BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2023
http://excel.fit.vutbr.cz

Deepfake Dataset for Evaluation of Human Capability on Deepfake Recognition

Karolína Radačovská*

Abstract

This paper is focused on the human ability to recognize audio synthetic media. It involves an experiment whose primary purpose is to determine whether a targeted group of people can distinguish a synthetic recording from an original one. Participants of the experiment will be presented with a survey containing a pairs of recordings. Synthetic audio recordings were clustered into groups with similar quality rating scores, identifying the recordings that are the most convincing. The results of the experiment were analyzed to measure participants' ability to accurately distinguish between the synthetic and original recordings across all quality groups.

*xradac00@stud.fit.vutbr.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

Synthetic media created by artificial intelligence has grown incredibly recently, providing many creations of fake images, videos, and audio recordings. The main problem of these media is the increasing numbers of cases of deepfake frauds. Hence, the question is: Can people recognize when they are being scammed?

Lots of experiments were conducted in this field of technology. In [1], researchers did an experiment comparing human abilities to detect synthetic speech with audio deepfake detectors. They found out that audio deepfake detectors miss some characteristics recognized by humans. Audio samples were in English; people with this language as their native language performed better at recognizing audio deepfakes than audio deepfake detectors. Another experiment [2] was conducted on college grounds with students. They were asked whether the audio clips were real or synthetic. They focused only on English speakers and various aspects of grammar in audio deepfakes. Overall, students were more likely to detect synthetic speech in complex and shorter sentences. However, these experiments focused on something other than deepfake quality according to objective measures that can evaluate the samples based on various characteristics.

The main goal of this research is to determine whether a targeted group of people can detect synthetic audio media while having the original one to hear the difference. Moreover, the provided synthetic audio samples were clustered into groups with similar quality score.

2. Deepfakes

Deepfakes are synthetic media created by AI using deep neural networks. They can appear as images, videos, or audio recordings. Audio deepfakes can be made by two methods. Text-to-speech (TTS) method presents creating synthetic speech from written text. Voice Conversion (VC) method transforms the existing original recording to a synthetic one with targeted voice [3]. For this work, VC method was chosen since there have to be a pair of recordings. Each pair has the same sentence and speaker, while the creation differs.

The original recording were taken from an existing audio library, Mozilla Common Voice dataset, which provides wide range of audio samples in many languages [4].

3. Chosen Language

The chosen languages for this work are Slovak and Czech languages. These languages are native to the most Brno University of Technology students. Inspired by many experiments, we are curious about people detecting synthetic recordings in their native language. Moreover, this work also seeks answers to

whether people can recognize audio deepfakes in a language similar to their native language.

Four speakers were selected from each language for the experiment - two women and two men.

4. Quality Measures

A dataset with pairs of recordings is evaluated using a proposed quality system. This categorization is supposed to help us find the best deepfake - the hardest one to detect.

For this purpose, objective measures for speech quality were studied. Ultimately, appropriate measures were chosen to estimate the rate of audio deepfake samples.

4.1 Log-likelihood Ratio

One of the best speech recognition software is Phonexia. For quality estimation of deepfake samples, a Phonexia Browser client application was used. This software applies speech processing technologies to recordings and visualizes the results [5, 6].

Phonexia uses a Log-likelihood Ratio (LLR) metric to determine a speaker's score. The likelihood ratio is widely used in signal processing and machine learning. It represents how many times more likely the data occur under one model than another. A higher score means a better fit between a given recording and a model. Lastly, the logarithm of the score is taken. Eventually, the system converts the LLR score to percentage [6].

4.2 Perceptual Evaluation of Speech Quality

Perceptual Evaluation of Speech Quality (PESQ) measure is a standardized objective method for evaluating the perceived quality of speech signals. This method is often used in telephone networks and codecs, and ITU-T recommends it [7]. This measure analyses the speech signals and tries to predict the subjective mean opinion score (MOS). It is computed as a linear combination of the average disturbance and the average asymmetrical disturbance [8]. The results range is typical $[-0.5, 4.5]$ with a higher score indicating better quality. For this work, PESQ implementation [9] was used.

