Speaker Anonymization via GAN-driven Artificial Embeddings and Prosody Transfer

Dumitru-Cosmin Melinte
Telecommunications Department
Technical University of Cluj-Napoca
Cluj-Napoca, Romania
Cosmin.Melinte2720@gmail.com

Mircea Giurgiu
Telecommunications Department
Technical University of Cluj-Napoca
Cluj-Napoca, Romania
Mircea.Giurgiu@com.utcluj.ro

***Abstract*—This paper introduces a novel approach for speaker anonymization based on cascaded speech-to-text (STT) and text-to-speech (TTS) Deep Neural Networks (DNNs), where the multi-speaker TTS component is driven by a Generative Adversarial Network (GAN) to anonymize the speakers' identity and by a prosody transfer module to preserve the original prosody. The GAN creates artificial speaker embeddings that do not correspond to any real identity, enabling a high diversity of synthesized voices. The prosody preservation and speaker embedding generation for the TTS component are clearly distinct approaches: the embeddings are progressively generated from a low and compact latent space, while the prosody is directly cloned onto the synthesized speech. The experimental results show low cosine similarity between original and anonymized embeddings, with some minor degradation in linguistic fidelity due to the performance of the STT and TTS components. Additionally, the developed visual interface supports comparative analysis of audio features before and after anonymization, contributing to a qualitative assessment of system performance.**

***Keywords*—Speaker anonymization, Speech-to-Text, Text-to-Speech, GAN embeddings, prosody transfer.**

I. INTRODUCTION

Speaker anonymization is the process by which a person's speech signal is modified so that their identity can no longer be recognized, while the linguistic content remains unchanged. The speech signal conveys more than the linguistic message: it contains personal information about the speaker such as identity, gender, emotions or mental state. All of these are sensitive data, so protecting the speaker's identity has become a major concern. In several applications, such as Automatic Speech Recognition (ASR), the speaker's specific voice is not relevant, because only the linguistic content matters. Thus, by transmitting data using an anonymized system, the speaker's privacy can be secured [1]. On the other hand, privacy protection increases when significant modifications are applied to the signal, but in most cases this negatively affects intelligibility, naturalness, and expressivity of the synthesized speech. Therefore, a compromise must be found between the level of privacy and the usefulness of the anonymization system [2].

This paper aims to develop a speaker anonymization system based on a speech-to-text-to-speech (S2T2S) architecture that preserves the naturalness and prosody of speech while maintaining a high level of privacy. It explores the integration of a Generative Adversarial Network (GAN) for creating anonymous embeddings and replicating the prosody of the original signal.

II. STATE OF THE ART

VoicePrivacy is an initiative launched internationally in 2020, aiming to unify efforts in the development of voice anonymization systems. The criteria defined by the *VoicePrivacy 2020 Challenge* include generating an anonymized speech signal that prevents the reconstruction of the speaker's identity, even when the anonymization algorithm is known. However, linguistic content and intelligibility must remain unaffected [3][4]. Typically, there are two main approaches for speaker anonymization:

(1) Signal processing methods. These methods directly alter acoustic attributes of speech: intonation, timbre or formants. Current approaches involve relatively simple techniques, such as modifying the McAdams coefficient, altering the fundamental frequency (F0), and changing spectral parameters in the speaker embedding (e.g., x-vector) [1][4][5].

(2) Artificial intelligence-based methods. They separate verbal content from speaker-specific features and replace them with arbitrary speech features before generating a new voice. Such methods are generally more robust and achieve better results in concealing speaker identity while maintaining intelligibility [1][2].

Most solutions developed under the *VoicePrivacy Challenge* initiative follow a speaker anonymization pipeline with typically three main stages: *(1) extracting linguistic information* using an automatic speech recognition (ASR) system or phonetic feature extractor; *(2) replacing the original speaker embedding* with an artificially generated one; and *(3) synthesizing a new voice* using the extracted linguistic content and a speech synthesis system [6][7].

An early method used x-vector embeddings, replacing the speaker's original embedding with one randomly selected from a large pool of speakers. The speech was then resynthesized using a neural acoustic model [4]. Voice conversion methods such as CycleGAN have also been explored to map the source voice to a different one without using transcription [8]. Although effective, these purely neural methods often produce ASR or TTS artifacts that degrade speech naturalness and may result in artificial-sounding voices [2]. Recent literature introduces a new approach based on Generative Adversarial Networks (GANs) to create realistic yet non-existent voices. This approach improves the quality of the anonymized voice without compromising privacy.

Meyer et al. [7] proposed training a GAN to generate artificial embeddings that follow the same distribution as real ones, but do not correspond to any actual speaker. These

979-8-3315-5896-3/25/$31.00 ©2025 IEEE

embeddings are used within a S2T2S pipeline. The original speech is transcribed with an ASR system, while the speaker embedding is replaced and the anonymized voice is synthesized. The results show that this method produces distinct and natural voices, outperforming prior approaches. Unlike randomly selected x-vectors [4] or CycleGAN methods [8], the method proposed in [7] generates synthetic embeddings from a continuous latent space, allowing for virtually unlimited voice diversity. The intermediate transcription stage helps reduce phonetic traits linked to speaker identity, which is difficult to achieve with direct voice conversion. This conceptual design improves privacy without sacrificing naturalness.

