An Open Dataset of Synthetic Speech

Artem Yaroshchuk*, Christoforos Papastergiopoulos†, Luca Cuccovillo*, Patrick Aichroth*, Konstantinos Votis†, and Dimitrios Tzovaras†

*Fraunhofer Institute for Digital Media Technology, Ilmenau, Germany
{artem.yaroshchuk@idmt.fraunhofer.de, luca.cuccovillo, patrick.aichroth}@idmt.fraunhofer.de

†Centre for Research and Technology Hellas, Thessaloniki, Greece
{papasterc, kvotis, dimitrios.tzovaras}@iti.gr

***Abstract*—This paper introduces a multilingual, multispeaker dataset composed of synthetic and natural speech, designed to foster research and benchmarking in synthetic speech detection. The dataset encompasses 18,993 audio utterances synthesized from text, alongside with their corresponding natural equivalents, representing approximately 17 hours of synthetic audio data. The dataset features synthetic speech generated by 156 voices spanning three languages, namely, English, German, and Spanish, with a balanced gender representation. It targets state-of-the-art synthesis methods, and has been released with a license allowing seamless extension and redistribution by the research community.**

***Index Terms*—datasets, neural networks, speech synthesis**

I. Introduction

In recent years, the application of deep learning has led to significant advances in synthetic speech generation, enabling the creation of speech that closely resembles real recordings. However, the advent of this technology has also raised concerns about the potential misuse of synthetic speech for malicious purposes, including the creation of fraudulent or harmful content. That, in return, requires new approach and technologies for synthetic speech detection.

The research community already started to pursue synthesis detection. However, a key hurdle in advancing these algorithms is the scarcity of diverse, high-quality datasets apt for training and testing. Many current datasets suffer from limited speaker diversity or use synthetic speech produced by outdated TTS models. This can result in overfitting and challenges in assessing the generalizability of detection algorithms. [1].

To address this challenge, we have developed a new synthetic speech recognition dataset. The main goal of this dataset is to provide researchers with a comprehensive and diverse collection of speech samples from different speakers, languages, and speaking styles, including both real speech samples and synthetic speech samples generated using up-to-date Text-To-Speech (TTS) speech synthesis models available in the literature. The dataset was generated from speech data aggregated from the open source TTS datasets which, to the best of our knowledge, should be GDPR-compliant, and it is covered by a permissive license that allows for extension and redistribution. In the following, it will be referred to as Open Dataset of Synthetic Speech (OSS).

This paper was supported by the EU H2020 AI4Media project (grant no 951911), and by the EU Horizon Europe vera.ai project (grant no 101070093).

The following sections are organized as follows: Section II provides an overview of the existing synthetic speech datasets, Section III describes the data collection process and an overview of the ODSS key characteristics, and finally, Section IV summarizes the paper and its contributions.

II. Related Work

The increased attention on synthetic speech detection in the research community has generated a demand for appropriate data. Consequently, numerous open datasets have been made available for the purposes of training and benchmarking. The following provides a description of the most prominent and widely-used synthetic speech datasets and their limitations, considering the open challenges of synthetic speech detection outlined in [1]. An overview of the available datasets is also reported in Table I.

A. The ASVspoof DF Challenge Dataset

The Automatic Speaker Verification Spoofing detection (ASVspoof) is a bi-annual challenge designed to promote the research of spoofing attacks against Automatic Speaker Verification (ASV) systems and their countermeasures. The first ASVspoof challenge took place in 2015, and has been updated over time to increase the difficulty of the tasks, reflecting the technological advances regarding synthesis. ASVspoof challenges provide datasets containing both bona fide and synthetic speech examples, which are generated using different TTS and Voice Conversion (VC) systems [3].

The 2021 version of the challenge featured three tasks: the Logical Access (LA) task focusing on detection of spoofed utterances injected into communication networks, the Physical Access (PA) task focusing on replay attacks and covering a variety of recording conditions, and the DeepFake (DF) task which targets synthetic speech detection, without specifically focusing on ASV scenarios.

The ASVspoof2021 DF dataset is derived from the Voice Cloning Toolkit (VCTK) corpus [8] and contains TTS and VC utterances distorted by 8 lossy codec compression configurations and an uncompressed one. It encompasses a wide range of voices and incorporates multiple synthesis algorithms, but due to the challenge-driven approach, it does not provide complete annotations regarding the synthesis and VC techniques

TABLE I
AVAILABLE DATASETS FOR THE DEVELOPMENT OF SPOOFING ATTACKS DETECTION MODELS.

