System Description for Voice Privacy Challenge 2024

Jeongae Lee[1,3], *Taeje Park*[1,3], *Yeawon You*[2,3]

[1]Sogang University, Seoul, Korea
[2]Ewha Womans University, Seoul, Korea
[3]Team V-beam

jalee3@sogang.ac.kr, xowp1020@sogang.ac.kr, yeawon@ewhain.net

Abstract

This paper details our innovative contributions to the VoicePrivacy Challenge 2024, focusing on the development and comparative evaluation of three distinct voice anonymization methods designed to enhance privacy while preserving the usability of speech data. Method 1 applies a novel signal post-processing technique that modifies the spectral properties of voice recordings to mask identifiable features. Method 2 utilizes an advanced machine learning algorithm to generate synthetic speech that retains the linguistic content but lacks individual-specific characteristics. Method 3 combines elements of acoustic transformation and artificial intelligence to obscure speaker identity effectively. Our experimental findings indicate that each method significantly enhances anonymization while preserving speech recognition accuracy and emotional expressiveness.

Index Terms: speech recognition, speaker anonymization, voice synthesis

1. Introduction

The increasing use of voice data in digital applications poses significant privacy risks, prompting the need for effective anonymization techniques. The VoicePrivacy Challenge has emerged as a pivotal initiative, promoting the development of methods that ensure the anonymity of voice data without compromising its utility for applications such as speech recognition and personal assistants. designed to improve voice data privacy while complying with strict regulatory standards such as the European General Data Protection Regulation (GDPR) and the Canadian Consumer Privacy Protection Act (CPPA).

Our research contributes to the VoicePrivacy Challenge by implementing three innovative methods tailored to voice anonymization. The first method utilizes adaptive signal processing to alter voice data's spectral characteristics, thus masking potential identifiers. The second method involves machine learning algorithms to generate de-identified synthetic speech outputs that retain essential linguistic qualities but lack personal traits. Our third approach integrates elements of both signal processing and artificial intelligence to create a robust anonymization layer that obscures personal identifiers effectively. Our experimental findings indicate that each method has its strong point improving certain metrics with novel approach.

2. Methodology

In this section, we introduce the methodologies employed to explore and evaluate the efficacy of differential privacy techniques in the context of voice anonymization. Our approach encompasses three distinct strategies, each designed to address the unique challenges posed by voice data. The first strategy seeks the most efficient method while reaching high security level. The second strategy focused on security level while keeping reasonable data utility. The third strategy propose method reaching high data utility without severe degradation on data security. Each of these strategies is systematically evaluated through a experiment, designed to measure their performance in terms of privacy guarantees and data utility.

2.1. Method 1

Our first method in algorithm 1 aims to achieve high efficacy while maintaining a high level of security in anonymized speech. Building on the baseline Model 2 (McAdams[1]) for audio anonymization, we enhance security by incorporating natural noise into the anonymized utterances. This addition is designed to mislead and interrupt attackers attempting to verify the speaker using automated speaker verification models.

The integration process consists of several distinct stages: the selection of natural noise, the segmentation of anonymized utterances, and the integration of noise with controlled parameters. In particular, natural noises are sourced from [2] by carefully selecting copyright-free and sufficiently lengthy audio files.

We begin by selecting 10 different natural noise soundwaves that do not contain any human voice. These soundwaves are carefully chosen to ensure they do not introduce any identifiable speech patterns. The selected noises include sounds from environments like trail walking, quiet libraries, and highways. The anonymized utterance A is then segmented into random chunks to increase variability and enhance security. The number of chunks k is randomly determined, ranging from 1 to 10. This randomness in segmentation adds an additional layer of complexity, making it more challenging for attackers to predict or model the noise patterns.

Each chunk of the anonymized utterance is randomly assigned one of the pre-selected natural noise files. The starting point within each natural noise file, which is approximately 15 seconds in length, is also set randomly. This ensures that the noise overlaps sufficiently with the speech chunks, effectively masking them without overpowering the original utterance. To maintain the intelligibility and naturalness of the anonymized speech, the loudness of the natural noise is adjusted to 60% of the original utterance's volume. This careful adjustment ensures that the noise serves its purpose of enhancing security without excessively disrupting the clarity of the speech.

