2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

Non-Target Conversion Based Speech Steganography for Secure Speech Communication System

Mingjun Zhang*, Yan Feng*, YuGao†, Longting Xu*‡
* College of Information Science and Technology, Donghua University, China
E-mail: xlt@dhu.edu.cn, zmj@mail.dhu.edu.cn
† AI Research Center, Midea Group (Shanghai) Co.,Ltd., Shanghai 201702, China
E-mail: gaoyu11@midea.com
‡ Corresponding Author

***Abstract*—The widespread application of speech data increases the risk of speaker identity being compromised during speech communication. To mitigate this risk and protect voice privacy, we propose a system aimed at ensuring the security of speech communication by avoiding the exposure of the actual speaker's identity. Our system comprises three main components. Firstly, the non-target voice conversion system based on generative adversarial network converts the original audio into the audio of a non-existent person while preserving the speaker embedding of the real audio. Secondly, during speech communication, we utilize speech steganography techniques to embed the actual speaker embedding into the converted audio. Finally, at the receiving end, we extract the actual speaker embedding from the transmitted converted audio and use it to reconstruct the original audio. Experimental results validate the effectiveness of our system, showcasing an innovative solution in the field of speech security.**

I. Introduction

With the integration of voice user interfaces into devices like AI speakers, speech data has become ubiquitous and widely used in areas such as smart homes and conversational AI, becoming an essential component of future lifestyles. However, speech signals contain a wealth of speaker-specific information, including sensitive attributes such as gender, identity, age, emotions, and sensations. If intercepted during transmission, these sensitive attributes can be extracted and used as biometric features or for malicious purposes like identity theft or voice spoofing [1]. The privacy protection of speech data has gained increasing attention, particularly after the introduction of the General Data Protection Regulation (GDPR) in the European Union [2], emphasizing the importance of restrictions on speech data usage and the implementation of security measures [3].

Voice conversion (VC) refers to the various modifications made to human speech, including the conversion of non-linguistic information such as voice timbre, prosody, or pitch, while keeping the linguistic information unchanged [4]–[6]. The primary purpose of voice conversion is to transform the identity of one speaker into another while retaining the linguistic content of the speech data. Speaker anonymization, on the other hand, is a speech data processing technique aimed at protecting the identity privacy of speakers [7], [8]. This technique involves steps such as content/speaker separation and voice conversion, to remove the speaker's identity information from the original speech while retaining other information such as linguistic content. By using the technique of speaker anonymization to convert the speaker's identity to a randomly generated non-existent identity, non-target voice conversion can be achieved. While the method performs well in removing speaker identity information, it is not sufficient to solely rely on it for protecting voice privacy in the process of speech communication. This is because when the receiving end perceives the received converted speech as mismatched with the known voice of the original speaker, doubts about the authenticity of the speech information arise.

Therefore, in the process of speech communication, the ability of the receiving end to reproduce the original speech is a crucial issue for protecting speaker privacy using non-target voice conversion techniques. To address this problem, speech steganography can play an important role. Speech steganography is a technique that embeds secret information into audio signals. It modifies the speech data to increase the capacity of hidden information, making it difficult for external observers to detect the presence of additional information. Speech steganography can be achieved through various conventional methods, including LSB substitution [9], echo hiding [10], and spread spectrum (SS) [11]. With the popularity of deep neural networks, many recent studies have utilized DNNs to embed and detect information [12]. By utilizing this technique, we embed the speaker embedding of the original speech data into the converted speech transmitted during the communication process, making it possible for the receiving end to reconstruct the original speech data.