Dataset	Real Utterances	Fake Utterances	Notes
FoR [2]	117,000	87,000	Gender balanced, class balanced and truncated versions of the original dataset are provided.
ASVSpoof [3]	5,128	25,096	Different datasets are provided to tackle three major forms of spoofing attacks, namely replay, voice conversion and speech synthesis.
WaveFake [4]	-	117,985	Fake utterances include a single female voice, resulting in data distribution bias.
ADD [5]	5,319	45,367	Low usability for systematic research due to restrictions in distribution.
Half-Truth [6]	26,554	26,554	Partially fake utterances are provided for the purpose of detecting manipulated real audio.
TIMIT-TTS [7]	-	79,120	Includes several multi-speaker synthesis methods while others are represented by a single female voice.
ODSS (proposed)	11,032	18,993	Provides multi-voice gender balanced synthetic speech utterances along with the corresponding bona fide counterparts.

involved, thus preventing a systematic study of the methods involved.

B. The Fake-or-Real (FoR) Dataset

The Fake-or-Real (FoR) [2] is a large corpora of pristine and synthetic English speech utterances generated by 7 commercial and open source TTS systems. The phrases and voices for the synthetic part are unrelated to the ones in the real recordings, which aggregated the Arctic [9], VoxForge [10], and LJ Speech [11] TTS corpora. The FoR dataset features a large variety of voices, however, synthetic ones do not correspond to the speakers included in the real subset. This disparity allows the possibility for a classifier model to learn features of the specific voices and classify based on the speaker identities, rather than on the synthesis traces. Therefore, whereas it could be used for testing purposes, it is not fit for training novel detection models. Furthermore, the dataset includes examples from autoregressive TTS models, but does not cover current state-of-the-art (SotA) synthesis methods which achieve similar or higher sample quality while having faster generation times.

C. The WaveFake Dataset

WaveFake is a collection of synthetic speech utterances generated using several flow-based and Generative Adversarial Network architectures [4]. While providing a TTS subset based on Mozilla Common Voice Corpus [12], the primary collection is sampled by extracting Mel spectrograms from LJSpeech [11] and JSUT [13] datasets and feeding them to the vocoder models trained on single female speaker voice provided in LJSpeech. Since the generated distribution covers only a single female voice, the dataset is unfortunately not suitable for training or benchmarking data-driven detection systems aiming at generalizing to unknown speaker voices.

D. The ADD Challenge Dataset

The Audio Deep Synthesis Detection (ADD) Challenge [5] adopts an approach similar to ASVspoof, but also consider a number of attack scenarios such as the addition of disturbances and background noises. The challenge provides associated datasets based on public multi-speaker Mandarin speech corpus AISHELL-3 [14], with non-overlapping speaker selections for training, dev, adaptation and test sets. While the corpora are featuring valuable properties such as large variety of voices, disjoint stage-sets and inclusion of noisy real and fake utterances, the gender, accent and age groups are not represented equally in the distributions. The main drawback, however, is that the dataset is not publicly available and can only be obtained by directly contacting the creators: its use is limited to non-commercial purposes and redistribution is not possible, neither in its original form nor in any derivative forms. The synthesis methods used for sampling have not been revealed by authors at present, and therefore the dataset presents low usability for systematic research.

E. The Half-Truth Dataset

Unlike the previous datasets, the Half-Truth dataset [6] focuses on detecting partially fake data, which is obtained by manipulating pristine speech with synthesized elements. In addition to the *Full* subset of completely synthetic utterances, it also includes a *Partial* subset obtained by modifying the original speech from the AISHELL-3 corpus through the splicing of synthetic audio segments. The dataset is designed to address a range of real-world scenarios that present a more difficult challenge than merely distinguishing between fully synthetic and entirely natural data. However, the synthetic data was collected from a *single* two-step TTS pipeline and thus does not cover a variety of synthesis algorithms.

F. The TIMIT-TTS Dataset

The TIMIT-TTS audio dataset [7] was devised for the needs of multimodal deepfake detection systems performing joint video and audio analysis. The authors concentrate on generating synthetic speech to accompany video data and propose a pipeline for converting a unimodal video deepfake dataset into a multimodal audio-video one.