The adjusted natural noise is overlaid onto each chunk of the anonymized utterance as shown in Figure 1. After the noise integration, the chunks are combined to form the final enhanced anonymized utterance A'. This process ensures that the resulting audio retains its linguistic integrity while being effectively

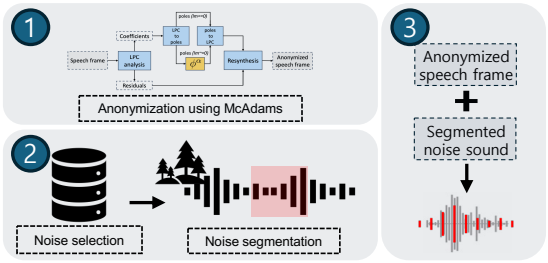

Figure 1: *Diagram for method 1*

Algorithm 1 Enhanced Security through Natural Noise Integration

Input:

- Anonymized utterance A
- A set of natural noise soundwaves $N = \{n_1, n_2, \ldots, n_{10}\}$
- Range of chunks k from 1 to 10
- Duration of noise segment: 15 seconds
- Loudness adjustment factor: 60% of the original utterance volume

Output: Enhanced anonymized utterance A'

Randomly determine the number of chunks k where $1 \leq k \leq 10$
Split the anonymized utterance A into k random chunks: $A = \{a_1, a_2, \ldots, a_k\}$
for $i = 1$ to k **do**
    Randomly select a natural noise file n_i from the set N
    Randomly select a starting point within the noise file n_i of approximately 15 seconds
    Adjust the loudness of the noise file n_i to 60% of the original utterance volume
    Overlay the adjusted noise n_i onto the chunk a_i
end for
Combine the chunks $\{a_1, a_2, \ldots, a_k\}$ to form the enhanced anonymized utterance A'
return A'

anonymized. By integrating these steps, our method significantly enhances the security of the anonymized speech. The use of natural noise, combined with the randomness in segmentation and noise assignment, makes it difficult for automated speaker verification models to accurately identify the speaker while preserving the overall quality and naturalness of the audio.

2.2. Method 2

As illustrated in Figure 2, our proposed anonymization methodology utilizes Speech-to-Text (STT) and Text-to-Speech (TTS) models within a structured three-stage process: feature extraction, feature integration, and speech synthesis. Initially, we extract fundamental frequency (F0) and textual data from the speech waveform. To augment the emotional expressiveness of the anonymized speech, emotion class prototype features, derived from labeled emotional states and corresponding F0 profiles, are integrated with the F0 features using an attention-based mechanism [3]. This integration enriches the F0 features with nuanced emotional content. Further refinement of this integration process is achieved through a Feature-wise Linear Modulation (FiLM) layer [4], which merges the emotionally enriched

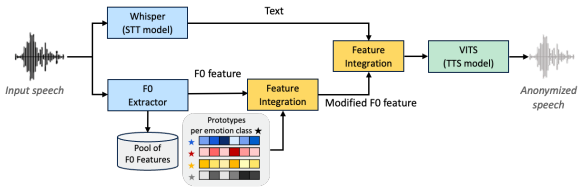

Figure 2: *Diagram for method 2*

F0 features with the textual data into a cohesive feature set. The final stage involves synthesizing the speech waveform from these integrated features using the TTS model. Detailed discussions of each stage are provided in the following subsections.

2.2.1. Feature Extraction

Fundamental frequency (F0), or pitch, is pivotal in speech processing, reflecting the vibration rate of the vocal folds [5]. We employ the Librosa library for its robust pitch analysis capabilities, essential for preserving the emotional depth of speech. For text extraction, the 'Whisper' STT model [6] is utilized for text extraction, known for its efficacy in varied acoustic settings, ensuring accurate capture and processing of linguistic content.

2.2.2. Integrating Emotion Class Prototypes with F0 Features

This section details the integration of emotion class prototype features into the F0 features extracted from the input speech. The "emotion class prototype" was derived by clustering based on emotion labels, where the average centroid for each emotion label was computed to define the prototypes. Employing an attention-based mechanism, this process ensures that the emotional nuances are effectively captured and represented in the synthesized speech. By aligning the F0 features with these emotion class prototypes, the emotional expressiveness of the output is significantly enhanced without compromising privacy.