In this paper, we propose a novel system, called the non-target Conversion based Speech Steganography System (CSSS), for protecting voice privacy in speech communication. At the sender's end, we employ a Generative Adversarial Network (GAN) [13], [14] based speaker anonymization system to create an artificial speaker embedding space and sample non-existent voices, thereby achieving conversion of the identity of the original speaker to a non-existent one. To reconstruct the original speech at the receiver end without leaking identity information, we employ a Convolutional Neural Network (CNN)-based speech steganography technique and introduce a graph encoder [15] based on Graph Convolutional Networks (GCN) [16] to embed the speaker embedding of the original

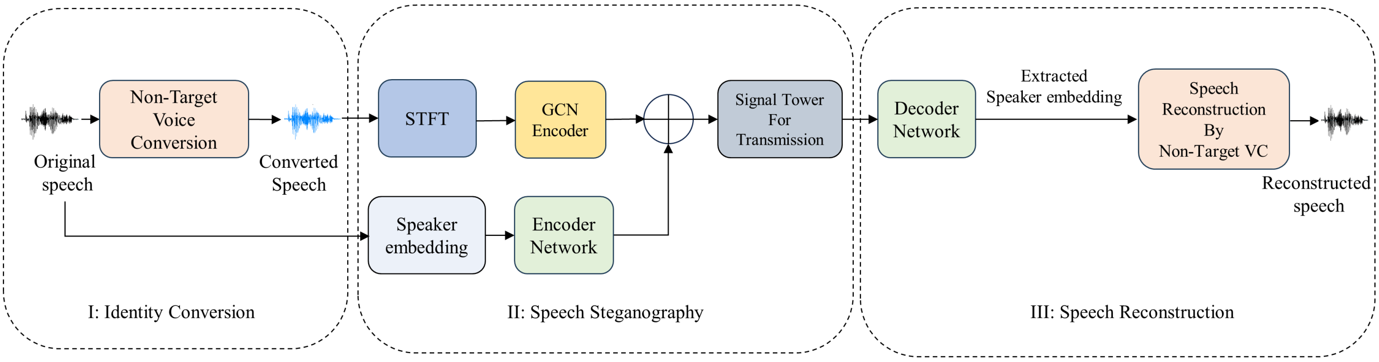

Fig. 1. Architecture of the proposed CSSS. The three sections from left to right correspond respectively to the transmitting end, the transmission process, and the receiving end of the speech signal.

speech into converted speech for transmission. At the receiver's end, the system extracts the hidden speaker embedding and combines it with the prosody information and speech-to-text information of the original speech data to reconstruct the original voice data.

The experimental results indicate that the speech processed by identity conversion and information steganography in CSSS has a high level of privacy, which can effectively prevent privacy attackers from identifying the real speaker. Additionally, the reconstructed speech is close to the original speech in terms of privacy and utility. Furthermore, the reconstructed speech in CSSS demonstrates high intelligibility, and naturalness, and maintains a high perceptual similarity to the original speech.

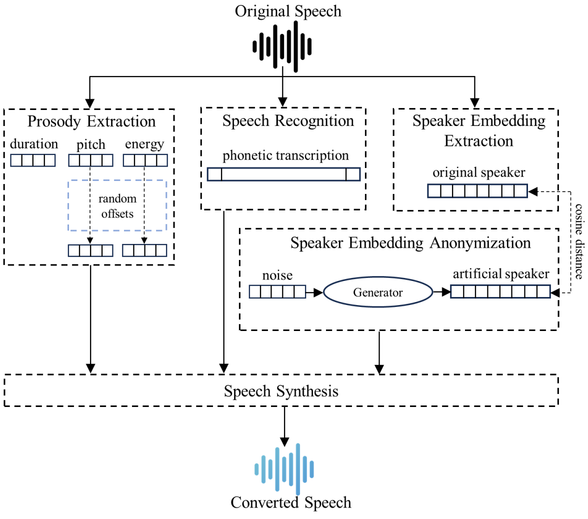

Fig. 2. Illustration of the pipeline for non-target voice conversion. By extracting prosody information, and speech content information, and utilizing cosine similarity to approximate the embedding of the original speaker, randomly generated non-existent speaker identity can be obtained.

II. Proposed System

In this section, we will provide a detailed description of the specific architectural design of CSSS. Fig. 1 illustrates the three main components of CSSS, namely the identity conversion, the speech steganography during speech communication, and the speech reconstruction.

