electronics

MDPI

Article

Comparison of the Ability of Neural Network Model and Humans to Detect a Cloned Voice

Krzysztof Milewski [1], Szymon Zaporowski [1,2,*] and Andrzej Czyżewski [1]

[1] Multimedia Systems Department, Gdansk University of Technology, Narutowicza 11/12, 80-233 Gdańsk, Poland
[2] Audio Acoustics Laboratory, Gdansk University of Technology, Narutowicza 11/12, 80-233 Gdańsk, Poland
* Correspondence: szyzapor@pg.edu.pl; Tel.: +48-790626729

Abstract: The vulnerability of the speaker identity verification system to attacks using voice cloning was examined. The research project assumed creating a model for verifying the speaker's identity based on voice biometrics and then testing its resistance to potential attacks using voice cloning. The Deep Speaker Neural Speaker Embedding System was trained, and the Real-Time Voice Cloning system was employed based on the SV2TTS, Tacotron, WaveRNN, and GE2E neural networks. The results of attacks using voice cloning were analyzed and discussed in the context of a subjective assessment of cloned voice fidelity. Subjective test results and attempts to authenticate speakers proved that the tested biometric identity verification system might resist voice cloning attacks even if humans cannot distinguish cloned samples from original ones.

Keywords: biometrics; deep learning; neural networks; transfer learning; speech processing; speaker verification; voice cloning; voice synthesis

check for updates

Citation: Milewski, K.; Zaporowski, S.; Czyżewski, A. Comparison of the Ability of Neural Network Model and Humans to Detect a Cloned Voice. *Electronics* **2023**, *12*, 4458. https://doi.org/10.3390/electronics12214458

Academic Editor: KC Santosh

Received: 18 September 2023
Revised: 25 October 2023
Accepted: 27 October 2023
Published: 30 October 2023

1. Introduction

With the rapid development of biometric authentication technologies, especially facial recognition-based ones, voice biometrics still lags behind in the application area. The main reason is its relatively low resistance to attacks by impostors. Attack techniques are constantly improving, not least because of the use of artificial intelligence to organize them. In this article, however, the authors wanted to show that the opposite can be true, i.e., whereby a properly trained speaker verification model can be highly resistant to such attacks.

The main objective of this research was to train and test a voice cloning model for use in attempts to confuse speaker verification systems and to test their resilience to attacks. Another objective of the study was to test the quality of the cloned voice samples in terms of the naturalness and quality of sound compared to the original recording. An average result of the subjective assessments provided by surveyed human subjects was used to benchmark the cloned voice quality.

The selected and used voice database is the "Common Voice" collection created by the Mozilla Foundation [1]. The features that especially influenced its selection were the transparency of the file with transcriptions of the recorded speech excerpts and the fact that it included information about the gender, age, and background of the speaker, which made it possible to manually select diverse recordings.

For this work, 100 samples of the cloned voice were generated based on the available neural network models included in the open repository [2], then employed in the simulation of an attack on the Deep Speaker speech recognition system [3] trained in our department, i.e., the Department of Electronics Telecommunications and Informatics at the Gdańsk University of Technology. It was decided to compile two questionnaires to investigate peoples' subjective opinions on the quality of the cloned voice samples to show to what extent listeners can distinguish between a human voice and a voice digitally generated in

the cloning process. The results were matched with the experiment results simulating an attack on the speaker identity verification system. Cloned voice samples and the method of subjective assessment of their quality are explained. In turn, the developed model for speaker authentication, its training, and validation is revealed. Finally, the experimental research results are presented and discussed in the chapter entitled, "Assessing the resilience of systems to voice cloning attacks" (Section 5), followed by a discussion and conclusions.

2. Speaker Authentication and Voice Cloning Methods

In speaker authentication, the key piece of information is the answer to the question, "Who is speaking?" so speech content is often meaningless in the context of recognition. The technology, therefore, focuses on using acoustic speech features that reflect learned language patterns and the anatomy of individuals. Speaker verification can be divided into two phases: registration and authentication. Registration involves recording the speaker's voice and extracting its features used within an accepted template or model; then, this information is stored in a database. These features can be various acoustic parameters of the voice, including spectrum [4], perceptual linear prediction (PLP) [5], or linear predictive coding (LPC) [6].

The recognition process is analogous to the registration process; however, once the features are extracted, they are matched against the data stored in memory. Then, if the top–down decision threshold adopted by the developer or the person implementing the system is exceeded, authentication ends with a positive result. A simplified diagram of the verification process is shown in Figure 1.

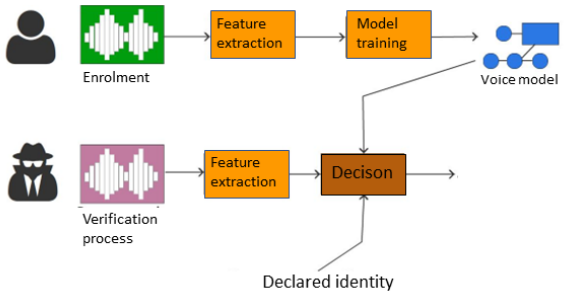


Figure 1. Diagram of the voice authentication system.

The first speaker verification application is the device patented in 1983 by Cavazza and Ciaramella to reduce noise in telecommunication networks by adaptively calculating the filtering threshold based on parameters obtained from sentences spoken by speakers [7]. This circuit juxtaposed the acoustic parameters obtained with the averaged values of these parameters stored in memory so that the adaptive noise threshold could be continuously adjusted.

In the following years, many systems using speaker recognition technology were developed. However, the most extensive progress occurred between 1990 and 2010, when technological developments spread access to relatively high computing power to the standard computer user. This was one of the main factors that made the evolution of scientific fields possible based on deep learning, a machine learning technique characterized by high efficiency and adaptability [8].

A. Contemporary speaker authentication solutions

Voice analysis is increasingly being used as a component of biometric security systems. Thanks to the increasing popularity of methods for identifying and verifying the speaker, backed by the high accuracy of these systems, biometric voice markers are beginning to be used on par with established handwriting recognition systems.

One popular solution used for speaker verification is MARF (Modular Audio Recognition Framework), an open-source software platform (currently available software version 0.3.0.6) consisting of a set of algorithms designed for preprocessing and processing audio files and speech recordings [9]. Both the learning and recognition processes are preceded

by preprocessing that involves passing the normalized audio file through a filter array, followed by feature extraction using Fast Fourier Transform (FFT) and Linear Predictive Coding (LPC), among others. Based on the features extracted in this way, a stochastic neural network model is created based on Markov models. Then, in recognizing a speaker, his or her declared identity is matched with the model using the cosine similarity method [9]. Solutions based on the MARF platform cannot infrequently be found in systems running on hardware with relatively low computing power, such as SBC (Single Board Computer) and less efficient machines. An example of a MARF system implementation of this type is PiWho [10], created to facilitate the creation of voice biometrics solutions on the Raspberry Pi platform.

One of the contemporary systems in use is Generalized end-to-end (GE2E) [11], created by Google LLC for its voice solutions. The distinguishing feature of this system is its training, which processes multiple utterances simultaneously in the form of packets consisting of the utterances of multiple speakers, where several utterances are related to each speaker. The extracted features are delivered to the input of a long short-term memory (LSTM) network that uses feedback to analyze the input data more efficiently in less time than other standard machine learning methods [12]. In addition, it returns embedding vectors, based on which a similarity matrix is built, which is used in the speaker verification process. The results of experiments conducted by the developers of this system suggest its advantage over other solutions [11].

On the other hand, there are systems like wave2vec developed by Facebook AI Research (FAIR), which represents a significant advancement in the field of self-supervised learning for speech representations [13]. This framework is designed to pre-train speech models using unlabeled audio data, which can then be fine-tuned for various downstream tasks, such as automatic speech recognition (ASR), with a small amount of labeled data. Wav2vec 2.0 introduces a novel self-supervised objective, which masks parts of the input audio and trains the model to predict the masked portions using context from the unmasked parts. The model leverages a multi-layer convolutional neural network (CNN) architecture, capturing intricate patterns in the audio signal. The embeddings generated by wav2vec 2.0 have demonstrated remarkable effectiveness, setting new benchmarks across various speech-processing tasks. Pre-training with wav2vec 2.0, followed by fine-tuning on labeled data, significantly outperforms fully supervised baselines, particularly in low-resource scenarios.

A different approach is to use an attention mechanism for speaker verification. This innovative technique aims to enhance the model's capability to focus on salient parts of the audio signal that are most indicative of a speaker's identity, thereby improving the accuracy and robustness of speaker recognition tasks. The model architecture leverages a self-attention mechanism that allows it to weigh different parts of the input sequence differently, providing a more flexible and adaptive way to capture speaker-specific characteristics. This is particularly crucial in variable conditions where the speaker's voice may be affected by noise, emotion, or other external factors. The authors demonstrated through their experiments that this approach yields significant improvements over traditional methods, especially in challenging conditions, establishing self-multi-head attention as a promising direction for future research and development in speaker recognition [14].

The Deep Speaker platform [3] is the backbone of modern neural speech analytics solutions. It maps extrapolated speech features into a hypersphere in which this similarity between speakers is measured using cosine similarity, thus achieving high speaker recognition. This system is used for identifying, verifying, and clustering voice recordings.

Using ResNet [15] facilitates training deep neural networks spliced through a stack of residual blocks (ResBlocks). Each of these blocks has several direct links between the outputs of a lower layer and the inputs of a higher layer. It allows the frequency dimension to be kept constant across all layers of the weave, improving the efficiency of the speaker identification process. In addition to ResNet, a GRU (Gated Recurrent Unit) network [15] is also used, whose machine learning performance is comparable to that of LSTM [16]. Speech

recognition experiments have shown that parameter values obtained by GRU trained faster and diverged less frequently [17].

Deep Speaker accepts raw audio as input in small batches of data, which then undergo preprocessing to extract audio features using DNN functions [18]. These data are then passed to the input of the ResCNN and DeepSpeech2 splicing networks, which includes a GRU network. The resulting frame data are further averaged at the utterance level, after which the link and normalization layers assign features to a given speaker.

In the final phase, the triplet loss layer takes three samples as input: an anchor (an utterance by a particular speaker), a positive example (another utterance by the same speaker), and a negative example (an utterance by another speaker). Based on these, corrections are made to increase the final cosine similarity value between the anchor and the positive example, and to decrease the same value for the anchor and the negative example.

B. Classical voice cloning methods

The issue of voice cloning is primarily related to the noticeable rise in the popularity of neural networks over the past decade. Speech synthesis is mainly based on the automatic translation of the written text into speech (Text-to-Speech). TTS systems are based on graphemes, letters, and groups of letters, transcribed into phonemes, the smallest unit of speech sound. Consequently, the primary resource in these systems is text and not directly sound. Synthesis of this type is performed by converting raw text into equivalent words. This process is called text normalization, preprocessing, or tokenization. In the next step, phonetic transcriptions are assigned to each word and are matched into groups corresponding to phrases and sentences. The next step is to convert the identified phonetic groups into linguistic representations and sounds [19]. The quality of a speech synthesizer is judged by its similarity to the human voice and the listeners' ability to understand it. The general scheme of operation of a typical TTS voice synthesizer is shown in Figure 2.

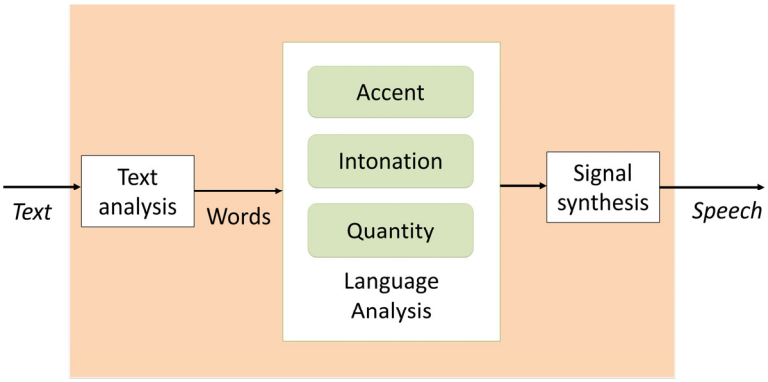


Figure 2. Operation diagram of a typical TTS voice synthesizer.

With the development of speech synthesis technology, it became possible to gain more and more control over characteristics other than intelligibility, such as the timbre itself and the prosody of the voice. It made the potential of synthesizing speech sounds similar in sound to any human voice, or voice cloning, a reality. Initially, the best results in this field could be achieved using concatenation methods, i.e., recording a speaker's voice to collect a database of speech segments. It is assumed that the speech segments contain as many speech phonemes as possible and then process and synthesize them to produce natural-sounding speech [20]. Today, the most popular synthesis and voice cloning solutions are based on deep learning technology, as are most speaker authentication systems.

C. Overview of modern solutions to voice cloning

The core of today's voice cloning technologies is a generative model for converting text to speech, known as Tacotron (TTS). The synthesis process has so far required breaking down the process into many separate steps, such as text analysis, acoustic modeling, and

audio synthesis. Since building these components required extensive expertise and could contain design flaws, Tacotron's developers decided to create a comprehensive system that synthesizes speech directly from the text. The synthesis is preceded by network training requiring only a pair of audio files and its transcription as input [21]. As a result of the continued development of the architecture above, developers have succeeded in creating a newer iteration of this system called Tacotron2. In addition, thanks to the WaveNet vocoder, which generates speech signals based on the Mel Spectrogram [22,23], recordings produced using this platform are much closer in sound to human voice recordings than the previous iteration of this software [24].

The SV2TTS tool [2] is a system that enables text-to-speech (TTS) conversion by generating speech sounds spoken by the voice of various speakers, including those whose speech samples were not included in the network training files [23]. The main components of SV2TTS are three independently trained components: a speaker encoder in the form of the GE2E network, trained for speaker verification, used to generate speaker embedding vectors, which is the template for the cloning process; a synthesizer based on the Tacotron2 network, which generates a Mel Spectrogram based on the input text and speaker embedding data; and the previously discussed WaveNet vocoder [24].

A standard among today's popular TTS systems is the use of Mel Spectrograms as intermediate representations of the data pipeline. They also have a separately trained acoustic model and vocoder [25]. However, despite featuring good prosody quality and audio clarity, solutions of this type are not optimal.

