 60 people pairs from SWAN database, we have tuned each model type. And then, we also generated several version videos for each of this model, where we use different masking, color correction, and other blending parameters. In total, we generated more than 20 deepfake variants for each video of each pair of people. These variations should allow to train and also test detection models that are invariant to the blending methods but instead focus on the distortions that are specific to deepfakes themselves, such as inconsistencies in accessories, issues with hair, eyes and teeth, geometrical facial distortions, etc.

To show how well different audio and visual deepfake generation methods preserve identity, we conducted an extensive vulnerability analysis using the ECAPA-TDNN-based state of the art speech recognition model from SpeechBrain[8], and MobileFaceNet [6], a popular pretrained PyTorch face recognition model[9].

To allow researchers to use the database in a transparent manner and verify and reproduce our vulnerability evaluations, we provide the generated audio and video samples, list of files and splits into subsets, source code for vulnerability analysis and a jupyter notebook with complete results and graphs as an open-source Python package[10]. Our Mobio-DF and LibriTTS-DF databases with examples of deepfake videos and voices can be found at the demo page[11].

[6] https://github.com/iperov/DeepFaceLab

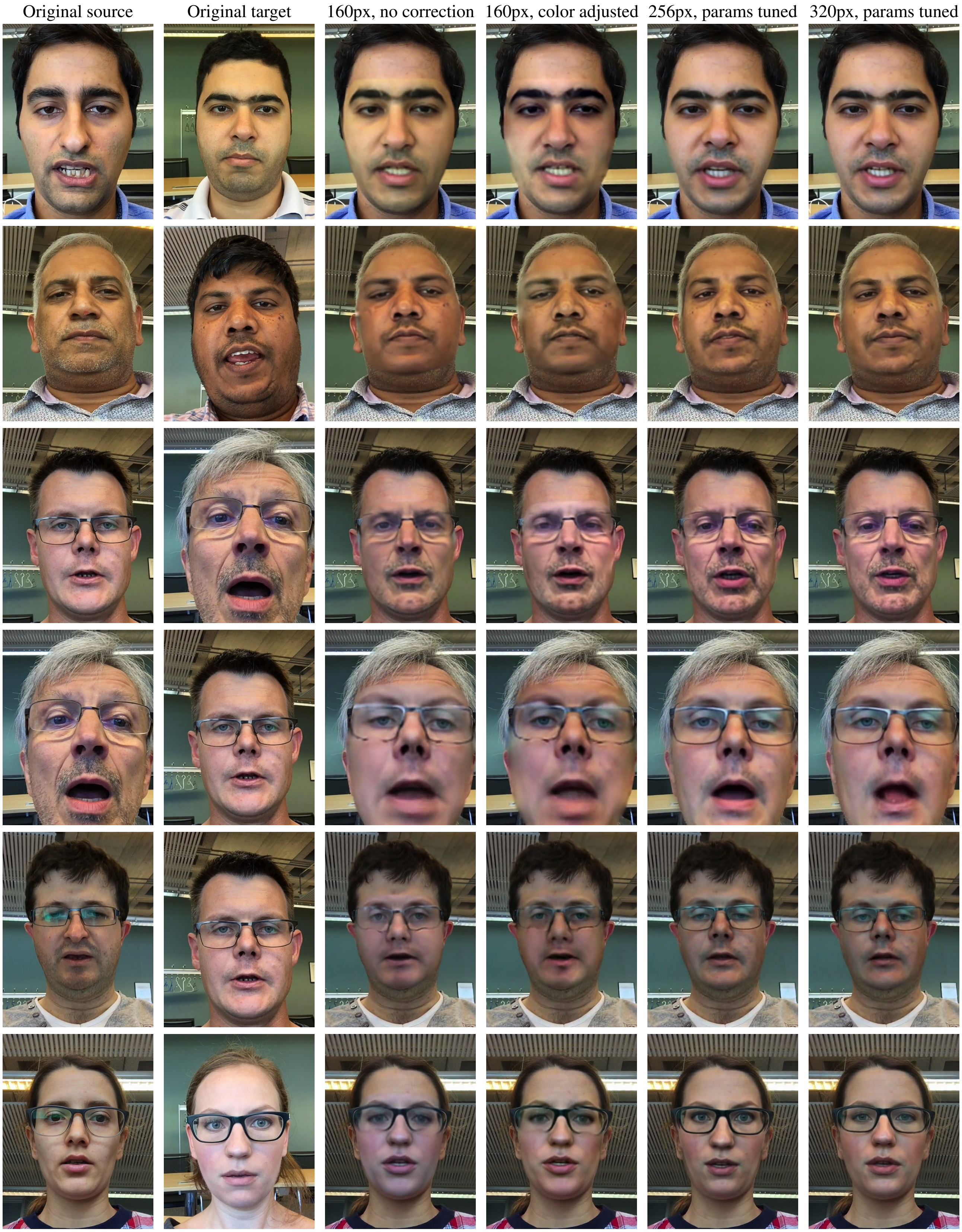
[7] https://github.com/neonbjb/tortoise-tts

[8] https://speechbrain.github.io/

[9] https://github.com/foamliu/MobileFaceNet

[10] Source code: https://gitlab.idiap.ch/bob/paper.ijcb2023.av-deepfakes

[11] Database: https://swan-df.github.io/



(a) Original source (b) Original target (c) 160px, no correction (d) 160px, color adjusted (e) 256px, params tuned (f) 320px, params tuned

Figure 1. Examples of deepfakes (cropped to face) from the SWAN-DF database. Face of the ‘target’ is placed into video of the ‘source’.

