Evaluating Deepfake Speech and ASV Systems on African Accents

Abstract

Automatic Speaker Verification (ASV) systems authenticate individuals who interact with digital systems using speech. Conversely, deep neural network (DNN)-based voice synthesis systems enable the creation of convincing human voice deepfake audio, capable of deceiving both people and ASV systems. Misuse of such deepfake audio poses identity risks and threatens ASV system security. This study presents experimental research on the impact of deepfake audio with African accents on ASV systems. The results indicate that modern ASV systems are less susceptible to deepfake audio in African accents.

1 Introduction

ASV systems have leveraged the unique characteristics of the human voice to provide convenient biometric authentication and verification [7]. Their widespread applications include voice differentiation in virtual assistants like Amazon's Alexa [4] and Google's Assistant [1, 2]. In Africa, where access to advanced technologies is limited, ASV systems offer a viable solution for user authentication in native languages, addressing the challenges of password usage. However, the reliability of ASV systems is threatened by deepfake attacks, which utilize advanced DNN-based speech synthesis to clone human voices and deceive both humans and ASV systems [16]. These deepfakes pose a significant challenge to identity verification methods, highlighting the need for robust ASV systems to defend against such attacks [23]. This work hypothesizes whether ASV systems can be fooled by deepfake speech generated on African accents. Prior studies primarily concentrated on native English speakers. This research centers on African English speakers who frequently interact with digital systems. Experiments assessed a selected DNN-based deepfake audio system and an ASV system, demonstrating that ASV systems are less susceptible to deepfake audio deception in African accents.[1]

2 Automatic Speaker Verification Systems

ASV systems have gained popularity as a cost-effective and convenient method for user authentication using voice biometrics [9]. These systems leverage built-in microphones in devices such as cell phones, eliminating the need for additional hardware [9]. Enrolling users involves recording and registering their voices, creating a unique voice embedding that captures their distinctive characteristics. During verification, features are extracted from the user's voice, and access is granted if a significant similarity is detected. This research focuses on Resemblyzer, an advanced open-source tool based on deep neural networks [23], as well as the Gaussian Mixture Model(GMM)-based architecture called Bob Spear [14].

2.1 Bob Spear, a GMM-based ASV system

Bob Spear is an open-source toolbox for ASV systems, providing a customizable framework [14]. It follows a standard speech recognition pipeline with preprocessing, feature extraction, modeling,

[1] Cloned samples are available for listening at `https://kwekuyamoah.github.io/deefake_demo/`

5th Deep Learning Indaba Conference (DLI 2023).

enrollment, and score computation stages [5]. Bob Spear supports various modeling tools, such as GMM [20], and offers score fusion techniques. However, newer systems like Microsoft Azure and Resemblyzer, which utilize DNN models, have gained more efficiency compared to Bob Spear [23].

2.2 Resemblyzer, an SV_2TTS based ASV system.

The Resemblyzer [3, 10], is an advanced open-source speech recognition system based on the SV_2TTS architecture. It utilizes a deepfake audio system and a generalized end-to-end loss function to enhance training and improve the Equal Error Rate(EER) [10, 22]. The system requires a minimum of 30 seconds of speech data for voice enrollment, generating a unique numerical representation called an embedding [10, 22]. During verification, the system compares the evaluated utterance's embedding with the saved embedding in the database, achieving an EER of 4.5% [10]. The Resemblyzer has gained significant popularity and is widely used in research and verification applications.

3 Deepfake Speech

Driven by advancements in deep neural networks, deepfake speech technology has become integral to various applications [15], including voice-driven interfaces, video games, and chatbots [13, 15, 18]. These systems combine generative models with speaker embeddings, text, audio, and speaker identity data during training to optimize parameters and minimize the difference between synthesized and ground-truth audio [6, 15]. Through inference, deepfake speech systems can generate audio for unseen speakers by retrieving their speaker characteristics from a collection of generated audio [6, 15]. Speaker encoding, as proposed by Arik et al. [6], is the predominant approach adopted by state-of-the-art deepfake audio systems, enabling the production of high-quality deepfake audio.

3.1 SV_2TTS

SV_2TTS is a zero-shot deepfake audio system that can generate natural speech for multiple speakers [11]. It comprises a Speaker Encoder, Synthesizer, and Vocoder [17, 21, 22]. The Speaker Encoder captures speaker-specific characteristics, enabling similarity in the embedding space [11]. The Synthesizer generates log-mel spectrograms using grapheme or phoneme [17] and utilizes an attention-based architecture for high-quality synthesis. The Vocoder, based on WaveNet, converts mel-spectrograms into waveforms [17]. The SV_2TTS model achieves zero-shot transfer by training on many speakers [11].