The process of generating spectrograms is based on the Fourier transform; thus, phase information is lost [26], while establishing vocoder training on these spectrograms can cause errors in the learning process, which reduce the quality of the resulting sound. A new comprehensive speech synthesis system with automatically learned speech representations and a jointly optimized acoustic model and vocoder called DelightfulTTS 2 is being developed [27]. DelightfulTTS 2 achieves better recording quality by using such a solution than competing solutions. The distinguishing feature of this system is that it replaces the mel-cepstrum-derived speech representation using a codec network. The codec network learns speech representations at the frame level based on vector-quantized auto-encoders with adversarial training (VQ-GAN), supported by a symmetric encoder-decoder network.

Several advanced voice recognition technologies partially resist attacks based on voice cloning. For example, they were applied to voice parameter analysis, where the voice authentication system analyzes parameters such as frequency, amplitude, and formant distribution to determine unique voice characteristics and authenticate the user. Biometric authentication uses voice and biometric features, such as speaking rhythm and speech rate, to authenticate the user. Deep learning models, such as neural networks, can detect subtle differences between natural and artificial voices and distinguish between them. Finally, two-factor authentication can be required to enhance system security, such as providing a password or single-use code via SMS. However, this option is less convenient for the user than a fully automatic authorization solution that is sufficiently immune to attacks by impostors.

It is worth noting that neither of these technologies is 100% immune to attacks. Still, their use may increase the security of voice authentication systems, reducing the risk of attack success.

3. Experiment Preparation

The project, the results of which are presented in this part of the article, was carried out employing a PC running the Ubuntu 20.04.3 LTS x86_64 distribution of the GNU/Linux operating system with the specifications listed in Table 1. The preparation and implementation of the experimental study are discussed in the remainder of this chapter.

Table 1. Specifications of the computing platform.

CPU	Intel i5—4690 K (4) @ 3900 GHz
GPU	NVIDIA GeForce GTX 970, 4 GB GDDR5
RAM	G.Skill Ares, DDR3, 16 GB (4 × 4 GB), 2400 MHz CL11

3.1. Selection of Voice Base and Recordings for Cloning

Following content analysis of the most popular voice databases, such as VoxCeleb [28,29], CSTR VCTK Corpus [30], and LibriSpeech [31], it was decided to choose the Common Voice corpora from the Mozilla Foundation. This choice was made because it includes additional information about the speakers in its collection, such as their age, gender, and possible ethnicity, which potentially provided important information for evaluating the quality of the chosen voice cloning tool.

For the study, 100 cloned samples were generated based on an equal number of recordings selected from the Common Voice collection. Factors influencing the selection of specific recordings were the sex of the speaker, age range, and, most importantly, the desire to avoid having a particular voice clip appear more than once. In the end, it was possible to obtain a set consisting equally of voice clips from speakers identifying themselves as male and female. Figure 3 shows the percentage of speakers in the respective age ranges and Figure 4 shows the same but divided by gender. In the Supplementary Materials, there is an additional chart showing the nationality structure of the speakers based on data available from the Common Voice database.

In the Supplementary Tables, Tables S1 and S2 show complete lists of the recordings selected for cloning, along with their transcriptions. Table S1 shows the selected recordings of the female speakers, while Table S2 shows the male ones.

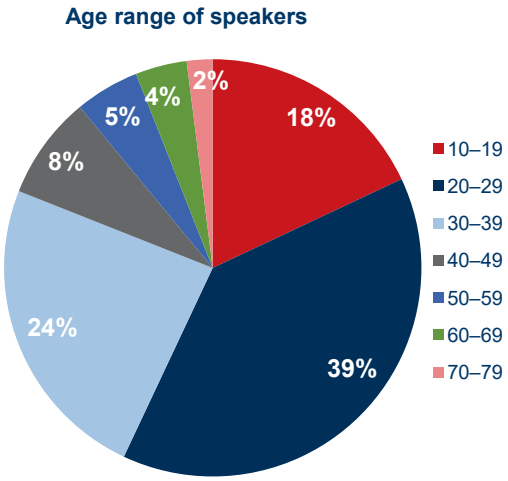


Figure 3. Percentage share of speakers in given age ranges in the selected set of recordings.

3.2. Preparation of Cloned Voice Samples

It was decided to use the previously mentioned Real-Time Voice Cloning repository created by Jemine [2], implementing the SV2TTS framework mentioned in the previous section, i.e., Tacotron synthesizer, WaveRNN vocoder, and GE2E encoder. This decision was based on the fact while the research project was being performed, during the pilot tests, this system ensured the subjectively best audio quality of the resulting files.

The demo_toolbox.py script made available as part of the repository above, written in Python, operates the voice cloning tool through a graphical interface, as shown in Figure 4. First, it is possible to select an audio file containing the speech of the speaker whose voice is to be copied. Then, typing the desired script in the text window initiates synthesis of an output file with the cloned voice. During this process, the tool displays a Mel Spectrogram

for the synthesized recording, making it easier to spot potential synthesis errors before exporting the output file. Operating the script is simple and intuitive. Generated sample lengths depend on input text length and are output as wave files, coded in PCM16, and sampled with a sampling rate of 16 kSa/s.

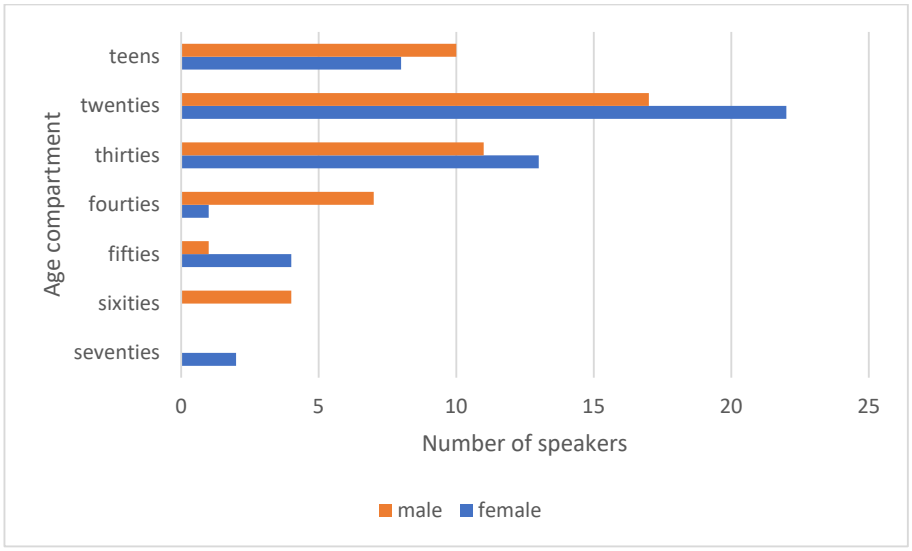


Figure 4. Number of speakers in given age ranges in the selected set of recordings divided by gender.

3.3. Preparation of Surveys

Given that the best proficiency in recognizing and analyzing the human voice is determined by physical skills developed in humans over evolution, it was decided to employ human listeners to assess the quality of voice excerpts generated by voice cloning systems. Information obtained in this way also provides research material for comparison with the results of the speaker identity verification carried out by a neural network.

The first survey on distinguishing a cloned sample from an original one aimed to estimate to what extent the resulting recordings resemble a natural human voice. For this purpose, it was decided to juxtapose pairs of recordings and ask the listeners which recording in the pair was generated with the voice cloning tool. To prepare for this survey, 13 recordings were selected from a previously prepared pool of recordings of female and male speakers. Table S3 in the Supplementary Tables presents a list of the recordings used in the survey. Each recording had two versions—the original and the cloned (generated). Some recordings were also repeated in the survey to verify the credibility of the survey participants.

Using Audacity [32], audio files were prepared as concatenations of the normalized components of the recordings in both the "clone–original" and "original–clone" configurations. White noise was also added to the resulting audio files at a level that hid any background noises in the original voice samples and minor sound artifacts in the cloned samples. This procedure was undertaken to avoid a situation where the listener is guided by secondary characteristics of the recordings instead of focusing on the voice fidelity.

Aiming to prepare a survey that did not require the supervision of a person in charge of the project, it was decided to make it easier for the survey participants to distinguish between samples in pairs using a visual cue. To this end, using the DaVinci Resolve program [33], the video files were compiled containing "1" and "2" visual cues following each other as the equivalent paired audio sample was played. Given the desire to obtain as many responses as possible to the survey, it was decided to draft it on the user-friendly Google Forms platform [34]. This procedure made preparing the question sheet convenient and the survey very accessible. It made it possible to sample a wider group of respondents than would have been possible when conducting it in a classic form.

The respondent's goal was to listen to a pair of recordings and then indicate which one was generated using a deep neural network. The answer was given as a single-choice value, defined as: 1—"Recording 1," 2—"Rather recording 1," 3—"Don't know," 4—"Rather recording 2," 5—"Recording 2."

To test the quality of the generated cloned audio files, it was decided to use the webMUSHRA tool [35], which provides listening tests in the MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor) methodology [36]. This methodology is typically used to evaluate the sound quality of audio files subjected to lossy audio compression algorithms. This study involves the interviewee listening to sets of recordings subjected to different compression algorithms and rating their quality on a scale of 0 to 100 against an uncompressed reference recording.

Audio files were selected to obtain results that would allow a subjective evaluation of the voice cloning system's capabilities regarding the quality of the samples generated, for which the absence of background noises and sound artifacts characterized the voice in the recordings. This selection was intended to facilitate the evaluation of the voice sound for those taking part in the survey. The set of recordings for each question and the original voice recording included two independently generated cloned samples and a voice recording. They were cloned with a sampling rate of 8 kS/s, including lower-sampled files in the set aimed to determine whether the interviewed subjects were credible in evaluating the listening samples. This procedure is used by default in the MUSHRA subjective testing methodology. The list of recordings selected for the survey is shown in Table S4 in the Supplementary Tables.

For the webMUSHRA version 1.4.3 tool to work correctly, it was necessary to prepare the files within each set so that they were all sampled at the same frequency and their duration was identical. The FFmpeg program was used to compose the sets and resample the recordings, while the Audacity [32] software version 3.2 was used to append digital silence to the recordings to pad them out to the desired length in seconds. The interface of the quality survey for cloning voice samples is shown in Figure 5. The survey examining the subjective evaluation of the quality of the recordings obtained was hosted on a private virtual server.

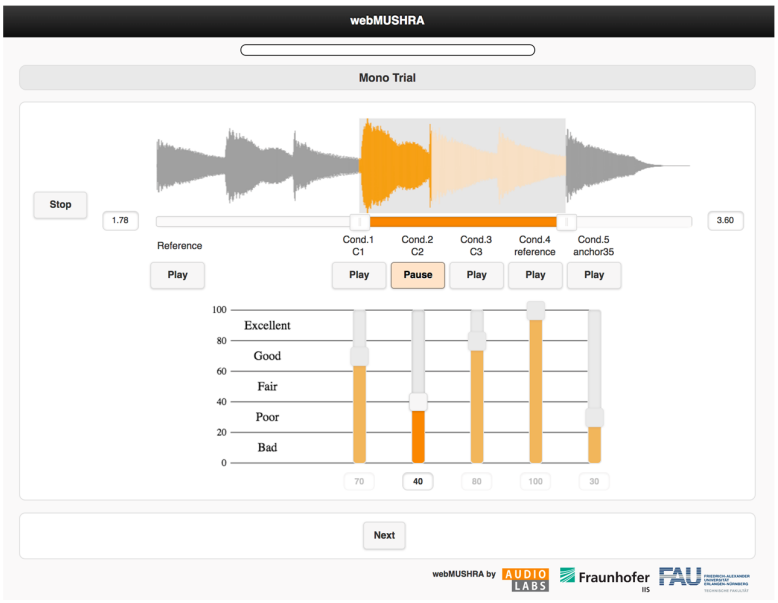


Figure 5. Voice sample cloning quality survey interface employing webMushra tool [33].

4. Applied Model for Speaker Authentication

The system that was employed to test resistance to a potential attack using cloned voice samples is the Deep Speaker platform. This platform was chosen because it is one of

the most modern, continuously developed security systems based on voice biometrics and it is very popular in professional solutions.

A network model previously trained on the LibriSpeech dataset and designed for scientific work conducted at the Department of Multimedia Systems at the Faculty of Electronics, Telecommunications, and Computer Science of the Gdansk University of Technology was used. Nearly all parts of the LibriSpeech corpus were used for training purposes, including the train-clean-100, train-clean-360, and train-other-500 parts. The dev-clean, dev-other, test-clean, and test-other parts were used for the evaluation process, also with the VCTK dataset. The data split for the training-test process was set at an 80–20 proportion, meaning nearly 770 h of recorded speech was used for training. Translating this into the number of speakers means that recordings of more than 2300 speakers were used. As for the sex breakdown of the data, the ratio split is near 1:1 (48% females and 52% males). Such a division excludes the possibility of undesirable phenomena associated with unbalanced data by sex; for example, a bias that provides better recognition of male voices than female ones. According to studies performed on the LibriSpeech dataset for ASR, even a disproportion of 30:70 in sex representation usually does not provide significant changes in terms of model performance [37].

4.1. Audio Preprocessing and Parameterization

The dataset was preprocessed, and the parameterization process was conducted before training. Unlike in the original paper on Deep Speaker, only Mel Frequency Cepstral Coefficients (MFCC) were used instead of combining filterbanks and MFCC as an audio signal parameterization method. The main reason for the change in the parameterization approach was to check how competitive MFCCs could be compared with the original filter bank and MFCC options.

The Librosa [38] and Python Speech Features [39] libraries were employed for the whole signal preprocessing. Librosa was used to pre-emphasize the signals and remove all sounds that did not represent speaking. For the preemphasis phase, Librosa's effect.preemphasis function was used with a coefficient value equal to 0.97, which refers to the HTK preemphasis filter employed for the MFCC calculation [40]. For audio trimming Librosa's effects.trim function was applied. This function removes leading and trailing silence with a given threshold (top_db parameter) from a given audio file. Also, in this stage, the frame length was equal to 1024 and the hop length to 512 samples. The MFCC calculation was performed using Python Speech Features. The Hamming window was utilized, with a length of 25 ms and with a 10 ms step between successive windows. The FFT size was 1024, with 40 calculated coefficients of MFCC. Also, 40 filters in the filterbank were employed, using Slaney's definition of mel filters [41,42]. Several experiments were conducted with different parameter values and using Librosa for generating MFCCs, but the previously mentioned parameters gave the best results. An essential operation that was changed compared to the original Python Speech Features approach was raising the log-mel amplitudes to a power of three before calculating the discrete cosine transform (DCT), which leads to reducing the influence of low-energy components [42].

