Voice synthesis in Polish and English – analyzing prediction differences in speaker verification systems

Joanna Gajewska[1], Alicja Martinek[1,2], Michał Ołowski[1], Ewelina Bartuzi-Trokielewicz[1]
{joanna.gajewska, alicja.martinek, michal.olowski, ewelina.bartuzi}@nask.pl

[1]NASK National Research Institute
ul. Kolska 12
01-045 Warsaw, Poland

[2]AGH University of Kraków
al. Mickiewicza 30
30-059 Kraków, Poland

Abstract

Deep learning has significantly enhanced voice synthesis, yielding realistic audio capable of mimicking individual voices. This progress, however, raises security concerns due to the potential misuse of audio deepfakes. Our research examines the effects of deepfakes on speaker recognition systems across English and Polish corpora, assessing both Text-to-Speech and Voice Conversion methods. We focus on the biometric similarity's role in the effectiveness of impersonations and find that synthetic voices can maintain personal traits, posing risks of unauthorized access. The study's key contributions include analyzing voice synthesis across languages, evaluating biometric resemblance in voice conversion, and contrasting Text-to-Speech and Voice Conversion paradigms. These insights emphasize the need for improved biometric security against audio deepfake threats.

1 Introduction

In times of rapid technological progress, voice synthesis has evolved significantly, bringing numerous conveniences as well as many threats. Audio deepfake, the synthetic generation of voice using advanced algorithms, can be employed to create high-quality false voice recordings convincing enough to deceive human ears and pose threat to speech-based biometric systems like Speaker Recognition (SR) or Automatic Speaker Verification (ASV). To meet these issues, spoofing challenges and countermeasures have been addressed for years (Wu et al., 2015; Delgado et al., 2018; Ren et al., 2019; weon Jung et al., 2022). While there are some results yielding good performance on unknown types of attacks during training process, research on the generalizability of different types of attacks is still an unresolved topic. In this article, we study the spoofing-uneware speaker recognition methods, confronted with speech synthesis methods, in order to test the relationship between the relative similarity of different people's voices and their synthetic counterparts.

This paper explores two primary methods of synthetic voice generation: Text-to-Speech (TTS) and Voice Conversion (VC). While these technologies have numerous positive applications, they can also be exploited for deceptive purposes (Brewster; Karimi). The aim of this article is not only to explore the two principal approaches to the voice generation, but also to examine the vulnerability of two Speaker Recognition (SR) systems, based on two different sets of utterances in two languages (English and Polish). Moreover, this work determines the impact of the similarity between source speaker's voice in voice conversion attacks on the effectiveness of the frauds. By analyzing these issues, the paper highlights the need for developing more effective biometric security measures capable of countering audio deepfakes. The contributions of this research are as follows:

- exploration of voice synthesis potential across two lingual speech corpora,
- evaluation of speakers' biometric similarity in voice conversion task with regards to SR systems predictions,
- comparison of two speaker recognition systems,
- contrast of Voice Conversion and Text-to-Speech paradigms.

2 Related Work

2.1 Text-to-Speech

Recent advances in Text-to-Speech technology have led to a wide array of approaches aimed at producing natural-sounding and intelligible synthetic speech. These systems are built upon text analysis, acoustic modeling, and vocoders (Tan et al.,

Proceedings of the 31st International Conference on Computational Linguistics, pages 9618–9629
January 19–24, 2025.

2021). The process begins with transforming text into linguistic features. This is followed by acoustic models, traditionally including HMM-based approaches (Fukada et al., 1992). The introduction of deep neural networks has substantially enhanced TTS, with generative models like WaveNet (van den Oord et al., 2016) directly modeling the waveform of speech, capturing its temporal dynamics. Subsequent innovations have marked significant milestones in TTS. Tacotron 2 (Shen et al., 2018), and DeepVoice-3 (Ping et al., 2018) have pushed the boundaries of speech naturalness and synthesis efficiency. FastSpeech 2 (Ren et al., 2022) addressed the expressiveness of speech by incorporating duration, pitch, and energy. Meanwhile, the VITS model combined Generative Adversarial Networks with Variational Autoencoders (Kim et al., 2021) to efficiently produce high-quality, diverse voice samples.

The incorporation of Large Language Models (LLMs) like TorToiSe (Betker, 2023) and VALL-E (X) (Zhang et al., 2023) has further advanced the field, enabling dynamic adjustment in tone and style for more expressive speech synthesis. Among the most popular solutions, due to their accessibility and high-quality voice generation, are open-source and commercial platforms such as Bark (Suno), Coqui (Casanova et al., 2024), and ElevenLabs TTS (ElevenLabs).

Adjacent to aforementioned approaches in TTS area there exist an end-to-end paradigm. E3 TTS (Gao et al., 2023) is based on diffusion model and does not rely on spectral features of the data. It is told to perform well on zero-shot tasks. NaturalSpeech (Tan et al., 2024) utilizes VAEs to achieve human-like quality of synthesized speech. A series of custom modifications, designed to facilitate audio processing were introduced, including phoneme pre-training or memory mechanism.

2.2 Voice Conversion

In the field of Voice Conversion (VC), a diverse range of models has significantly advanced voice synthesis capabilities. Variational Autoencoders (VAEs) stand out as a foundational framework. These models excel in encoding spectral features of a source voice. A comprehensive exploration by Lian (Lian et al., 2022) highlights the significant role of VAEs, demonstrating their capacity to handle the complex task of voice transformation while maintaining the integrity of the original linguistic message. Further developments brought about by StarGAN-v2-VC (Li et al., 2021) and AutoVC (Qian et al., 2019) introduced new features such as changing many voices at once without direct pairing and better separating the speaker's identity from the speech content, respectively. These innovations have made voice conversion more versatile and high-quality.