A. Identity Conversion

Prosody is closely related to speaker identity, affecting an individual's pitch and speaking speed. Pitch (F_0) and speaking speed not only reveal information about the speaker but also provide essential features for most machine learning models that process speech. Therefore, in privacy-preserving related tasks, maintaining a constant F_0 could potentially lead to identity leakage [17]. Ref. [18] overcame the trade-off between privacy and utility of prosody by implementing prosody cloning, and achieved an anonymization system with high privacy and high pitch relevance. We adopt the speaker anonymization pipeline proposed by [18] to achieve non-target voice conversion, where the converted speech has a high pitch similarity to the original speech while having a non-existent speaker identity, as shown in the structure in Fig. 2.

The pipeline extracts three types of information from the input speech: (i) prosodic information as pitch, energy, and duration sequences; (ii) the linguistic content in the form of phonetic sequences by using an Automatic Speech Recognition (ASR) model; and (iii) the speaker information as speaker embedding. Pitch and energy can be adjusted by multiplying them with random offsets to further enhance privacy. Additionally, in the Speaker Embedding Anonymization section depicted in Fig. 2, a GAN is employed to transform random noise into speaker embeddings that are unrelated to the original speaker, with the cosine distance ensuring sufficient similarity between the generated vector and the vector corresponding to the input speaker. Finally, the pipeline utilizes a Text-to-Speech (TTS) model to re-synthesize the audio based on language, prosody, and generated speaker information. The resulting converted audio maintains the same language content and prosody features as the original audio.

B. Speech Steganography

1) CNN-based Encoder Network: The traditional speech steganography techniques are mainly divided into the time domain and transform domain methods. These methods, however, encounter a couple of challenges. To begin with, the process of manually selecting appropriate redundancy to conceal secret information requires a considerable investment of manpower, resources, and time. Additionally, the manual selection of redundancy for embedding secret information may result in noticeable alterations in the carrier, leaving behind discernible traces.

Steganographic methods based on the Encoder-Decoder structure of neural networks only require training the adopted model. Utilizing a trained DNN for steganography and information extraction in speech carriers can effectively overcome the shortcomings of traditional methods. In the second part of CSSS, we propose a CNN-based steganographic structure that can embed vector data into audio, utilizing it as an Encoder-Decoder network. The specific structure is shown in Fig. 3.

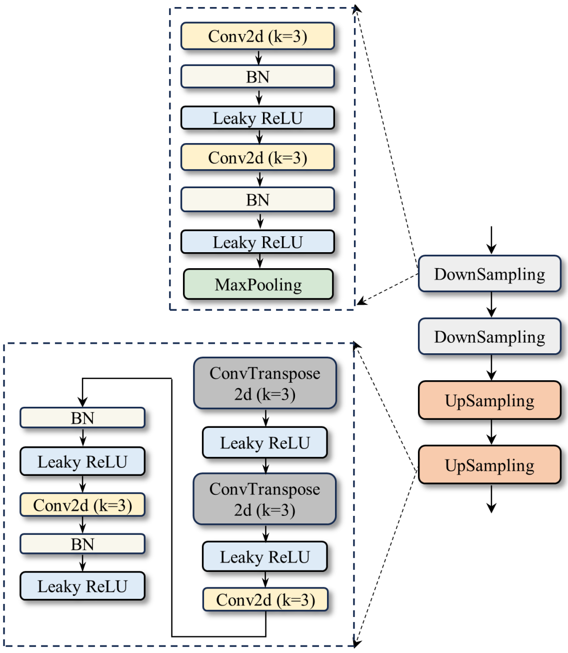

Fig. 3. The structure of the encoder network. BN stands for Batch Normalization.

