Deepfake Speech on African Accents: How do Modern Systems Perform?

Kweku Andoh Yamoah [1] Hussein Fuseini [1] Dennis Asamoah Owusu [1] David Adjepon-Yamoah, Ph.D. [1]

[1]Department of Computer Science and Information Systems, Ashesi University, Ghana

Motivation and Objectives

- Africa faces challenges in implementing advanced technology, creating a unique opportunity to explore features like the human voice.
- Despite the potential of using voice as a biometric feature, its effectiveness is hindered by the existence of speech/audio deepfakes[MSS15].
- Deepfake audio technology has reached a level of sophistication capable of deceiving both humans and Automatic Speaker Verification Systems (ASVs) [MSS15].
- Previous studies on human perception of deepfake speech consistently show that deepfakes failed to deceive humans, but in trusted settings deception is achieved[Wen+21].
- However, the majority of research has focused on Western Speakers of English, raising the question of how modern deepfake systems will perform on African Accents.
- The main objectives of this study are to investigate the performance of these systems on African Accents and assess the vulnerability of African speakers to such emerging threats.

Deepfake Systems

This study explores the creation of deepfakes using DNN-based architectures, which involve manipulating a person's image, audio, or video to resemble someone else [Ngu+21]. Deepfakes leverage advanced machine learning and artificial intelligence techniques to deceive individuals [Ngu+21]. The key strength of deepfakes lies in their utilization of DNNs, inspired by the structure and function of brain neurons [Kie+20]. These interconnected neural networks perform complex non-linear calculations, enabling diverse applications like speech recognition and semantic word embedding [Kie+20; JM].

The generation of deepfakes involves three primary approaches: the encoder-decoder approach, the deep autoencoder approach, and the generative adversarial network (GAN) approach. However, this study specifically focuses on highlighting the significance of the encoder-decoder approach. The reason for this choice lies in its capacity to produce realistic and compelling results, as evidenced by multiple studies [Kie+20; Ngu+21; Wen+21], and its extensive adoption in practical applications through the use of open-source technologies[Jia+19; Kie+20; Ngu+21; Qia+19; Wan+18].

Encoder-Decoder Approach

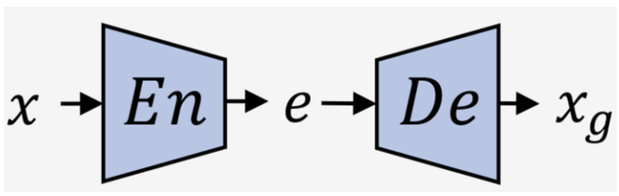

Figure 1. Encoder Decoder Architecture adapted from [ML21]

Deepfake Audio Systems

In this section, we concentrate on deepfake audio generation systems that utilize non-parallel data to achieve many-to-many speaker conversion. We specifically investigate state-of-the-art systems capable of zero-shot conversion, enabling the transformation of the voice of unseen speakers with only a limited number of their utterances [ZHZ20].

1. **SV_2TTS** The SV_2TTS is a zero-shot deepfake audio system comprising of a Speaker Encoder, Synthesizer, and Vocoder[Jia+19], with the Speaker Encoder capturing speaker-specific characteristics to ensure similarity in the embedding space [Jia+19]. The Synthesizer utilizes an attention-based architecture can generate log-mel spectrograms using grapheme/phoneme information. The Vocoder, converts mel-spectrograms into waveforms [Oor+16].
2. **AutoVC** - AutoVC is a zero-shot and text-independent deepfake audio system that utilizes an autoencoder network [Qia+19]. The architecture comprises a speaker encoder, content encoder, and decoder, with the speaker encoder generating consistent speaker embeddings, and the content encoder combining mel-spectrograms and speaker embeddings to create content embeddings[Qia+19].
3. **Gazev** GAZEV is an extension of the StarGAN-VC framework enabling zero-shot voice conversion by introducing adaptive instance normalization and an additional speaker embedding loss [ZHZ20]. GAZEV's architecture includes a Generator for speech generation and a Discriminator that uses gender information to assess authenticity, resulting in high-quality voice conversions [ZHZ20].

Experimental Setup

To investigate the performance of deepfake speech systems on African accents, we structure our experiment as follows: an attacker, denoted as $\boldsymbol{A}$, attempts to create deepfakes on a set of speech samples, denoted as $\boldsymbol{S}$, for a target individual, denoted as $\boldsymbol{T}$. The attacker can only obtain a specific set of speech samples from the target, referred to as $\boldsymbol{S_T}$. The experimental setup is based on the following assumptions:

1. An attacker requires less than 5 minutes of speech samples from the target ($\boldsymbol{T}$).
2. An attacker utilizes a DNN-based deepfake system optimized for African-accented speech samples.
3. The attacker's goal is to generate audio data convincing enough to deceive others into believing it is the target ($\boldsymbol{T}$).

