Emotion-Enhanced Speaker Anonymisation Using the FreeVC Framework

Yuqi Li, Yuanzhong Zheng, Jingyi Fang, Jinming Chen

Abstract

With the rapid development of artificial intelligence voice communication, people are paying more attention to the protection of voice privacy. Maintaining the intelligibility and naturalness of speech while effectively anonymizing it is a major challenge. Excessive modifications to the voice may make the speech difficult to understand or sound unnatural, which reduces the usability of the anonymized speech. Inspired by voice conversion technology which can be used for voice anonymization. For instance, altering a speaker's voice features to another unrecognizable voice can serve the purpose of anonymization.We proposed EESA, based on the FreeVC framework, which is a high-quality, one-shot voice conversion scheme. EESA enhances the FreeVC framework by adding emotional feature extraction and the selection of targets for voice anonymization. We demonstrate the potential of EESA by applying the evaluation framework of the Voice Privacy Challenge 2024.The evaluation results exceed several current baselines.

Index Terms: voice conversion, voice anonymization, FreeVC

1. Introduction

In recent years, the field of voice anonymization has emerged as a critical area of research, driven by growing concerns over privacy protection and data security in digital communications. Voice anonymization seeks to modify a speaker's vocal attributes to prevent the identification of the speaker while maintaining the intelligibility and naturalness of the speech.

Voice anonymization technology plays a vital role in a myriad of applications, from ensuring individual privacy in public databases to protecting sources in journalistic pursuits. One of the significant applications is in mobile health data management, where privacy-preserving voice-based search schemes enhance the efficiency and privacy of in-home healthcare systems. For example, the research by Hadian et al. [1] proposes a system where encrypted voice data can be securely stored and queried by healthcare providers without compromising patient privacy. This system uses homomorphic encryption to allow searches on encrypted voice data, ensuring that sensitive health information remains secure against any unauthorized access.

Additionally, voice anonymization is crucial in environments where speech data can be easily exploited for surveillance or unauthorized data mining. The study by Tomashenko et al. [2] addresses concerns in speech translation systems where voice anonymization can prevent the identification of speakers while maintaining the intelligibility and quality of translated speech. This is particularly relevant in sensitive communication scenarios, such as cross-border journalism or multinational business meetings, where the identity of the speaker must be protected.

Moreover, voice anonymization technologies find applications in enhancing the security of voice-enabled IoT devices, particularly those used in smart homes and other consumer-focused technologies. As explored by Nautsch et al. [3] , these technologies help safeguard users from potential eavesdropping and personal data exploitation by anonymizing voice inputs before they are processed or stored. This approach not only protects the privacy of users but also secures the voice data from potential threats in digital environments where data breaches and cyber-attacks are increasingly common.

One of the foundational approaches in this field is outlined in the study "Speaker Anonymization Using X-vector and Neural Waveform Models," [4] which introduces a method that leverages X-vector and neural waveform models to anonymize speech. This approach underscores the possibility of altering vocal features effectively, thus shielding the speaker's identity without significantly compromising the quality of speech communication.

Building on community-driven research efforts, the "VoicePrivacy 2020 Challenge" [5] has significantly propelled the field forward. This initiative has fostered the development of innovative voice anonymization techniques designed to counteract the advanced capabilities of automatic speaker recognition systems, marking a pivotal step in collaborative scientific inquiry.

Further explorations using Artificial Neural Networks (ANNs) for speaker anonymization have demonstrated that deep learning techniques can be adeptly applied to modify audio data, thereby preventing speaker identification by automatic systems. Such as the use of Wasserstein GANs to create artificial speaker embeddings that are effectively untraceable to any real speaker data, enhancing voice privacy significantly. [6] Another approach includes manipulating the fundamental frequency (F0) using deep neural networks, which preserves speech naturalness and intelligibility while anonymizing the speaker. [7] Additionally, neural audio codec language models have been employed to achieve high levels of voice privacy, proving the practicality of these models in real-world applications. [8] These findings not only enhance the technological toolkit available for voice privacy but also validate the efficacy of ANNs in real-world applications.

However, the application of voice anonymization technologies introduces new challenges, particularly concerning the balance between privacy and speech intelligibility. "The Effect of Voice Anonymization on Privacy and Speech Intelligibility" explores these trade-offs, emphasizing the critical need to maintain clear communication while achieving robust anonymization.