Another important conceptual aspect addressed by Meyer et al. [9] is the preservation of original prosodic features. This hypothesis is explored by extending a S2T2S anonymization pipeline and cloning the original prosody to the anonymized voice. Their results suggest that prosodic information from the original speech can be added to the new voice without revealing the identity of the original speaker.

Similarly, Lee et al. [10] proposed that, in addition to prosody, emotional characteristics should also be explicitly integrated. Their findings indicate that enriching the input representation with emotional state vectors increases the level of anonymization while preserving intelligibility. However, there may be a slight reduction in the accuracy of emotion recognition in the resulting voice.

III. PROPOSED METHODOLOGY AND ARCHITECTURE IMPLEMENTATION

A. Innovative Aspects of the Proposed Solution

A.1. Using a S2T2S processing pipeline. The proposed solution is built around a S2T2S architecture, which is inspired by recent work aimed at separating linguistic content from speaker characteristics through intermediate transcription and reintegration during synthesis [6][7]. Our research adopts a similar idea that is explicitly developed for Romanian, a language with scarce text and speech datasets freely available for research. All components are configured and evaluated on Romanian data, which distinguishes the system from general approaches focused primarily on English. By transferring the speaker's prosody to the synthesized speech, the system aims to maintain naturalness without compromising identity confidentiality.

A.2. Adopting a prosody transfer module to maintain speech naturalness. The prosody preservation component is inspired by findings showing that prosody can be transferred without reintroducing speaker identity [9][10]. In this work, the original prosody (e.g., pitch, energy, duration, and pauses) is directly transferred onto a fully synthesized identity generated by a dedicated GAN, rather than to a real or an embedding-derived identity. This approach preserves the temporal and intonational structure while ensuring that the vocal signature remains artificial.

A.3. Development of GAN-based embeddings for the anonymized multi-speaker TTS. The speaker embedding generation component is inspired by GAN-based approaches from the *VoicePrivacy Challenge* [7][8], but differs significantly in the generator's architecture. Instead of concatenating x-vector/ECAPA embeddings with high dimensionality, the proposed architecture uses a compact generator starting from a 64-dimensional latent space, producing 192-dimensional embeddings through a dynamic path of 64–256–512–192 dimensions. This simplification reduces model complexity, facilitates training on the available data, and ensures that the speaker's identity is fully synthetic, with no reference to real voices. Training is stabilized using a Wasserstein loss with gradient penalty, combined with a diversity loss to prevent excessive similarity between samples [2][7][8]. The embedding-level analysis component is inspired by speaker verification practices, but extended here by explicitly reporting cosine similarity, both between the original and anonymized embeddings and among anonymized embeddings themselves, over sets of increasing size. The presented results demonstrate both the necessary mismatch for anonymization and the coherence of the synthetic distribution.

A.4. Development of visualization tools. In addition, the prosody visualization component is inspired by standard signal inspection tools but is distinguished by an interface that compares the pitch contours of the original and anonymized speech files side by side and displays key values such as duration, average energy, and cosine distance. This tool transforms the proposed complex processing pipeline into a functional evaluation environment.

B. Speech Processing Pipeline

The sequence of speech processing steps is illustrated in Fig. 1 and covers the full chain from audio preprocessing to the generation of the anonymized synthetic voice, preserving the linguistic content and prosodic features of the original signal. In this implementation, prosody is cloned to a GAN-synthesized identity. This is a key feature because it clearly decouples temporal-intonational naturalness from vocal identity.

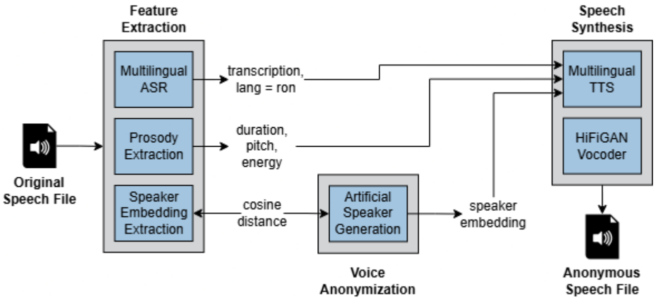


Fig. 1. Block diagram of the anonymzation pipeline.

B.1. Speech transcription using ASR. Automatic speech recognition (ASR) is performed with OpenAI Whisper models for Romanian [11]. The provided Whisper models have been previously tested on Romanian, achieving high accuracy performance in controlled environments. The audio input is standardized to mono, with a sampling frequency of 16 kHz. The model consumes log-Mel spectrograms and autoregressively outputs Romanian transcripts. We do not apply pause removal or non-speech filtering at this stage. This text representation isolates linguistic content from speaker-specific acoustics, which is essential before resynthesis from anonymized identities.

B.2. Prosody extraction. From the original signal, the fundamental frequency (F0), energy, and duration are extracted over successive frames to obtain the prosodic contour that controls temporal and intonational modulation during synthesis. In this work, these features are used to transfer the prosody to the synthetic output so that rhythm, pauses, and stress patterns are preserved in the anonymized version. For qualitative inspection, the developed graphical interface displays the pitch contours of the original and anonymized signals side by side, along with aggregated values for duration, energy, and cosine distance, thus supporting the evaluation of the cloning effect. Prosodic feature extraction is implemented using the Librosa and Parselmouth speech processing libraries.