Additionally, specific technologies like SoftVC (van Niekerk et al., 2022) have made strides in keeping more of the original speech content intact, while models like VALL-E (Wang et al., 2023) show the promise of using large language models to make synthetic voices sound even more natural and expressive. The field has also expanded to include specialized applications, like singing voice conversion (SVC) and the VITS model, which use advanced techniques to ensure the synthetic voice maintains the quality of the original. Both commercial and open-source platforms, such as ElevenLabs VC and SoftVC VITS SVC (SVC Develop Team) have made these powerful voice conversion tools more accessible to a wider audience. End-to-end approach is also present within Voice Conversion frameworks. Many-to-many or zero-shot conversion is possible to perform with NVC-Net (Nguyen and Cardinaux, 2022). Key characteristic of this approach lies in fully convolutional nature of network architecture, which in this case, ensures fast inference. Utilization of information perturbation found its application in (Xie et al., 2022). Algorithm proposed in this work builds on 3 distinct encoders, each for: content, speaker and pitch. Information perturbation is applied through series of functions embedded in content encoder, in order to strip speaker-related signal and focus on linguistic features. LVC-VC (Kang et al., 2023) presents another angle at end-to-end systems for Voice Conversion. Local variable convolutions facilitate high benchmark performance, whilst keeping the model size compact. The novelty and critical importance of our investigation lie in elucidating the implications of synthetic speech on the integrity and security of biometric authentication processes. Our study aims to provide a holistic understanding of how Text-to-Speech and Voice Conversion technologies perform across different languages and genders, and how biometric similarities can affect the efficacy of voice conversion techniques.

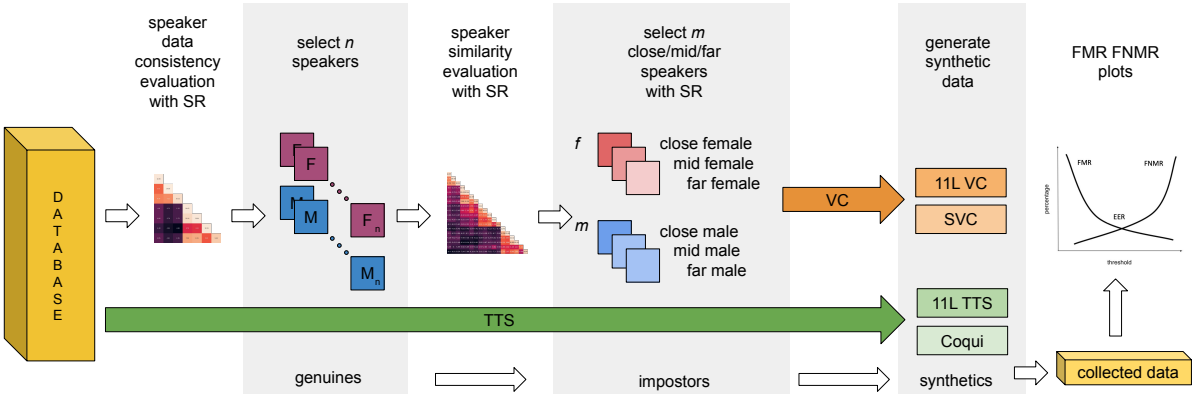

Figure 1: Diagram of data processing pipeline. In this research $n = 10$ and $m = 2$. 2 databases were used for separate experiments, CLARIN PL and LibriSpeech, and 2 SR methods: Resemblyzer and SpeechBrain.

3 Methodology

Our approach involved analysing the distinctive characteristic features within the generated voice samples and assessing the resilience of biometric systems against potential attacks leveraging these synthetics. The goal was to identify which speech synthesis technologies are most effective in mimicking individual characteristics, which is crucial for understanding the extent to which synthetic voices are consistent with authentic data and how biometric systems can be vulnerable to modern voice attack methods. Importantly, we analysed two databases with speakers from different languages and conducted gender-based calculations to check impact of these factors on voice synthesis and biometric security. The entire experimental pipeline is depicted in Figure 1, which will be described in details in the following subsections.

3.1 Speaker Recognition

To evaluate synthetically generated voice materials, we used two open-source systems for speaker recognition:

• **Resemblyzer** (Resemble AI, 2020) is an open-source deep neural network speaker encoder that has gained widespread recognition in recent literature for its robust speaker verification and identification capabilities. Trained on extensive datasets such as VoxCeleb1, VoxCeleb2 and LibriSpeech-train using the generalized end-to-end loss (Wan et al., 2020), it achieves high accuracy in capturing the unique characteristics of a speaker's voice. The system enrolls each speaker with approximately from 5 to 30 seconds of their speech, creating a detailed embedding that encapsulates their vocal identity. For speaker recognition, Resemblyzer computes the embedding of the incoming speech and utilizes a dot product to compare it against existing embeddings in the database, effectively determining the speaker's identity with a high degree of reliability.

• **SpeechBrain** (Ravanelli et al., 2021) is an open-source platform designed for speech processing and machine learning research, which includes a powerful tool for speaker verification using a pretrained ResNet TDNN model. The system excels in extracting speaker embeddings, trained on the extensive VoxCeleb1 and VoxCeleb2 datasets, ensuring a broad learning scope from diverse vocal characteristics. It utilizes Additive Margin Softmax Loss to enhance the discriminative power of the speaker embeddings, crucial for accurate speaker verification. Verification is conducted by measuring the cosine distance between embeddings, providing a reliable method for determining speaker similarity based on their voice.