2. Related work

The generation approaches of synthetic faces can be split in four different categories i) completely synthetically generated images (identity is usually not preserved) using StyleGANs [22, 23], ii) morphed images when faces of two people are morphed[12] [11, 51], iii) face swapping based video deepfakes [34, 8, 50], and iv) reenactment based deepfake videos [21, 13, 9], which grew out of the idea of using a recurrent network to synthesize mouth texture directly from the voice [52].

Methods for detecting visual fakes range from those based on simple visual or facial features [62, 60, 2, 37], binary classifiers trained on fake images [55, 17] and videos [50, 41, 40], to the methods that try to generalize to new deepfake methods or various post-processing blending techniques [3, 33, 29].

The state of the art in text to speech and voice conversion is represented by probabilistic generative models, particularly those based on diffusion [20, 14], but also flow [48, 47, 16] or a combination or both [58, 26]. Such techniques originated in the image processing or computer vision literature. They are characterized by an iterative conversion from a simple distribution (that lends itself to sampling) to a complicated distribution representative of speech (but difficult to sample from). At each iteration, a DNN is used to guide the conversion; in flow it is a transformation, in diffusion a denoizing process.

Similar to the work on detecting visual deepfakes, methods for detection of synthetic speech are also struggling with generalization to unseen attacks as is evident from the latest ASVspoof challenge and related work [59, 56]. Although the latest detection methods more and more rely on the end-to-end systems for feature extraction and modeling, the earlier work was often based on acoustic features [53] and classical GMM-like modeling [30, 31].

3. Generative methods used in the database

Our main goal is to create an audio-visual database of people speaking on camera where both video and audio channels are completely generated and which would look and sound as realistic as possible. We explored different methods for generating fake speech and fake faces and settled on the models by DeepFaceLab[6] [42] for fake face swapping and FreeVC [36] for voice conversion. In this section, we describe the methods and their variations that we used to generate audio-visual deepfakes for our SWAN-DF database. In addition, we describe the other speech generative methods that we could only use to create audio deepfakes resulted in LibriTTS-DF dataset.

3.1. Video deepfakes

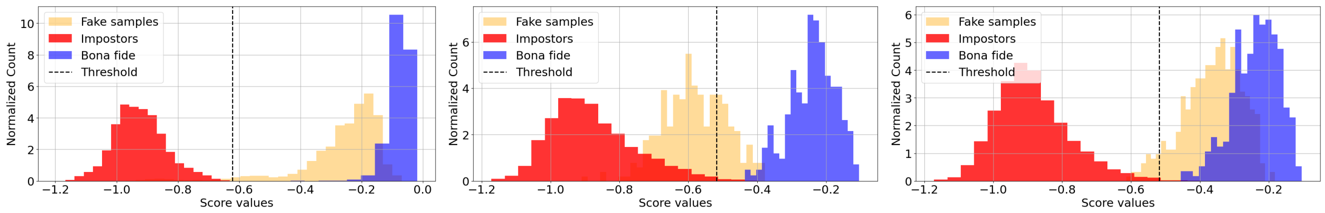
As a source of original videos, we selected 46 different identities from a publicly available SWAN database [45], which was recorded in 2019. The videos in HD (resolution 720×1280) include a person looking into iPhone or iPad Pro frontal camera and saying a set of phrases. From these 46 identities, we manually matches 60 pairs of people for face swapping process. In the selection process, we tried to match accessories, such as eye glasses, head and facial hair styles, skin colors, and genders. A well matched pair of faces typically leads to a visually more realistic deepfake. Since SWAN dataset has 16 videos with sound per each person, swapping faces for 60 pairs of people, results in $16 \times 60 = 960$ of the deepfakes per a given model architecture and a blending process (see frames extracted from the original and deepfake videos in Figure 1 or view the videos on the demo page[11]).

To generate video deepfakes, we used a well known open source repository DeepFaceLab[6], which implemented two main GAN-based architectures the authors call DF and LIAE [42]. We used three pretrained models provided by the DeepFaceLab community that can generate faces of 160×160 (DF architecture), 256×256 (LIAE architecture), and 320×320 (LIAE architecture) resolutions. The models are pretrained on the large datasets of 'whole faces' (in DeepFaceLab terminology a facial area that includes chin and the half of a forehead) of several identities that allows models to learn the generic structure of a face and reduces the time required to tune the model to a specific pair of identities.