4.2. Model Architecture, Training, and Validation

The utilized model implemented the Deep Speaker model developed by Baidu [3]. A modified implementation of the available model was employed [43]. We chose the Deep Speaker model because of its distinctive architecture; it uses both ResNet-type blocks and convolutional layers. In addition, training using distance metric learning methods in the form of a triplet loss function is also used. This architecture is characterized by low values of the Equal Error Rate metric, which is 2.33% for the model presented in the publication [3]. Deep Speaker architecture was used also because it is an end-to-end solution that allows quick implementation. The training was performed employing a workstation with the specifications shown in Table 2.

Table 2. Specifications of workstation employed for training models.

CPU	AMD Ryzen Threadripper 2990 WX
RAM	128 GB
GPU	2× Nvidia Titan RTX
Storage space	500 GB m.2 PCIe NVMe SSD + 2 × 4 TB HDD
Operating system	Ubuntu 18.04.05 LTS

Two training approaches were utilized: 1. Softmax pretraining plus triplet-loss training, and 2. Softmax training, only. First, TensorBoard [44] was employed as a framework for tracking the training, then changed to Weight & Biases [45]. Figures 6 and 7 depict the sample training loss and validation loss for one of the Deep Speaker model triplet trainings. Softmax training should be treated as a pretraining phase because of the better performance of triplet-loss trained models when comparing both approaches regarding EER and accuracy. In general, Softmax training was performed for 100–200 epochs; triplet-loss training was used.

Figure 6. Example of training loss function for Deep Speaker model triplet-loss training.

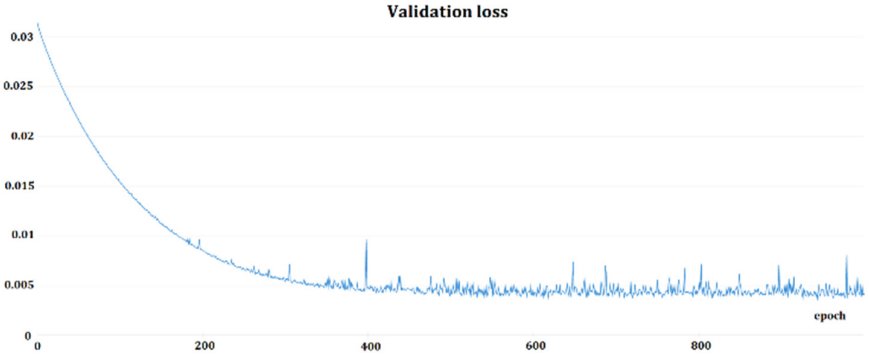


Figure 7. Example of validation loss function for Deep Speaker model triplet-loss training.

Triplet loss represents a specific training method; in some cases, the training leads to overfitting, and in other cases, it provides better generalization. In the presented approach, triplet loss was performed by selecting 1 positive sample, 1 anchor, and 99 negative samples, similar to the article describing the Deep Speaker architecture [3].

The models employed in experiments were trained as follows:

ResCNN_checkopint_299—200 softmax epochs and 99 triplet epochs;
ResCNN_checkopint_674—200 softmax epochs and 474 triplet epochs;
ResCNN_checkopint_706—150 softmax epochs and 556 triplet epochs;
ResCNN_checkopint_959—200 softmax epochs and 759 triplet epochs;
ResCNN_checkopint_984—200 softmax epochs and 784 triplet epochs.

Evaluation of selected models showed that the lowest accuracy rate was 96% for ResCNN_checkopint_299 with EER 5.6%, and the highest accuracy was 99.6% for model

ResCNN_checkopint_959 with EER 2.5%. The rest of the models had an EER and accuracy value between the given numbers.

5. Assessing Resilience of Systems to Voice Cloning Attacks

The cosine similarity between the original voice recording and the recording generated by cloning was used as a decision parameter. Following the results of repeated pilot experiments, the value of 0.75 was chosen as the decision threshold representing the mean value for all selected models indicated by the Equal Error Rate (EER) parameter.

A cosine similarity factor was calculated using the following equation (Equation (1)):

$$C_s = \frac{A \cdot B}{||A|| \cdot ||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}} \quad (1)$$

where A and B are two n-dimensional vectors for the biometric voice sample, and A_i and B_i are the i-th components of the A and B vectors, respectively. The cosine similarity factor was calculated for each biometric voice sample, comparing the round-robin series. The next stage is used to find the threshold value of cosine similarity, which determines the authenticity of a person in the whole system. To obtain this value, the value of the threshold was iteratively changed from the range taken by the cosine similarity (from 0 to 1) starting from 0 and ending with 1, and it was checked how the False Acceptance Rate (FAR) and False Rejection Rate (FRR) metrics described by the formulas (Equations (2) and (3)) changed depending on the chosen value. The FRR and FAR are crucial metrics in biometric systems used to evaluate their performance and user experience. FRR measures the rate at which the system incorrectly rejects a legitimate user; i.e., it quantifies the system's likelihood of mistakenly denying access to an authorized individual. FRR helps in setting an appropriate threshold that balances convenience and security. By adjusting this threshold, system administrators can influence the FRR and FAR to meet specific security and usability requirements. FAR measures the rate at which the system incorrectly identifies an unauthorized user as a legitimate one; i.e., it measures the system's likelihood to incorrectly accept an imposter. In other words, FAR quantifies the probability of a false-positive identification in a biometric system. EER is a specific point on the ROC (Receiver Operating Characteristic) curve where FAR and FRR intersect, meaning they are equal. The EER represents the threshold setting where the rates of false acceptance and false rejection are balanced. The EER is a critical point of interest because it provides an optimal trade-off between FAR and FRR. It represents the point where, for a given system, the two error rates are minimized and are approximately equal. Achieving a lower EER is often the goal of system optimization. The presented metrics are commonly used in biometric systems.

An example of the FAR and FRR curves used to calculate the Equal Error Rates is presented in Figure 8. Each dot represents one iteration of the searching loop—one value for FAR (depicted in red) and FRR (blue). The EER is represented as a golden dot at the intersection of the FAR and FRR curves; thus, the EER is the value for which the False Acceptance Rate and False Rejection Rate values are the same.

$$FAR = \frac{impostors\ cosine\ similarity\ score\ exceeding\ threshold}{all\ impostors\ scores} == \frac{FP}{(FP + TN)} \quad (2)$$

$$FRR = \frac{genuine\ cosine\ similarity\ scores\ falling\ below\ threshold}{all\ genuine\ scores} = \frac{FN}{(FN + TP)} \quad (3)$$

where FP = False Positives, TN = True Negatives, FN = False Negatives, and TP = True Positives. The threshold value used for the conducted experiments was calculated using the following formula (Equation (4)):

$$T_{thr} = \frac{1}{n} \sum_{i=1}^{n} T_i \quad (4)$$

where Ti is the value of the threshold (cosine similarity) for which the EER was obtained for consecutive used models.

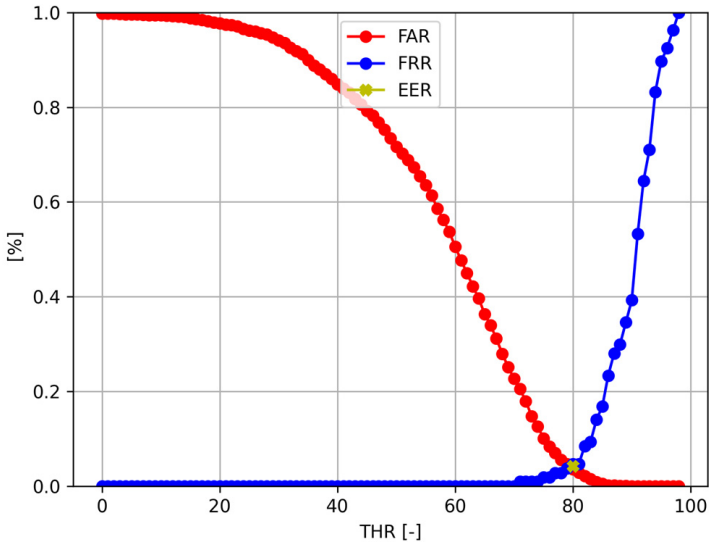


Figure 8. Example of FAR and FRR curves used to calculate EER depending on the threshold applied to the cosine similarity value T_{thr}.

Figure 9 includes graphs showing quantitatively the ranges of the obtained verification results for the cloned samples, by network models, sequentially after 299, 674, 706, 959, and 984 training cycles. The sum of recordings used in this experiment was 100, and complete information about the employed audio is given in Supplementary Tables (Tables S1 and S2).

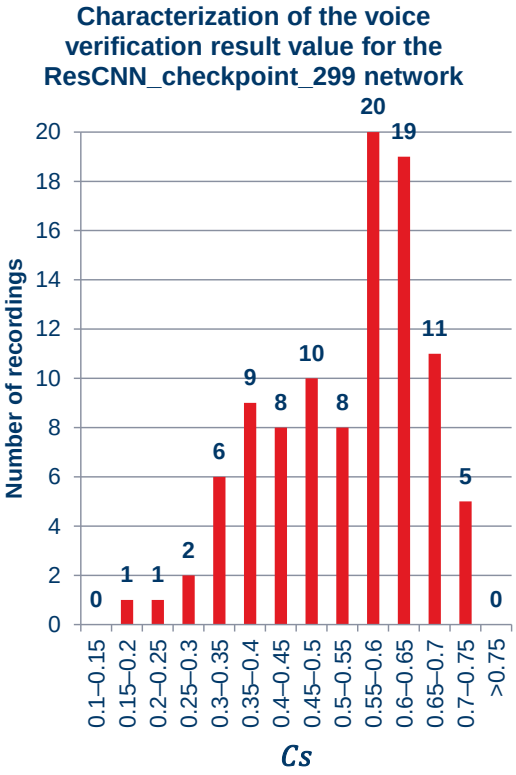
As can be observed from the charts above and in the summary charts in Figures 10 and 11, the values resulting from the verification process do not vary much under the different models. This means that when using the platform selected for the project, the distinction among networks may be negligible.

The number of attack attempts that succeeded, exceeding the decision threshold (the value computed using Equation (2)), is relatively low but worth discussing. The only commonality shared between all of them is the subjectively good sound quality of the recordings, both the voice statements and those obtained through the file cloning process. However, these results should not necessarily be seen as a weakness of the Deep Speaker platform.

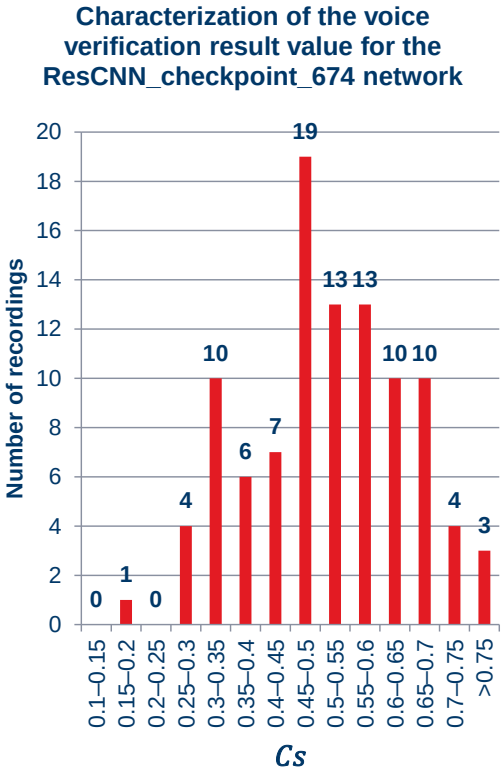
The cloned sample of the common_voice_en_21773347.mp3 recording obtained the highest similarity value with the original C_s = 0.78590524 for the ResCNN_checkpoint_674 network, suggesting that one way to improve the robustness of the system could be to raise the acceptance threshold T_{thr} to, for example, T_{thr} = 0.8.

Another factor contributing to these samples passing the verification is that single files were used, with a short duration of no more than a few seconds. The resulting speaker models have a severely limited amount of information, which is also reflected in the results. It should also be taken into account that the positive verification of the cloned samples depends to some extent on randomness. Recordings that were able to exceed the decision threshold T_{thr} and, thus, cheat the system are listed in Table 3.

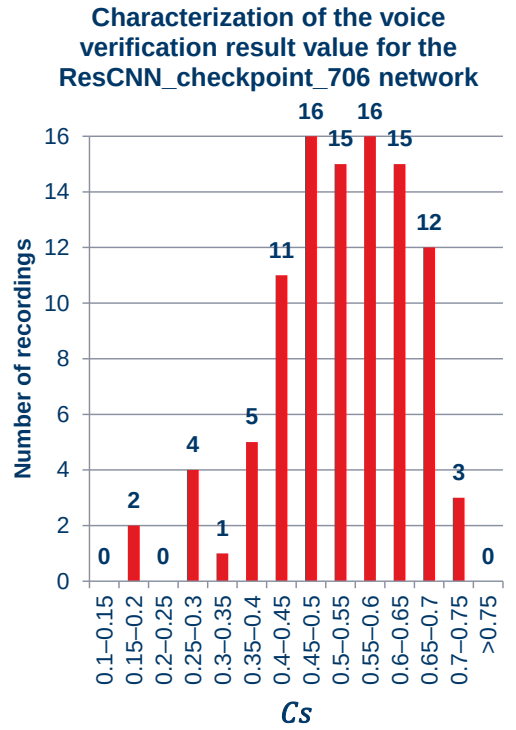
Based on the results, it can be seen that the Deep Speaker platform can resist attacks by voice cloning methods. This means that the systems that synthesize cloned voice recordings (configured, e.g., as is presented in the introduction of the work) are not sufficiently effective to pose a significant threat to the effectiveness of security measures based on voice biometrics.