Speaker verification biometric systems were employed to evaluate synthetic data, including the extent to which voice synthesis methods convey individual characteristics, and to investigate the intra-class consistency of the analyzed databases in accordance with a data-centric concept for detecting potential errors within the collections. Verification was conducted by measuring the cosine distance or calculating the dot product between embeddings respectively for Speechbrain and Resemblyzer. Additionally, these systems were utilized to identify individuals with closely resembling characteristics, pinpoint generic voices, and differentiate those distant in the personal voice feature space.

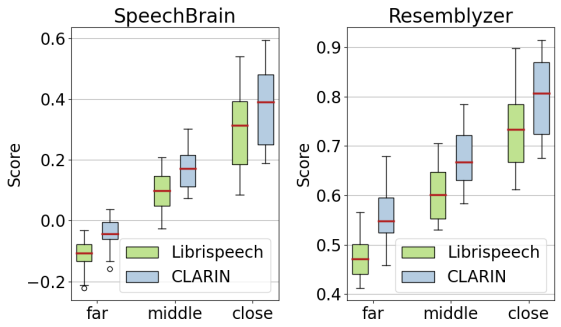

Figure 2: Boxplots illustrating the score spread for SpeechBrain and Resemblyzer for both analyzed databases, categorized into three biometric consistency classes: far, middle, and close.

3.2 Experimental dataset

In our research, we specifically chose to focus on two databases: CLARIN-PL (Marasek et al., 2015), which contains Polish speech, and Librispeech (Panayotov et al., 2015), which is based on English data. Our goal was to compare a phonetically complex language, with a reference language that is widely used in scientific research. The CLARIN-PL database, with over 250 speakers and an average recording length exceeding 5 minutes per person, provided an optimal balance, meeting the requirements of a representative number of speakers and available audio length per speaker while ensuring data consistency. The selection of a subset of the LibriSpeech database for English was driven by the need to align it in terms of speaker count and recording type, as well as optimizing audio length for training voice conversion (VC) models.

Based on preliminary analyses of the calculated intra-class similarity of speakers from speaker recognition systems, it was found that folders designated for individual speakers frequently contained recordings from multiple individuals. Therefore, the data was appropriately reorganized to include samples originating from one speaker per folder. This procedure was necessary to ensure the accuracy and reliability of our experiments, thereby minimizing data-related errors and optimizing the training environment for speaker recognition and voice cloning methods.

Our analysis focused on minimizing intra-class differences to enhance data consistency and reliability for voice synthesis. It was done via selecting files containing no more than 30% of silence, sifting out speakers with less than 6 minutes of data. In order to guarantee speaker consistency, files were manually checked to eliminate mixed-up ones within speaker data. The final data subsets were structured as follows:

- CLARIN-PL: 239 speakers (151 females, 88 males),
- LibriSpeech: 240 speakers (123 females, 117 males), from the test and development subsets, with additional data from the training set incorporated to minimize any potential biases.

3.3 Generating synthetic voices

3.3.1 Text-to-Speech

For each individual from the analyzed voice databases, synthetic voices were generated using two TTS methods, Coqui XTTS-V2 (Casanova et al., 2024) and ElevenLabs TTS (Multilingual v2 with default hyperparameters: Stability: 50%, Similarity: 75%, Style Exaggeration: 0, Speaker boost: True) (ElevenLabs), allowing for the generation of speech across multiple languages. Training was conducted with around 6 minutes of recordings for both methods. Speech outputs were generated using texts not restricted by copyright. For voice samples from the database containing Polish speakers, a fragment from *"Na marne: szkic powieściowy"* by H. Sienkiewicz was utilized. Whereas, for voice samples from the database with English-speaking individuals, a passage from Chapter 3 of *"Winnie the Pooh"* by A. Milne was used.

3.3.2 Voice Conversion

Experiments related to voice conversion methods were also carried out with two algorithms, which enable generating audio in many languages: Soft Conditional Variational Autoencoder with Adversarial Learning for Singing Voice Conversion (voicepaw GitHub user) – VC-SVC and Multilingual-v2 "Speech to Speech" module form ElevenLabs (ElevenLabs) – VC-11Labs, with default hyperparameters: Stability: 50%, Similarity: 75%, Style Exaggeration: 0, Speaker boost: True.

The methodology for selecting data to falsify biometric systems through VC methods commenced with the identification of biometrically consistent classes, characterized by the closest average intra-class comparison indicators. For our study, we concentrated on identifying the ten most biometrically consistent females and males for each speaker recognition system, with respect to each database.

For these individuals, two closely matching counterparts were identified based on their biometric similarities (close). These selections were based on proximity in the feature space, aiming to find subjects whose biometric signatures exhibit minimal divergence when analyzed through the lens of the chosen SR system. This process ensures that the selected individuals for comparison bear a high degree of resemblance in their biometric traits, providing a stringent test for the system's resilience against impersonation. Additionally, for each of ten most biometrically consistent females and males we selected data representing the average of the biometric feature space (middle) to depict the typical population characteristics within each dataset. This middle category serves as a baseline against which the performance of the VC methods can be gauged under normal operational conditions. Finally, individuals representing the edges of the biometric feature distribution (far) were identified for each selected females and males speakers. These subjects exhibit traits that are significantly distinct from the central tendency of the dataset. Including these extreme cases is crucial for assessing the VC methods' capability to handle the full spectrum of variability inherent in human biometric data and for understanding the limits of system security against potential voice conversion attacks. To sum up, 10 male and 10 female speakers with the highest consistency were selected according to both speaker recognition systems. For each of these reference speakers, two intra-gender speakers were selected within the respective similarity groups of voices: close, middle and far. The distribution of results across these groups (far, middle, close) is depicted in Figure 2, providing a visual representation of the variability.

Through this approach, the study encompasses a comprehensive range of biometric variability, from the most similar to the most divergent voice samples, offering a robust framework for evaluating the effectiveness of VC methods in generating synthetic voices that could potentially bypass biometric security measures.