TIMIT-TTS is based on 5 multi-voice and 7 single-voice synthesis systems, and considers as real counterpart the entire VidTIMIT corpus [15] of 430 video and audio recordings of 43 people. The authors provide several versions of the dataset which are obtined by applying 5 data augmentation techniques.

The TIMIT-TTS corpus covers a broad range of synthesis systems, some of which are represented by speech utterances from multiple voices. It also provides more challenging data by employing a number of post-processing operations, including dynamic time warping for lip synchronisation.

The main drawback of the dataset is that the generated data is not paired with a bona-fide counterpart and original voices corresponding to the synthetic ones can not be found in the reference dataset used by authors. Therefore, similarly as for the FoR dataset, the TIMIT-TTS corpus cannot be used reliably to train data-driven models for speech synthesis detection, but is rather limited to their test.

III. The ODSS dataset

In this section, we will introduce the methodology applied to generate ODSS. After describing the primary requirements, we will detail the collection process of real utterances, their pre-processing, the synthesis algorithms included in the final dataset, and therefore the final composition and related limitations.

A. Primary requirements

A primary requirement addressed by ODSS is the inclusion of samples generated by the most recent speech synthesis methods. This is of primary importance, since the latest Generative Adversarial Network (GAN)-based synthesis algorithms are demonstrating quality that is comparable to the one of the corresponding ground truth samples, while also achieving low inference latency. The ability to generate synthetic speech in real-time enables a large number of misuse scenarios, making it essential to incorporate utterances from such novel systems into the dataset. Most existing corpora, however, do not address this requirement since they have not been updated over the course of the years.

Furthermore, ODSS ensures congruence between synthetic and natural data. This requirement addresses the challenge that detection systems may learn hidden regularities introduced by *unintended* discrepancies between the distributions of classes – e.g., speaker identities. We therefore decided to pair synthetic data with corresponding real / natural data to which it will be compared: Each synthetic utterance has a bona fide equivalent, so that the same balance in real utterances is also found in their synthetic counterparts.

B. Real audio data collection

The first step in the data collection process involves identifying potential sources of natural speech data, including consideration of several aspects:

- The distribution of the real data needs to represent the general target population, which means it has to cover a variety of voice characteristics by including a diverse range of speakers. To ensure that the collected data is not biased towards a particular speaker identity, it should be balanced in terms of gender so that each speaker contributes an equal amount of content. Ideally, all dialects or accents of the target language should also be equally represented.
- Since the real data needs to be paired with its synthetic counterpart, the real speech subset must be suitable for training text-to-speech models. That means it must include a sufficient amount of transcribed real speech utterances representing each voice.
- Another important requirement for the collection process was to adhere to legal constraints related to data protection, and to the best of our knowledge, the underlying real data selected should be compliant with the European General Data Protection Regulation (GDPR). Moreover, we aimed at selecting a content license that allows for modification and redistribution, while ensuring its compliance with the usage rights of the underlying datasets used.

Given the data protection and pre-processing requirements, as well as the prerequisites imposed by the needs of model training, public real language datasets were the most appropriate data sources. Over the years, many corpora have been proposed with different licenses, access policies, and copyright status. These data collections can be divided into two groups:

- The first category consists of highly curated datasets obtained under controlled laboratory conditions. These corpora tend to offer high quality, but often at the cost of a limited data quantity. Additionally, they sometimes come with licenses that prevent redistribution in original or derivative forms.
- The second category consists of ad-hoc collections of data acquired in uncontrolled environments.

To address data scarcity, many independent research institutions have created multiple corpora based on annotated speech recordings made by volunteers. These volunteers read public domain books using their own devices, resulting in a wide variety of undocumented acoustic conditions, hardware qualities, recording pipelines, and post-processing operations (see: LibriVox and the related ProjectGutenberg). The resulting datasets were constructed by curating the collected recordings, which were normalized, for example, by removing leading and trailing audio, aligning the transcript, and evening out the volume. To encourage re-use and adoption, all of these corpora were released with permissive licenses (Public Domain, CC0, or CC-BY). However, they had to contend with the uncontrolled (and sometimes amateur) setup of the audio material, as well as with possible error in the associated transcriptions.