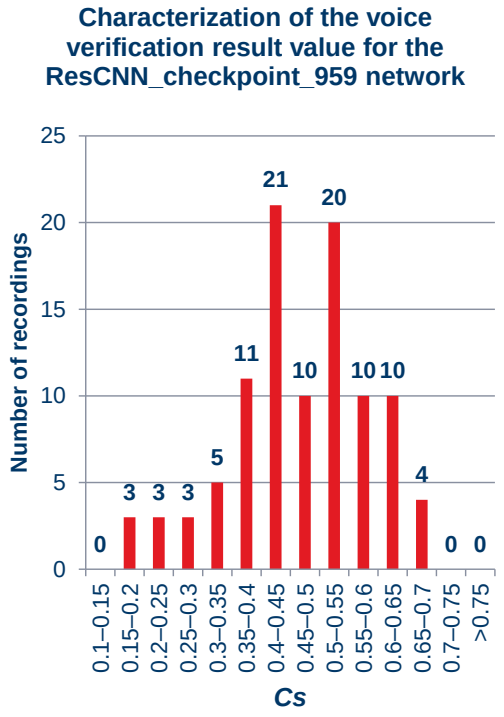
(**a**)



(**b**)

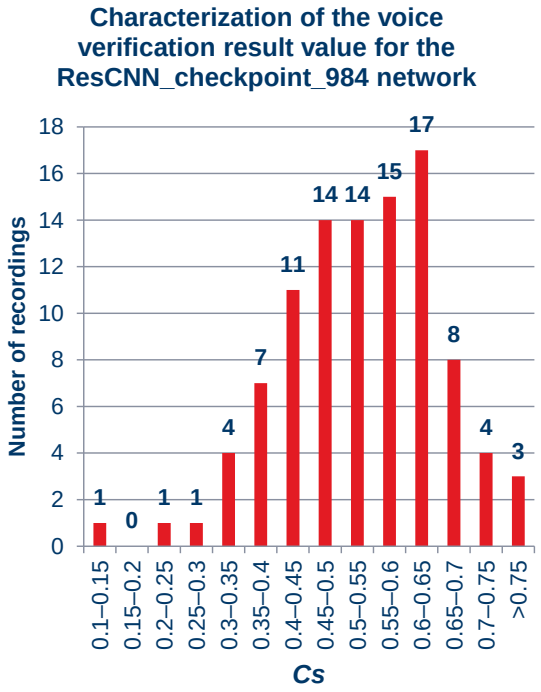


(**c**)



(**d**)

Figure 9. *Cont.*



(**e**)

Figure 9. Verification result of sets containing a different number of recordings obtained for cloned samples: (**a**) ResCNN_checkpoint_299, (**b**) ResCNN_checkpoint_674, (**c**) ResCNN_checkpoint_706, (**d**) ResCNN_checkpoint_959, (**e**) ResCNN_checkpoint_984.

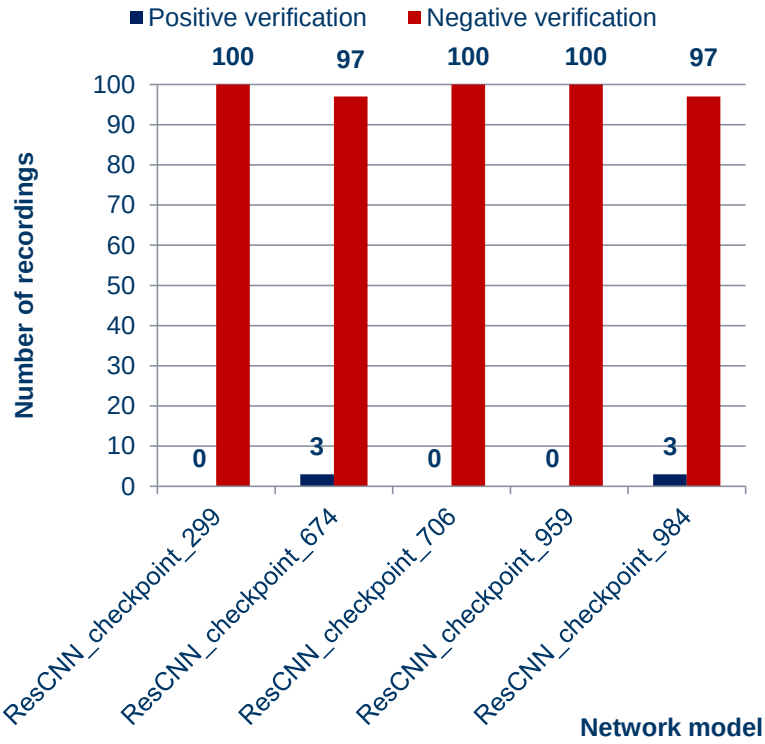


Figure 10. Quantitative characteristics of speaker authentication results. Positive verification means failure—whereby the impostor succeeded in impersonating the speaker with a cloned voice.

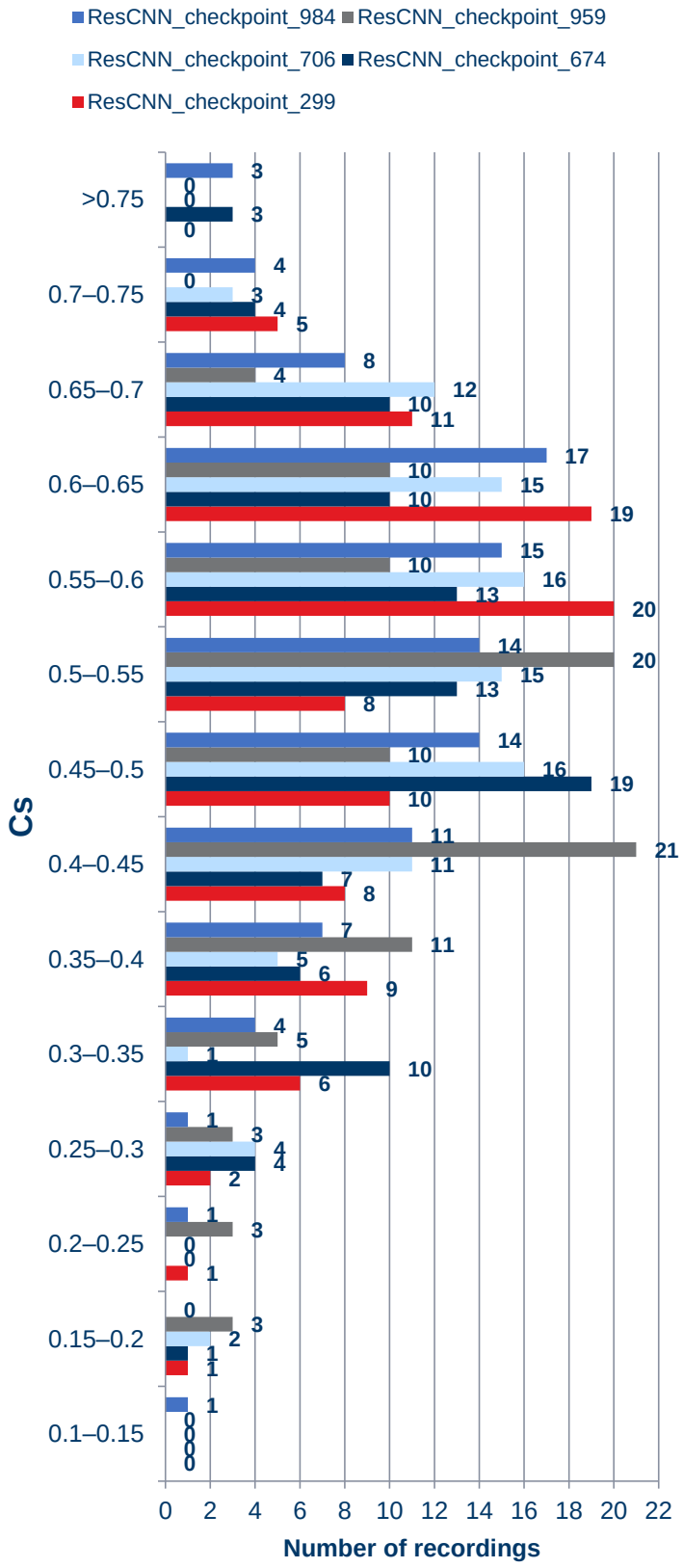


Figure 11. Summary characteristics of verification score values obtained for cloned samples by the Deep Speaker authentication system.

Table 3. List of recordings that exceeded the decision threshold T_{thr} in the speaker verification system.

Lp	Id	Transcription	Voice Type
1.	common_voice_en_21773347	Magnesium fluoride is transparent over an extremely wide range of wavelengths.	Male
2.	common_voice_en_18614630	Come, come, we must not fence and parry now.	Female
3.	common_voice_en_19888382	The island's main economy is tourism, as cruises from the mainland are regular.	Male
4.	common_voice_en_22221017	He was promoted to temporary general.	Male
5.	common_voice_en_22183913	He gave it a score of three out of four.	Female

Survey Method and Results

The comparison of speech samples with listeners was conducted using a modified subjective preference test method based on pairwise sample comparison. The paired comparison testing method involves presenting the test subject with two or more audio samples and encouraging him or her to choose the preferred option without quantifying the magnitude of the difference. This method is often used for subjective research to understand what attributes are most important to listeners and what patterns are most desirable.

The procedure is then repeated for different pairs of samples, and the results are analyzed to obtain information about the surveyed person's preferences. This can be performed by counting the number of choices for each sample or by using more advanced statistical methods, such as latent preference modeling. However, even a relatively simple statistical test quickly shows the range of differences obtained. If the differences are clearly insignificant, one can conclude that the samples under study are practically indistinguishable. The critical point is that listening during the surveys was conducted in a quiet environment using good-quality, closed-back studio headphones to reduce possible bias from noisy environments or using different headphone types.

Survey 1—distinguishing between original and cloned recordings

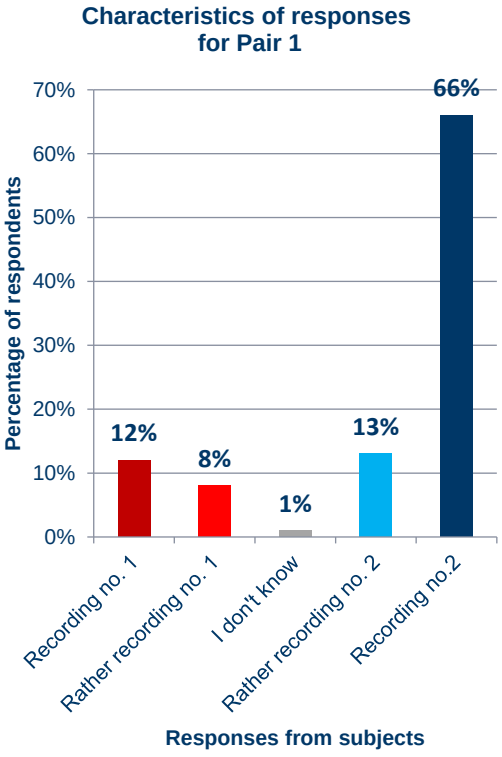
A survey was conducted to study the distinguishability between voice recordings and voice clones. The goal of the respondents was to answer the question, "Which recording was generated using a voice cloning system?" This represents a kind of modification of the paired comparison test, which customarily asks the "choose the better" question. However, the respondents were instructed to consider the sample that they judged the natural one, not the cloned one. Another modification of the test is the expansion of the set of answers to 5 variants: "recording no. 1 is better", "recording no. 1 is a bit better", "recording no. 2 is better", "recording no. 2 is a bit better", and "I don't know". Therefore, the method is more of a choice survey than a classic paired comparison test. We used this approach to extract more information from the listeners.

The question pertained to one or several different sets of recordings. There were 17 pairs in the presented survey. In 14 cases, pairs contained the original sample from the Common Voice dataset and a cloned one. Three times, both recordings were cloned samples. The names of all recordings used in this survey to create each pair are shown in the Supplementary Tables in Table S3. The response statistics were grouped into two subsets, with the first set containing the recordings that appeared once within the survey (shown in Table S5) and the second set consisting of the recordings that were repeated for control purposes (available in Table S6).

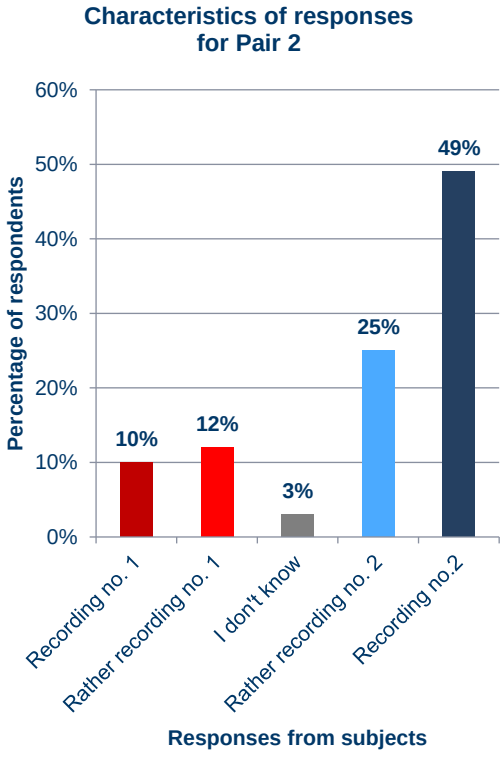
The survey results are shown in Figure 12. The survey made it possible to collect answers from 99 respondents. To improve the legibility of the graphs, the color scheme was chosen so that the dark blue strongly corresponds to a correct answer (cloned recording), while red strongly corresponds to an incorrect answer (original recording). Pale shades of red and blue mean that the respondents were not strongly sure that the recording was,

respectively—original (red) or cloned (blue). For the two blue shades in one Figure, both given samples were original recordings from the Common Voice dataset.