The process begins with the transformation of both the class prototype and F0 features to a common feature space to facilitate effective integration. This transformation is performed using neural network layers with ReLU activations to introduce non-linearity, enhancing the model's ability to capture complex patterns:

$$\mathbf{p} = ReLU(W_p \cdot Prototype) \tag{1}$$

$$\mathbf{f} = ReLU(W_f \cdot F0) \tag{2}$$

where W_p and W_f are the weight matrices of the transformation layers for the prototypes and F0 features, respectively.

Upon transforming these features, we compute the similarity between the transformed prototype and F0 features using a dot product, which serves as the basis for our attention mechanism. The softmax function is then applied to these similarity scores to derive attention weights, ensuring a normalized distribution of these weights:

$$Similarity = \mathbf{p}^\top \mathbf{f} \tag{3}$$

$$AttentionWeights = Softmax(Similarity) \tag{4}$$

These attention weights are subsequently used to modulate the transformed F0 features ($\mathbf{f}$), allowing the emotion class prototype features to effectively guide the emotional content of the synthesized speech. The final integration is achieved through a weighted summation of the features:

$$F_{integrated} = AttentionWeights \times \mathbf{f} \tag{5}$$

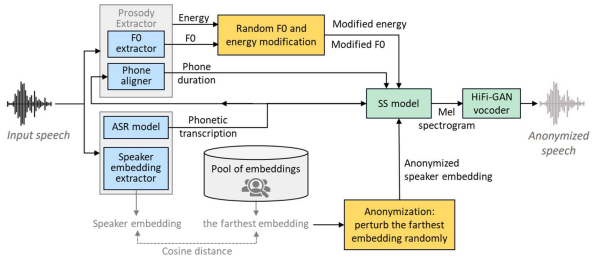

Figure 3: *Diagram for method 3*

This methodology ensures that the feature not only retains its linguistic integrity but also accurately reflects the intended emotional cues, achieving a balance between anonymity and expressiveness.

2.2.3. Feature Integration of Text and F0 Features Using FiLM Layer

Following the initial extraction and emotional enrichment of F0 features, the next stage involves their integration with text data using a Feature-wise Linear Modulation (FiLM) layer.

The FiLM layer modulates the features by applying element-wise affine transformations, specifically calculating the scaling factor γ derived from the enriched F0 features as follows:

$$\gamma_n = 1 + Linear(Normalize(F_{integrated})) \tag{6}$$

where Normalize standardizes the enriched F0 features and Linear adjusts these features to the appropriate dimensions. The scaling factor is then applied to the text embeddings feature map T_n using:

$$FiLM(T_n|\gamma_n) = \gamma_n \odot T_n \tag{7}$$

where $\odot$ denotes the Hadamard (element-wise) product, $T_n$is the feature map from the text embeddings, and γ represents the scaling factors applied element-wise to T_n. This modulation adjusts the activation levels within the feature map, effectively aligning them with the prosodic cues provided by the enriched F0 features.

2.3. Method 3

Figure 3 illustrates the entire anonymization pipeline for Method 3. Our appproach is an enhanced version of the anonymization module from the baseline model 3. As described in [7], the overall system is divided into four main processes: (1) Feature extraction, (2) Pitch and energy modification, (3) Speaker embedding anonymization, and (4) Speech synthesis. While baseline model 3 employs artificial embeddings generated by GAN, Our proposed model randomizes embeddings using representative embedding sets generated through k-means clustering from an external pool. This highlights the primary distinction between baseline model 3 and our proposed method.

During the feature extraction process, various features such as phonetic transcription, fundamental frequency (F0), energy, phone duration, and speaker embedding $\mathbf{x}$ are extracted from the input speech. For the speaker embedding, we use style embeddings with a dimension of 256, which was introduced in [8].

In the pitch and energy modification process, pitch and energy are modified using randomly generated values.

In the speaker embedding anonymization process, the speaker embedding of the input speech is anonymized using a pool of embeddings. In order to create a pool, k-means clustering is applied to the embedding vectors extracted from a pool of speech data. For k-means clustering, cosine distance is used as the distance metric, where cosine distance between two vectors $\mathbf{u}$ and $\mathbf{v}$ are defined by

$$d(\mathbf{u}, \mathbf{v}) = \mathbf{1} - \frac{\mathbf{u} \cdot \mathbf{v}}{|\mathbf{u}||\mathbf{v}|}. \tag{8}$$

From the pool of embeddings, the one with the largest cosine distance $\mathbf{x}_{\text{perturb}}$ from the speaker embedding of the input speech $\mathbf{x}$ is chosen. Also, $N_{\text{candidate}}$ embeddings with the largest cosine distance from $\mathbf{x}$ are selected. Among these $N_{\text{candidate}}$ embeddings, N_{anon} embeddings are selected randomly for perturbation. $\mathbf{x}_{\text{perturb}}$ is randomized according to a randomize probability p, with entries replaced by those from the N_{anon} embeddings. For our experiments, we created the pool of embeddings using the Librispeech train-clean-360 dataset.