In order not to limit ODSS to English, we extended the search to try to cover at least two other European languages, as this could be useful to distinguish language-specific from language-independent synthesis features and to evaluate the advantages and disadvantages of multilingual models.

The aforementioned requirements were eventually fulfilled by the real speech corpora that are described below.

1) VCTK: The CSTR VCTK Corpus (Centre for Speech Technology Voice Cloning Toolkit) [8] contains speech utterances from 109 native English speakers from various age groups and regional accents. The audio was recorded in a controlled environment using an omnidirectional microphone and a small diaphragm condenser microphone with a very wide bandwidth. Original recordings were brought to a uniform 48 kHz sample rate and encoded with 16-bit FLAC format. The

text content read by each speaker was selected from Herald Glasgow newspapers, with each speaker reading the same Rainbow Passage and elicitation paragraph to provide more homogeneous data for training TTS systems. The dataset was published under "Attribution 4.0 International — CC BY 4.0" license.

2) Hi-Fi TTS: The Hi-Fi TTS dataset [16] consists of speech data of 6 female and 4 male English speakers gathered from LibriVox audiobooks and Project Gutenberg. The speakers were selected based on bandwidth and SNR estimation for the most recently added book. The corpus contains two subsets: *clean* with at least 40 dB SNR and *other* with at least 32 dB SNR. The corpus is licensed under CC BY 4.0 license and available publicly released at Open Speech and Language Resources (OpenSLR) repository.

3) HUI-ACG: HUI (Hof University Institute) Audio Corpus German [17] is a German language corpus featuring five speakers with 32 - 96 hours of audio and 97 hours of audio from additional 117 speakers sourced from LibriVox. The recordings were split into snippets with duration ranging from 5 to 40 seconds. The dataset is licensed under the "Creative Commons Attribution-ShareAlike" (CC BY SA 4.0) license.

4) SLR-ES: The Latin American Spanish OpenSLR collection [18] is a multidialectal corpus featuring various Latin American dialects of Spanish. The collection includes speech datasets for Argentinian, Chilean, Colombian, Peruvian, Puerto Rican, and Venezuelan Spanish with almost 40 hours of audio recorded from 174 male and female speakers in total. The transcripts are not normalized and partially speaker-wise duplicated, also audio utterances contain periods of silence as well as noise spikes of a non-speech nature. The corpora are published under the "Creative Commons BY-SA 4.0" license and hosted on OpenSLR repository.

C. Pre-processing

Most of the natural speech data captured was already pre-processed by the original authors for text-to-speech purposes. However, in some instances, the examples contain silence at the beginning and at the end of the utterance. Moreover, in the case of the Spanish OpenSLR datasets, some recordings include hardware noise spikes at their conclusion, making performance-based trimming inapplicable. Because of that, we used the Voice Activity Detection (VAD) MarbleNet model implemented by NVIDIA [19] to trim the beginning and the end of each utterance. A summary of the utilized datasets after trimming is provided in Table II.

TABLE II
PRE-PROCESSED TTS DATASETS.

Dataset	VCTK	HiFiTTS	SLR-ES	HUI-ACG
Language	English	English	Spanish	German
Number of utterances	43,475	323,978	23,820	65,379
Total duration	40.6h	287.6h	20.6h	141.1h
Average duration	3.36s	3.2s	3.11s	7.77s
Number of speakers (m/f)	46/61	4/6	80/89	3/2
Speaker-wise content (avg)	22.7m	28.8h	7.3m	28.2h

D. Speech Synthesis

Most of the neural network based TTS methods follow a two-stage pipeline. The first stage models the acoustic features (predominantly mel spectrogram) of speech based on text, optionally including grapheme-to-phoneme (G2P) conversion. The second stage synthesizes a high-fidelity raw waveform from the intermediate representation generated in the first stage.

The first TTS methods that achieved sample quality close to human speech were autoregressive architectures; however, these models generate a single sample per forward pass, resulting in very slow synthesis processes.

In the past few years, the application of advanced generative Deep Neural Network (DNN) techniques introduced TTS methods capable of generating speech of a quality comparable to human speech faster than real-time. To acquire synthetic utterances, we utilize two pipelines based on SotA 2-stage and end-to-end models, as described below.