In general, the results obtained from the questionnaires show that the respondents demonstrated various levels of difficulty while indicating the recording of the synthesized voice, depending on the question and the recording. For example, when asked about pair 4 (Figure 12d) and pair 16 (Figure 12p), most respondents incorrectly indicated the voice recording as a cloned voice. Responses to the question about pair 12 (Figure 12l) also had a high percentage of incorrect answers. One potential reason for these choices is the high quality of the synthesized recordings relative to the original ones. But, in the case of pair 16 (Figure 12p), the answers given to a preceding question about pair 13 (Figure 12m), in which both recordings were synthesized utterances by the same speaker, should be taken into account; this may have erroneously suggested the answer to the respondents in the last question. The uneven distribution of votes is an interesting observation regarding the responses to questions for which the two recordings in the pair were cloned samples (pairs 9 (Figure 12i), 13 (Figure 12m), and 15 (Figure 12o)), despite the prior scrutiny of the answers given for those filled in at random. This is particularly evident for pair 13 (Figure 12m). Figure 13 shows the value of the standard error depending on the original–clone pair used in the survey. The standard error for all pairs is similar, ranging from approximately 0.124 to 0.170.

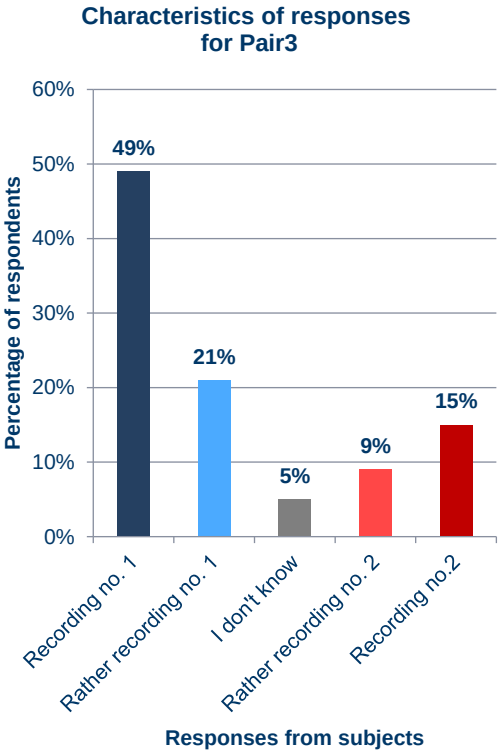


(**a**)

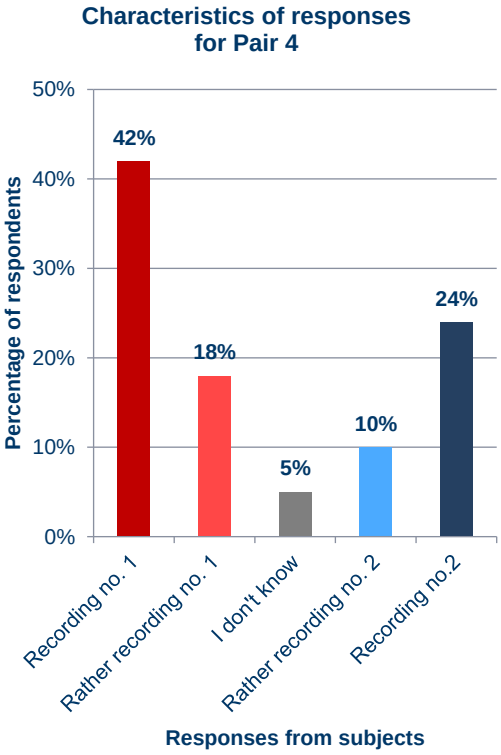


(**b**)

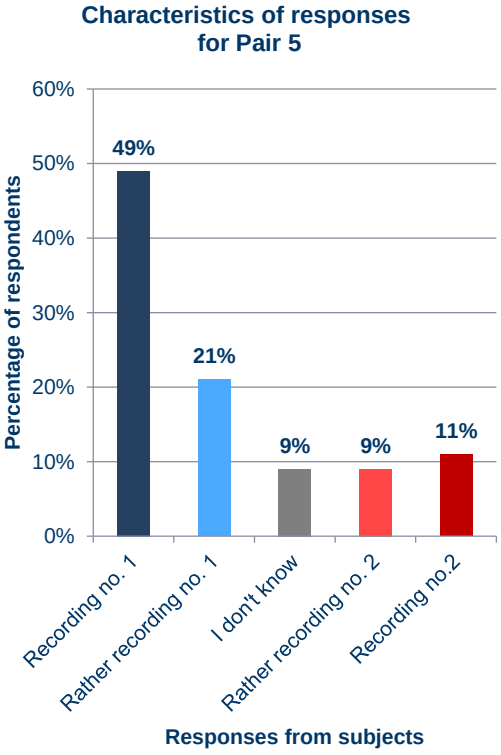
Figure 12. *Cont.*



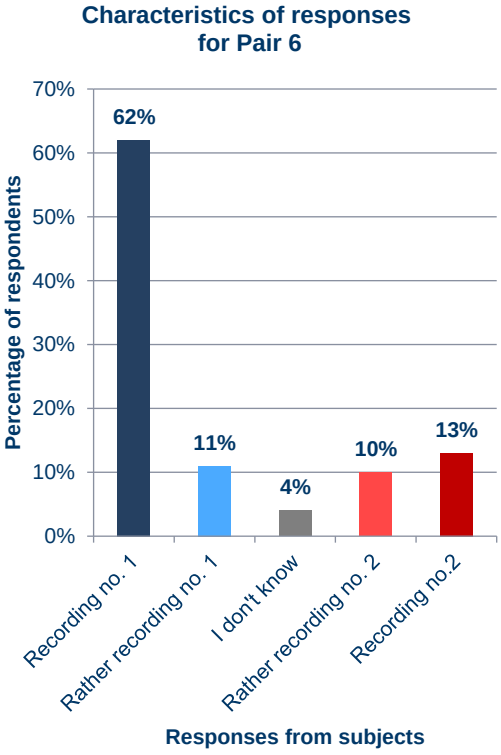
(**c**)



(**d**)

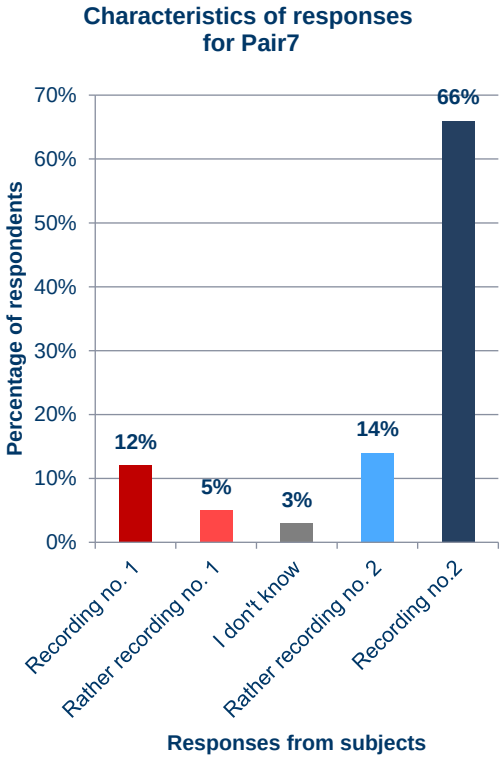


(**e**)

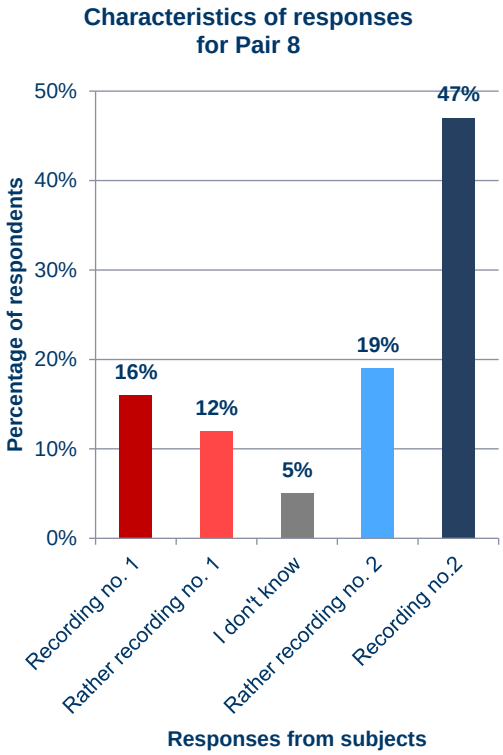


(**f**)

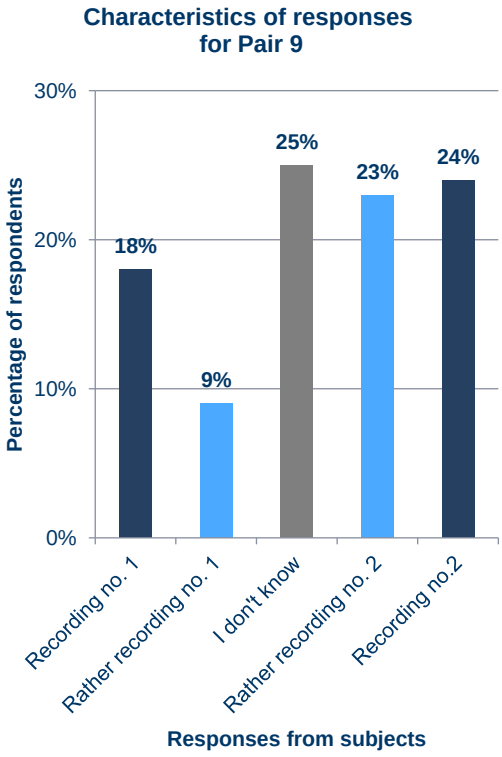
Figure 12. *Cont.*



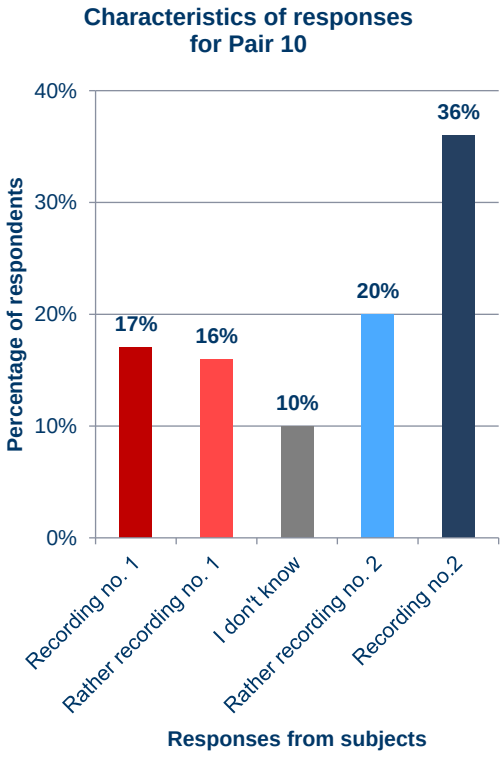
(**g**)



(**h**)

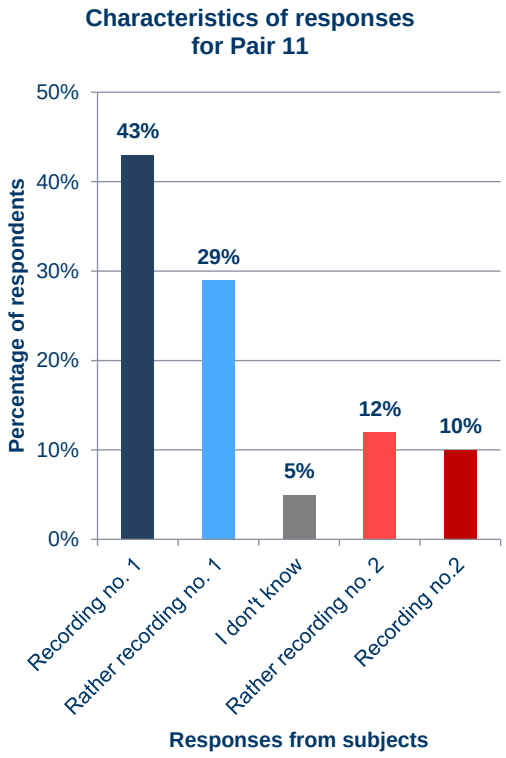


(**i**)

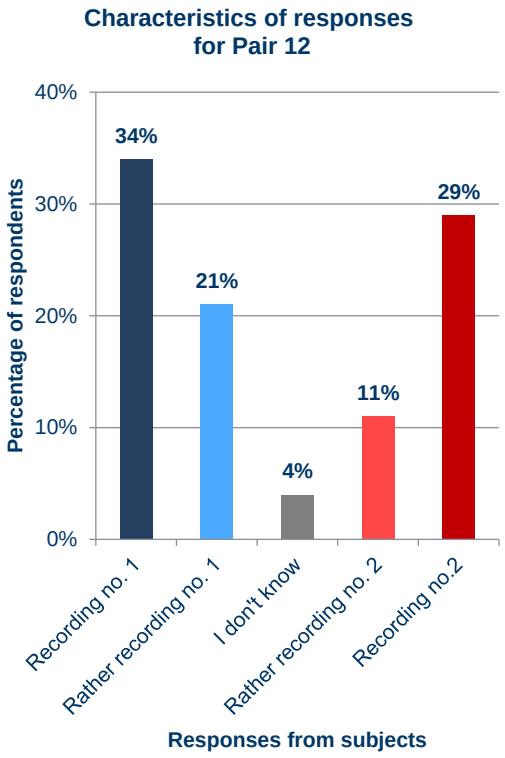


(**j**)