B.3. Speaker embedding extraction. A numerical representation of the speaker's identity, called a speaker embedding, is extracted from the same reference input utterance. This vector acts as a *signature*, which is replaced by an anonymized embedding generated by a GAN. This is a key mechanism that breaks the link between the transcribed content and the original identity. The ECAPA-TDNN model is used for embedding extraction via the SpeechBrain framework [12]. The embeddings used to train the GAN were computed from recordings in the Romanian Common Voice 21.0 speech dataset.

B.4. GAN training for synthetic embedding generation. A GAN is trained on real speaker embeddings to learn their distribution and generate diverse, non-identifiable embeddings. Unlike previous architectures based on x-vector/ECAPA concatenation, this system uses a compact generator that starts from 64-dimensional latent noise vectors and outputs 192-dimensional embeddings through layers of size 64–256–512–192. Stability is enforced using Wasserstein loss with gradient penalty, and a diversity loss is used to prevent mode collapse. This design simplifies training and supports the goal of producing fully synthetic identities. GAN training is implemented in PyTorch using WGAN-GP and diversity loss.

The proposed WGAN-GP architecture has a MLP (Multi-Layer Perceptron) generator with a critic acting as a discriminator. The role of the critic is to evaluate the generated embeddings against the real ones. The generator takes as input a Gaussian latent vector $z \in R^{64}$ and consists of three linear layers (64–256–512–192) with Rectified Linear Unit (ReLU) activations after the first two layers. The output is a 192-dimensional embedding. The critic receives a 192-dimensional embedding and has three linear layers (192–512–256–1) with Leaky Rectified Linear Unit (LeakyReLU, slope 0.2) activations after the first two layers. At the data level, per-sample standardization is applied: the mean is subtracted and the result is divided by the standard deviation, with a fallback if $\sigma < 10^{-9}$. Training uses the WGAN-GP loss with gradient penalty $\lambda_{gp} = 10$, Adam optimizer $\left(lr = 10^{-4}, \beta = (0.5,0.9)\right)$, batch size 64, 50 epochs, and five critic updates are used per generator update. The generator also includes a diversity loss term based on the mean cosine similarity within the batch, weighted by $\lambda_{div} = 5$.

Speaker embeddings for training were extracted from the Romanian subset of the Common Voice Corpus 21.0 (release: March 19, 2025; license: CC-0; audio format: MP3; validated hours: 22 h; size: 1.02 GB; number of voices: 443) [13]. The corpus was used solely to learn the embedding distribution. The synthesized identities are entirely artificial and unrelated to real speakers.

B.5. Anonymous embedding generation with the GAN. After training, the generator receives a random latent noise vector and outputs a 192-dimensional anonymous embedding that does not correspond to any real speaker, while remaining coherent within the learned embedding space. This vector replaces the original embedding during synthesis, thereby anonymizing the speaker's identity.

B.6. Anonymous speech synthesis with prosody transfer. A multi-speaker TTS model combined with a neural vocoder generates the anonymized speech signal using the transcribed text and the anonymous embedding. The original prosody defined by pitch contour, energy, duration, and silences is then transferred to the output signal to preserve the original speaker's intonation and rhythm while replacing the identity. While this approach aligns with recent trends in prosody preservation, it is methodologically distinct. We apply prosody to an identity entirely synthesized by a GAN, which, to our knowledge, represents a key difference from many other implementations in the literature. Voice synthesis is performed using the IMS-Toucan TTS model and the HiFi-GAN vocoder [14].

IV. Experimental Results

A. Experimental Results on GAN Training

The GAN model was trained in two separate sessions of 50 epochs each, using real speaker embeddings. The goal was to generate synthetic embeddings that match the distribution of the original ones while remaining diverse and anonymous. During training, three metrics were monitored at each epoch: generator loss (*Loss G*), discriminator loss (*Loss D*), and diversity loss. The average values are shown in Table I. Their evolution indicates a gradual stabilization of the GAN and a progressive improvement in diversity.

Following the two training sessions, the generator loss increased notably, indicating that the generator became more effective at fooling the discriminator. The discriminator loss remained relatively stable, suggesting a balance between the two network components. The similar values of the diversity loss metric indicate good consistency in generating varied embeddings.

TABLE I. METRICS IN GAN TRAINING

GAN Training	Loss D	Loss G	Diversity Loss
Session 1	-4.99	9.91	0.15
Session 2	-4.83	11.36	0.15

B. Evaluation of GAN-generated Embeddings

To assess the quality of the embeddings generated by the trained GAN, two types of similarity were analyzed using the cosine similarity coefficient*: (1)* between the original embedding and its corresponding anonymized embedding; and *(2)* between all generated anonymized embeddings, excluding self-similarity on the main diagonal.