Finally, in the speech synthesis process, the anonymized embedding is synthesized with the modified energy, modified F0, and phone duration to create a mel spectrogram. It is then fed into a HiFi-GAN vocoder to generate the anonymized speech.

Algorithm 2 Pool of embeddings generation

Input:
- A pool of speech data $S = \{s_1, s_2, \ldots, s_{N_{\text{total}}}\}$
- The number of vectors for a pool of embedding: N_{pool}

Output: A pool of embeddings $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{N_{\text{pool}}}\}$

Initialize: $U = \emptyset$
for $i = 1 : N_{\text{total}}$ **do**
    Extract the embedding $\mathbf{u}_i$ from the speech data s_i
    Add $\mathbf{u}_i$ to the set U
end for
Apply k-means clustering to U with the cosine distance metric to obtain N_{pool} centroids: $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{N_{\text{pool}}}$
Return $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{N_{\text{pool}}}\}$

Algorithm 3 Embedding perturbation

Input:
- An embedding vector $\mathbf{x} \in \mathbb{R}^{N_{\text{embed}}}$
- A pool of embeddings $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{N_{\text{pool}}}\}$
- randomize probability $p \in [0, 1]$

Output: A perturbed embedding $\mathbf{x}_{\text{perturb}}$

Initialize: $U = \emptyset$
for $i = 1$ to N_{pool} **do**
    Calculate the distance d_i between $\mathbf{x}$ and $\mathbf{v}_i$ using (8)
end for
Find the index $i_{\max} = \underset{i \in \{1,2,\ldots,N_{\text{pool}}\}}{\arg\max} \; d_i$
Let $\mathbf{x}_{\text{perturb}} = \mathbf{v}_{i_{\max}}$
for $i = 1$ to N_{embed} **do**
    Generate $u \sim \mathcal{U}(0, 1)$
    if $u \leq p$ **then**
        Generate $j \sim \text{Uniform}\{1, 2, \ldots, N_{\text{pool}}\}$
        $\mathbf{x}_{\text{perturb},i} = \mathbf{v}_{j,i}$
    end if
end for
Return $\mathbf{x}_{\text{perturb}}$

Parameter	Value
N_{total}	1000
N_{pool}	500
$N_{\text{candidate}}$	200
N_{anon}	100
p	0.2

Table 1: *Parameters for Method 3*

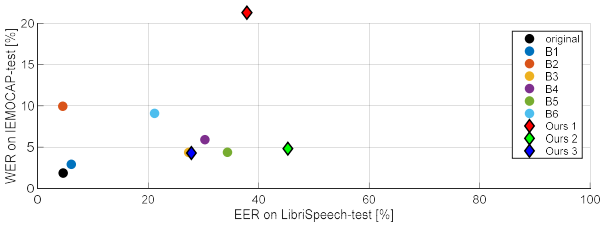

Figure 4: *EER-WER on LibriSpeech-test data*

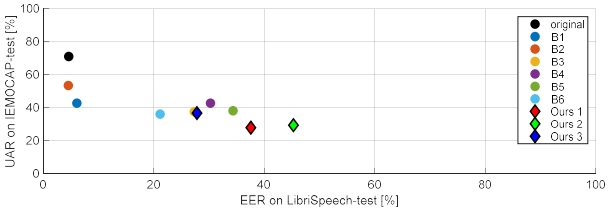

Figure 5: *EER-UAR on IEMOCAP-test data*

3. Experiments

3.1. Datasets

For our experiments, we utilized two primary datasets: LibriSpeech and IEMOCAP. LibriSpeech, which consists of 960 hours of read English speech derived from audiobooks, was used for Automatic Speaker Verification (ASV) and Automatic Speech Recognition (ASR) evaluations. The IEMOCAP dataset, an emotional audio-visual corpus, was used for Speech Emotion Recognition (SER) evaluation. To accommodate the small number of speakers and data, we adopted a leave-one-conversation-out cross-validation strategy for SER, where eight sessions are used for training and the remaining sessions for development and evaluation. Performance was assessed using the evaluation methodologies provided by VoicePrivacy for both LibriSpeech and IEMOCAP's evaluation sets. In the case of Method 2, as per the workshop guidelines, we utilized authorized pretrained STT and TTS models, thus bypassing a separate training process.