Lastly, "Anonymizing Speech: Evaluating and Designing

Speaker Anonymization Techniques" [6] provides a comprehensive review of current methods, including speech synthesis and voice conversion. This paper discusses the strengths and weaknesses of various techniques, offering insights into their practical implications and setting the stage for future innovations in the field. This thesis identifies key challenges in evaluating privacy protection, examines common voice conversion-based systems, proposes new transformation methods to minimize speaker PII, and introduces a novel attack method to test the robustness of anonymization systems.

The collective insights from these studies underscore the dynamic nature of voice anonymization research and its crucial role in addressing both the technical and ethical challenges associated with privacy in digital communication. As this field evolves, it continues to push the boundaries of what is possible in protecting individual privacy while ensuring that communication remains effective and natural.

The experiment is to develop a voice anonymization system for speech data that conceals the speaker's voice identity while protecting linguistic content and emotional states according to the requirements of "The VoicePrivacy 2024 Challenge Evaluation Plan" [9]

2. Related Work

FreeVC [10] is a text-free one-shot voice conversion (VC) system that leverages the VITS framework for high-quality waveform reconstruction. This approach is significant in the realm of VC as it focuses on disentangling content information without relying on text annotations. Instead, it utilizes self-supervised learning techniques to extract content information, offering a potential solution to the limitations posed by text-based VC systems that require large annotated datasets.

It distinguishes itself by employing an end-to-end training strategy, using a combination of a Conditional Variational Autoencoder (CVAE) augmented with GAN training. The model architecture includes unique components like a bottleneck extractor and a normalizing flow, which help in refining the content information extraction process to enhance the purity and disentanglement of speaker-independent features.

Furthermore, the FreeVC model introduces spectrogram-resize based data augmentation, which manipulates the speaker information without altering the content, thereby strengthening the model's ability to disentangle the two. This method shows promising results in improving the robustness and generalizability of the model, making it effective even with minimal data from target speakers.

As illustrated in the 1 the backbone of EESA is inherited from FreeVC, which is a CVAE augmented with GAN training.Different from FreeVC,EESA module extractes a novel emotion embedding, denoted as e. This embedding, in conjunction with the speaker embedding s extrated from the speaker encoder, is concatenated together to form a composite embedding represented by g.

3. Proposed Method

The overall architecture of our anonymization model is illustrated in 1, where the Posterior Encoder and Decoder follow the VITS [11] architecture. Here, we specifically focus on the bottleneck extractor, emotion encoder, and speaker encoder.

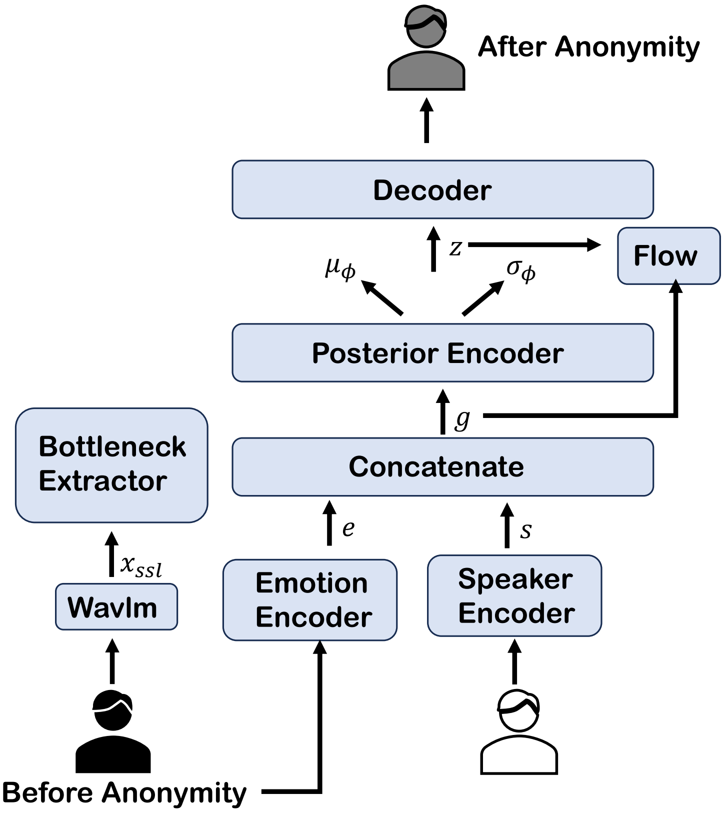

Figure 1: *Anonymization System Diagram: x_{ssl} represents the source speaker's speech content, e represents the source speaker's emotion embedding, and s represents the target speaker's speaker embedding.*