3.4 Metrics

In the evaluation of speaker verification systems with synthetic data, we employed several key metrics to provide a detailed assessment of system performance and security. The False Match Rate (FMR) is an indicator that measures the frequency at which the system erroneously accepts impostors or synthetic voice samples as genuine.

$$FMR = \frac{\text{number of false matches}}{\text{total number of impostors/synthetics}}$$

The False Non-Match Rate (FNMR) evaluates the system's tendency to incorrectly reject genuine attempts.

$$FNMR = \frac{\text{number of false non-matches}}{\text{total number of genuines}}$$

The Equal Error Rate (EER) provides a balanced view of the system's performance, indicating the point where FMR equals FNMR. We employed a 10-fold cross-validation approach to estimate metric averages and standard deviations, enhancing the robustness and reliability of our evaluation.

Parallel to SR assessment, data about recordings facilitating Mean Opinion Score (MOS) calculations was gathered. Detailed description of mentioned social study can be found in Appendix A.

Table 1: MOS for 22 audio clips, grading their naturalness in scale from 1 to 5, where 5 means fully natural one.

	Real	SVC	11labsVC	11labsTTS	Coqui
POL	3.80	3.33	3.62	3.60	2.14
ENG	3.44	3.17	3.17	3.91	2.85

4 Experimental Results

This study assesses voice cloning via EER and compares FNMR against FMR, illustrated in Figure 3. Analyzing results across TTS and VC, we aim to assess their success in preserving original speaker vocal traits. Findings show VC methods significantly outperform TTS in generating high-quality synthetic voices. Table 2 presents the results for the speaker verification method Resemblyzer. This approach, when working with authentic data, achieves an EER of 1.11% and 3.42% for CLARIN-PL and LibriSpeech databases, respectively. The elevated EER for LibriSpeech may result from character impersonations by speakers, impacting both verification accuracy and VC method evaluation. For CLARIN-PL database, TTS-Coqui method is well detected, with an EER of 4.46%. Data generated by TTS-11Labs allows for the transfer of significantly more individual-specific information, resulting in an EER of 28.53%. A significant difference in results is observed between data derived from Polish and English databases. In the case of English, the TTS methods show a much closer

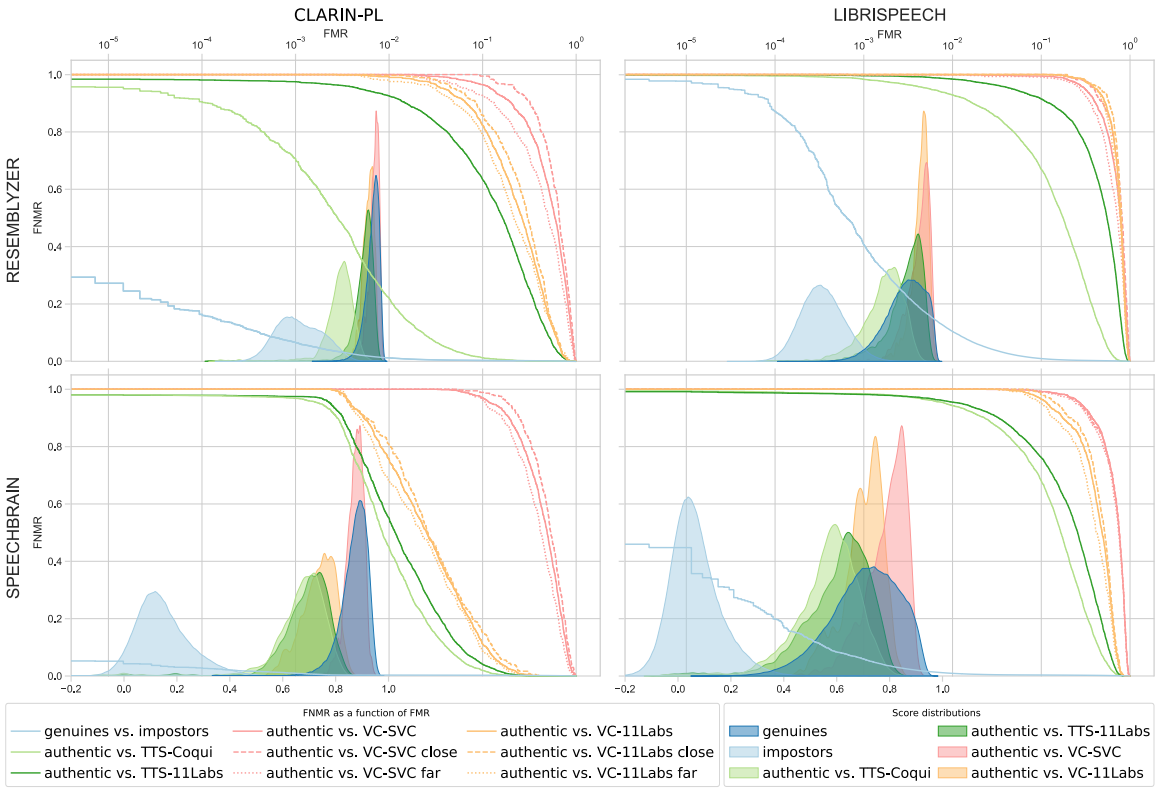

Figure 3: Results of FNMR-FMR analysis and similarity scores distributions with respect to used databases and SR systems.