3.2 AutoVC

AutoVC is a zero-shot and text-independent deepfake audio system that utilizes an autoencoder network [19]. Unlike other systems, it employs an information bottleneck within the autoencoder to preserve content while discarding speaker/style information [19]. The architecture consists of a speaker encoder, a content encoder, and a decoder [19]. The speaker encoder generates speaker embeddings for consistent representations of the same speaker [19]. The content encoder combines mel-spectrograms and speaker embeddings to generate content embeddings [19]. The decoder utilizes the WaveNet vocoder to convert spectrograms back into audio waveforms [19].

3.3 GAZEV

GAZEV, an extension of the StarGAN-VC framework, facilitates zero-shot voice conversion by introducing the adaptive instance normalization operator and an additional speaker embedding loss [24]. The adaptive operator adjusts normalization parameters based on speaker identity, improving the model's ability to generate voice conversions [24]. The speaker embedding loss ensures proximity between generated and target audio embeddings, resulting in accurate and convincing conversions [24]. GAZEV's architecture includes a Generator for generating speech and a Discriminator that uses gender information to assess authenticity, enhancing the system's effectiveness in producing high-quality voice conversions [24].

4 Experimental Setup

According to Wenger et al., [23], modern ASV systems are vulnerable to deepfake speech attacks; raising concerns about the potential threats posed by African-accented deepfake audios to ASV systems. To investigate this, the study assumes that an attacker has access to a limited set of speech samples from a target individual and aims to deceive ASVs into believing the target has been successfully verified. Based on this, we conduct empirical measurements to verify that ASV systems can successfully verify African users. A user study was used to evaluate the perceptual quality of deepfake audio generated with a DNN-based deepfake audio system on African accents. The experiments will be validated using Mean Opinion Score(MOS) (details are highlighted at A.1 in Appendix) [12], and EER as metrics [8].

4.1 Speech Data Collection and Implementation

The study collected 214 English voice recordings from 71 English-speaking African individuals from 17 African countries, comprising 33 females and 38 males aged 18-24 years. The total duration of the recordings was 1 hour; the average duration of recordings for the deepfake audio system was 9 seconds, while for the ASV system, it was 2 seconds. Participants were instructed to read an English sentence and repeat the phrase *"carry the water"* five times.[2] The recordings were saved in *.m4a* format using a voice recorder app on a Google Pixel 5A 5G phone and later converted to *.wav*. The files were anonymized by giving each participant a unique id, eg. *m_sv_p000.wav*. The files were divided into two datasets: one for the deepfake system, selected for its phonetic robustness, and the other for the ASV system, consisting of recordings of the phrase "carry the water."

4.2 Selected Deepfake Audio System and ASV System

The SV_2TTS system, a deepfake audio system, was chosen for this research due to its superior performance(MOS of 4.5) compared to the other models (AutoVC and GAZEV). SV_2TTS is an open-source implementation that is highly effective in generating deepfake speech. To complement the selection of SV_2TTS, the Resemblyzer ASV system was chosen as it was trained on similar datasets and with the same loss function as SV_2TTS, ensuring compatibility between the two systems.

5 Experiments and Results

The experiments for this research involve evaluating the performance of the SV_2TTS model in generating deepfake audio and assessing the effectiveness of the Resemblyzer model in identifying and authenticating speakers with African accents.

5.1 SV_2TTS on African Accents

This experiment evaluated speech quality, audio perception, and voice equatability on the SV_2TTS deepfake audio. We experimented on 119 synthesized speech instances targeting 71 speakers(33 females and 38 males). The English sentences used for this experiment were chosen because of their phonetic robustness and used in prior works like [23]; the sentence can capture most of the phenomes present in the English language. Feedback was collected through 12 surveys, with participants answering questions for each test. Only individuals who volunteered to participate in this study answered to each survey (all participants came from the Ashesi community). 50% of our participants identified as males, whereas the other half identified as females. All participants were over the age of 18, and no one was compensated.

5.2 Results from SV_2TTS on African Accents

The study found that the SV_2TTS deepfake audio system fell short of baseline scores for speech naturalness. The average MOS for generating deepfake audio on African accents was 2.83, 1.17 points lower than previous work [11]. Approximately 50% of participants perceived the deepfake audios as fake, 28% as real, and 23% undecided. The voice equatability analysis showed that around

[2]The selected English sentences read by participants can be referenced at `https://bit.ly/3K9MHdt`

Table 1: Summary of Results for SV2TTS on African Accents

Nature of Experiments	Results		
Voice Naturalness	**MOS:** 2.84 ± 0.032		
Audio Perception	*Fake:*49.6%	*Real:*27.7%	*Undecided:*22.7%
Voice Equatability	*Yes:*16.8%	*No:*79.8%	*Undecided:*3.4%

Table 2: Summary of Results for Resemblyzer on African Accents. Scores range between 0-1

Nature of Experiments	Similarity Score
Different Speaker	0.5 ± 0.1
Same Speaker	0.81 ± 0.1

80% of participants reported that the deepfake and genuine recordings did not belong to the same speaker. These results indicate the challenges in producing convincing deepfake audio for African accents using SV_2TTS. The results are summarized in table 1.