For each of the three model architectures and for each pair of identities, we tuned the pretrained model for $50K$ iterations, which resulted in about 4 hours for 160px resolution model, 13 hours for 256px resolution, and 20 hours for 320px resolution on Tesla P40 GPU. Through the trial and error, we have selected some of the specific training parameters for each of the model and we provide these parameters in our open source packages[10]. For the model of 160px resolution, we trained three different variants, including i) training face together with its face mask and with color correction on, ii) no masked training and no color correction, and iii) with mask training but no color correction. For the other two resolutions, we only trained model that included mask training and had no color correction switched on. In total, we tuned five different types of models for each of the 60 swap-pairs.

Arguably, an important part of what constitutes a deepfake and makes it different from the real image, from a forensics point of view, is the blending technique that was used to place-in the generated face of a target into the original video frame of a source (see Figure 1). A DeepFaceLab GAN model typically generates a square image, in addition, it also learns mask of a face. This mask is used to cut off

[12] https://github.com/yaopang/FaceMorpher



(a) Face deepfakes with 256px resolution model (b) Voice deepfakes with YourTTS (c) Voice deepfakes with FreeVC

Figure 2. Score histograms of MobileFaceNet face and SpeechBrain speaker recognition models evaluated on the real videos (SWAN database) and a variant of generated deepfakes (SWAN-DF). The dotted vertical line marks the threshold computed on the validation set.

the face from the generated image and replace the source face with it. During the replacement process, the smoothing, blurring, warping, and color correction techniques can be applied to make the generated face look naturally fitting into the destination frame. These techniques, which we refer to as *blending*, change the appearance of the resulted frame, introduce some unique distortions, and therefore impact the methods trained to detect the deepfakes. One could argue that the existing deepfake detection methods detect mostly the residues from the blending techniques rather than the patterns left by the GANs used to generate faces [55]. Therefore, to offset the lack of the variation of blending techniques in the existing deepfake datasets, we have used more than 20 variants of deepfakes using different sets of blending parameters for the five models that we have trained for each swap pair. Each variation results in a differently looking deepfake face and we believe such variability in the dataset will be useful for the research community.

For ethical reasons, we selected not to publish open source code that makes it easier to create deepfakes, besides what is already available in DeepFaceLab repository[6], and also not to publish our trained models. We do however provide all of the parameters for training and blending that we have used in the process[10].

3.2. Audio deepfakes

We generated speech deepfakes using four voice conversion methods: YourTTS [4], HiFiVC [24], DiffVC [43], and FreeVC [36] and two text to speech methods: Adaspeech [5] and TorToiSe TTS[7]. We did not use text to speech methods for our video deepfakes, since the speech they produce is not synchronized with the lip movements in the video. There are efforts to correct this issue by using an additional model, e.g., Wav2Lip[13] [44], to synchronize lips in video with a given speech, but they suffer from many visible artifacts and often produce unconvincing results. One notable effort that used the combination of a TTS method and Wav2Lip is the FakeAVCeleb [25] audio-visual dataset, but the quality of the resulting videos is questionable. Some commercial systems use text to speech and then speech to video approaches to generate AI assistants, notably Synthesia[14], but the synchronization issue persists there as well.

Therefore, we used voice conversion methods (YourTTS, HiFiVC, DiffVC, and FreeVC) to generate fake speech for the SWAN-DF dataset, but we used text to speech methods (Adaspeech and TorToiSe TTS) with only one YourTTS voice conversion to generate a separate dataset based on LibriTTS [61]. We used the test-clean subset of LibriTTS dataset with 39 speakers to generate deepfakes. We took a random 30 utterances from a speaker to either tune a model (Adaspeech) or compute speaker embeddings (TorToiSe or YourTTS). We then generated fake samples from the same utterances used for tuning.

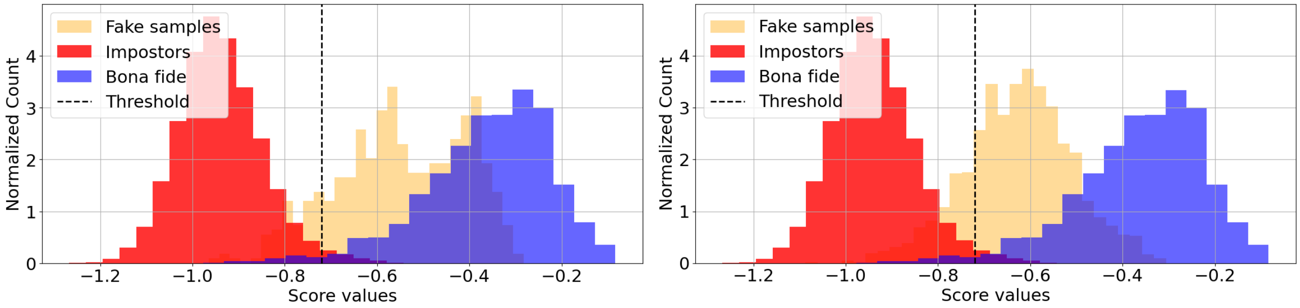
A brief summary and the parameters of the speech deepfake methods used are as follows:

- **Adaspeech [5]:** a text to speech model specifically designed to be adapted to a custom voice and acoustic conditions. This model was further modified by us to allow adapting to a new voice with the aim of preserving the speaker identity. The model was pretrained to generate spectrograms from text on the popular VCTK corpus [54] of speech from 109 speakers reading 400 sentences for $300K$ iterations. This pretrained model was adapted (tuned) for each of 39 speakers from LibriTTS [61] dataset using 30 utterances of that speaker for $4K$ iterations. The adapted model is then used to generate the user-specific spectrograms from the provided text. HifiGan [28] vocoder pretrained on LJSpeech [18] database is then used to generate the final synthetic speech samples.

- **TorToiSe TTS[7]:** a text to speech model is based on the autoregressive and diffusion encoders and is inspired by and very similar to DALLE [46] for images. The author of this zero-shot generative model argues that it requires just a few seconds of reference speech to generate a high fidelity speech from any textual input. The model is pretrained on a set of speech databases,

[13] https://github.com/Rudrabha/Wav2Lip

[14] https://www.synthesia.io/tools/text-to-speech-video-maker



(a) TorToiSe TTS

(b) YourTTS

Figure 3. Vulnerability results of SpeechBrain speaker recognition model when evaluated on the original LibriTTS data and voice deepfakes generated by TorToiSe TTS and YourTTS models. The dotted vertical line marks the threshold computed on the validation set.

including a private one collected by the author, totaling about a million hours of speech. To generate our LibriTTS-DF database, we used 30 utterances per speaker to compute the latent vectors of the model and then we produce the same but synthetic 30 samples from the corresponding text. The preset 'fast' of the model was used during the generation process.

- **YourTTS [4]:** originally a text to speech model based on the end-to-end VITS [27] but with an addition of a separate speaker embedding (from a speaker recognition model [15]) to encode speaker identity. The inclusion of speaker encoding allows the use of YourTTS in a zero-shot voice conversion manner by simply substituting the embedding of one speaker with the embedding of another. We used the provided model pretrained on VCTK [54] and LibriTTS [61] datasets. For our LibriTTS-DF dataset, we converted each speaker to a randomly chosen 5 other speakers, using only 10 utterances for speaker encoding. For SWAN-DF dataset, we used all 16 available utterances per each speaker for the encoding and converted the voices for the same pairs of speakers as in video deepfake swapping (see Section 3.1 for details).

- **HiFiVC [24]:** a zero-shot many-to-many voice conversion system that relies on automated speech recognition (ASR) features, pitch tracking inspired by PPG-VC [39], and their version of the waveform prediction model that extends HiFi GAN [28]. We used provided model pretrained on VCTK dataset [54] without any tuning. We used the same number of utterances as for YourTTS method to compute the speaker embeddings during the conversion.

- **DiffVC [43]:** another zero-shot many-to-many voice conversion method designed for the general case when source and target speakers do not belong to the training dataset. Since the authors emphasized that no tuning is required and for the provided model pretrained on LibriTTS dataset [61], we used this method *as is*. Please note that we used this method to generate deepfakes for SWAN-DF dataset, which is very different from LibriTTS. We used the same number of utterances as for YourTTS method to compute the speaker embeddings during the conversion.

- **FreeVC [36]:** an end-to-end model for voice conversion based on the approach proposed in VITS [27]. The model relies on WavLM features [7] and a computationally heavy augmentation technique based on the resizing of spectrograms to several spectral bands, which allow to exclude noise from the data when learning speech characteristics. The provided model is pretrained on VCTK speech corpus [54] and we have adapted it to the acoustic domain of our database by tuning the model on the mixture of subset from VCTK and data from SWAN dataset. We then convert voices for SWAN-DF using the same swap pairs as we did for video deepfakes (see Section 3.1 for details). By using different tuning parameters of the FreeVC model, we produced 5 different variants of data.

We obtain the complete audio-visual deepfakes by combining the videos produced by face swapping and the speech generated using voice conversion methods. We can match all face with all voice deepfakes, thus obtaining a very large set of videos where visual and voice channels are different. With 960 videos/utterances in SWAN-DF for one deepfake variant, with more than 20 variations of face swapping, and 8 voice conversion variants, we can get more than $150K$ video combinations.

4. Vulnerability assessment of deepfakes

To evaluate how realistic the generated deepfakes are, we used the ECAPA-TDNN-based model from SpeechBrain[9], one of the best performing speaker recognition systems, and MobileFaceNet [6], which is one of the popular and practical face recognition models[10].