1) FastPitch: FastPitch [20] is a parallel feed-forward network designed to generate mel-scale spectrograms for TTS based on Feed-Forward Transformer (FFTr) blocks introduced by its predecessor FastSpeech [21], and conditioned on fundamental frequency contours. An FFTr block is composed of a self-attention mechanism and 2-layer 1D convolutional network with rectified linear unit (ReLU) activation. FastPitch utilizes an FFTr block to produce a hidden representation from the input tokens, which is then forwarded to pitch and duration predictor modules. The second FFTr block transforms the hidden representation into the output spectrogram frame sequence. The primary improvement of FastPitch over its predecessor is its conditioning on the fundamental frequency estimated for each input token, enhancing the quality to the level of state-of-the-art autoregressive models. During inference, it is possible to modify a predicted pitch contour to control the speech prosody, and therefore raise expressiveness of the generated signal.

2) HiFiGAN: HiFiGAN [22] is a method that leverages a GAN [23] architecture and uses a mel-spectrogram as an input to produce raw waveforms. GAN approach is based on the simultaneous training of two models: a generator, which synthesizes new examples, and a discriminator, whose task is to classify examples as real (originating from the original training data) or generated. With each training step, the discriminator compares real recordings with the generator outputs, and updates its weights to maximise the probability of assigning a correct label to both. The generator employs the discriminator loss but has an objective to maximise it, consequently synthesizing examples more similar to the training data. While GAN architectures have been successful for image generation tasks, their application in the speech synthesis domain is more challenging [24]. MelGAN [25] introduced a multi-scale architecture that utilizes multiple discriminator models which operate on downsampled audio and advanced training techniques. These were further improved in HiFiGAN, which we included in ODSS, to achieve a sample quality comparable to autoregressive counterparts.

3) VITS: VITS [26] is a parallel end-to-end TTS architecture based on variational autoencoder (VAE) architecture and an adversarial training process adopted from HiFiGAN. VAE is a generative model that consists of two main components: an encoder, which compresses the input into a latent representation, and a decoder, which reconstructs the input from the latent representation. During the training process, the encoder refines its transformation to better preserve the structural information of the data and the decoder learns to generate the data from the encoded representations. VITS employs two encoders: a posterior encoder that transforms ground truth linear spectrograms and is used only during training, and a prior encoder that produces latent representations from text tokens. The decoder model is a HiFiGAN, which operates on the internal representations transformed by a normalized flow [27]. According to experimental results based on the Mean Opinion Score (MOS), VITS surpasses two-stage TTS methods and achieves scores close to those of ground truth.

E. Dataset composition

To provide the utmost naturalness of the synthetic data, we chose utterances that were longer than 2 seconds but capped their duration at a maximum of 10 seconds for the dataset.

We then balanced the dataset by gender, selecting an equal number of male and female speakers per dataset and discarding those with a smaller amount of content. Since some of the original corpora contain the same texts spoken by different speakers, we select unique text examples to avoid data loss in cross-speaker splits. As the original datasets vary significantly in terms of the number of speakers and the amount of content per speaker, we also attempted to equalize the total speech duration for each language and sample a nearly equal amount of content per voice within each dataset. Finally, all included audio utterances were re-sampled and converted to a common PCM WAV format with a sampling rate of 16 kHz.

To obtain synthetic utterances, we employed the NVIDIA open-source NeMo framework for FastPitch and HiFiGAN, and used pre-trained model checkpoints hosted at the NVIDIA NGC Catalog [28]. We selected models trained and finetuned on HiFi-TTS, Spanish OpenSLR and HUI-ACG datasets. For VITS, we used a checkpoint trained on HiFi-TTS to retrieve corresponding examples.

To collect VCTK, HUI-ACG and Spanish OpenSLR utterances generated by VITS, we instead used the CoquiTTS framework [29], which is licensed under the Mozilla Public License 2.0 and hosted at a public open source repository. We trained three models on each of the mentioned datasets for 1000 epochs each, using the 22050Hz recipe provided in the repository. Notably, all the models used were trained on phonemes using eSpeak NG backend [30] to normalize texts. The proposed dataset can be accessed online[1].