resemblance to authentic samples, yielding EERs of 29.38% for TTS-Coqui and 52.93% for TTS-11Labs, correspondingly. Regarding voice conversion methods, it was also observed that data from LibriSpeech database resembles more authentic samples. The EER for the English database using the VC-SVC method is 67.21%, while for the VC-11Labs method, it is 70.71%. Gender differences in synthesizer performance are also noticeable, with female voices showing a 4-5% closer resemblance to originals than male voices. In the case of CLARIN-PL database, both voice conversion methods transfer more biometric characteristics for males, resulting in an EER of 57.57% for VC-SVC and 39.07% for VC-11Labs. For females, this rate is nearly 4% lower for VC-SVC, at 53.29%. For VC-11Labs, the difference is significantly larger (over 6%) and stands at 32.40%. Experiments were conducted to investigate the impact of the biometric similarity of individuals used in voice synthesis on the performance of SR systems. It was observed that the closer the "impostor's" voice is to the "victim's" in the biometric feature space, the more individual-specific information the synthetic samples convey. This is particularly significant for women. Cloning a woman's voice using VC-SVC on CLARIN-PL database yields a 61.67% EER, which drops to 46.23% with less similar voice samples, a 15% decrease. On LibriSpeech, the difference is slightly over 5%, with EERs for close and distant individuals being 72.05% and 66.67%, respectively. For men, larger differences are noticeable for English speech - about 10%.

Table 3 presents comparison results for the second analyzed verification system - SpeechBrain. This system significantly better verifies speakers, achieving EER of 0.37% for CLARIN-PL database containing Polish speakers and 3.42% for the English-speaker database. In this case, a greater ability to differentiate synthetic samples using TTS-Coqui method over TTS-11Labs is also observed, with errors of 5.36% and 6.81% for CLARIN-PL database, respectively. Individual voice characteristics are better transferred in LibriSpeech, yielding EERs of 27.79% and 34.87% for TTS-Coqui and TTS-11Labs, respectively. The VC-SVC generates high-quality voices in both languages (with EER of 49.37% for CLARIN-PL and 62.33% for LibriSpeech). In the SpeechBrain system, the impact of similarity between the individuals used for voice generation is noticeable but lower than for Resemblyzer. For the most similar individuals in

CLARIN-PL database, EER is 51.88%, while for those most distant, it is 49.95%. For LibriSpeech database, these rates are 66.33% for conversions of the most biometrically similar individuals and 64.60% for the least similar voices.

An additional study was conducted to evaluate the naturalness and authenticity of the generated samples using the Mean Opinion Score (MOS), with results presented in Table 1. The study was conducted in Poland, therefore, it can be assumed that the native language of the participants is the local one. In Polish, the highest MOS scores, after real samples, were achieved by the 11labsTTS and 11labsVC samples. In English, the 11labsTTS method attained the highest MOS score, even surpassing real recordings, which may be due to English not being the native language of the experiment participants. The VC methods received equal scores, while samples generated by the Coqui method were consistently rated as the least natural in both languages.

The influence of gender is noticeable, male voices in the English database are characterized by significantly better quality (EER for male voices is 71.09%, females - 62.44%), while for CLARIN-PL database, the reverse is true (EER for female voices is 51.55%, males - 45.87%). With the second VC method VC-11Labs, synthetic samples in Polish were not biometrically similar to the original voices, resulting in EER of 8.43%. Synthetic voices produced from English voices were significantly better, resulting in EER of 48.03%. For the VC-11Labs method in both databases, male voices were of better quality, resulting in EER of 9.44% (for females, 2.59 % lower) for CLARIN-PL database and 54.44% for LibriSpeech database (for females, it was 41.50%). In the case of SR using the SpeechBrain tool, the impact of the similarity of individuals selected for voice conversion on the results was noticeable. For similar individuals, EER was 9.44%, while for those distant in the feature space, it dropped to 7.78% for CLARIN-PL database (with larger differences in the group of male voices), similarly for LibriSpeech 50.25 and 44.59% (with a greater impact among female voices). Additionally, analysis of results for both SR systems showed that both TTS methods yield lower quality of records for Polish women. This pattern was not observable in LibriSpeech.

5 Conclusions

In our study, we utilized pre-trained models to analyze the distinctive characteristic features within the generated voice samples and assess the resilience of biometric systems against potential attacks leveraging these synthetics. Our evaluation protocol was designed to determine the effectiveness of various speech synthesis technologies in mimicking individual vocal characteristics, which is crucial for understanding how closely synthetic voices can replicate authentic data and the extent to which biometric systems are vulnerable to advanced voice spoofing methods.

The experiments show differences in the operation of speaker verification systems for different voice synthesis methods, taking into account the analysis of sex and language of the speakers. Similarity of genuines and impostors correlates with credibility of generated synthetic audio. The closer the biometric features of the "impostor's" voice to those of the "victim's," the more individual-specific information the synthetic samples conveyed. Multiple tests of SR systems showed that Resemblyzer is more prone to be fooled by deepfakes than SpeechBrain. However, experiments presented vulnerability of both methods. Counter-intuitively, VC methods appeared to be more frequently accepted by SR systems than TTS approach. This may be attributed to the VC's ability to retain certain speech style characteristics of the original speaker, thereby presenting more identifiable biometric features to SR systems. However, the MOS evaluation, aimed at capturing the naturalness and authenticity of the generated samples, does not indicate a trend that VC models produce superior deepfakes. While this aspect was not the primary focus of the earlier discussion, it is an important consideration that suggests VC methods may be more effective in preserving the speaker's unique vocal attributes. Additionally, the study identified variable outcomes in the efficacy of VC methods related to gender-specific differences. The transfer of biometric features and the resultant EERs were not consistent and varied depending on the database and speaker recognition methods used. This suggests that the relationship between gender and the performance of VC methods is complex and not universally applicable. Further comprehensive research is required to elucidate these observations and understand the underlying factors contributing to the variability in VC efficacy across genders.