Model, training params	Blending Method	IAPMR
160px, no mask	no blending	96.27
160px, mask training	seamless, mlk color	95.22
160px, mask training + color	overlay, no color	96.43
256px, mask training	params tuned	96.78
256px, mask training	overlay, no color	**97.36**

Table 1. Vulnerability of face recognition to selected variants of video deepfakes from SWAN-DF dataset.

4.1. Evaluation protocol

We assess the vulnerability of the speaker and face recognition to the deepfakes in the same way as the vulnerability of the biometric systems is assessed to the presentation attacks, as per the recommendation presented in the standard [1]. Therefore, we report false match rate (FMR), which is similar to false positive rate (FPR), and false non-match rate (FNMR), which is similar to false negative rate (FNR), and impostor attack presentation match rate (IAPMR), which is the proportion of attacks that are incorrectly accepted as genuine samples by a biometric system (for details, see ISO/IEC 30107-3 standard [1]).

We split the deepfakes of our SWAN-DF and LibrtiTTS-DF databases into development and evaluation subsets roughly equal in size. We also ensured that the identities in the different subsets do not overlap. To compute the metrics, we define the threshold on the development set that corresponds to equal error rate (EER) computed on the real original data. We use this threshold when computing FMR and FNMR on the scores of the real data from the evaluation set, and also to compute IAPMR rate, when instead of the scores for zero-effort impostors, we use the scores corresponding to deepfakes.

To demonstrate the accuracy of the selected recognition systems, we evaluated them on the real video data assuming no deepfake attacks are present. For both systems, we used the pretrained models provided by the respective repositories. We used only two real samples to enroll each identity and the rest of the samples from the identity were used for computing the error rates. Using the EER threshold from the development set, SpeechBrain on the real audio from evaluation set resulted in very low 0.3% FMR and 0.0% FNMR values for the SWAN-DF dataset and 1.99% FMR and 1.82% FNMR values for the LibriTTS-DF dataset. Similarly, MobileFaceNet resulted in low 0.0% FMR and 0.05% FNMR values when computed on the real videos from evaluation set of SWAN-DF database. The red (zero-effort impostors) and blue-colored (bona fide samples) histograms in Figure 2 and Figure 3 illustrate well the low FMR and FNMR values, since these histograms are clearly separated.

Approaches	IAPMR
HiFiVC	0.00
DiffVC	8.09
YourTTS	27.43
FreeVC, not tuned	15.44
FreeVC, tuned 70K iterations	92.59
FreeVC, tuned 109K iterations	**94.21**

Table 2. Vulnerability of speaker recognition to selected voice conversion based deepfakes from SWAN-DF dataset.

4.2. Vulnerability to SWAN-DF

Table 1 shows the IAPMR rates, computed for each video frame separately, using MobileFaceNet [6] on the selected face deepfake variants we generated for SWAN-DF database. The high IAPMR rates in the table mean that more than 95% of the deepfake frames were recognized by MobileFaceNet as corresponding to the claimed real identity. This demonstrates that the evaluated face recognition model is highly vulnerable to the generated deepfake videos. Figure 2a also illustrates this result by showing how the scores (yellow histogram) for deepfake variant generated using 256px model (see the fourth row of Table 1) are next to the bona fide scores (blue histogram) and almost completely on the right side of the threshold (the dotted vertical line).

Table 2 shows similar IAPMR rates for the voice deepfakes generated with several voice conversion algorithms. The results in this table are quite different from those for the face deepfakes. The table shows that some voice conversion algorithms, notably HiFiVC and DiffVC, pose no threat to SpeechBrain speaker recognition system as it did not confuse the speech generated by these algorithms with the claimed real identities. FreeVC without any tuning and zero-shot YourTTS show IAPMR rates above 15%, which are not that small, considering that FMR and FNMR for real voices are very close to zero (see Section 4.1). It appears that without tuning, the identities do not transfer well to the generated speech and do not pose a great threat to the recognition system. However, if we tune the model, such as the domain transfer we have done for FreeVC model, we can achieve the vulnerability level comparable to the face deepfakes with IAPMR rates higher than 92%. Also, the last two rows of the Table 2 show that the longer tuning leads to the higher IAPMR rate.

Figure 2b and Figure 2c illustrate the differences between zero-shot approaches like YourTTS and when we use the tuned model like FreeVC (tuned for 109K iterations). These figures show that for the tuned model, the histogram of the deepfake scores shifts very near to the blue histogram of the bona fide scores.

Approaches	IAPMR
YourTTS	80.06
Adaspeech	83.51
TorToiSe	**86.61**

Table 3. Vulnerability of speaker recognition to voice conversion based deepfakes from LibriTTS-DF dataset.

4.3. Vulnerability to LibriTTS-DF

Vulnerability evaluation on LibriTTS-DF database is interesting in a sense that it allows us to observe two important points i) the differences between voice deepfakes generated using text to speech and zero-shot voice conversion methods and ii) the different between using audio dataset recorded in the room with acoustic isolation like LibriTTS and the dataset recorded in a standard noisy office environment like SWAN.