[1]A. Yaroshchuk *et al.*, *ODSS: An open dataset of synthetic speech*, Zenodo, 2023. DOI: 10.5281/zenodo.8370668 [31]

F. Limitations

The source corpora of real data on which ODSS is based showed a significant imbalance: Some had a low number of speakers with several hours of content each, while others had a high number of speakers with just a few minutes each, as previously reported in Table II. Therefore, the content was balanced with respect to original subset, to have the same number of speakers of each gender and the same amount of content per speaker. However, this also means that we could not achieve balance in terms of the amount of content per speaker globally. This additional balancing is reserved for future work as more source corpora become available.

For this initial version of ODSS, we were also unable to address is the impact of the recording equipment used to acquire real data. Corpora based on data collected from volunteers introduce a wide variety of undocumented acoustic conditions and hardware qualities, which may lead to unexpected biases in the underlying distribution. Our hope is that the amount of participants involved determines such a large variability that the bias due to mismatching recording conditions is minimum.

Lastly, even though we monitored the quality of the synthetic utterances within ODSS to avoid the presence of audible artifacts, we were not able to quantify the overall quality my means of mean opinion scores (MOS). This information might become part of future releases of the dataset.

G. Overview

The proposed dataset is based on a pristine speech sample with the key characteristics summarized in Table III and includes synthetic speech utterances generated by SotA neural TTS algorithms. Table IV presents the VITS models subset and Table V describes synthetic data corresponding to two-step pipelines of Fastpitch spectrogram generator and HiFi-GAN vocoder.

TABLE III
PRISTINE SPEECH SAMPLE

TTS dataset	HiFiTTS	HUI-ACG	SLR-ES	VCTK
Language	English	German	Spanish	English
Number of utterances	3,778	1,926	2,257	3,071
Total duration	8,160s	8,160s	8,161s	8,160s
Average duration	2.16s	4.24s	3.62s	2.66s
Number of speakers	8	4	96	48
Speaker-wise content (avg)	1020s	2040s	85s	170s

TABLE IV
VITS SYNTHESIZED UTTERANCES

Training dataset	HiFiTTS	HUI-ACG	SLR-ES	VCTK
Language	English	German	Spanish	English
Number of utterances	3,778	1,926	2,257	3,071
Total duration	9,002s	9,492s	8,855s	10,093s
Average duration	2.38s	4.93s	3.92s	3.29s
Number of speakers	8	4	96	48
Speaker-wise content (avg)	1125s	2373s	92s	210s

TABLE V
FastPitch/HiFiGAN synthesized utterances

Training dataset	HiFiTTS	HUI-ACG	SLR-ES
Language	English	German	Spanish
Number of utterances	3,778	1,926	2,257
Total duration	8,562s	8,038s	7,382s
Average duration	2.27s	4.17s	3.27s
Number of speakers	8	4	96
Speaker-wise content (avg)	1070s	2009s	77s

In total, the proposed dataset includes 18,993 audio utterances, synthesized from text, along with their corresponding natural counterparts. This results in a total of 17.06 hours of audio data. The corpora includes synthetic speech generated using 156 voices in three languages: English, German and Spanish, and is balanced by gender.

IV. Conclusion

In this work, we proposed the ODSS dataset: a multilingual, multispeaker dataset for synthetic speech detection that targets up-to-date SotA TTS methods. For each synthetic utterance, we provide a bona fide counterpart to ensure a balance between natural and generated speech. The dataset is gender-balanced and features three languages: English, Spanish and German.

This first release of the ODSS dataset is limited by the currently available TTS datasets, open-source implementations of the synthesis methods and publicly-available models. In future developments, we plan to incorporate a larger amount of synthesis pipelines and real speech datasets, and to investigate the impact of excluding text examples used in training synthesis models, to determine the performance of detection approaches on unseen data.

The ODSS dataset is covered by a CC BY-SA 4.0 license: We would gladly welcome contributions by additional researchers and research institutions, and we encourage a large-scale distribution of the entire corpus. Likewise, we welcome any future benchmark of synthesis detection upon this dataset, and plan to include the results in its accompanying metadata.

References

[1] L. Cuccovillo *et al.*, "Open challenges in synthetic speech detection," in *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2022, pp. 1–6.

[2] R. Reimao and V. Tzerpos, "FoR: A dataset for synthetic speech detection," in *IEEE International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, 2019, pp. 1–10.

[3] H. Delgado *et al.*, *ASVspoof 2021 challenge – speech deepfake database*, version 1.0, 2021. DOI: 10.5281/zenodo.4835108.