3.1. BottleNeck Extractor

In our proposed model, the BottleNeck Extractor, following the FreeVC [10] framework, effectively compresses the input features into a condensed yet informative representation suitable for subsequent processing stages. This module is specifically designed to reduce the dimensionality of the input features, which initially are 1024-dimensional x_{ssl}, to a more compact size of $2 * d$ dimensions, where d is significantly smaller than 1024. For instance, setting $d = 128$ results in an output of 256 dimensions. This reduction not only streamlines the processing pipeline by reducing computational load but also aids in distilling the most salient features necessary for the task at hand.

The primary purpose of this bottleneck configuration is to retain the essential content information of the voice input while effectively stripping away speaker-specific characteristics that could lead to overfitting or privacy issues. By carefully designing the convolutional layers of the bootleneck extractor, including strategic choices in kernel size, dilation rate, and layer depth, the module captures a broad context of the acoustic input without binding tightly to the idiosyncrasies of individual speaker voices.

To achieve this, the BottleNeck Extractor utilizes a series of convolutional layers with a dilated structure that enhances the model's ability to aggregate information over larger temporal spans of the input data. The dilation allows the network to abstract higher-level features that are more indicative of the linguistic content rather than the speaker's identity. Following the convolutional layers, the projection layer splits the encoded features into two parts: the mean and the log variance. These components represent a parameterized space where the content of the speech is preserved in the mean, and variations potentially related to the speaker characteristics are captured in the log variance. This design enables the subsequent stages of the model to focus on reconstructing or transforming the speech content based on the mean, utilizing the variance part to add necessary variations during generation or transformation processes, thereby maintaining the integrity of the linguistic content while minimizing speaker-dependent features.

3.2. Emotion Encoder

In the emotion encoder section, we adopted two approaches, with the subsequent results section presenting a comparison of their outcomes. The first method involves using the openSMILE [12] software to extract features from audio files in WAV format, supplemented by LSTM for further feature extraction. The second method utilizes the wav2vec 2.0 and large-robust model [13] fine-tuned on emotion dataset MSP-Podcast[14], specifically wav2vec2-large-robust12-ft-emotion-mspdim, to extract audio features.

3.2.1. Based on openSMILE

In our research, we employ openSMILE, a versatile feature extraction tool, to capture a rich set of acoustic properties from audio samples, specifically for the task of emotion recognition. The openSMILE toolkit allows for the extraction of a comprehensive array of features, including Root Mean Square (RMS) energy, Mel-Frequency Cepstral Coefficients (MFCCs), Zero-Crossing Rate (ZCR), voice probability, and fundamental frequency (F0), among others. Each attribute, such as `pcm_RMSenergy_sma_max`, `pcm_RMSenergy_sma_min`, and `pcm_fftMag_mfcc_sma[1-12]_max`, represents various statistical characteristics of these acoustic features, capturing aspects such as maximum, minimum, range, and mean values across short-term windows.

These extracted features are pivotal for understanding the nuanced emotional undertones within human speech, which can vary widely in their acoustic texture and intensity. For instance, RMS energy metrics reflect the audio's loudness dynamics, essential for distinguishing between different levels of excitement or calmness, each associated with specific emotional states. Similarly, MFCCs provide insights into the spectral characteristics that are crucial for identifying the timbral qualities of speech, influencing the emotional tone perceived by listeners. These spectral features capture variations in vocal tract configurations that correspond to different emotions, making them valuable for emotion recognition tasks.