The encoder network is a two-dimensional fully convolutional neural network based on the U-Net [19] architecture. It consists of a contraction part and an expansion part, which are responsible for the downsampling and upsampling steps. The contraction part comprises two downsampling modules, each containing two 3×3 convolutional layers. Batch normalization and Leaky ReLU [20] activation function are applied after each convolutional layer. The last layer of the downsampling module is a max-pooling layer, used to reduce the spatial dimensions of the speaker embeddings and capture important features. The expansion part consists of two upsampling modules, each primarily composed of two transpose convolutional layers and two 3×3 convolutional layers. Transpose convolutional layers are used to increase the spatial dimensions of the speaker embeddings and recover fine-grained information. In addition, each speaker embedding used for steganography is a matrix of size 128×1.

2) Graph Encoder: To embed richer and more accurate speaker embeddings in converted speech, we introduce the graph encoder proposed in [15]. Specifically, we construct a graph representation for each converted speech and encode its structural details using a dual-layer GCN encoder. The encoded graph features will complement the original feature representation, thereby enhancing the accuracy and richness of the overall feature representation.

The detailed structure of the graph encoder is shown in Fig. 4. After the converted speech undergoes Short-Time Fourier Transform (STFT) operation, a sequence X is obtained. Each element in the sequence X can be regarded as a node in the graph and the connections between nodes can be characterized by graph relationships.

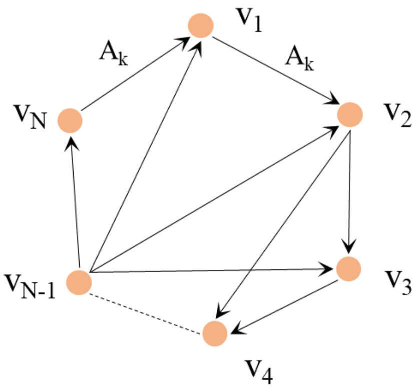

Fig. 4. Illustration of the graph encoder. Each orange dot represents an element in the sequence X, while the value of A_k on each line reflects the presence of structural details between two points.

Specifically, the graph encoder establishes a graph representation $G_{(t)} = (\mathcal{V}_t, \mathbf{A}_k, \mathbf{A}_k)$ for each sequence $X \in \mathbb{R}^{N \times T}$, where $\mathcal{V}_t$ is the set of vertices. Each element in the sequence is treated as a graph vertex, with its value representing the signal residing on the vertex, i.e., the identifier of vertex X at that location. The graph encoder maps the sequence X into the graph domain and examines the structural details among the elements. Assuming there exist structural details between the i_{th} element indexed by vertex v_i and the j_{th} element indexed by vertex v_j, $\mathbf{A}_k(i, j)$ is set to 1; otherwise, $\mathbf{A}_k(i, j) = 0$.

C. Speech Reconstruction

At the receiving end, the speech reconstruction system extracts the hidden speaker embedding from the transmitted converted speech, which is achieved through the decoder module. It is worth noting that the structure of the decoder module is identical to the encoder module used in speech steganography.

Besides, the architecture of the reconstruction module is essentially identical to that of the pipeline in the first part of CSSS. The key distinction lies in the reconstruction module's omission of GAN-generated artificial speaker embeddings.

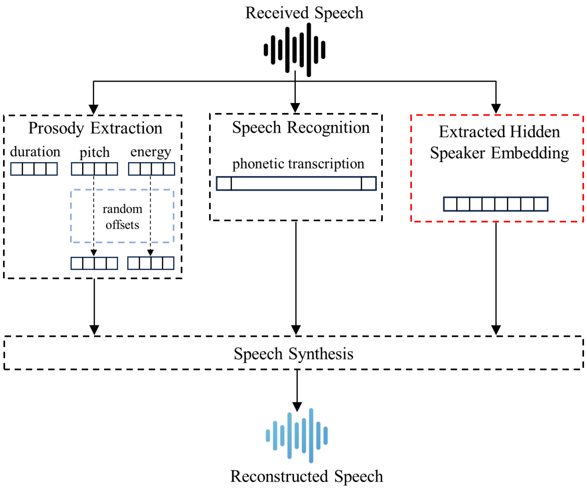

Fig. 5. Illustration of the speech reconstruction module. The part highlighted in red box indicates the difference from Fig. 2.