Fig. 2 illustrates the dual histograms for each sample for original–anonymous and between–anonymous cosine similarities. The results indicate low cosine similarity between original and anonymized embeddings, with values

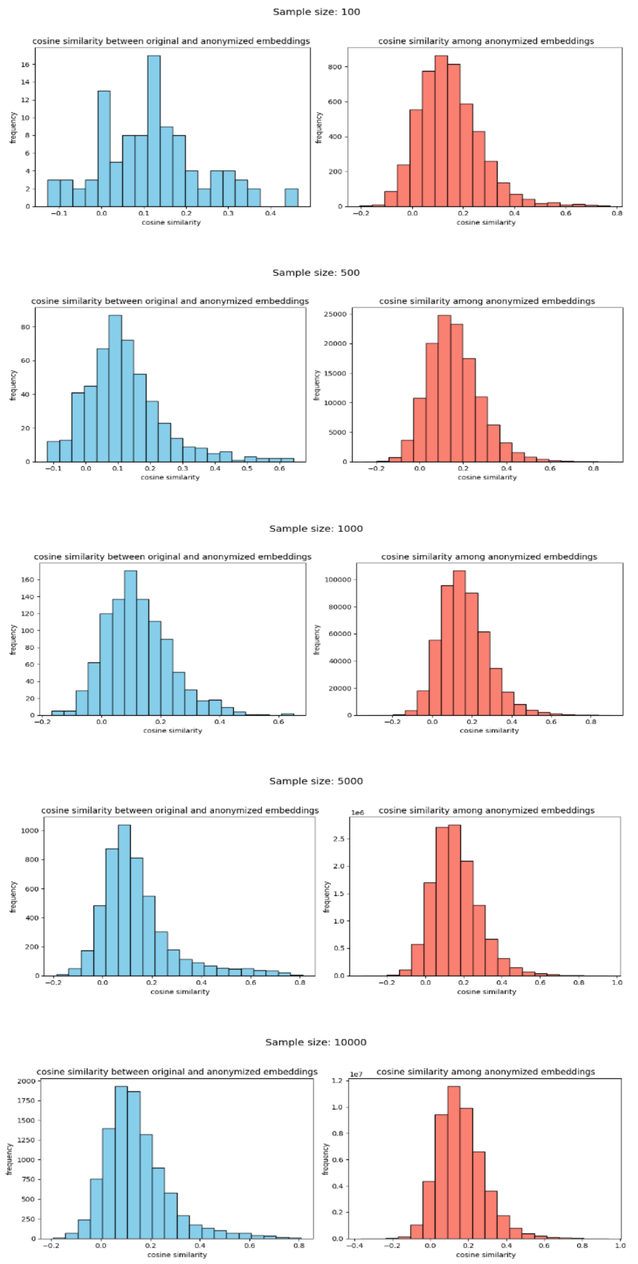


Fig. 2. Dual histogram for each sample (original vs. anonymous and between-anonymous embeddings).

ranging from 0.12 to 0.14, demonstrating the necessary mismatch required for effective speaker anonymization. The mean similarity between anonymized embeddings remains relatively stable, between 0.14 and 0.16, indicating a good balance between diversity and coherence. The evaluation was performed on five datasets of different sizes: 100, 500, 1,000, 5,000, and 10,000.

The results are summarized in Table II. It is remarked that the GAN model successfully generates embeddings that are sufficiently different from the originals while remaining coherently distributed in the speaker embedding space, thus supporting anonymization without fully compromising realism.

TABLE II. SIMILARITY OF GAN-GENERATED EMBEDDINGS

No. of samples	Mean cosine similarity (original-anonymous)	Mean cosine similarity (anonymous–anonymous)
100	0.1199	0.1451
500	0.1196	0.1570
1,000	0.1239	0.1611
5,000	0.1374	0.1594
10,000	0.1440	0.1606

C. Comparing Original and Anonymized Embeddings

To evaluate the GAN model's ability to generate anonymous embeddings that differ from the original ones, an analysis based on cosine similarity was performed. This metric measures how close two vectors are in the embedding space, with a value of 1 indicating perfect identity and 0 indicating full orthogonality. Negative values occur when vectors point in opposite directions. The evaluation compared each original embedding with its corresponding anonymized embedding. Three datasets were tested, containing 500, 3,500, and 10,000 embeddings. For each dataset, the full distribution of cosine similarity values was computed. The aggregated results are presented in Table III. Fig. 3 shows the distribution of cosine similarity for each dataset size.

The cosine similarity evaluation confirms that the model effectively removes distinctive features of the original voice without compromising the acoustic coherence of the resulting embeddings. The low similarity values between original and anonymized embeddings validate identity separation and demonstrate the effectiveness of the proposed GAN architecture in preventing overlap between real and synthetic distributions. The generated voices sound natural in informal listening tests but cannot be linked to any real speaker. Moreover, the consistent distribution of anonymized embeddings suggests that the model does not produce isolated samples but successfully spans a coherent space of synthetic voices. This is essential for practical applications that require scalable and stable diversity.

Overall, the results confirm that the system achieves strong anonymization with preserved prosody, although intelligibility requires improvement.

TABLE III. COSINE SIMILARITY OF ORIGINAL VS. ANONYMOUS EMBEDDINGS

No. of Samples	Mean Similarity	Minimum	Maximum	Similarity < 0.5
500	0.1194	-0.1206	0.6708	98.8%
3,500	0.1290	-0.2265	0.7981	97.23%
10,000	0.1466	-0.2125	0.8081	97.13%

The distribution plot for each set in Fig. 3 shows a concentration of values around the 0.1–0.2 range, with a rapid decrease towards the extremes. This confirms that the anonymous embeddings generated by the GAN are sufficiently distant from the original ones, supporting the anonymization objective. At the same time, the preservation of moderate positive similarity values indicates that the generated embeddings remain phonetically coherent, without becoming random or unstable.