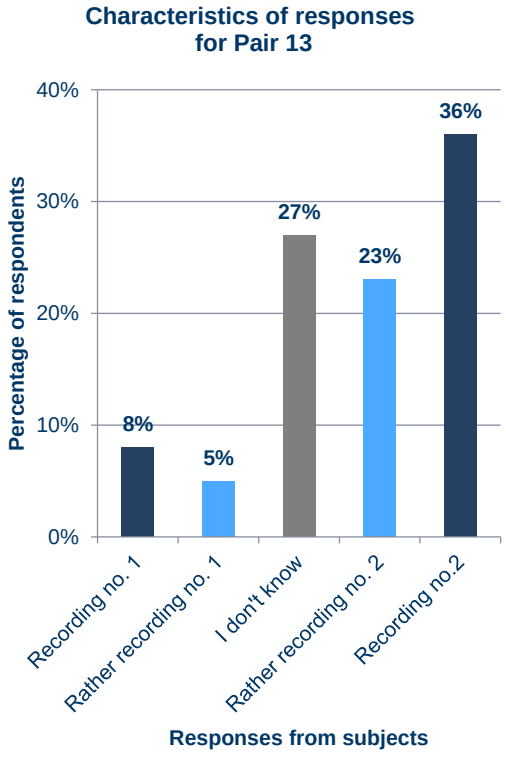
Figure 12. *Cont.*



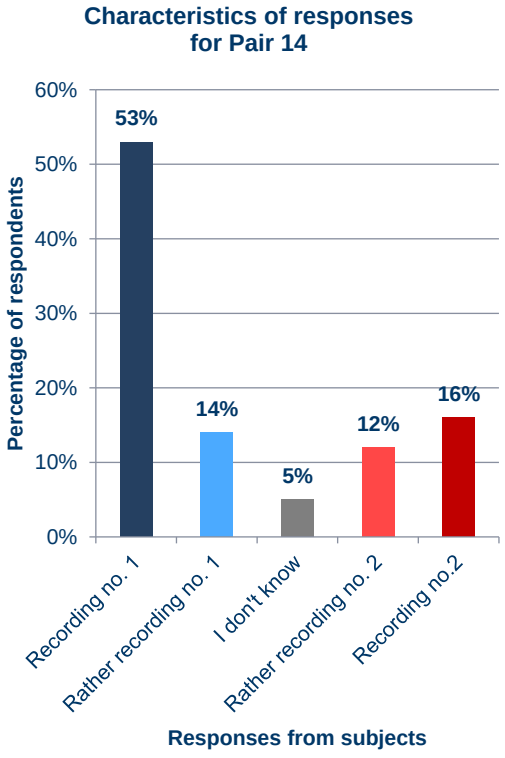
(**k**)



(**l**)

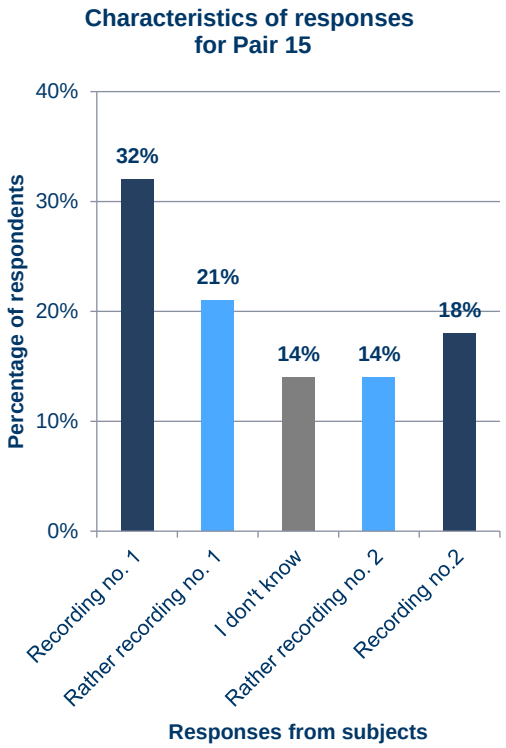


(**m**)

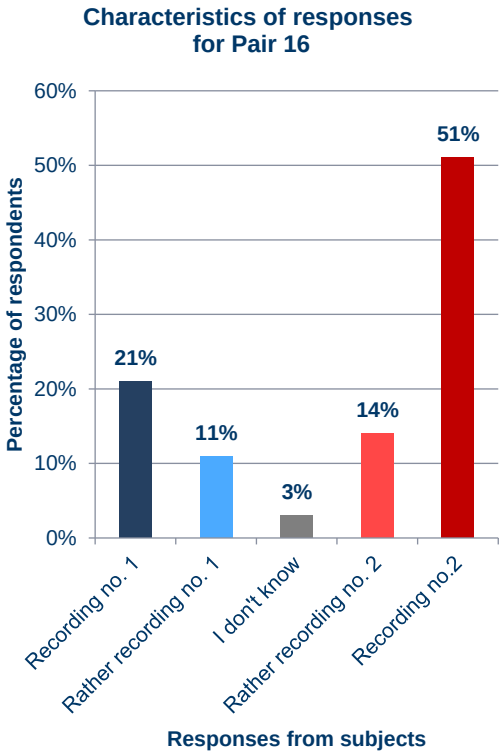


(**n**)

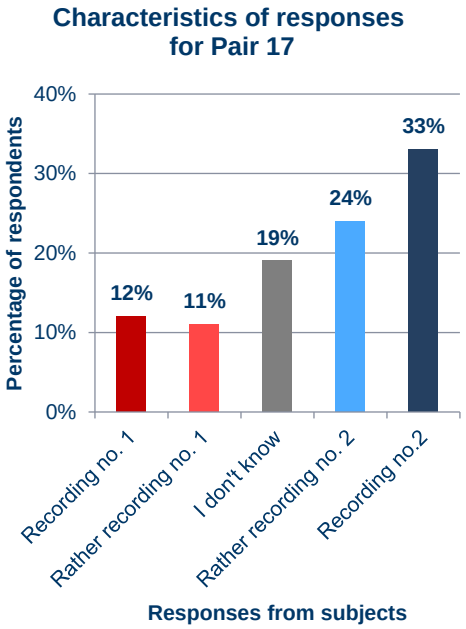
Figure 12. *Cont.*



(**o**)



(**p**)



(**q**)

Figure 12. Statistics of responses to questions about each pair from the survey; (**a**) Pair 1; (**b**) Pair 2; (**c**) Pair 3; (**d**) Pair 4; (**e**) Pair 5; (**f**) Pair 6; (**g**) Pair 7; (**h**) Pair 8; (**i**) Pair 9; (**j**) Pair 10; (**k**) Pair 11; (**l**) Pair 12; (**m**) Pair 13; (**n**) Pair 14; (**o**) Pair 15; (**p**) Pair 16; (**q**) Pair 17.

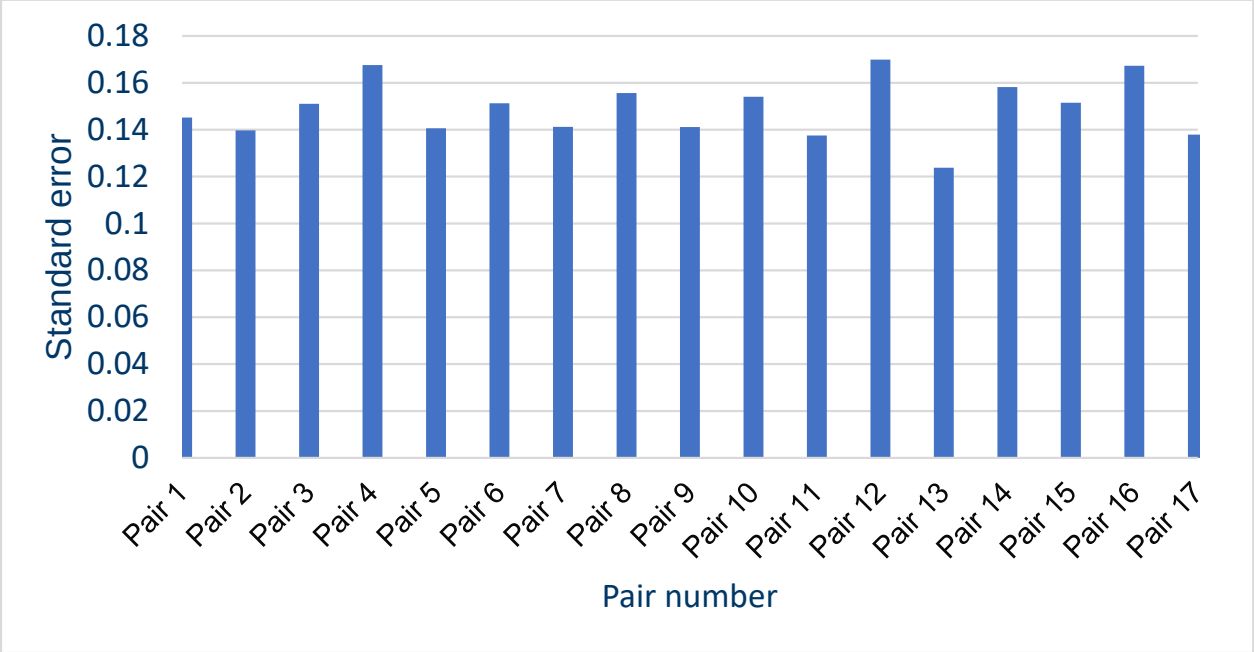


Figure 13. Standard error value for each of Survey 1 audio recordings pairs.

Likewise, the accuracy of people participating in Survey 1 for each pair was calculated, which is shown in Figure 14. Following the data shown in Figure 15, the minimum value of the accuracy measure was 32.32% for pair number 16, with a maximum value of 83.8% for pair number 15, which was the anchor pair (both samples were clones). The mean value of the accuracy measure was 63.9%. In contrast, if the anchorage pairs were not included, the mean value would be 62%.

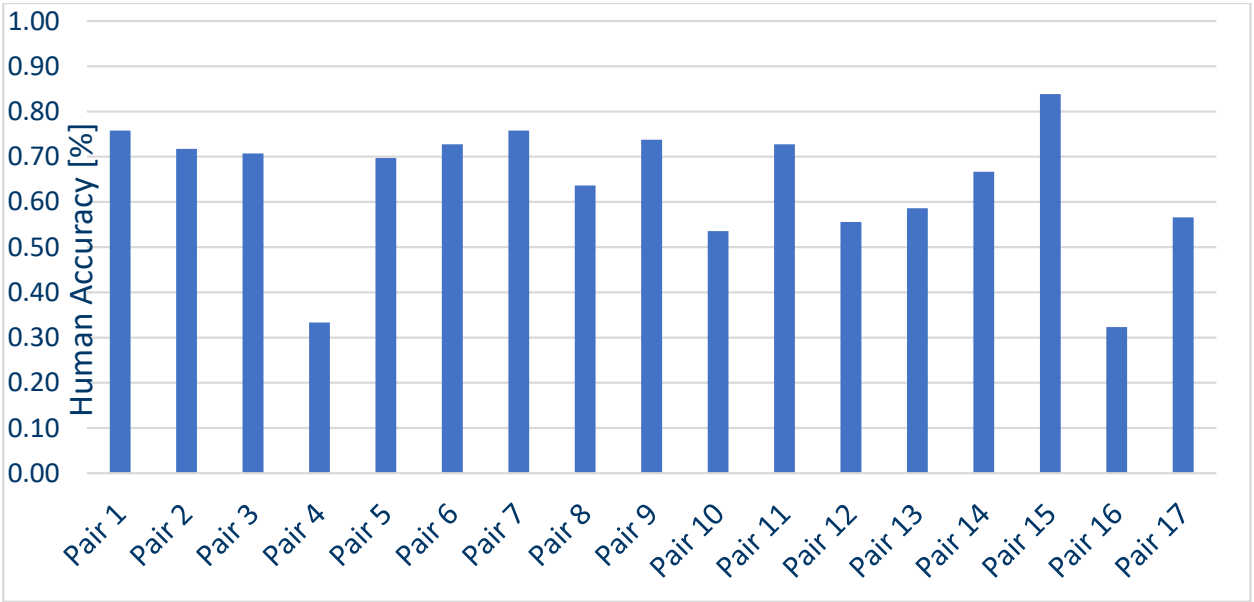
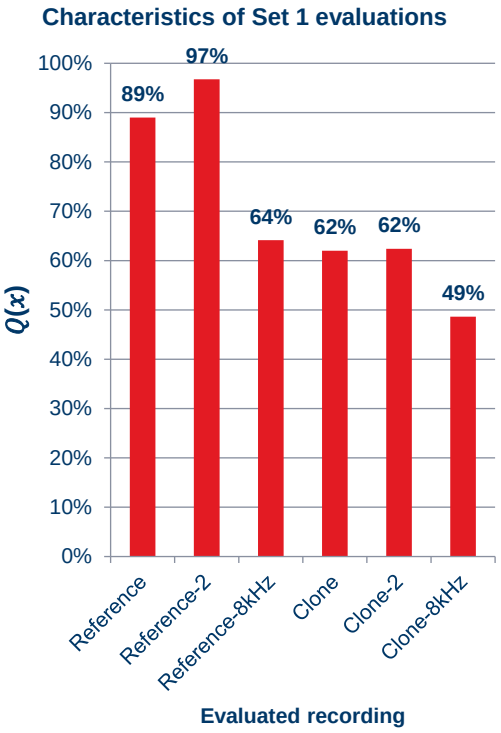
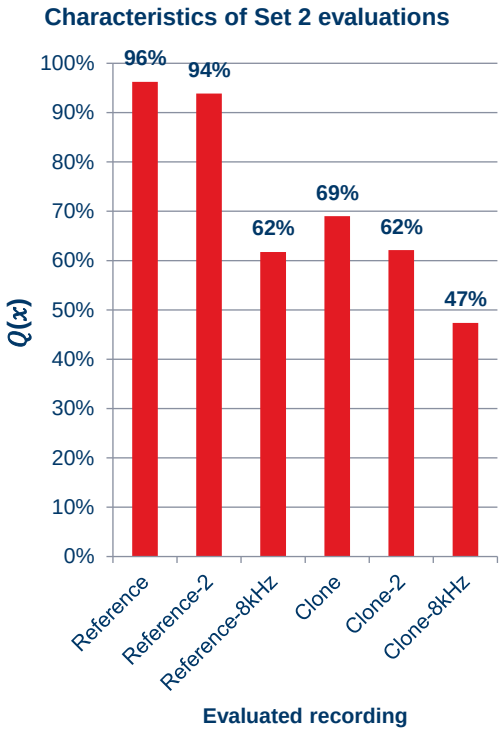


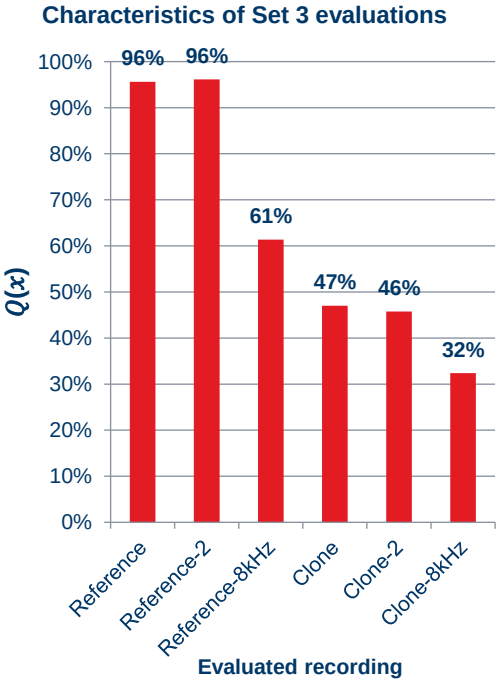
Figure 14. Accuracy calculated for each audio recording pair in Survey 1—How participants have dealt with voice clone recognition.