Table 2: Summary of EER[%] showcasing the mean and std deviation for speaker verification using Resemblyzer across two databases.

Comparison type	**CLARIN-PL**			**LIBRISPEECH**		
	all	**female**	**male**	**all**	**female**	**male**
authentics - **genuines vs. impostors**	1.11 (0.00)	1.46 (0.00)	1.25 (0.00)	3.42 (0.01)	4.88 (0.01)	4.44 (0.01)
auth. vs. synth - **TTS-Coqui**	4.46 (0.02)	3.06 (0.02)	6.90 (0.03)	29.38 (0.04)	28.26 (0.04)	30.47 (0.04)
auth. vs. synth - **TTS-11Labs**	28.53 (0.05)	26.95 (0.07)	30.90 (0.05)	52.93 (0.06)	53.10 (0.07)	52.38 (0.07)
auth. vs. synth - **VC-SVC - all**	55.13 (0.10)	53.29 (0.12)	57.57 (0.11)	67.21 (0.07)	70.37 (0.07)	64.54 (0.08)
auth. vs. synth - **VC-SVC - close**	60.16 (0.11)	61.67 (0.10)	58.73 (0.15)	71.09 (0.08)	72.05 (0.09)	69.93 (0.08)
auth. vs. synth - **VC-SVC - middle**	54.21 (0.10)	51.67 (0.16)	58.00 (0.10)	68.52 (0.07)	70.32 (0.08)	66.43 (0.13)
auth. vs. synth - **VC-SVC - far**	50.27 (0.14)	46.23 (0.15)	54.84 (0.16)	62.26 (0.08)	66.67 (0.13)	58.75 (0.08)
auth. vs. synth - **VC-11Labs - all**	35.51 (0.08)	32.40 (0.10)	39.07 (0.11)	70.71 (0.08)	72.57 (0.10)	68.46 (0.11)
auth. vs. synth - **VC-11Labs - close**	37.12 (0.09)	32.78 (0.11)	41.29 (0.15)	72.43 (0.07)	73.24 (0.10)	70.11 (0.11)
auth. vs. synth - **VC-11Labs - middle**	33.33 (0.09)	30.33 (0.14)	39.30 (0.11)	70.68 (0.09)	73.37 (0.09)	67.61 (0.10)
auth. vs. synth - **VC-11Labs - far**	32.76 (0.10)	27.22 (0.09)	36.49 (0.14)	68.67 (0.09)	70.74 (0.12)	66.63 (0.11)

Table 3: Summary of EER[%] showcasing the mean and std deviation for speaker verification using SpeechBrain across two databases.

Comparison type	**CLARIN-PL**			**LIBRISPEECH**		
	all	**female**	**male**	**all**	**female**	**male**
authentics - **genuines vs. impostors**	0.37 (0.00)	0.02 (0.00)	0.00 (0.00)	0.95 (0.00)	1.46 (0.00)	1.23 (0.01)
auth. vs. synth - **TTS-Coqui**	5.36 (0.02)	4.49 (0.02)	6.13 (0.04)	27.79 (0.05)	28.83 (0.04)	25.85 (0.06)
auth. vs. synth - **TTS-11Labs**	6.81 (0.02)	5.34 (0.02)	8.87 (0.05)	34.87 (0.05)	33.02 (0.05)	36.20 (0.08)
auth. vs. synth - **VC-SVC - all**	49.37 (0.12)	51.55 (0.15)	45.87 (0.13)	66.33 (0.08)	62.44 (0.05)	71.09 (0.07)
auth. vs. synth - **VC-SVC - close**	51.88 (0.10)	54.80 (0.16)	49.44 (0.14)	66.23 (0.10)	62.56 (0.06)	69.93 (0.08)
auth. vs. synth - **VC-SVC - middle**	49.35 (0.14)	52.78 (0.17)	44.24 (0.15)	67.14 (0.07)	64.95 (0.05)	70.67 (0.08)
auth. vs. synth - **VC-SVC - far**	49.95 (0.12)	46.13 (0.16)	43.89 (0.14)	64.60 (0.08)	60.34 (0.10)	72.93 (0.10)
auth. vs. synth - **VC-11Labs - all**	8.43 (0.05)	6.85 (0.05)	9.44 (0.08)	48.03 (0.07)	41.50 (0.09)	54.44 (0.11)
auth. vs. synth - **VC-11Labs - close**	9.44 (0.06)	6.67 (0.06)	10.56 (0.09)	50.25 (0.08)	42.58 (0.10)	55.00 (0.10)
auth. vs. synth - **VC-11Labs - middle**	8.06 (0.05)	6.46 (0.06)	8.33 (0.07)	49.44 (0.08)	43.11 (0.12)	54.44 (0.13)
auth. vs. synth - **VC-11Labs - far**	7.78 (0.04)	6.11 (0.05)	7.78 (0.09)	44.59 (0.08)	35.82 (0.09)	51.67 (0.13)

There are many challenges in the field of biometric voice recognition, particularly from the point of view of the current level of development of generative methods that enable voice manipulation. Further work may include a thorough analysis of the scale of the vulnerability presented, by increasing the number of languages and SR models, as well as proposing methods to counter such attacks.

6 Limitations

Our approach utilized the novel tools of speech cloning, both commercially available and open-source. These solutions became famous for the good quality of the generated materials (Walczyna and Piotrowski, 2023) and thus impose a significant threat for Automatic Speaker Verification (ASV) applications. On the other hand the relevance of our study may decrease over time due to continuous advances and updates to speech synthesis and speech recognition technology, which is already very effective. The databases used, CLARIN-PL and LibriSpeech, required substantial cleanup due to inconsistencies and errors, and our requirement for a significant number of minimum audio length speakers limited our dataset options. Additionally, we maintained gender balance by selecting a similar number of male and female speakers, a total of 200 per database, which may limit sample diversity.