Table 3 shows IAPMR rates for zero-shot YourTTS, for zero-shot advanced diffusion-based TorToiSe TTS model and for Adaspeech, where the pretrained TTS model was tuned for each speaker. Surprisingly, the results show high IAPMR rates for all methods, with TorToiSe, despite being zero-shot model, being the most threatening to SpeechBrain speaker recognition system. Also note that from a personal subjective experience, TorToiSe TTS model produces the most realistic and pleasant to the ear audio utterances. Figure 3a shows how much more the scores for deepfakes generated by TorToiSe are close to the bona fide scores. Comparing with Figure 3b, it can be noted that, although IAPMR values are comparable, the scores for deepfakes of YourTTS model are noticeably nearer to the red distribution than the scores of TorToiSe deepfakes.

Comparing the result for YourTTS in Table 3 with the results in Table 2, we can notice that IAPMR value 80.06% for LibriTTS-DF is significantly higher than 27.43% value for SWAN-DF. Figure 2b and Figure 3b illustrate this phenomena as well. In both cases, we used the same pretrained YourTTS model and similar approach on how we have generated voice conversion deepfakes. Therefore, the difference can be explained by the fact that the original SWAN dataset was recorded in an office environment using consumer smartphone and tablet, which had a significant effect on the resulted generative speech. Subjectively, the utterances produced by YourTTS for SWAN-DF dataset are much more noisy with a lot of metallic sounds compared to the same in LibriTTS-DF.

5. Conclusion

In this paper, we presented SWAN-DF database of high quality audio-visual deepfake that pose a high threat to state of the art speaker and face recognition systems. We also generated more than 20 face swapping variants using the combination of different models and blending techniques. We used 8 different voice conversion methods from 5 different models to generate voice deepfakes. Such large variation of the deepfakes in both audio and visual domains will allow researchers to develop multimodal, more robust, and generalizable methods for deepfakes detection. In addition to the database of audio-visual deepfakes, we released a LibriTTS-DF database of only voice deepfakes, which we generated using both voice conversion and text to speech method, including TorToiSe, one of the latest diffusion-based models.

The deepfakes will remain to be a serious threat to the identity protection systems and a challenge for an automatic and human detection. Therefore, such high fidelity datasets like SWAN-DF and LibriTTS-DF play a key role in helping to overcome these challenges.

Acknowledgements

This work was funded by Swiss Center for Biometrics Research and Testing and NAST: Neural Architectures for Speech Technology, Swiss National Science Foundation grant[15]. The authors would also like to thank Christophe Ecbert for the help with setting up face recognition system and Kevin Carta for the advice on which tuning parameters of DeepFaceLab deepfake generation models to use.

References

[1] I. F. 30107-3:2017. Information technology — Biometric presentation attack detection — Part 3: Testing and reporting. Standard, International Organization for Standardization, Geneva, Switzerland, 09 2017. 7

[2] A. Agarwal, R. Singh, M. Vatsa, and A. Noore. SWAPPED! digital face presentation attack detection via weighted local magnitude pattern. In *IEEE International Joint Conference on Biometrics (IJCB)*, pages 659–665, Oct. 2017. 4

[3] S. Aneja and M. Nießner. Generalized zero and few-shot transfer for facial forgery detection. *arXiv preprint*, 2020. 4

[4] E. Casanova, J. Weber, C. D. Shulby, A. C. Júnior, E. Gölge, and M. A. Ponti. YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone. In *International Conference on Machine Learning, ICML, Baltimore, Maryland, USA*, volume 162, pages 2709–2720. PMLR, 2022. 2, 5, 6

[5] M. Chen, X. Tan, B. Li, Y. Liu, T. Qin, S. Zhao, and T. Liu. AdaSpeech: Adaptive text to speech for custom voice. In *International Conference on Learning Representations, ICLR*. OpenReview.net, May 2021. 2, 5

[6] S. Chen, Y. Liu, X. Gao, and Z. Han. MobileFaceNets: Efficient CNNs for accurate real-time face verification on mobile devices. In *Biometric Recognition*, pages 428–438, Cham, 2018. Springer International Publishing. 2, 6, 7

[15] https://data.snf.ch/grants/grant/185010

[7] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, and F. Wei. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *arXiv preprint*, 2021. 6

[8] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer. The deepfake detection challenge dataset. *arXiv preprint*, 2020. 1, 4

[9] N. Drobyshev, J. Chelishev, T. Khakhulin, A. Ivakhnenko, V. Lempitsky, and E. Zakharov. MegaPortraits: One-shot megapixel neural head avatars. In *Proceedings of the ACM International Conference on Multimedia*, MM'22, pages 2663–2671, New York, NY, USA, 2022. Association for Computing Machinery. 4

[10] M. Du, S. Pentyala, Y. Li, and X. Hu. Towards generalizable deepfake detection with locality-aware autoencoder. In *Proceedings of the ACM International Conference on Information; Knowledge Management*, CIKM'20, pages 325–334, New York, NY, USA, 2020. 2