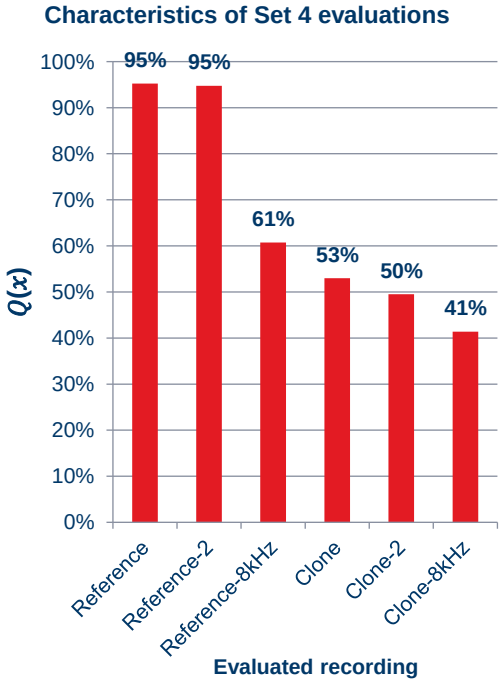
(**a**)



(**b**)

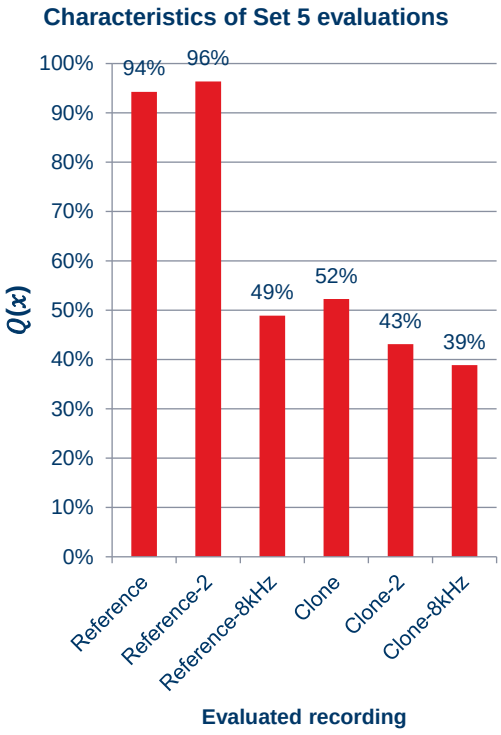


(**c**)

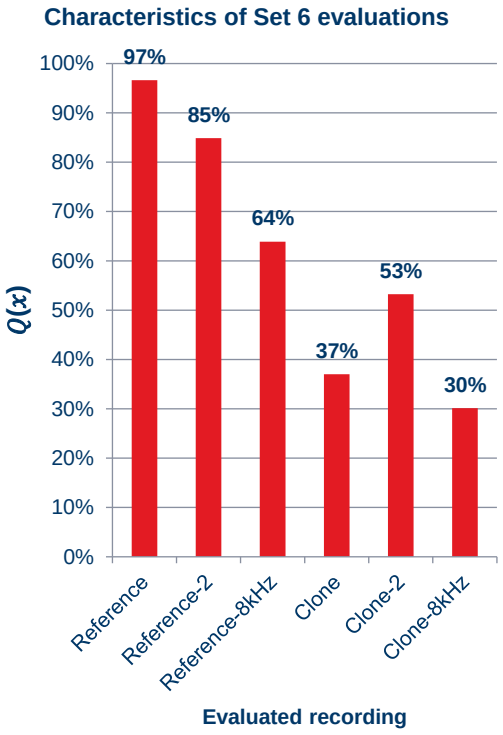


(**d**)

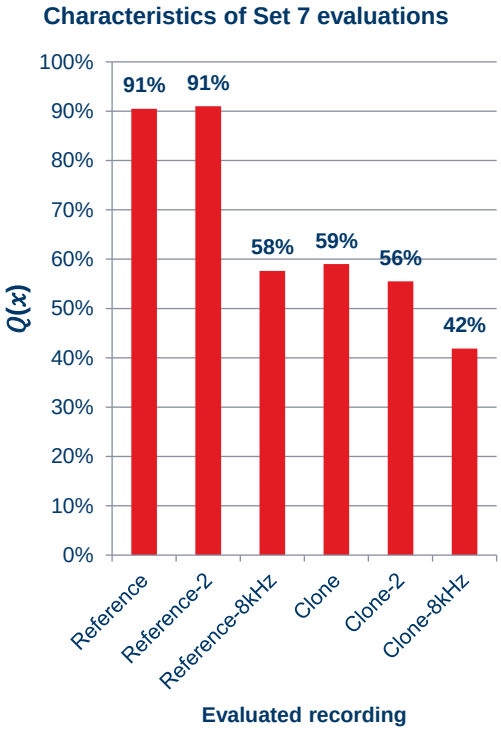
Figure 15. *Cont.*



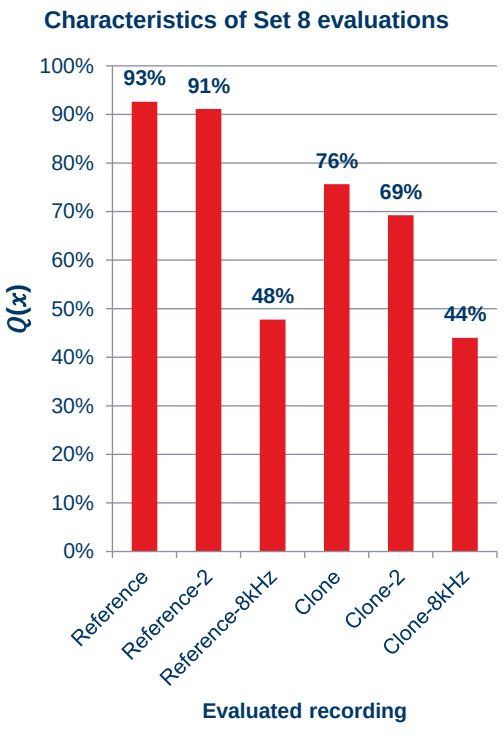
(**e**)



(**f**)

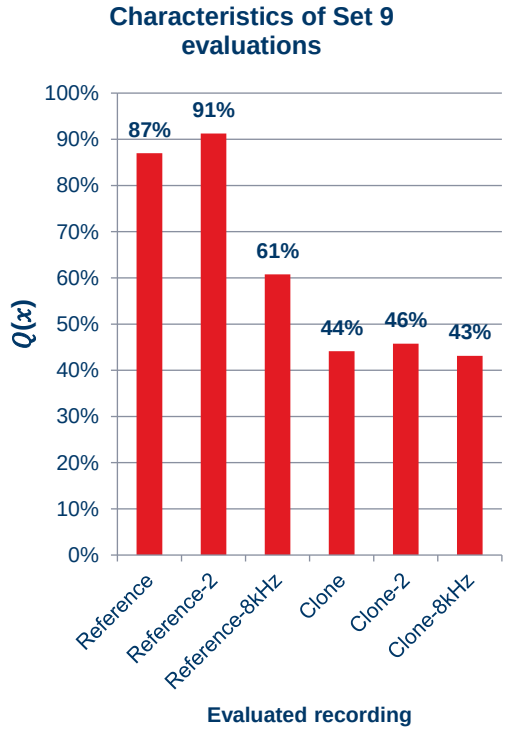


(**g**)



(**h**)

Figure 15. *Cont.*



(**i**)

Figure 15. Characteristics of quality ratings for sets of recordings: (**a**) set 1; (**b**) set 2; (**c**) set 3; (**d**) set 4; (**e**) set 5; (**f**) set 6; (**g**) set 7; (**h**) set 8; (**i**) set 9.

Survey 2—subjective evaluation of the quality of cloned recordings

As part of the survey of the subjective assessment of the quality of the cloned samples, 23 people were polled. All of them were instructed on the rules governing the survey before responding and remained under the project executor's constant supervision. Each of the subsets consisted of six recordings, including two reference recordings from the Common Voice corpus, sampled as 48 kSa/s; two cloned recordings, generated using the previously mentioned cloning framework with a sample rate of 16 kSa/s; and one reference recording and one cloned recording, both downsampled to 8 kHz. Each subset used a different file from the Common Voice dataset, utilized as the references and cloned samples. The names of the files used in the survey are shown in the Supplementary Tables in Table S4. The percentage scale ranges describing the quality assessment against the original recording based on ITU-R BS.1534 recommendation [34] are provided below (Equation (5)):

$$Q(x) = \begin{cases} Excellent\ if\ 80\% < x \leq 100\% \\ Good\ if\ 60\% < x \leq 80\% \\ Fair\ if\ \ 40\% < x \leq 60\% \\ Poor\ \ if\ 20\% < x \leq 40\% \\ Bad\ \ if\ 0\% \leq x \leq 20\% \end{cases} \quad (5)$$

The results are summarized in Figures 14 and 15.

As can be observed in the graphs shown in Figure 15 and the summary graph shown in Figure 16, those surveyed rated the quality of the recordings containing cloned voice samples noticeably lower than the quality of the reference recordings, and in Figure 15a,c,d,f,i, lower even than the voice recordings with an 8 kSa/s sampling rate. When asked after the survey, the reasons for assigning such scores were commented by subjects as the unnatural sound of the voice or the presence of sound artifacts. The unnaturalness was related to several factors: they concerned the speed at which the excerpts were spoken; the prosody of

the sentences differing from the original recording; and the absence or presence of audible imperfections in human speech, such as breathing.

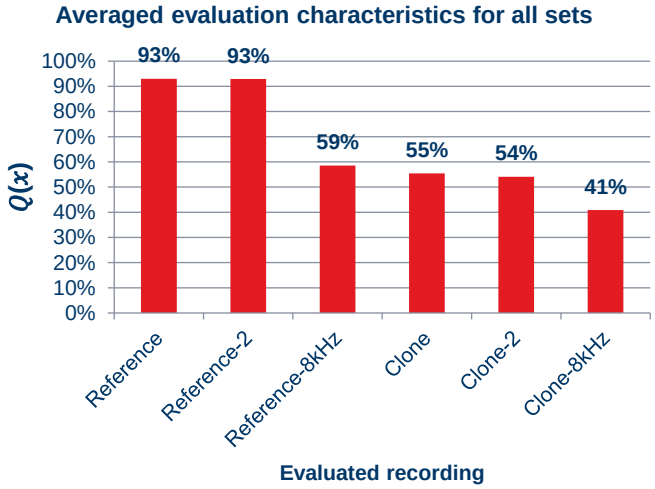


Figure 16. Averaged characteristics of quality ratings for all sets of recordings.

6. Discission

A study of the vulnerability of the Deep Speaker model-based system to attack attempts using voice cloning tools was carried out. It was shown through testing that the developed speech authentication system has a good level of security against such security breach attempts, achieving an effectiveness of 98.8% (averaged over the tested models).

By surveying the distinguishability between original voice recordings and their cloned counterparts and assessing the sound quality of the generated audio files, it can be assumed that speech cloning technology is generally not ready yet to synthesize a fully natural-sounding voice. However, the many individual recordings that misled the speaker verification system and survey respondents are a disturbing indication of this technology's growth potential.

In our research results, this is visible in the graphs showing that respondents who considered the cloned voice to be real are represented by prominent bronze bars related to the max cosine similarity obtained by Deep Speaker, i.e.,

- Figure 12d (common_voice_en_19739363.mp3; max cosine similarity ~0.44);
- Figure 12l (common_voice_en_19687887.mp3; max cosine similarity ~0.72);
- Figure 12p (common_voice_en_20781783.mp3; max cosine similarity ~0.69).

According to the subjective tests, the average quality assessment score against the original record indicates no significant difference between the voice sample and the synthesized clone for downsampled recordings, carrying mainly the bandwidth of 4 kHz that is typical of telephone channels. This may lead to the conclusion that it is difficult for a human to find much difference between a cloned voice and the original for such a narrowed bandwidth. Conducting further subjective tests on samples with different bandwidths could provide more information on the human perception of cloned voice samples.

On the other hand, the artificial neural network-based model was not susceptible to cloned samples since, for 500 attempts, only 6 were successful, meaning that 1.2% of all attack attempts were effective. It could be argued whether such a number is acceptable in a practical biometric system. However, it is not easy to meet the conditions that helped to successfully mislead the voice verification model. Another aspect that should be noted is that female voices seem more challenging to recognize between the clone and the original. The presented experiments are worth considering as a proof-of-concept for further work related to the analysis of cloned recordings and their perception by humans, as well as the differentiation of human and cloned voices by artificial neural networks.

When comparing the accuracy measures obtained in the experiments to distinguish the clone from the original recording, the neural network models perform significantly better, achieving results close to 100% for the experiment presented, as mentioned before. On average, people taking part in the listening tests achieved just over 60% efficiency. It should be noted here that the two experiments were not identical. In the case of the model test, 100 samples were used; in the case of the humans, 26 people took part in the test and made 18 comparisons across 9 questions. Despite this difference, which could favor the listening test due to the smaller number of samples and good listening conditions, the network achieves a much better result.

Analyzing the above findings, it can be assumed that voice cloning in its current form does not directly threaten monitored voice authentication systems, e.g., in bank branches. Still, it can be potentially dangerous when used in a telecommunication channel, e.g., at a call center, generally in places where the so-called "human factor" can fail. However, it is essential to keep in mind the sudden increase in interest in voice cloning technologies, reflected in the development of increasingly perfect voice cloning techniques, which some startups [46] and large corporations [47] are using. With the developed approaches, they are able to create not only a perfect copy of a person's voice but also passages related to the intonation of speech and the transmission of emotions, which could lead to much more difficulty in distinguishing a cloned sample from the actual voice of the person. The approach can be appreciated in which, to prevent the misuse of a created voice cloning solution, only an article and a demo of the solution are made available, without sharing the source code of the complete solution [48]. This may indicate that, for the time being, with the use of the new generation of voice cloning frameworks, there is a potential risk of impersonating people without their knowledge. Meanwhile, voice biometric verification systems and people cannot be fully prepared to distinguish a sample generated in this way from even a short excerpt of an actual voice.