Our study included only Polish and English languages, which limited the applicability of our findings to these language groups. Expanding the study to other languages or accent varieties could provide broader knowledge, but such expansion would pose challenges related to the availability of balanced, comprehensive datasets. Thus, while our findings provide valuable information on the vulnerability of biometric systems to voice spoofing, they are limited by time, data, and scope constraints of our study setup.

References

James Betker. 2023. Better speech synthesis through scaling. *Preprint*, arXiv:2305.07243.

Thomas Brewster. Huge Bank Fraud Uses Deep Fake Voice Tech To Steal Millions. www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/. Accessed: 2023-10-01.

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, and Julian Weber. 2024. Xtts: a massively multilingual zero-shot text-to-speech model. In *Interspeech 2024*, pages 4978–4982.

Héctor Delgado, Massimiliano Todisco, Md Sahidullah, Nicholas Evans, Tomi Kinnunen, Kong Aik Lee, and Junichi Yamagishi. 2018. ASVspoof 2017 Version 2.0: meta-data analysis and baseline enhancements . In *Proc. The Speaker and Language Recognition Workshop (Odyssey 2018)*, pages 296–303.

ElevenLabs. ElevenLabs website. https://elevenlabs.io. Accessed: 2024-03-01.

T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai. 1992. An adaptive algorithm for mel-cepstral analysis of speech. In *ICASSP 1992 - 1992 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 137–140 vol.1.

Yuan Gao, Nobuyuki Morioka, Yu Zhang, and Nanxin Chen. 2023. E3 tts: Easy end-to-end diffusion-based text to speech. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8.

Wonjune Kang, Mark Hasegawa-Johnson, and Deb Roy. 2023. End-to-End Zero-Shot Voice Conversion with Location-Variable Convolutions. In *Proc. INTERSPEECH 2023*, pages 2303–2307.

Faith Karimi. 'Mom, these bad men have me': She believes scammers cloned her daughter's voice in a fake kidnapping. https://edition.cnn.com/2023/04/29/us/ai-scam-calls-kidnapping-cec/index.html. Accessed: 2023-10-01.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5530–5540. PMLR.

Yinghao Aaron Li, Ali Zare, and Nima Mesgarani. 2021. Starganv2-vc: A diverse, unsupervised, non-parallel framework for natural-sounding voice conversion. In *Interspeech 2021*, pages 1349–1353.

Jiachen Lian, Chunlei Zhang, and Dong Yu. 2022. Robust disentangled variational speech representation learning for zero-shot voice conversion. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6572–6576.

Krzysztof Marasek, Danijel Koržinek, Łukasz Brocki, and Kamila Jankowska-Lorek. 2015. Clarin-PL studio corpus (EMU). http://hdl.handle.net/11321/236. CLARIN-PL digital repository.

Bac Nguyen and Fabien Cardinaux. 2022. Nvc-net: End-to-end adversarial voice conversion. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7012–7016.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *ICASSP 2015 - 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O. Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. 2018. Deep voice 3: Scaling text-to-speech with convolutional sequence learning. *Preprint*, arXiv:1710.07654.

Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson. 2019. AutoVC: Zero-shot voice style transfer with only autoencoder loss. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5210–5219. PMLR.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. 2021. SpeechBrain: A general-purpose speech toolkit. *Preprint*, arXiv:2106.04624. ArXiv:2106.04624.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2022. Fastspeech 2: Fast and high-quality end-to-end text to speech. *Preprint*, arXiv:2006.04558.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. Fastspeech: Fast, robust and controllable text to speech. *Preprint*, arXiv:1905.09263.

Resemble AI. 2020. Resemblyzer Github Repository. `https://github.com/resemble-ai/Resemblyzer`. Accessed: 2023-10-21.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, Rif A. Saurous, Yannis Agiomvrgiannakis, and Yonghui Wu. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783.

Suno. BARK Github Repository. `https://github.com/suno-ai/bark`. Accessed: 2024-02-25.

SVC Develop Team. SoftVC VITS Singing Voice Conversion Github Repository. `https://github.com/svc-develop-team/so-vits-svc`. Accessed: 2023-10-01.

Xu Tan, Jiawei Chen, Haohe Liu, Jian Cong, Chen Zhang, Yanqing Liu, Xi Wang, Yichong Leng, Yuanhao Yi, Lei He, Sheng Zhao, Tao Qin, Frank Soong, and Tie-Yan Liu. 2024. NaturalSpeech: End-to-End Text-to-Speech Synthesis With Human-Level Quality . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 46(06):4234–4245.

Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu. 2021. A survey on neural speech synthesis. *Preprint*, arXiv:2106.15561.

Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. In *9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, page 125.

Benjamin van Niekerk, Marc-André Carbonneau, Julian Zaïdi, Matthew Baas, Hugo Seuté, and Herman Kamper. 2022. A comparison of discrete and soft speech units for improved voice conversion. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

voicepaw GitHub user. So-VITS-SVC Github Repository. `https://github.com/voicepaw/so-vits-svc-fork`. Accessed: 2023-11-13.

Tomasz Walczyna and Zbigniew Piotrowski. 2023. Overview of voice conversion methods based on deep learning. *Applied Sciences*, 13(5).

Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. 2020. Generalized end-to-end loss for speaker verification. *Preprint*, arXiv:1710.10467.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2023. Neural codec language models are zero-shot text to speech synthesizers. *Preprint*, arXiv:2301.02111.