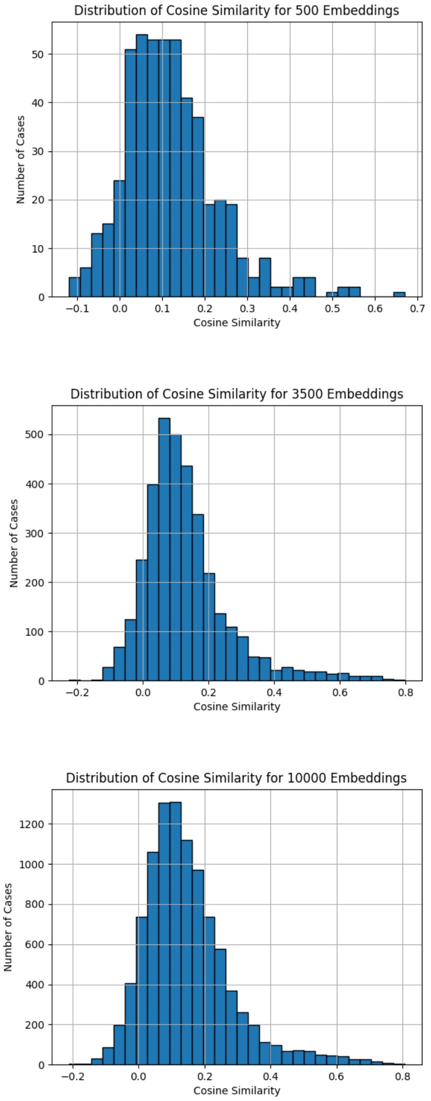


Fig. 3. Cosine Similarity Distribution: Original Embedding vs Anonymous.

The distributions also demonstrate good consistency in the results: the values do not vary abruptly as dataset size increases, and the proportion of embeddings with similarity values below 0.5 remains above 97% in all tested cases.

D. Evaluation of ASR Errors of Anonymized Speech

To assess the extent to which linguistic content is affected by the anonymization process, the Character Error Rate (CER) and Word Error Rate (WER) metrics were used with the ASR. Two comparisons are shown in Table IV: (i) between the original reference text and the ASR transcription of the original speech, and (ii) between the ASR transcription of the original speech and the ASR transcription of the anonymized speech. The speech dataset used in this test is the same as that used for objective evaluation (see Section IV.F).

The ASR baseline was highly accurate (CER = 1.12%, WER = 5.45%), given the reduced size of the training dataset (e.g., only 22 h in Romanian compared with 360 h in English [19]). After anonymization, “machine intelligibility” degraded substantially, with a CER of 14.42% and a WER of 37.74%. Therefore, while the anonymization method successfully altered speaker identity, it also introduced significant distortion for automatic speech recognition.

TABLE IV. CER AND WER RESULTS FOR ASR

Performance metric	Original speech	Anonymized speech
CER (%)	1.12	14.42
WER (%)	5.45	37.74

Compared to other results for Romanian, our ASR baseline provides similar performance given the type of the model and the size of the training data [20]. However, with respect to ASR on anonymized speech, there is room for improvement, as current state-of-the-art systems report WER below 10% [19]. It is still important to note that such systems are trained with hundreds of hours of speech and the models are much larger than the one used in this research.

The CER and WER values indicate that, while anonymous embeddings are effective for modifying vocal identity, speech reconstruction does not fully preserve linguistic content. The causes of the errors include transcription errors in the initial ASR stage, phonetic artifacts introduced by synthetic embeddings, and limitations of the speech synthesis model used. It is important to note that these results do not invalidate the system's primary objective of speaker anonymization. However, they show that the current configuration penalizes speech intelligibility and naturalness. Improving the TTS component and refining the synthetic embeddings are essential directions for optimizing overall performance.

In this research, the system was optimized for identity disruption and prosody preservation instead of lexical intelligibility. Consequently, WER was not a primary optimization target. We expect WER to decrease as we integrate stronger Romanian ASR and TTS models and apply basic post-synthesis corrections.

E. Subjective Evaluation of Anonymized Speech

A Mean Opinion Score (MOS) test was conducted on the anonymized samples following the ITU-T P.800 absolute category rating (ACR) methodology, using a 5-point scale where 1 = “bad” and 5 = “excellent”. Twenty-five participants evaluated the intelligibility of ten anonymized audio files generated with the proposed system. The overall MOS, computed over all 250 ratings, was 3.59 ± 0.14 (95% confidence interval (CI)), which corresponds to intelligibility between “fair” and “good”. This indicates that anonymized speech remains understandable.

However, performance varied across stimuli: the highest MOS was 4.36 (CI [4.13–4.59]), while the lowest score was 2.48 (CI [2.05–2.91]). This variability suggests that although anonymization preserves intelligibility in most cases, some input utterances are more affected by the transformation process. Overall, the results confirm that the proposed anonymization method maintains intelligibility at an acceptable level for human listeners, while also highlighting opportunities for improvement in robustness across different speech inputs.