Upon ingesting pre-processed feature vectors, the EmotionEncoder utilizes LSTM layers to analyze temporal sequences and encapsulate emotional states from varying speech patterns and prosodic features. The LSTM captures dynamic changes in the sequence of acoustic feature vectors, crucial for identifying subtle emotional nuances in human speech. The LSTM outputs are then transformed by a linear layer into a lower-dimensional embedding space representing emotional states. A ReLU activation function refines these embeddings, enhancing non-linearity for complex emotion recognition. Normalization to unit length ensures consistent magnitudes across samples, facilitating robust comparison and aggregation of emotional states. Additionally, the EmotionEncoder can handle varying speech input lengths by computing embeddings over partial segments, maintaining accuracy across diverse input conditions. This design effectively captures and encodes the emotional essence from rich acoustic features, showcasing the combination of traditional feature extraction with modern neural architectures for enhanced emotion recognition.

3.2.2. Based on Wav2vec 2.0

In the development of our emotion recognition system, we employed a sophisticated approach wav2vec2-large-robust-12-ft-emotion-msp-dim [15] utilizing the Wav2Vec 2.0 [16] framework, specifically tailored for Dimensional Speech Emotion Recognition. The model, pre-trained on the MSP-Podcast [14] dataset, leverages deep learning to predict emotional dimensions such as arousal, dominance, and valence, ranging approximately from 0 to 1. Prior to fine-tuning, the model was streamlined by pruning from 24 to 12 transformer layers, enhancing processing efficiency without compromising the predictive accuracy.

This adapted model integrates a regression head that projects the extracted features onto a space representing the emotional dimensions. The architecture encompasses a sequence of operations including dropout for regularization, dense layers for feature transformation, and a final projection to the output space. During inference, the model processes raw audio inputs, which are first normalized and converted into a batch of input values by the Wav2Vec2Processor. The emotional content is then decoded, producing not only the predicted emotional states but also the pooled states of the last transformer layer, which are valuable for detailed emotional analysis.

Through this model, we extract emotion embeddings of dimensions [1,1024] from input audio files sampled at 16kHz.

4. Experiments

4.1. Dataset

4.1.1. VCTK

The VCTK [17]corpus is a key speech dataset with recordings from 110 English speakers of various accents, including British and American. Each speaker reads about 400 newspaper sentences, offering a rich phonetic diversity for training and testing speech synthesis and voice conversion models.

4.1.2. RADVESS

The RAVDESS [18] dataset features audio and video recordings from 24 professional actors expressing emotions like happiness, sadness, and anger through speech and song. Each of the 1,248 samples is rated for emotional validity and intensity, making RAVDESS valuable for developing algorithms in speech emotion recognition and cross-modal emotional recognition tasks.

4.1.3. Librispeech

LibriSpeech [19] is a corpus of around 1,000 hours of English speech from LibriVox audiobooks, known for its large scale and diversity in American accents. It includes male and female speakers and is essential for training and evaluating automatic speech recognition (ASR) systems. The dataset is structured into training, development, and test sets, making it a standard resource in speech recognition benchmarks and competitions.

Table 1: *Objective evaluation results over Original data, baselines and proposed EESA on EER, WER and UAR,EESA(Wv) represents an emotion encoder that is based on Wav2vec 2.0, while EESA(oS) represents an emotion encoder based on openSMILE.*

Dataset & Metrics	**Gender**	**Orig.**	**B1**	**B2**	**B3**	**B4**	**B5**	**B6**	**EESA(Wv)**	**EESA(oS)**
LibriSpeech-dev (EER, %) ↑	female	10.51	10.94	12.91	28.43	34.37	35.82	25.14	38.92	38.49
	male	0.93	7.45	2.05	22.04	31.06	32.92	20.96	36.65	38.20
	Average dev	5.72	9.2	7.48	25.24	32.71	34.37	23.05	37.79	38.35
LibriSpeech-test (EER, %) ↑	female	8.76	7.47	7.48	27.92	29.37	33.95	21.15	37.96	38.85
	male	0.42	4.68	1.56	26.72	31.16	34.73	21.14	35.64	35.19
	Average dev	4.59	6.07	4.52	27.32	30.26	34.34	21.14	36.23	37.02
LibriSpeech-dev (WER, %) ↓	-	1.8	3.07	10.44	4.29	6.15	4.73	9.69	4.45	4.79
LibriSpeech-test (WER, %) ↓	-	1.85	2.91	9.95	4.35	5.90	4.37	9.09	4.38	4.34
IEMOCAP-dev (UAR, %) ↑	-	69.08	42.71	55.61	38.09	41.97	38.08	36.39	45.09	43.52
IEMOCAP-test (UAR, %) ↑	-	71.06	42.78	53.49	37.57	42.78	38.17	36.13	42.74	37.37

4.1.4. IEMOCAP

The IEMOCAP [20] database provides approximately 12 hours of audiovisual data from 10 actors in both scripted and improvised scenarios, capturing emotions like happiness, sadness, and anger. With detailed annotations by multiple annotators, IEMOCAP is an essential resource for developing emotion recognition systems in speech synthesis and human-computer interaction.