5.3 Resemblyzer on African Accents

The study collected 95 audio samples from 19 individuals(11 females and 8 males) to evaluate the Resemblyzer ASV system's effectiveness in enrolling and verifying speakers with African accents who spoke English. Participants recorded five utterances of the phrase *"carry the water"*; chosen randomly due to the varied ways one can say it out loud. Two experiments were conducted, one comparing utterances from the same individual (ground truth and authentic test utterances) and another comparing utterances from different individuals (ground truth and fake test utterances). The purpose was to assess the system's ability to accurately distinguish between authentic and fake utterances and also determine the similarity between utterances from the same individual and different individuals with African accents.

5.4 Results for Resemblyzer on African Accents

The first experiment involved assessing the Resemblyzer's performance by comparing ground truth utterances from one target against fake test utterances from different targets. This experimentation yielded an average similarity score of 0.50 ± 0.1, signifying that the system possesses a 50% likelihood of being deceived by an African accent distinct from that of the enrolled speaker. In the second experiment, the Resemblyzer successfully verified ground truth utterances against test utterances from the same target, achieving an average verification score of 0.81 ± 0.1, closely approximating the established threshold score of 0.84. These results highlight the Resemblyzer's accuracy in handling African-accented speech for verification purposes. A concise summary of these findings is presented in Table 2.

6 Conclusion

This study aimed to investigate the susceptibility of ASV systems to deepfake speech generated on African accents. Two main experiments were conducted using the SV_2TTS model for deepfake audio generation and the Resemblyzer ASV system. A custom dataset of 214 audio samples from 71 speakers was used, with 119 samples for deepfake generation and 95 for Resemblyzer's evaluation. The study found that modern deepfake systems struggle to generate high-quality audio for African accents, achieving a MOS of 2.83. However, the Resemblyzer effectively enrolled and authenticated African-accented speakers. These results highlight the limitations of current deepfake technology for African accents and emphasize the Resemblyzer's accuracy in ASV systems. Consequently, the research concludes that ASV systems are less susceptible to deepfake audio attacks on African accents. Future research should focus on developing enhanced deepfake systems tailored for generating realistic audio specifically for African accents.

Acknowledgments and Disclosure of Funding

We acknowledge the assistance and support rendered by the Computer Science and Information Systems Department of Ashesi University.

Statement on Ethics

The user study and data collection process strictly adhered to ethical guidelines set by the institutional IRB board to protect participant privacy and well-being. Informed consent was obtained from participants who willingly chose to take part. Stringent measures were implemented to anonymize audio recordings and safeguard participant identities. The recordings were securely stored on restricted-access cloud servers. After the study concludes, all collected data will be permanently deleted and destroyed to maintain participant privacy. The research prioritizes protecting personal information and follows strict data management protocols to ensure data confidentiality and security.

References

[1] [n. d.]. Google Assistant, your own personal Google. https://assistant.google.com/

[2] [n. d.]. Link your voice to your devices with Voice Match - Android - Google Assistant Help. https://support.google.com/assistant/answer/9071681

[3] [n. d.]. Resemblyzer. GitHub. Retrieved November 22, 2021 from https://github.com/resemble-ai/Resemblyzer. https://github.com/resemble-ai/Resemblyzer

[4] Amazon. [n. d.]. What Is Alexa Voice ID? https://www.amazon.com/gp/help/customer/display.html?nodeId=GYCXKY2AB2QWZT2X. publisher: Amazon.

[5] André Anjos, Laurent El-Shafey, Roy Wallace, Manuel Günther, Christopher McCool, and Sébastien Marcel. 2012. Bob: a free signal processing and machine learning toolbox for researchers. In *Proceedings of the 20th ACM international conference on Multimedia - MM '12*. ACM Press, Nara, Japan, 1449. https://doi.org/10.1145/2393347.2396517

[6] Sercan O. Arik, Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou. 2018. Neural voice cloning with a few samples. *arXiv:1802.06006 [cs, eess]* (Oct. 2018). http://arxiv.org/abs/1802.06006 arXiv: 1802.06006.