In addition to MOS ratings, participants provided free-text comments for each anonymized stimulus. The qualitative analysis revealed consistent patterns across evaluations. Most listeners confirmed that the anonymized speech remained intelligible, describing them as “understandable”. However, several issues were repeatedly mentioned. A frequent observation was the “few omissions or distortions of phonemes”, which occasionally made words harder to recognize. Some utterances were perceived as “produced with a foreign accent” (e.g., “French-like intonation”) or with “robotic quality”. In certain cases, listeners reported that “the beginning or ending of sentences was less clear”, with “missing sounds” or “abrupt cutoffs”.

Overall, this analysis indicates that while the anonymization preserves intelligibility in most cases, there is perceptible degradation in pronunciation consistency and voice naturalness, which may explain the variability across stimuli.

F. *Objective Metrics for Evaluation of Anonymized Speech*

This evaluation was conducted using 50 high-quality studio recordings collected from 10 speakers (5 females, 5 males, 5 sentences per speaker) with the aim to analyze the prosodic aspects revealed by the fundamental frequency (F0) contours of original and anonymized speech, as well as objective metrics PESQ (Perceptual Evaluation of Speech Quality) and STOI (Short-Time Objective Intelligibility). While PESQ estimates, on a scale from 1 to 5, how the speech would sound to a human listener, STOI predicts, on a scale from 0 to 1, how well a listener can understand the spoken words.

Fig. 4 shows the objective evaluation of prosody features for a sample sentence. The following aspects can be observed: a) the overall duration of the sentence is preserved after anonymization; b) the anonymized sentence maintains the syllabic structure and the accents (maxima in F0); c) the pitch (F0) is lower, on average by 25 Hz, corresponding to the anonymization algorithm, while the overall decay intonation is preserved.

The PESQ estimation (Fig. 5) indicates a score of 4.5 for the original signals and a very compact distribution of the PESQ values for all anonymized samples around 2.7. This reflects a homogenous quality of the anonymized speech, independent of the speaker, prosody, or the semantic content. The obtained values are comparable with those reported in [15], but lower than the average value (e.g., 3.43) reported in [16].

The STOI parameter (Fig. 6) shows a very high intelligibility score of 1.0 for all original signals and a range of variation between 0.55 and 0.95 for the anonymized signals. The analysis of these values reveals significant correlation between STOI and ASR performance (e.g., errors in character recognition) and only partial correlation with TTS quality. The results are comparable with those reported in [17] in terms of range of variation, but they are lower than the typical score of 0.95 reported in [18].

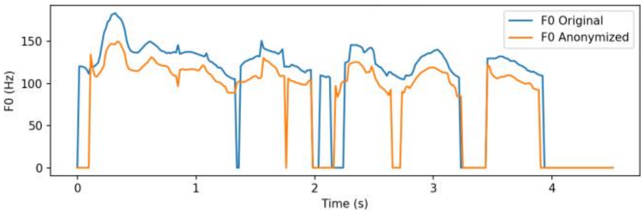


Fig. 4. Fundamental frequency of the original vs. anonymized speech.

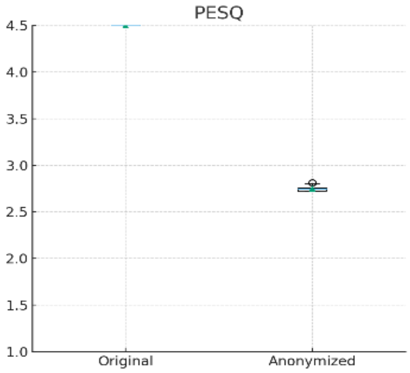


Fig. 5. Compact PESQ distribution for anonymized speech (right).

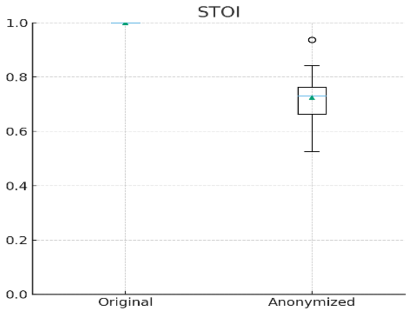


Fig. 6. Variation of STOI values for anonymized speech.

G. *Comparison of Original and Anonymized Embeddings*

By comparing original and anonymized speech embeddings, it is shown that the proposed GAN architecture is not only simpler than existing solutions, but also competitive. Fig. 7 illustrates that the original and anonymized embeddings of the 10 speakers from the test set are quite different in the two-dimensional projected t-SNE space: some speakers are strongly anonymized, while for the others the anonymization is only mild.

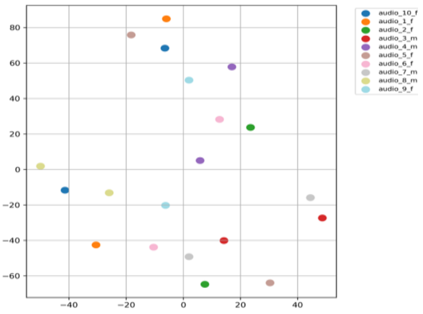
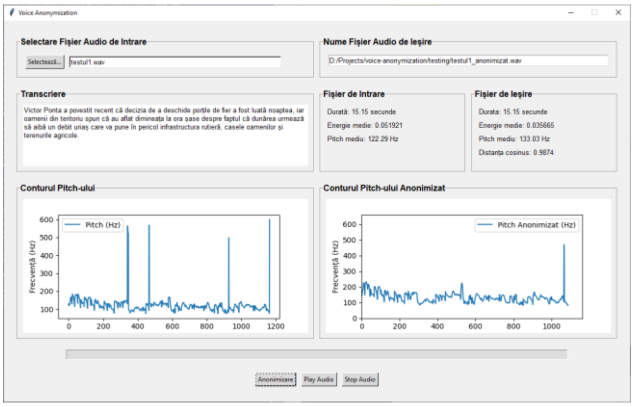


Fig. 7. t-SNE plots for original and anonymized embeddings.