[4] J. Frank and L. Schönherr, "WaveFake: a data set to facilitate audio deepfake detection," in *NeurIPS – Datasets track*, 2021.

[5] J. Yi *et al.*, "ADD 2022: The first audio deep synthesis detection challenge," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 9216–9220.

[6] J. Yi *et al.*, "Half-Truth: A partially fake audio detection dataset," in *ISCA Interspeech*, 2021, pp. 1654–1658.

[7] D. Salvi *et al.*, "TIMIT-TTS: A text-to-speech dataset for multimodal synthetic media detection," *IEEE Access*, vol. 11, pp. 50 851–50 866, 2023.

[8] J. Yamagishi *et al.*, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," 2019, University of Edinburgh. The Centre for Speech Technology Research (CSTR). DOI: 10.7488/ds/2645.

[9] J. Kominek and A. W. Black, "The CMU arctic speech databases," in *ISCA Speech Synthesis Workshop*, 2004, pp. 223–224.

[10] "VoxForge." (2006), [Online]. Available: http://www.voxforge.org (visited on 09/29/2023).

[11] K. Ito and L. Johnson. "The LJ Speech dataset." (2017), [Online]. Available: https://keithito.com/LJ-Speech-Dataset/ (visited on 09/29/2023).

[12] "Common Voice." (2019), [Online]. Available: https://commonvoice.mozilla.org (visited on 09/29/2023).

[13] S. Takamichi *et al.*, "JSUT and JVS: Free Japanese voice corpora for accelerating speech synthesis research," *Acoustical Science and Technology*, vol. 41, no. 5, pp. 761–768, 2020.

[14] Y. Shi *et al.*, "AISHELL-3: A multi-speaker Mandarin TTS corpus," in *ISCA Interspeech*, 2021, pp. 2756–2760.

[15] C. Sanderson and B. C. Lovell, "Multi-region probabilistic histograms for robust and scalable identity inference," in *Lecture Notes in Computer Science (LNCS)*, vol. 5558, 2009, pp. 199–208.

[16] E. Bakhturina *et al.*, "Hi-fi multi-speaker english tts dataset," in *ISCA Interspeech*, 2021.

[17] P. Puchtler, J. Wirth, and R. Peinl, "Hui-audio-corpus-german: A high quality tts dataset," in *KI 2021: Advances in Artificial Intelligence*, 2021, pp. 204–216.

[18] A. Guevara-Rukoz *et al.*, "Crowdsourcing latin american spanish for low-resource text-to-speech," in *Language Resources and Evaluation Conference (LREC 2020)*, 2020, pp. 6504–6513.

[19] F. Jia, S. Majumdar, and B. Ginsburg, "MarbleNet: Deep 1D time-channel separable convolutional neural network for voice activity detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6818–6822.

[20] A. Łańcucki, "Fastpitch: Parallel text-to-speech with pitch prediction," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6588–6592.

[21] Y. Ren *et al.*, "Fastspeech: Fast, robust and controllable text to speech," in *NeurIPS*, 2019.

[22] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *NeurIPS*, 2020, pp. 17 022–17 033.

[23] I. Goodfellow *et al.*, "Generative adversarial nets," in *NeurIPS*, 2014, pp. 2672–2680.

[24] C. Donahue, J. McAuley, and M. Puckette, "Adversarial audio synthesis," in *International Conference on Learning Representations*, 2019.

[25] K. Kumar *et al.*, "MelGAN: Generative adversarial networks for conditional waveform synthesis," in *NeurIPS*, 2019.

[26] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning*, 2021, pp. 5530–5540.

[27] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *International Conference on Machine Learning*, 2015, pp. 1530–1538.

[28] E. Harper *et al.*, *NeMo: A toolkit for conversational AI and large language models*. [Online]. Available: https://github.com/NVIDIA/NeMo (visited on 09/29/2023).

[29] G. Eren and The Coqui TTS Team, *Coqui TTS*, version 1.4, 2021. DOI: 10.5281/zenodo.8009420.

[30] *Espeak ng text-to-speech*. [Online]. Available: https://github.com/espeak-ng/espeak-ng (visited on 09/29/2023).

[31] A. Yaroshchuk, C. Papastergiopoulos, and L. Cuccovillo, *ODSS: An open dataset of synthetic speech*, Zenodo, 2023. DOI: 10.5281/zenodo.8370668.