[11] M. Ferrara, A. Franco, and D. Maltoni. The magic passport. In *IEEE International Joint Conference on Biometrics*, pages 1–7, 2014. 4

[12] J. Frank and L. Schönherr. WaveFake: A data set to facilitate audio deepfake detection. In *Conference on Neural Information Processing Systems, NeurIPS, Datasets and Benchmarks Track*, 2021. 2

[13] P. Grassal, M. Prinzler, T. Leistner, C. Rother, M. Nießner, and J. Thies. Neural head avatars from monocular RGB videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 4

[14] S. Gu, D. Chen, J. Bao, F. Wen, B. Zhang, D. Chen, L. Yuan, and B. Guo. Vector quantized diffusion model for text-to-image synthesis. *CoRR*, abs/2111.14822, 2021. 4

[15] H. S. Heo, B.-J. Lee, J. Huh, and J. S. Chung. Clova baseline system for the VoxCeleb speaker recognition challenge 2020. *arXiv preprint*, 2020. 6

[16] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Conference on Neural Information Processing Systems, NeurIPS*, 2020. 4

[17] N. Hulzebosch, S. Ibrahimi, and M. Worring. Detecting CNN-generated facial images in real-world scenarios. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 4

[18] K. Ito and L. Johnson. The LJ Speech dataset. https://keithito.com/LJ-Speech-Dataset/, 2017. 2, 5

[19] A. Jain, P. Korshunov, and S. Marcel. Improving generalization of deepfake detection by training for attribution. In *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6, 2021. 1

[20] M. Jeong, H. Kim, S. J. Cheon, B. J. Choi, and N. S. Kim. Diff-TTS: A denoising diffusion model for text-to-speech. In *Conference of the International Speech Communication Association, Interspeech*, pages 3605–3609. ISCA, Sept. 2021. 4

[21] L. Jiang, R. Li, W. Wu, C. Qian, and C. C. Loy. DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 4

[22] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405, 2019. 4

[23] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of StyleGAN. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8107–8116, 2020. 4

[24] A. Kashkin, I. Karpukhin, and S. Shishkin. HiFi-VC: High quality ASR-based voice conversion. In *12th Speech Synthesis Workshop (SSW) 2023*, 2023. 5, 6

[25] H. Khalid, S. Tariq, M. Kim, and S. S. Woo. FakeAVCeleb: A novel audio-video multimodal deepfake dataset. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021. 5

[26] H. Kim, S. Kim, and S. Yoon. Guided-TTS: A diffusion model for text-to-speech via classifier guidance. In *International Conference on Machine Learning, ICML*, volume 162, pages 11119–11133. PMLR, 2022. 4

[27] J. Kim, J. Kong, and J. Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *Proceedings of the International Conference on Machine Learning, ICML*, volume 139, pages 5530–5540. PMLR, 2021. 6

[28] J. Kong, J. Kim, and J. Bae. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Conference on Neural Information Processing Systems (NeurIPS 2020)*, Dec. 2020. 5, 6

[29] P. Korshunov, A. Jain, and S. Marcel. Custom attribution loss for improving generalization and interpretability of deepfake detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8972–8976, 2022. 2, 4

[30] P. Korshunov and S. Marcel. Impact of score fusion on voice biometrics and presentation attack detection in cross-database evaluations. *IEEE Journal of Selected Topics in Signal Processing*, 11(4):695–705, 2017. 4

[31] P. Korshunov and S. Marcel. *Handbook of Biometric Anti-Spoofing: Presentation Attack Detection*, chapter A Cross-Database Study of Voice Presentation Attack Detection, pages 363–389. Springer International Publishing, 2019. 4

[32] P. Korshunov and S. Marcel. Vulnerability assessment and detection of Deepfake videos. In *International Conference on Biometrics (ICB 2019)*, Crete, Greece, June 2019. 1

[33] P. Korshunov and S. Marcel. Improving generalization of deepfake detection with data farming and few-shot learning. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, Jan. 2022. 2, 4

[34] I. Korshunova, W. Shi, J. Dambre, and L. Theis. Fast face-swap using convolutional neural networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3697–3705, Oct. 2017. 4

[35] P. Kwon, J. You, G. Nam, S. Park, and G. Chae. KoDF: A large-scale korean deepfake detection dataset. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10724–10733, Los Alamitos, CA, USA, oct 2021. IEEE Computer Society. 1

[36] J. Li, W. Tu, and L. Xiao. FreeVC: Towards high-quality text-free one-shot voice conversion. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. 2, 4, 5, 6

[37] Y. Li, M. Chang, and S. Lyu. In Ictu Oculi: Exposing AI created fake videos by detecting eye blinking. In *IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7, 2018. 4

[38] Y. Li, P. Sun, H. Qi, and S. Lyu. Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics. In *IEEE Conference on Computer Vision and Patten Recognition (CVPR)*, Seattle, WA, United States, 2020. 1