3.2. Experimental Setups

Our experimental framework is based on baselines from the VoicePrivacy 2024 Challenge. The setup includes two NVIDIA GeForce RTX 3090 GPUs for handling computationally intensive tasks. Key software components include Python 3.11, CUDA toolkit 11.7 for GPU acceleration, and PyTorch 2.0.1 for neural network operations. Software dependencies are meticulously managed to ensure consistent interactions and updates, facilitated by a version control system with automatic triggers.

4. Results

Our study uses three metrics for evaluation[9]: the Equal Error Rate (EER) for privacy, measuring speaker similarity using cosine distances in an automatic speaker verification system; the Word Error Rate (WER) for linguistic accuracy, gauging transcription discrepancies in anonymized speech; and the Unweighted Average Recall (UAR) for emotional expressiveness, assessing emotion recognition accuracy in speech emotion recognition systems.

Our results, as detailed in Table 2, Table 3, and Table 4, and visually represented in Figure 4 and Figure 5, reveal a nuanced balance between enhancing speaker anonymity and maintaining utility in speech output. For method 3, the specific parameter values are shown in Table 1. Table 2 showcases the EER for our methods (Ours 1, Ours 2, and Ours 3), demonstrating a substantial increase in EER compared to the original data and baselines, indicating enhanced speaker anonymity.

However, detailed examination in Table 3 reveals significant variations among the methods. Method 1, in particular, exhibits a Word Error Rate (WER) that is substantially higher than both the original recordings and any of the baseline models. This severe degradation in speech intelligibility for Method 1 contrasts sharply with the results from Method 3, which shows a WER closer to baseline values, suggesting a better balance between anonymity and speech intelligibility. The UAR scores in Table 4, provide insight into the emotional expressiveness of each method. Although Methods 1 and 2 resulted in lower UAR scores compared to baselines, reflecting challenges in emotional conveyance, these findings are instructive for identifying specific enhancements needed in anonymization processes. Method 3 exhibited higher UAR scores, nearing baseline performance, underscoring its capability to retain more emotional nuances, thus validating its potential as a more effective approach to speech anonymization.

Figure 4 and Figure 5 graphically illustrate these trade-offs, where the shift towards higher EER and lower WER and UAR across our methods visually underscores the impact of our anonymization techniques. These figures effectively highlight the relationship between increased privacy protection and its consequential effects on speech utility.

Despite these findings, it is essential to consider the trade-offs inherent in anonymization technologies. Our approach significantly obscures speaker identity, a critical aspect of privacy protection. However, this enhancement in privacy comes at the cost of reduced utility metrics. We posit that the elevated EER values, reflecting strong anonymization performance, are a desirable outcome in scenarios where privacy is paramount. The trade-offs observed highlight the complex interplay between achieving robust privacy protection and retaining high utility in anonymized speech. We believe our methods achieve a commendable balance, providing substantial privacy gains while maintaining acceptable levels of speech intelligibility and emotional expressiveness, as evidenced by our systematic evaluations.

This balance underscores the effectiveness of our innovative approach in settings where both privacy and utility are critical considerations, positioning our methods as a viable solution for applications demanding high degrees of speaker anonymization.

Dataset	Gender	Orig.	Ours 1	Ours 2	Ours 3	B1	B2	B3	B4	B5	B6
LibriSpeech-dev	female	10.51	40.03	44.32	28.81	10.94	12.91	28.43	34.37	35.82	25.14
	male	0.93	35.22	47.52	22.18	7.45	2.05	22.04	31.06	32.92	20.96
	Average	5.72	37.63	45.92	25.50	9.20	7.48	25.24	32.71	34.37	23.05
LibriSpeech-test	female	8.76	39.78	45.05	28.46	7.47	7.48	27.92	29.37	33.95	21.15
	male	0.42	35.36	44.77	27.16	4.68	1.56	26.72	31.16	34.73	21.14
	Average	4.59	37.57	44.91	27.81	6.07	4.52	27.32	30.26	34.34	21.14