[7] Pascal Belin, Patricia E. G. Bestelmeyer, Marianne Latinus, and Rebecca Watson. 2011. Understanding Voice Perception: Understanding voice perception. *British Journal of Psychology* 102, 4 (Nov. 2011), 711–725. https://doi.org/10.1111/j.2044-8295.2011.02041.x

[8] Amira Boulmaiz, Noureddine Doghmane, Saliha Harize, Nasreddine Kouadria, and Djemil Messadeg. 2020. The use of WSN (wireless sensor network) in the surveillance of endangered bird species. In *Advances in Ubiquitous Computing*. Elsevier, 261–306. https://doi.org/10.1016/B978-0-12-816801-1.00009-8

[9] Anil Jain, Lin Hong, and Sharath Pankanti. 2000. Biometric identification. *Commun. ACM* 43, 2 (Feb. 2000), 90–98. https://doi.org/10.1145/328236.328110

[10] Corentin Jemine and Université de Liège > Bac sc Info. 2019. Master thesis : Real-Time Voice Cloning. (June 2019). https://matheo.uliege.be/handle/2268.2/6801

[11] Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, and Yonghui Wu. 2019. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *arXiv:1806.04558 [cs, eess]* (Jan. 2019). http://arxiv.org/abs/1806.04558 arXiv: 1806.04558.

[12] Biing Hwang Juang, M.Mohan Sondhi, and Lawrence R. Rabiner. 2003. Digital Speech Processing. In *Encyclopedia of Physical Science and Technology*. Elsevier, 485–500. https://doi.org/10.1016/B0-12-227410-5/00178-2

[13] Dan Jurafsky and James H Martin. [n. d.]. *Speech and language processing* (3rd edition draft ed.). https://web.stanford.edu/~jurafsky/slp3/

[14] Elie Khoury, Laurent El Shafey, and Sebastien Marcel. 2014. Spear: An open source toolbox for speaker recognition based on Bob. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Florence, Italy, 1655–1659. https://doi.org/10.1109/ICASSP.2014.6853879

[15] Hafiz Malilk and Raghavendar Changalvala. 2019. Fighting ai with ai: fake speech detection using deep learning. In *Audio Engineering Society*. Journal of the Audio Engineering Society, Porto, Portugal. https://par.nsf.gov/servlets/purl/10109075

[16] Dibya Mukhopadhyay, Maliheh Shirvanian, and Nitesh Saxena. 2015. All Your Voices are Belong to Us: Stealing Voices to Fool Humans and Machines. In *Computer Security – ESORICS 2015*, Günther Pernul, Peter Y A Ryan, and Edgar Weippl (Eds.). Vol. 9327. Springer International Publishing, Cham, 599–621. https://doi.org/10.1007/978-3-319-24177-7_30

[17] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: a generative model for raw audio. *arXiv:1609.03499 [cs]* (Sept. 2016). http://arxiv.org/abs/1609.03499 arXiv: 1609.03499.

[18] Pavol Partila, Jaromir Tovarek, Gokhan Hakki Ilk, Jan Rozhon, and Miroslav Voznak. 2020. Deep learning serves voice cloning: how vulnerable are automatic speaker verification systems to spoofing trials? *IEEE Communications Magazine* 58, 2 (Feb. 2020), 100–105. https://doi.org/10.1109/MCOM.001.1900396

[19] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson. 2019. Autovc: zero-shot voice style transfer with only autoencoder loss. *arXiv:1905.05879 [cs, eess, stat]* (June 2019). http://arxiv.org/abs/1905.05879 arXiv: 1905.05879.

[20] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn. 2000. Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing* 10, 1-3 (Jan. 2000), 19–41. https://doi.org/10.1006/dspr.1999.0361

[21] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, R. J. Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. 2018. Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions. *arXiv:1712.05884 [cs]* (Feb. 2018). http://arxiv.org/abs/1712.05884 arXiv: 1712.05884.

[22] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. 2018. Generalized End-to-End Loss for Speaker Verification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Calgary, AB, 4879–4883. https://doi.org/10.1109/ICASSP.2018.8462665

[23] Emily Wenger, Max Bronckers, Christian Cianfarani, Jenna Cryan, Angela Sha, Haitao Zheng, and Ben Y. Zhao. 2021. "Hello, it's me": deep learning-based speech synthesis attacks in the real world. *arXiv:2109.09598 [cs, eess]* (Sept. 2021). https://doi.org/10.1145/3460120.3484742 arXiv: 2109.09598.

[24] Zining Zhang, Bingsheng He, and Zhenjie Zhang. 2020. GAZEV: GAN-Based Zero-Shot Voice Conversion over Non-parallel Speech Corpus. *arXiv:2010.12788 [cs, eess]* (Oct. 2020). http://arxiv.org/abs/2010.12788 arXiv: 2010.12788.

A Appendix

A.1 Mean-Opinion Scale

Table 3: Overview of the Mean-Opinion Score.

Score	Quality	Listening Effort
5	Excellent	No effort required
4	Good	No appreciable effort required
3	Fair	Moderate effort required
2	Poor	Considerable effort required
1	Bad	No meaning understood with reasonable effort