The experiments will follow this approach:

- Conduct empirical measurements to verify that deepfake audio systems can successfully generate synthetic speech for all accents.
- Perform a user study to evaluate the perceptual quality of deepfake audio generated with a DNN-based deepfake audio system on African accents.

To validate these experiments, the research utilizes the Mean Opinion Score (MOS), a subjective test that assesses the naturalness and similarity of deepfaked audio compared to the true audio [JSR03].

Table 1. 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

Experimental Setup Cont

Speech Dataset

1. **Custom data**: 119 English voice recordings from 71 English-speaking African students and faculty members at the institution.
2. **Total duration**: 1 hour (60 minutes), with each recording averaging 9 seconds (0.15 minutes) in length.
3. **Data collection**: Participants repeated sentences from a provided list, with repetitions ranging from one to ten per participant.
4. **Recording method**: Voices recorded using a voice recorder app on a Google Pixel 5A 5G phone, saved in *.m4a* format.
5. **Anonymization**: Recorded audio files anonymized through renaming.

Selected Deepfake Audio System

1. Experiments focused on the SV_2TTS system, a deepfake audio system.
2. SV_2TTS exhibited exceptional performance with an MOS of 4.5, outperforming AutoVC and GAZEV models.
3. The system is an open-source implementation, utilizing the encoder-decoder approach for highly effective deepfake speech generation.

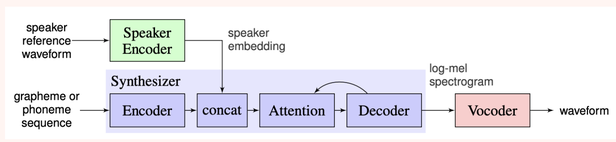

Figure 2. Architecture of SV_2TTS adapted from [Jia+19]

Ethics

The user study and data collection followed ethical guidelines set by the institutional IRB board, ensuring participant privacy and well-being. Informed consent was obtained from willing participants. Stringent measures, including audio anonymization and secure storage on restricted-access servers, were employed to safeguard identities. After the study, all data will be permanently deleted to maintain participant privacy. The research prioritizes data confidentiality and security through strict data management protocols.

Experiments and Results

Experiments

The study focused on assessing the quality of speech generated by the SV_2TTS deepfake audio system using target speech samples from individuals of African ancestry who speak English. Three main aspects were evaluated:

- **Speech Quality**: Determining the naturalness and listening effort of SV_2TTS-generated speech and comparing it to established baseline scores.
- **Perception of Audio**: Assessing whether listeners perceived the deepfake audio as genuine or fake, aiming to gauge its convincing and authentic quality.
- **Voice Equatability**: Examining whether listeners could distinguish between the deepfake audio and the original speaker's voice, measuring similarity.

The experiment involved 71 target speakers, and each speech sample was used as input to the SV_2TTS model to generate deepfake audios. In a survey,participants listened to the audios and rated their quality, identified if they were real or fake, and determined if they came from the same speaker.

Results

- Speech Quality

Table 2. The overall MOS achieved in this study for deepfake audio generation with SV_2TTS was 2.84, falling 1.17 points below the baseline scores reported by [Jia+19], indicating difficulties in producing satisfactory quality audio for African accent.

Dataset	Deepfake System	MOS
Custom Dataset	SV_2TTS	2.84 ± 0.032

- Perception of Audio

Table 3. Approximately 50% of participants perceived SV_2TTS-generated deepfake audios on African accents as fake, indicating challenges in producing clear audios for this speech type.

	Participant Perception		
	Fake	Real	Undecided
Custom Dataset	49.6%	27.7%	22.7%

- Voice Equatability

Table 4. Approximately 80% of participants noted the voice samples as not belonging to the same speaker, highlighting SV_2TTS's difficulty in creating convincing deepfake audio for African accents.

	Participant Perception		
	Yes	No	Maybe
Custom Dataset	16.8%	79.8%	3.4%

Summary and Conclusion

- **Aim**: Investigate deepfake systems' performance in African accents and assess African speakers' vulnerability to emerging threats.
- **Experiments**: Three main experiments using the SV_2TTS model with a custom dataset of 119 audio samples from 71 speakers.
- **Findings**: Modern deepfake systems struggle to produce high-quality audio for African accents (MOS of 2.84); voices identified as fake and not belonging to the same person.
- **Conclusion**: The study concludes that modern deepfake audio systems exhibit limited performance when generating audio for African accents, resulting in a reduced threat to Africans at present. However, it emphasizes the potential vulnerability that may arise in the future with advancements in technology. Continued research and improved methods are essential to monitor and address this evolving challenge effectively.