Table 2: *EER (%) achieved on data anonymized by ours and baselines vs. original (Orig.) data*

Dataset	Orig.	Ours 1	Ours 2	Ours 3	B1	B2	B3	B4	B5	B6
LibriSpeech-dev	1.80	98.15	5.02	4.14	3.07	10.44	4.29	6.15	4.73	9.69
LibriSpeech-test	1.85	97.88	4.78	4.28	2.91	9.95	4.35	5.90	4.37	9.09

Table 3: *WER (%) achieved on data processed by ours and baselines vs. original (Orig.) data*

Dataset	Orig.	Ours 1	Ours 2	Ours 3	B1	B2	B3	B4	B5	B6
IEMOCAP-dev	69.08	28.02	28.45	38.09	42.71	55.61	38.09	41.97	38.08	36.39
IEMOCAP-test	71.06	27.93	29.44	37.57	42.78	53.49	37.57	42.78	38.17	36.13

Table 4: *UAR (%) achieved on data processed by ours and baselines vs. original (Orig.) data*

5. Discussion

The findings of this report underscore the potential and challenges of applying differential privacy techniques to voice anonymization. Our experimental results demonstrate that the proposed methods, particularly the ones incorporating natural noise and emotion class prototype features, significantly enhance privacy protection, as indicated by the higher Equal Error Rate scores. These methods effectively obscure the speaker's identity, making it more difficult for adversarial attacks to re-identify individuals based on their voice data. This is crucial in an era where voice-activated technologies are becoming ubiquitous, posing increased risks of privacy breaches.

However, the trade-offs between privacy and data utility remain a critical concern. While our methods achieved higher privacy metrics, they also exhibited a noticeable decline in utility, particularly in tasks requiring emotional content recognition. The decrease in Unweighted Average Recall indicates that the anonymized data loses some of its richness, which is vital for applications like emotion detection and nuanced speech analysis. This degradation can impact the effectiveness of voice-based systems in real-world applications, where maintaining both privacy and high data utility is essential.

Looking forward, it is evident that further research is needed to strike a better balance between privacy and utility. Enhancing the sophistication of differential privacy algorithms to retain more of the original data's characteristics without compromising privacy is a promising direction. Additionally, exploring adaptive noise addition techniques that can dynamically adjust based on the context and sensitivity of the data might offer a viable solution. As privacy regulations continue to evolve, developing robust, privacy-preserving techniques that comply with legal standards while meeting the practical needs of data utility will be paramount. The insights from this report lay a foundation for such advancements, contributing to the broader goal of secure and effective data anonymization in the digital age.

6. Conclusion

We proposed three anonymization methods for voice, one of unstructured data, and evaluated their effectiveness in terms of both privacy and utility. Compared to the original voice and a baseline model, our methods exhibited higher EER performance in terms of privacy. However, they showed performance degradation in utility, particularly in emotion recognition task. Considering these aspects, as future work, we plan to research and investigate factors capable of capturing emotional content in voices. We will also explore methods to enhance utility for various contents beyond emotion. Furthermore, we aim to explore approaches for simplifying or lightweighting models suitable for practical system applications.

7. References

[1] J. Patino, N. Tomashenko, M. Todisco, A. Nautsch, and N. Evans, "Speaker anonymisation using the mcadams coefficient," *arXiv preprint arXiv:2011.01130*, 2020.

[2] F. Font, G. Roma, and X. Serra, "Freesound technical demo," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 411–412.

[3] C.-W. Kuo, C.-Y. Ma, J.-B. Huang, and Z. Kira, "Featmatch: Feature-based augmentation for semi-supervised learning," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*. Springer, 2020, pp. 479–495.

[4] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[5] B. Gold, N. Morgan, and D. Ellis, *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*, 1st ed. United States: John Wiley Sons Inc., 1999.

[6] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022.

[7] 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)*, 2023, pp. 1–5.

[8] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International conference on machine learning (ICML)*. PMLR, 2018, pp. 5180–5189.

[9] N. Tomashenko, X. Miao, P. Champion, S. Meyer, X. Wang, E. Vincent, M. Panariello, N. Evans, J. Yamagishi, and M. Todisco, "The voiceprivacy 2024 challenge evaluation plan," *arXiv preprint arXiv:2404.02677*, 2024.