H. *Supporting Software Tools for Results Visualization*

Throughout this research, the graphical interface of the application was expanded to improve usability and clarity in evaluating results. The improvements were functional as well as aesthetic, contributing to a streamlined interpretation of the anonymization system's behavior.

The interface (Fig. 8) allowed users to select an input file and specify the output filename. After processing, the application displayed a summary of the input signal, including its automatic transcription, duration, average energy, pitch contour, and cosine distance to the anonymous embedding.



testul1.wav

D:/Projects/voice-anonymization/testing/testul1_anonimizat.wav

Victor Ponta a povestit recent că decizia de a deschide porțile de fier a fost luată noaptea, iar oamenii din teritoriu spun că au aflat dimineața la ora șase despre faptul că dunărea urmează să aibă un debit uriaș care va pune în pericol infrastructura rutieră, casele oamenilor și terenurile agricole.

Durată: 15.15 secunde
Energie medie: 0.051921
Pitch mediu: 122.29 Hz

Durata: 15.15 secunde
Energie medie: 0.035665
Pitch mediu: 133.03 Hz
Distanța cosinus: 0.9874

Fig. 8. Graphical user interface of the visualization tool.

The interface integrates a parallel display system for both input and output file characteristics, providing a clear view of the changes introduced by the anonymization process. Each file is shown in a separate section, with values calculated independently for duration, average energy, and cosine distance.

Additionally, the interface includes two graphs of the pitch contour for the original and anonymized files. These graphs provide an intuitive visualization of intonation differences and help to illustrate how vocal identity is altered. All information is presented in a clear format. Interactive elements such as the Play Audio, Stop Audio, and Anonymize buttons have been retained alongside the progress bar to facilitate user interaction with the program. Users can observe in real-time the impact of the generated embedding on the audio signal's structure, which supports qualitative analysis of the proposed method's effectiveness.

I. *Discussion of results and future improvements*

While the system presented in this paper effectively anonymizes voices and maintains high embedding diversity, the WER results could be improved. Our evaluation emphasizes privacy and prosody control. The system was not tuned for WER in the iteration, and intelligibility improvements are deferred to future work. This limitation is primarily caused by the ASR and TTS models currently employed. For future system development, we will consider integrating more advanced models as they become available in our research group. Additionally, transcription correction or post-synthesis strategies could improve semantic clarity. In the long term, a perceptual evaluation component will be considered to validate the impact of modifications on human perception.

Another future direction involves developing an AI-based automated verifier capable of assessing the quality and efficiency of the anonymization process. Its role would be to automatically analyze anonymized audio files, estimating similarity scores to the original voice, the level of prosody preservation, and the intelligibility of the message. By integrating specialized neural classifiers, the system could detect significant degradations of linguistic content. Thus, the entire voice anonymization process could be automated, allowing continuous monitoring and improvement of synthesized voice quality without human intervention.

In terms of runtime analysis, the system was initially developed on an Ubuntu server with 4 RTX 2080 GPUs, where speech segments of 5 seconds are processed in real-time. However, when run on a Windows laptop with RTX 3060 GPU, the same speech segments were anonymized in 7.5 seconds, leaving room for few optimizations to reach near real time processing.

V. CONCLUSIONS

This paper presented the development and evaluation of a speaker anonymization pipeline based on a S2T2S architecture and the generation of synthetic embeddings using a simplified GAN model. The proposed approach aims to separate linguistic content from speaker-specific features, providing a solution that prioritizes identity protection and prosody preservation. Lexical intelligibility is a secondary objective in this version and will be addressed in future work.

Experimental evaluations showed that the artificially generated embeddings have a low cosine similarity to the original ones, averaging between 0.12 and 0.14, indicating the misalignment required for effective anonymization. At the same time, the anonymized embeddings remain coherently distributed in the acoustic space, with average similarity between 0.14 and 0.16, suggesting a balance between diversity and consistency. These results confirm the GAN ability to produce acoustic vectors close to those of real speakers, without being linkable to an actual vocal identity. The associated visualization tool provides an intuitive framework for comparing audio characteristics before and after anonymization.

In conclusion, the proposed solution represents a step toward speaker identity privacy, leveraging adversarial generation to create unlimited embeddings and synthesize natural but anonymous voices. However, there remains a trade-off between the level of anonymization and the semantic quality of speech, which requires further refinement.

Future work will focus on: (1) integrating advanced ASR and TTS models; (2) developing automatic and objective metrics for perceptual evaluations; (3) introducing an adaptive feedback mechanism to dynamically adjust GAN or TTS model parameters in real time; and (4) testing the system under varied environmental conditions. This combination of controlled generation and phonetic fidelity makes the proposed approach effective in protecting vocal identity and adaptable to practical contexts where clarity and naturalness of communication remain essential.