7. Conclusions

The studies conducted and shown in the article indicate that people may easily become confused and point out cloned recordings as an original sample. In some instances, nearly 50% of respondents pointed to the wrong answer, confusing the synthesized recording with the original one. When evaluating the quality of cloned and real recordings after changing the sampling frequency, people in most cases indicated that the quality of the cloned and original recordings was similar (Good or Fair according to the ITU-R BS.1534 recommendation) and in some cases even better than the original (graded as Good for the synthesized sample and Fair for original one).

Meanwhile, during attempts to attack the developed biometric system, nearly all verification attempts with cloned samples failed (98.8% of samples were rejected). This proves that biometric systems based on deep neural networks are able to identify synthesized samples with greater efficiency in cases where humans have problems distinguishing between them. Therefore, it can be expected that, with the rapid improvements in voice cloning algorithms, the problem of differentiating synthesized samples from voice recordings will become more complex for humans and neural networks.

8. Future Works

One possible area of further development of the work is to retest the immunity to voice cloning attacks using neural network models trained in different languages to check the behavior of the verification system when the language changes. Another potential development direction is to retrain an existing model under different acoustic conditions to simulate a hypothetical attempt to "steal" a person's voice. Transfer learning [25] could shorten the training model process without the need to prepare a new model from scratch. Also, testing the robustness of voice biometric verification systems against the new generation of voice cloning systems seems a logical path for further research.

With the ongoing development of neural network algorithms, it would be advisable to periodically repeat the test on newer security systems based on voice biometrics against potential attacks using voice cloning software, which is developing rapidly.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/electronics12214458/s1, Table S1. List of selected recordings of female speakers; Table S2. List of selected recordings of male speakers; Table S3. Selected recordings used in the survey on distinguishing cloned from original sample; Table S4. List of recordings used in the study of the subjective sound quality of cloned recordings; Table S5. Recordings that appeared once in the survey; Table S6. Records that appeared twice in the survey for control purposes; Figure S1. Distribution of speakers' nationalities in the dataset employed for survey no. 1.

Author Contributions: Conceptualization, S.Z.; Methodology, S.Z.; Software, K.M. and S.Z.; Validation, K.M. and S.Z.; Investigation, K.M.; Resources, S.Z.; Data curation, K.M.; Writing—original draft, K.M.; Writing—review & editing, S.Z. and A.C.; Visualization, K.M.; Supervision, A.C.; Project administration, A.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded from the budget of project No. POIR.01.01.01-0092/19 entitled: "BIOPUAP—a biometric cloud authentication system" subsidized by the Polish National Centre for Research and Development (NCBR) from the European Regional Development Fund and funds of Electronics, Telecommunications, and Informatics Faculty, Gdańsk University of Technology.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Available online: https://commonvoice.mozilla.org/en (accessed on 11 October 2019).
2. CorentinJ. Real Time Voice Cloning. Available online: https://github.com/CorentinJ/Real-Time-Voice-Cloning (accessed on 17 January 2023).
3. Li, C.; Ma, X.; Jiang, B.; Li, X.; Zhang, X.; Liu, X.; Cao, Y.; Kannan, A.; Zhu, Z. Deep Speaker: An End-to-End Neural Speaker Embedding System. 2017. Available online: https://arxiv.org/abs/1705.02304 (accessed on 17 January 2023).
4. Liu, X.; Sahidullah, M.; Kinnunen, T.A. Comparative Re-Assessment of Feature Extractors for Deep Speaker Embeddings. In Proceedings of the Interspeech, Shanghai, China, 25–29 October 2020; pp. 3221–3225. [CrossRef]
5. Hermansky, H. Perceptual linear predictive (PLP) analysis of speech. *J. Acoust. Soc. Am.* **1990**, *87*, 1738–1752. [CrossRef] [PubMed]
6. Chauhan, N.; Isshiki, T.; Li, D. Speaker Recognition Using LPC, MFCC, ZCR Features with ANN and SVM Classifier for Large Input Database. In Proceedings of the 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS), Singapore, 23–25 February 2019; pp. 130–133. [CrossRef]
7. Cavazza, M.; Ciaramella, A. Device for speakers's verification. *J. Acoust. Soc. Am.* **1990**, *87*, 1831. [CrossRef]
8. Cholet, F. *Deep Learning. Working with the Python Language and the Keras Library*; Helion SA: Gliwice, Poland, 2019.
9. Mokhov, S.A.; Sinclair, S.; Clément, I.; Nicolacopoulos, D. The Modular Audio Recognition Framework (MARF) and its Applications: Scientific and Software Engineering Notes. *arXiv* **2009**, arXiv:0905.1235.
10. Aditya, K. PiWho. Available online: https://github.com/Adirockzz95/Piwho (accessed on 17 January 2023).
11. Wan, L.; Wang, Q.; Papir, A.; Moreno, I.L. Generalized end-to-end loss for speaker verification. In Proceedings of the ICASSP 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, Calgary, AB, Canada, 15–20 April 2018; pp. 4879–4883.
12. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
13. Baevski, A.; Zhou, H.; Mohamed, A.; Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. In Proceedings of the Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, Virtual, 6–12 December 2020; pp. 1–19.
14. India, M.; Safari, P.; Hernando, J. Self multi-head attention for speaker recognition. In Proceedings of the 20th Annual Conference of the International Speech Communication Association, INTERSPEECH, Graz, Austria, 15–19 September 2019; pp. 4305–4309.
15. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
16. Cho, K.; van Merriënboer, B.; Bahdanau, D.; Bengio, Y. On the properties of neural machine translation: Encoder–decoder approaches. In Proceedings of the SSST 2014—8th Workshop on Syntax, Semantics and Structure in Statistical Translation, Doha, Qatar, 25 October 2014; pp. 103–111. [CrossRef]
17. Jozefowicz, R.; Zaremba, W.; Sutskever, I. An Empirical Exploration of Recurrent Network Architectures. In Proceedings of the 32nd International Conference on Machine Learning PMLR, Lille, France, 6–11 July 2015; Volume 37, pp. 2342–2350.

18. Amodei, D.; Anubhai, R.; Battenberg, E.; Case, C.; Casper, J.; Catanzaro, B.; Chen, J.; Chrzanowski, M.; Coates, A.; Diamos, G.; et al. Deep speech 2: End-to-end speech recognition in English and Mandarin. In Proceedings of the 33rd International Conference on Machine Learning ICML 2016, New York, NY, USA, 20–22 June 2016; pp. 312–321.
19. Chapuzet, A. Speech Synthesis (TTS), How to Use It and Why Is It So Important? Available online: https://vivoka.com/how-to-speech-synthesis-tts (accessed on 17 January 2023).
20. Schwarz, D. Current Research in Concatenative Sound Synthesis. In Proceedings of the 2005 International Computer Music Conference, ICMC 2005, Barcelona, Spain, 4–10 September 2005. Available online: https://quod.lib.umich.edu/i/icmc/bbp2372.2005.045/1 (accessed on 17 January 2023).
21. Feedforward Deep Learning Models. 33rd AFIT Data Science Lab R Programming Guide. Available online: https://afit-r.github.io/feedforward_DNN (accessed on 17 January 2023).
22. Wang, Y.; Skerry-Ryan, R.J.; Stanton, D.; Wu, Y.; Weiss, R.J.; Jaitly, N.; Yang, Z.; Xiao, Y.; Chen, Z.; Bengio, S.; et al. Tacotron: Towards end-To-end speech synthesis. In Proceedings of the Annual Conference of the International Speech Communication Association INTERSPEECH 2017, Stockholm, Sweden, 20–24 August 2017; pp. 4006–4010.
23. Shen, J.; Pang, R.; Weiss, R.J.; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerrv-Ryan, R. Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 4779–4783.
24. Tamamori, A.; Hayashi, T.; Kobayashi, K.; Takeda, K.; Toda, T. Speaker-Dependent WaveNet Vocoder. In Proceedings of the Interspeech 2017, Stockholm, Sweden, 20–24 August 2017; pp. 1118–1122. [CrossRef]
25. Jia, Y.; Zhang, Y.; Weiss, R.J.; Wang, Q.; Shen, J.; Ren, F.; Chen, Z.; Nguyen, P.; Pang, R.; Moreno, I.L.; et al. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In Proceedings of the Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, Montréal, QC, Canada, 3–8 December 2018; pp. 4480–4490.
26. Tan, X.; Qin, T.; Soong, F.; Liu, T.-Y. A Survey on Neural Speech Synthesis. 2021. Available online: https://arxiv.org/abs/2106.15561 (accessed on 17 January 2023).
27. Liu, Y.; Xue, R.; He, L.; Tan, X.; Zhao, S. DelightfulTTS 2: End-to-End Speech Synthesis with Adversarial Vector-Quantized Auto-Encoders. In Proceedings of the 23rd Annual Conference of the International Speech Communication Association INTERSPEECH 2022, Incheon, Republic of Korea, 18–22 September 2022; pp. 1581–1585.
28. Nagrani, A.; Chung, J.S.; Xie, W.; Zisserman, A. Voxceleb: Large-scale speaker verification in the wild. *Comput. Speech Lang.* **2019**, *60*, 101027. [CrossRef]
29. Chung, J.S.; Nagrani, A.; Zisserman, A. VoxceleB2: Deep speaker recognition. In Proceedings of the 19th Annual Conference of the International Speech Communication Association INTERSPEECH 2018, Hyderabad, India, 2–6 September 2018; pp. 1086–1090.
30. Yamagishi, J.; Veaux, C.; MacDonald, K. CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92). University of Edinburgh. The Centre for Speech Technology Research (CSTR). Available online: https://datashare.ed.ac.uk/handle/10283/3443 (accessed on 17 January 2023). [CrossRef]
31. Panayotov, V.; Chen, G.; Povey, D.; Khudanpur, S. Librispeech: An ASR corpus based on public domain audio books. In Proceedings of the ICASSP, 2015 IEEE International Conference on Acoustics, Speech and Signal Processing, South Brisbane, Australia, 19–24 April 2015; pp. 5206–5210.
32. Audacity Team. Audacity. Available online: https://www.audacityteam.org/ (accessed on 17 January 2023).
33. BlackmagicDesign. DaVinci Resolve 18. Available online: https://www.blackmagicdesign.com/en/products/davinciresolve/ (accessed on 13 August 2022).
34. Google. Google Forms. Available online: https://www.google.com/forms/about/ (accessed on 15 August 2022).
35. Schoeffler, M.; Bartoschek, S.; Stöter, F.-R.; Roess, M.; Westphal, S.; Edler, B.; Herre, J. webMUSHRA—A Comprehensive Framework for Web-based Listening Tests. *J. Open Res. Softw.* **2018**, *6*, 8. [CrossRef]
36. *ITU-R BS.1116*; ITU-R BS.1534-3: Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems. International Telecommunication Union: Geneva, Switzerland, 2015; p. 34.
37. Garnerin, M.; Rossato, S.; Besacier, L. Investigating the Impact of Gender Representation in ASR Training Data: A Case Study on Librispeech. In Proceedings of the GeBNLP 2021—3rd Workshop on Gender Bias in Natural Language Processing, Bangkok, Thailand, 5 August 2021; pp. 86–92. [CrossRef]
38. Ellis, D.; Raffael, D.; Liang, D.; Ellis, D.P.W.; McVicar, M.; Battenberg, E.; Nieto, O. librosa: Audio and Music Signal Analysis in Python. In Proceedings of the 14th Python in Science Conferences, Austin, TX, USA, 9–15 July 2018. [CrossRef]
39. Lyons, J.; Wang, D.Y.-B.; Gianluca, A.; Shteingart, H.; Mavrinac, E.; Gaurkar, Y.; Watcharawisetkul, W.; Birch, S.; Zhihe, L.; Hölzl, J.; et al. Python Speech Features: Release v0.6.1. 2020. Available online: https://github.com/jameslyons/python_speech_features/tree/0.6.1 (accessed on 17 January 2023).
40. Young, S.; Gales, M.; Liu, X.A.; Povey, D.; Woodland, P. *The HTK Book*, version 3.5a; Department of Engineering, University of Cambridge: Cambridge, UK, 2015. Available online: https://www.researchgate.net/publication/289354717_The_HTK_Book_version_35a (accessed on 14 April 2023).
41. Slaney, M. *Auditory Toolbox: A Matlab Toolbox for Auditory Modeling Work*; Technical Report 10; Interval Research Corporation: Palo Alto, CA, USA, 1998. Available online: https://engineering.purdue.edu/~malcolm/interval/1998-010/AuditoryToolboxTechReport.pdf (accessed on 17 January 2023).

42. Tyagi, V.; Wellekens, C. On desensitizing the Mel-cepstrum to spurious spectral components for robust speech recognition. In Proceedings of the ICASSP '05, IEEE International Conference on Acoustics, Speech, and Signal Processing, Philadelphia, PA, USA, 18–23 March 2005; Volume 1, pp. I/529–I/532. [CrossRef]
43. Rémy, P. Deep Speaker: An End-to-End Neural Speaker Embedding System—Unofficial Tensorflow/Keras Implementation of Deep Speaker. Available online: https://github.com/philipperemy/deep-speaker (accessed on 17 January 2020).
44. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M.; et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. 2015. Available online: https://arxiv.org/abs/1603.04467 (accessed on 15 January 2023).
45. Biewald, L. Experiment Tracking with Weights and Biases. 2020. Available online: https://www.wandb.com/ (accessed on 9 July 2021).
46. Eleven Labs. Eleven Labs—Voice Lab. Available online: https://beta.elevenlabs.io/ (accessed on 9 February 2023).
47. Wang, C.; Chen, S.; Wu, Y.; Zhang, Z.; Zhou, L.; Liu, S.; Chen, Z.; Liu, Y.; Wang, H.; Li, J.; et al. Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers. Available online: https://arxiv.org/abs/2301.02111 (accessed on 9 February 2023).
48. Microsoft. VALL-E. Available online: https://valle-demo.github.io/ (accessed on 9 February 2023).