4.1.5. Conclusion

Our training process is based on the VCTK dataset. Since the VCTK dataset lacks a rich variety of emotions, we fine-tuned our model on the RAVDESS dataset. The trained checkpoints were then used to anonymize the LibriSpeech and IEMOCAP datasets. Finally, we evaluated the performance using the assessment scripts provided by Voice Privacy Challenge 2024.

4.2. Preprocess:Content Extraction

We developed a systematic preprocessing routine to extract content features from audio files using the WavLM [21] model, a state-of-the-art neural network optimized for audio analysis. Initially, we set up the environment to handle the audio files by specifying parameters such as the input and output directories and the desired sampling rate. Subsequently, the WavLM model was initialized on a GPU to leverage accelerated computing capabilities, ensuring efficient processing of large datasets.

The preprocessing involved loading audio files from the specified directory, where each file was resampled to a uniform sampling rate of 16 kHz to maintain consistency across the dataset. This resampling is critical as it standardizes the input for neural processing, which is sensitive to variations in input data format. Each audio file was then converted into a PyTorch tensor and processed through the WavLM model to extract high-dimensional content features that encapsulate the characteristics of the spoken content without retaining speaker-specific information.

These extracted features were saved in a structured directory format, organized by speaker identifiers extracted from the filenames, which facilitates easy access and management of data for subsequent machine learning tasks. This preprocessing pipeline not only enhances the quality of the input data for our models but also streamlines the workflow for handling large-scale audio datasets, making it a crucial step in our research methodology.

This step in the process extracts the linguistic content of the source speaker's voice, without including any identifiable or emotional information.

5. Results

The baseline system is based on the research findings from the following articles [22, 23, 24, 8, 6] ,with results data cited from[25].

For the LibriSpeech dataset, the EESA system notably improves WER, indicating better speech recognition performance, especially in the test set where the WER for male voices has significantly decreased. In terms of privacy, an increased EER suggests the system's effectiveness in enhancing speaker anonymization by making speaker identification more challenging.

Using the IEMOCAP dataset for emotional recognition, the EESA maintains high UAR scores. Although these do not exceed the original scores, they demonstrate the system's ability to accurately recognize and classify emotions without losing emotional content, even with enhanced anonymization.

Overall, the EESA system effectively balances privacy improvements and utility in speech and emotion recognition, highlighting its potential for secure and sensitive voice-driven applications.

6. Conclusions

Our newly developed EESA anonymization system excels in key metrics—Word Error Rate (WER), Equal Error Rate (EER), and Speech Emotion Recognition (SER), surpassing most baseline systems. It effectively maintains privacy and the natural clarity of speech. Unlike typical systems that compromise emotional characteristics for greater anonymization, EESA achieves a balanced approach, enhancing speech privacy without losing emotional expressiveness. This marks a significant advancement in our anonymization techniques.

User feedback from hearing tests indicates that speech processed by the EESA system retains a very natural sound quality. Looking ahead, we plan to expand our research to include a variety of subjects for anonymization, moving beyond single individual targets. This expansion aims to refine the system's effectiveness and broaden its real-world applications.

7. References

[1] M. Hadian, T. Altuwaiyan, X. Liang, and W. Li, "Efficient and privacy-preserving voice-based search over mhealth data," in *2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*. IEEE, 2017, pp. 96–103. [Online]. Available: https://ieeexplore.ieee.org/document/8010720

[2] N. Tomashenko, X. Wang, and E. Vincent, "Voice anonymization in speech translation," *arXiv preprint arXiv:2007.15064*, 2020. [Online]. Available: https://arxiv.org/abs/2007.15064

[3] A. Nautsch *et al.*, "Preserving privacy in speaker and speech characterization," *Computer Speech & Language*, vol. 58, pp. 441–480, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230818303875

[4] F. Fang, X. Wang, J. Yamagishi, I. Echizen, M. Todisco, N. Evans, and J.-F. Bonastre, "Speaker Anonymization Using X-vector and Neural Waveform Models," May 2019, arXiv:1905.13561 [cs, eess, stat].