[39] S. Liu, Y. Cao, D. Wang, X. Wu, X. Liu, and H. Meng. Any-to-many voice conversion with location-relative sequence-to-sequence modeling. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29:1717–1728, apr 2021. 6

[40] D. M. Montserrat, H. Hao, S. K. Yarlagadda, S. Baireddy, R. Shao, J. Horvath, E. Bartusiak, J. Yang, D. Güera, F. Zhu, and E. J. Delp. Deepfakes detection with automatic face weighting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2851–2859, 2020. 4

[41] H. Nguyen, J. Yamagishi, and I. Echizen. Capsule-forensics: using capsule networks to detect forged images and videos. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2307–2311, 2019. 4

[42] I. Perov, D. Gao, N. Chervoniy, K. Liu, S. Marangonda, C. Umé, M. Dpfks, C. S. Facenheim, L. RP, J. Jiang, S. Zhang, P. Wu, B. Zhou, and W. Zhang. DeepFaceLab: A simple, flexible and extensible face swapping framework. *arXiv preprint*, 2020. 2, 4

[43] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, M. S. Kudinov, and J. Wei. Diffusion-based voice conversion with fast maximum likelihood sampling scheme. In *International Conference on Learning Representations (ICLR)*. OpenReview.net, 2022. 5, 6

[44] K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the ACM International Conference on Multimedia*, MM'20, pages 484–492, New York, NY, USA, 2020. 5

[45] R. Ramachandra, M. Stokkenes, A. Mohammadi, S. Venkatesh, K. B. Raja, P. Wasnik, E. Poiret, S. Marcel, and C. Busch. Smartphone multi-modal biometric authentication: Database and evaluation. *arXiv preprint*, 2019. 2, 4

[46] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. Zero-shot text-to-image generation. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 139, pages 8821–8831. PMLR, July 2021. 5

[47] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. In *International Conference on Learning Representations, ICLR 2021*. OpenReview.net, May 2021. 4

[48] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Liu. FastSpeech: Fast, robust and controllable text to speech. In *Conference on Neural Information Processing Systems, NeurIPS 2019*, pages 3165–3174, Dec. 2019. 4

[49] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. FaceForensics: A large-scale video dataset for forgery detection in human faces. *arXiv.org*, 2018. 1

[50] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. FaceForensics++: Learning to detect manipulated facial images. In *International Conference on Computer Vision (ICCV)*, 2019. 1, 4

[51] E. Sarkar, P. Korshunov, L. Colbois, and S. Marcel. Are GAN-based morphs threatening face recognition? In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2959–2963, 2022. 4

[52] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman. Synthesizing Obama: Learning lip sync from audio. *ACM Trans. Graph.*, 36(4):95:1–95:13, July 2017. 4

[53] M. Todisco, H. Delgado, and N. Evans. Constant q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. *Computer Speech & Language*, 45:516–535, 2017. 4

[54] C. Veaux, J. Yamagishi, and K. MacDonald. CSTR VCTK corpus: English multi-speaker corpus for cstr voice cloning toolkit. In *Public dataset*, 2017. 5, 6

[55] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros. CNN-generated images are surprisingly easy to spot...for now. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 4, 5

[56] X. Wang, X. Qin, T. Zhu, C. Wang, S. Zhang, and M. Li. The DKU-CMRI System for the ASVspoof 2021 Challenge: Vocoder based Replay Channel Response Estimation. In *Proc. ASVspoof Challenge workshop*, pages 16–21, 2021. 4

[57] Y. Wang, X. Chen, J. Zhu, W. Chu, Y. Tai, C. Wang, J. Li, Y. Wu, F. Huang, and R. Ji. HifiFace: 3D shape and semantic prior guided high fidelity face swapping. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1136–1142, Aug. 2021. 1

[58] Y. Wu, X. Tan, B. Li, L. He, S. Zhao, R. Song, T. Qin, and T. Liu. Adaspeech 4: Adaptive text to speech in zero-shot scenarios. *CoRR*, abs/2204.00436, 2022. 4

[59] J. Yamagishi, X. Wang, M. Todisco, M. Sahidullah, J. Patino, A. Nautsch, X. Liu, K. A. Lee, T. Kinnunen, N. Evans, and H. Delgado. ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection. In *Proc. ASVspoof Challenge workshop*, pages 47–54, 2021. 2, 4

[60] X. Yang, Y. Li, H. Qi, and S. Lyu. Exposing GAN-synthesized faces using landmark locations. In *ACM Workshop on Information Hiding and Multimedia Security*, pages 113–118, June 2019. 4

[61] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu. LibriTTS: A corpus derived from librispeech for text-to-speech. In *Conference of the International Speech Communication Association, Interspeech*, pages 1526–1530. ISCA, Sept. 2019. 2, 5, 6

[62] Y. Zhang, L. Zheng, and V. L. L. Thing. Automated face swapping and its detection. In *IEEE International Conference on Signal and Image Processing (ICSIP)*, pages 15–19, Aug. 2017. 4