ACKNOWLEDGEMENT

This work is supported by the project "*Romanian Hub for Artificial Intelligence - HRIA*", Smart Growth, Digitization and Financial Instruments Program, MySMIS no. 334906.

REFERENCES

[1] S. Meyer, P. Tilli, P. Denisov, F. Lux, J. Koch, and N. T. Vu, "Anonymizing speech with generative adversarial networks to preserve speaker privacy," in Proc. IEEE Spoken Language Technology Workshop (SLT), 2022.

[2] H. Turner, G. Lovisotto, and I. Martinovic, "Speaker anonymization with distribution-preserving x-vector generation for the VoicePrivacy Challenge 2020," in Proc. VoicePrivacy Workshop, 2020.

[3] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J. Patino, J.-F. Bonastre, P.-G. Noé, and M. Todisco, "Introducing the VoicePrivacy initiative," in Proc. Interspeech, 2020, pp. 1693–1697, doi: 10.21437/Interspeech.2020-1333.

[4] B. M. L. Srivastava, N. Tomashenko, X. Wang, E. Vincent, J. Yamagishi, M. Maouche, A. Bellet, and M. Tommasi, "Design choices

for x-vector based speaker anonymization," in Proc. Interspeech, 2020, pp. 1713–1717, doi: 10.21437/Interspeech.2020-2692.

[5] C. O. Mawalim, K. Galajit, J. Karnjana, S. Kidani, and M. Unoki, "Speaker anonymization by modifying fundamental frequency and x-vector singular value," Comput. Speech Lang., vol. 73, p. 101326, 2022.

[6] S. Meyer, F. Lux, P. Denisov, J. Koch, P. Tilli, and N. T. Vu, "Speaker anonymization with phonetic intermediate representations," in Proc. Interspeech, 2022, pp. 4925–4929, doi: 10.21437/Interspeech.2022-10703.

[7] S. Meyer, P. Tilli, F. Lux, P. Denisov, J. Koch, and N. T. Vu, "Cascade of phonetic speech recognition, speaker embeddings GAN and multispeaker speech synthesis for the VoicePrivacy 2022 Challenge," in Proc. 2nd Symp. on Security and Privacy in Speech Communication, 2022.

[8] G. P. Prajapati, D. K. Singh, P. P. Amin, and H. A. Patil, "Voice privacy through x-vector and CycleGAN-based anonymization," in Proc. Interspeech, 2021, pp. 1684–1688, doi: 10.21437/Interspeech.2021-1573.

[9] S. Meyer, F. Lux, J. Koch, P. Denisov, P. Tilli, and N. T. Vu, "Prosody is not identity: A speaker anonymization approach using prosody cloning," in Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Rhodes, Greece, 2023, pp. 1–5.

[10] J. Lee, T. Park, and Y. You, "Voice anonymization using emotion-enriched feature integration with STT and TTS models," in Proc. 4th Symp. on Security and Privacy in Speech Communication, 2024, pp. 50–54, doi: 10.21437/SPSC.2024-9.

[11] T. Gigant, "Automatic speech recognition – Whisper model for Romanian", 13 Sept 2023. https://huggingface.co/gigant/whisper-medium-romanian.

[12] M. Ravanelli, T. Parcollet, A. Moumen, S. de Langen, C. Subakan, P. Plantinga, et al., "Open-Source Conversational AI with SpeechBrain 1.0", J,ournal of Machine Learning Research, 2024. [Online]. Available: http://jmlr.org/papers/v25/24-0991.html

[13] Mozilla Foundation, "Common Voice Corpus", 2024. [Online] https://commonvoice.mozilla.org/datasets.

[14] P. Eibl, and M. Wöllmer, "IMS-Toucan: A neural text-to-speech toolkit", 2023. [Online]. Available: https://toucantts.com/.

[15] U. E. Gaznepoglu and N. Peters, " Evaluation of the Speech Resynthesis Capabilities of the VoicePrivacy Challenge Baseline B1," in Voice Privacy Challenge 2023. [Online]. Available https://arxiv.org/abs/2308.11337

[16] L. Chen, and K.A. Lee, "Any-to-any Speaker Attribute Perturbation for Asynchronous Voice Anonymization," in IEEE Transactions on Information Forensics and Security PP(99):1-1, pp. 7736-7747, 2025. [Online]. Available: https://www.arxiv.org/abs/2508.15565

[17] J. Pohlhausen, F. Nespoli, J. Bitzer, "Enhancing Speech Privacy with LPC Modifications," in 4th Symposium on Security and Privacy in Speech Communication", 6 Sept. 2024, Kos, Greece. [Online] https://www.isca-archive.org/spsc_2024/pohlhausen24_spsc.pdf

[18] B. Tura-Vecion, et al., "Universal Semantic Disentangled Privacy-preserving Speech Representation Learning," in Proc. of Interspeech 2025, pp. 3633-3637. [Online]. Available: https://arxiv.org/abs/2505.13085

[19] N. Tomashenko, and al., "The VoicePrivacy 2024 Challenge Evaluation Plan," in VoicePrivacy Challenge Competition, June 2024. [Online]. Available: https://arxiv.org/abs/2404.02677

[20] V. Păis, et al., "Under-Represented Speech Dataset from Open Data: Case Study on the Romanian Language." in Applied Sciences, 2024, 14(19), 9043.[Online].Available:https://doi.org/10.3390/app14199043