Instead, it utilizes the speaker embedding extracted by the decoder module to synthesize the original speech by incorporating the prosody and textual information of the converted speech obtained through transmission. The structure is shown in Fig. 5.

III. Experimentals And Results

A. Setup

The training of the pipeline in the first part of CSSS follows the experimental setup proposed in [18]. The ASR model is trained using the VCTK dataset. We utilize SpeechBrain [21] to train the embedding extractor on approximately 2,800 hours of speaker-validated data from VoxCeleb 1 and 2 [22]–[24]. To synthesize speech, we train the TTS system using the LibriTTS corpus [25]. For the GAN model, we augment the LibriTTS data with the RAVDESS [26] and Emotion Speech Dataset (ESD) [27] to obtain a larger and more diverse corpus for training purposes. For the second part of CSSS, we use VoxCeleb2 to train our speech steganography network. For the third part of CSSS, the speech reconstruction module utilizes the pre-trained model from the non-target voice conversion pipeline.

We validate the performance of CSSS using a subset of LibriSpeech and VCTK as a test set and divided into trial and enrollment data. The privacy of the transmitted speech used for speech communication in CSSS is evaluated using the Automatic Speaker Verification (ASV) attacker of Voice Privacy Challenge 2020 (VPC'20) and measured in terms of Equal Error Rate (EER). The utility of CSSS is measured using the Word Error Rate (WER) of the Voice Privacy Challenge 2022 (VPC'22) ASR model. We compared the CSSS to the original speech and to the main baselines of VPC'20 (BL 1) and VPC'22 (BL 1.b).

B. Evaluation metrics

1) Objective Evaluation: Table 1 lists the EER results for the privacy metric in the first column. For the first four items, the closer the EER score is to 50%, the more difficult it is for attackers to identify the true speaker of the audio. Compared with the original data and the two baseline systems of VPC, the steganographic converted speech used for speech communication in CSSS achieved high privacy standards for both datasets. As for the fifth item, the privacy metric of the reconstructed speech is close to that of the original data, indicating that CSSS can indeed recover the speaker's identity well.

TABLE I
Objective evaluation results compared to baseline and proposed systems. In the first four items, the closer the EER is to 50%, the better, and the lower the WER, the better. The best score in each column is highlighted in bold. For the fifth item, it is compared separately to the original, and the lower the EER and WER, the better, indicating the proximity of the reconstructed speech by CSSS to the original speech.

model	EER(%)		WER(%)	
	Libri	VCTK	Libri	VCTK
Original	4.39	3.19	4.15	12.82
VPC'20 - BL 1	34.44	31.30	6.73	15.23
VPC'22 - BL 1.b	31.80	24.11	6.08	15.60
CSSS' Steg	**48.92**	**51.74**	**5.93**	**11.07**
CSSS' Recon	5.66	3.70	4.21	11.47

The second column of Table 1 shows that in the first four items, steganographic converted speech still outperforms both baseline systems in terms of the language content (WER) metric. On the VCTK dataset, the WER is even reduced by 13.65% compared to the original data. In addition, the WER metrics of the fifth reconstructed speech are very close to the original data. This indicates that the CSSS-processed transmitted speech and reconstructed speech still have high utility while achieving privacy preservation and restoring the original speaker identity, respectively.

On the other hand, the accuracy of extracting the hidden speaker embedding from the receiving end is very important for the whole system. We use cosine similarity to measure the similarity between the speaker embedding extracted by the decoder module and the original speaker embedding. The results in Table 2 show that the average cosine similarity over the entire test set reaches 0.9999 in the system introducing the graph encoder, with almost no loss of speaker embedding information. However, the system with the introduction of the graph encoder has a reduced SNR metric compared to the system without it, but as the audio quality is still high, and availability of accurate original speaker embedding at the receiver is more critical for reconstructing the original speech. Therefore, in order to achieve almost lossless transmission of speaker embedding, the introduction of graph encoders is necessary for CSSS.