Jee weon Jung, Hemlata Tak, Hye jin Shim, Hee-Soo Heo, Bong-Jin Lee, Soo-Whan Chung, Ha-Jin Yu, Nicholas Evans, and Tomi Kinnunen. 2022. Sasv 2022: The first spoofing-aware speaker verification challenge. In *Interspeech 2022*, pages 2893–2897.

Zhizheng Wu, Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, Cemal Hanilçi, Md. Sahidullah, and Aleksandr Sizov. 2015. ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. In *Proc. Interspeech 2015*, pages 2037–2041.

Qicong Xie, Shan Yang, Yi Lei, Lei Xie, and Dan Su. 2022. End-to-end voice conversion with information perturbation. In *2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 91–95.

Ziqiang Zhang, Long Zhou, Chengyi Wang, Sanyuan Chen, Yu Wu, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2023. Speak foreign languages with your own voice: Cross-lingual neural codec language modeling. *Preprint*, arXiv:2303.03926.

A Social Study

A study regarding naturalness and authenticity of speech samples was carried out via on-line anonymous questionnaire. Such data is a source for calculating Mean Opinion Score (MOS) metric, which is widely used to asses quality of synthetic audio.

The form had two sets of questions, followed by demographic inquiry. In the first set, respondents were presented series of 22 audio clips and were tasked with grading their naturalness in scale from 1 to 5, where 5 means fully natural one. Experimental data was anonymous (with regards to its source) in order to reduce possible bias. Files selected for assessment covered real samples, TTS-generated by coqui and 11Labs, as well as VC (taking into consideration distinction between close, middle and far relations of speaker-attacker pair) created with both, SVC and 11Labs. Source languages were equally present in this evaluation. Representation of generative methods was balanced, with 3 files per VC method; 3 TTS samples and 2 real files, summing up to 11 clips per language.

Prior to the questionnaire completion, respondents were informed that audio clips contain both real and fake examples. Such an approach was necessary to gather data about the human subjective perception, but on the other hand it evoked more caution and made people more focused on the listening part. They were also informed that there were no wrong answers and their subjective opinion is crucial part of this study.

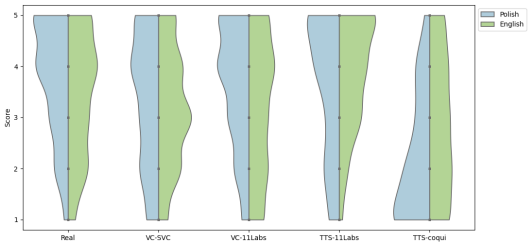

Figure A.1: Distributions of scores.

The interviewers were mainly from large cities (69%) with a higher education (89%). The gender of the study group is well balanced (Figure A.2), and the age distribution is described by the Figure A.3.

A.1 Mean Opinion Score

Answers gathered from 71 respondents, allowed calculating the Mean Opinion Score (MOS), which

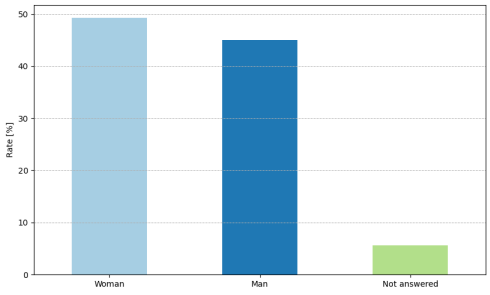

Figure A.2: Gender distribution of the interviewers.

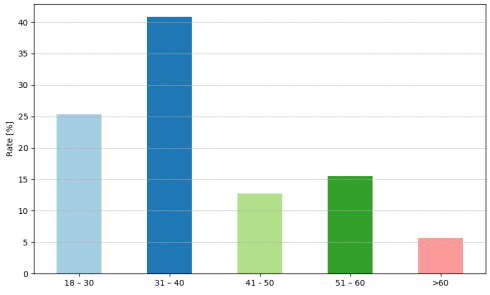

Figure A.3: Age distribution of the interviewers.

is the arithmetic mean over all individual grades. The results of this study are presented on the Figure A.4. In Polish, the 11labsTTS and 11labsVC samples received the highest MOS scores, second only to the real recordings. In English, the 11labsTTS method achieved the top MOS score, even outperforming real samples, possibly it may be caused by English was not the participants' native language. The VC methods were rated equally, while the Coqui samples were consistently judged as the least natural in both languages.

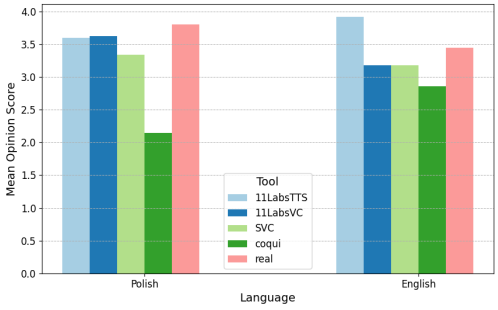

Figure A.4: Naturalness grading in scale from 1 to 5, where 5 means fully natural one.

A.2 Study conclusion

The warning about synthetic speech, along with the quality of the recordings and the way the speakers delivered the text, may have made respondents overly critical in their evaluations. This is evident from the response distribution for the bonafide samples and the MOS metric: real samples sometimes received the same or even lower scores than certain synthetic voice methods.

The survey result shows that the difference between synthetic and real voice escapes human perception and the preferred method is not clearly visible. This conclusion does not cover TTS-Coqui. This method has proven to be the distinctive worst.

A.3 Limitations

Due to the nature of the survey, it was not possible to standardize the acoustic conditions experienced by participants. This lack of control may have introduced bias into the scoring, potentially obscuring differences between the samples being compared.

The survey was conducted with a relatively small sample of 71 participants, primarily educated individuals from large cities with a declared familiarity with new technologies. A more comprehensive study would benefit from a more diverse research group.