4.3 Mel Cepstral Distortion

Another commonly used measure in assessing the quality of synthesized speech is Mel Cepstral Distortion (MCD). It measures how different two sequences of mel cepstral coefficients are [10, 11]. In this work, this MCD implementation [12] was used. This implementation requires precalculated mel cepstral coefficients and then computes MCD. One of the variants expects a possible difference in timing and uses dynamic time warping. This method can "align" sequences in time in case they do not match. The results in this implementation are in the range of $[4, 8]$ where the higher score means more distortion.

5. Quality Evaluation

All audio samples' quality evaluations were converted to a percentage and averaged. The collected evaluations were clustered into groups using a cluster algorithm k-means.

The k-means algorithm is an iterative algorithm that randomly selects k data points representing centroids. Then the data points are assigned to the nearest centroid. The algorithm calculates the distance between data points and centroids and assigns the points to the nearest centroids. Then the centroids are recalculated as the mean value of the data in clusters. The algorithm repeats these steps until the centroids are not moving [13].

6. Experiment

6.1 The Dataset

So far, the dataset contains 121 recordings from 8 speakers, clustered in 4 groups of the quality system. However, the future version of dataset will contain more speakers to create a bigger dataset.

6.2 Experiment Design

The experiment contains demography questions like age, gender, native language, and previous experiences with deepfakes (if there are any). The central part of the survey is pair recordings from which people are asked to try and detect the synthetic one. The targeted group is students, considering they have more experience with internet as they spend hours on social media, which means they are more likely to be familiar with deepfake media.

7. Conclusion

This paper studies the ability of humans to recognize original and synthetic audio recordings. The experiment involved giving participants pairs of audio recordings and asking them to identify the synthetic one. The deepfakes were rated based on quality from objective measures, and the results were analyzed to determine participants' ability to differentiate between the two types of recordings.

Acknowledgements

I would like to thank my supervisor Ing. Anton Firc for providing me with help and guidance throughout my work. Provided feedback and direction helped me a lot to make decisions about this project.

References

[1] Nicolas M Müller, Karla Pizzi, and Jennifer Williams. Human perception of audio deepfakes. In *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*, pages 85–91, 2022.

[2] Gabrielle Watson, Zahra Khanjani, and Vandana P. Janeja. Audio deepfake perceptions in college going populations. *CoRR*, abs/2112.03351, 2021.

[3] Yipin Zhou and Ser-Nam Lim. Joint audio-visual deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14800–14809, 2021.

[4] Common Voice. Common voice. online.

[5] Phonexia. We are phonexia. https://www.phonexia.com/about-us/, 2023.

[6] Phonexia. *Phonexia Browser*. Phonexia s.r.o., Chaloupkova 3002/1a, 612 00 Brno-Královo Pole.

[7] Philipos C Loizou. Speech quality assessment. *Multimedia analysis, processing and communications*, pages 623–654, 2011.

[8] Yi Hu and Philipos C Loizou. Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on audio, speech, and language processing*, 16(1):229–238, 2007.

[9] Rafael G. Dantas Miao Wang, Christoph Boeddeker and ananda seelan. Pesq (perceptual evaluation of speech quality) wrapper for python users, May 2022.

[10] John Kominek, Tanja Schultz, and Alan W Black. Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion. In *SLTU*, pages 63–68, 2008.

[11] Erika Brandt, Frank Zimmerer, Bistra Andreeva, and Bernd Möbius. Mel-cepstral distortion of german vowels in different information density contexts. In *Interspeech*, pages 2993–2997, 2017.

[12] Matt Shannon. Mel cepstral distortion (mcd) computations in python., 2014.

[13] František V. Zbořil. Základy umelé inteligence, 2022.