2) Subjective Evaluation: In our task, human perception of the original speech and the final reconstructed speech at the

TABLE II
RESULTS ARE DIVIDED INTO TWO CATEGORIES: CSSS WITH AND WITHOUT GRAPH ENCODER. WE MEASURED THE AVERAGE SNR OF THE STEGANOGRAPHIC SPEECH OBTAINED BY DIFFERENT SYSTEMS ON THE ENTIRE TEST SET, AS WELL AS THE AVERAGE COSINE SIMILARITY BETWEEN THE SPEAKER EMBEDDINGS EXTRACTED BY THE RECEIVING END AND THE ORIGINAL SPEAKER EMBEDDINGS.

System	SNR of Speech	Cos Similarity
with Graph Encoder	38.514	**0.9999**
w/o Graph Encoder	**55.557**	0.9948

receiving end is crucial. This is because if the reconstructed speech at the receiving end diverges significantly from the original speech, users will doubt the authenticity of the transmitted speech information. Therefore, we randomly selected 20 utterances from 10 different speakers from the test set. Each utterance was presented along with the corresponding output from different systems, to evaluate the perceptual similarity, naturalness, and intelligibility of the speech obtained by different systems compared to the original speech. Then, at least 7 participants rated each metric on a 1-5 Likert scale.

TABLE III
MOS MEASURES THE NATURALNESS, INTELLIGIBILITY, AND PERCEPTUAL SIMILARITY BETWEEN THE ORIGINAL UTTERANCE AND THE OUTPUT UTTERANCE OF DIFFERENT SYSTEMS ON A SCALE OF 1-5. THE OUTPUTS OF THE SECOND THROUGH FOURTH SYSTEMS ARE USED FOR PRIVACY PROTECTION, AND THE SIMILARITY METRICS SHOULD BE AS LOW AS POSSIBLE AND NOT COMPARABLE TO THE RECONSTRUCTED SPEECH, SO THE THIRD COLUMN DOES NOT HAVE THE DATA HIGHLIGHTED IN BOLD.

Model	naturalness		intelligibility		similarity	
	MOS	σ	MOS	σ	MOS	σ
original	*4.64*	±0.75	*4.58*	±0.43	-	-
VPC'20 - BL 1	4.17	±0.67	4.29	±0.79	1.47	±0.32
VPC'22 - BL 1.b	3.71	±0.87	4.40	±0.56	1.55	±0.44
CSSS' Steg	4.32	±0.67	**4.67**	±0.72	1.43	±0.41
CSSS' Recon	**4.48**	±0.56	4.65	±0.54	4.23	±0.61

Based on the Mean Opinion Score (MOS) and its standard deviation σ shown in Table 3, it can be seen that compared to other systems, the steganographic and reconstructed speech from CSSS both have high intelligibility and naturalness, and the reconstructed speech also achieves high perceptual similarity with the original speech. This further confirms the effectiveness of CSSS in restoring the original speech.

IV. CONCLUSION

In this paper, we present a system called CSSS (non-target Conversion based Speech Steganography System), designed to prevent the leakage of speaker identity information during speech communication. The CSSS system is capable of effectively concealing the speaker's identity and embedding the original speaker's embedding information into converted speech for transmission. Additionally, by extracting the hidden speaker embedding information at the receiving end, CSSS can successfully reconstruct the original speech. With these designs, the CSSS not only avoids the leakage of speaker identity information during speech communication but also prevents suspicions regarding the authenticity of the speech information due to a mismatch between the converted speech and the known sender's voice. Through experimental validation and manual evaluation, it is confirmed that the proposed CSSS has an excellent performance in speaker embedding steganography, and the speech reconstructed by the system has high intelligibility, naturalness, and perceptual similarity with the original speech.

REFERENCES

[1] Z. Wu, J. Yamagishi, T. Kinnunen, *et al.*, "Asvspoof: The automatic speaker verification spoofing and countermeasures challenge," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 588–604, 2017.

[2] P. Regulation, "Regulation (eu) 2016/679 of the european parliament and of the council," *Regulation (eu)*, vol. 679, p. 2016, 2016.

[3] A. Nautsch, C. Jasserand, E. Kindt, M. Todisco, I. Trancoso, and N. Evans, "The gdpr & speech data: Reflections of legal and technology communities, first steps towards a common understanding," *arXiv preprint arXiv:1907.03458*, 2019.

[4] H. Lu, X. Wu, H. Guo, S. Liu, Z. Wu, and H. Meng, "Unifying one-shot voice conversion and cloning with disentangled speech representations," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 11 141–11 145. DOI: 10.1109/ICASSP48485.2024.10446296.

[5] J. Lim and K. Kim, "Wav2vec-vc: Voice conversion via hidden representations of wav2vec 2.0," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 10 326–10 330. DOI: 10.1109/ICASSP48485.2024.10447984.

[6] Z. Łatka, J. Gałka, and B. Ziółko, "Cross-gender voice conversion with constant f0-ratio and average background conversion model," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6825–6829. DOI: 10.1109/ICASSP.2019.8683369.

[7] I.-C. Yoo, K. Lee, S. Leem, H. Oh, B. Ko, and D. Yook, "Speaker anonymization for personal information protection using voice conversion techniques," *IEEE Access*, vol. 8, pp. 198 637–198 645, 2020. DOI: 10.1109/ACCESS.2020.3035416.

[8] M. Matassoni, S. Fong, and A. Brutti, "Speaker anonymization: Disentangling speaker features from pre-trained speech embeddings for voice conversion," *Applied Sciences*, vol. 14, no. 9, 2024, ISSN: 2076-3417. DOI: 10.3390/app14093876. [Online]. Available: https://www.mdpi.com/2076-3417/14/9/3876.

[9] M. Zamani, A. A. Manaf, R. Ahmad, A. M. Zeki, and S. Abdullah, "A genetic-algorithm-based approach for audio steganography," *World Academy of Science Engineering & Technology*, no. 54, p. 360, 2009.

[10] D. Gruhl, "Echo hiding, proceedings of information hiding," in *International Workshop*, 1996.

[11] R. M. Nugraha, "Implementation of direct sequence spread spectrum steganography on audio data," in *International Conference on Electrical Engineering & Informatics*, 2011.

[12] S. Agarwal, R. Soun, R. Shivani, V. Varanasi, N. Gill, and R. Sawhney, "Hypersteg: Hyperbolic learning for deep steganography," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[14] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *International conference on machine learning*, PMLR, 2019, pp. 7354–7363.

[15] T. Wang, Z. Pan, M. Ge, Z. Yang, and H. Li, "Time-domain speech separation networks with graph encoding auxiliary," *IEEE Signal Processing Letters*, vol. 30, pp. 110–114, 2023.

[16] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016.

[17] N. Tomashenko, X. Wang, E. Vincent, *et al.*, "The voiceprivacy 2020 challenge: Results and findings," *Computer Speech & Language*, vol. 74, p. 101 362, 2022.

[18] 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 *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[19] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, Springer, 2015, pp. 234–241.

[20] J. Xu, Z. Li, B. Du, M. Zhang, and J. Liu, "Reluplex made more practical: Leaky relu," in *2020 IEEE Symposium on Computers and communications (ISCC)*, IEEE, 2020, pp. 1–7.

[21] M. Ravanelli, T. Parcollet, P. Plantinga, *et al.*, "Speechbrain: A general-purpose speech toolkit," *arXiv preprint arXiv:2106.04624*, 2021.

[22] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, vol. 60, p. 101 027, 2020.

[23] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[24] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[25] H. Zen, V. Dang, R. Clark, *et al.*, "Libritts: A corpus derived from librispeech for text-to-speech," *arXiv preprint arXiv:1904.02882*, 2019.

[26] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, e0196391, 2018.

[27] K. Zhou, B. Sisman, R. Liu, and H. Li, "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 920–924.