[5] N. Tomashenko, X. Wang, X. Miao, H. Nourtel, P. Champion, M. Todisco, E. Vincent, N. Evans, J. Yamagishi, and J.-F. Bonastre, "The VoicePrivacy 2022 Challenge evaluation plan," 2022. [Online]. Available: https://www.voiceprivacychallenge.org/docs/VoicePrivacy_2022_Eval_Plan_v1.0.pdf

[6] P. Champion, "Anonymizing Speech: Evaluating and Designing Speaker Anonymization Techniques," Mar. 2024, arXiv:2308.04455 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2308.04455

[7] E. Gaznepoglu and N. Peters, "Deep Learning-based F0 Synthesis for Speaker Anonymization," Jun. 2023, arXiv:2306.16860 [eess].

[8] M. Panariello, F. Nespoli, M. Todisco, and N. Evans, "Speaker anonymization using neural audio codec language models," Jan. 2024, arXiv:2309.14129 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2309.14129

[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," 2024.

[10] J. li, W. tu, and L. xiao, "FreeVC: Towards High-Quality Text-Free One-Shot Voice Conversion," Oct. 2022, arXiv:2210.15418 [cs, eess].

[11] J. Kim, J. Kong, and J. Son, "Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech," Jun. 2021, arXiv:2106.06103 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2106.06103

[12] F. Eyben, M. Wöllmer, and B. Schuller, "opensmile - the munich versatile and fast open-source audio feature extractor," in *Proceedings of the ACM Multimedia (MM)*. Florence, Italy: ACM, 2010, pp. 1459–1462, 25.-29.10.2010.

[13] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems (NeurIPS)*, Vancouver, BC, Canada, 2020, pp. 12 449–12 460.

[14] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, 2017.

[15] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–13, 2023.

[16] W.-N. Hsu, A. Sriram, A. Baevski, T. Likhomanenko, Q. Xu, V. Pratap, J. Kahn, A. Lee, R. Collobert, G. Synnaeve, and M. Auli, "Robust wav2vec 2.0: Analyzing Domain Shift in Self-Supervised Pre-Training," Sep. 2021, arXiv:2104.01027 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2104.01027

[17] C. Veaux, J. Yamagishi, and K. MacDonald, "SUPERSEDED - CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit," *The Rainbow Passage which the speakers read out can be found in the International Dialects of English Archive: (http://web.ku.edu/˜idea/readings/rainbow.htm).*, Apr. 2017, accepted: 2017-04-04T09:21:53Z Publisher: University of Edinburgh. The Centre for Speech Technology Research (CSTR). [Online]. Available: https://datashare.ed.ac.uk/handle/10283/2651

[18] R. F. Livingstone SR, "The ryerson audio-visual database of emotional speech and song (ravdess)," in *A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391*, 2018.

[19] M. Korvas, O. Plátek, O. Dušek, L. Žilka, and F. Jurčíček, "Free English and Czech telephone speech corpus shared under the CC-BY-SA 3.0 license," in *Proceedings of the Eigth International Conference on Language Resources and Evaluation (LREC 2014)*, 2014, p. To Appear.

[20] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, Dec. 2008.

[21] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, and F. Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," 2021.

[22] 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."

[23] J. Patino, N. Tomashenko, M. Todisco, A. Nautsch, and N. Evans, "Speaker anonymisation using the McAdams coefficient," Sep. 2021, arXiv:2011.01130 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2011.01130

[24] 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)*, Jun. 2023, pp. 1–5, iSSN: 2379-190X. [Online]. Available: https://ieeexplore.ieee.org/document/10096607

[25] N. Tomashenko, X. Wang, E. Vincent, J. Patino, B. M. L. Srivastava, P.-G. Noé, A. Nautsch, N. Evans, J. Yamagishi, B. O'Brien, A. Chanclu, J.-F. Bonastre, M. Todisco, and M. Maouche, "The VoicePrivacy 2020 Challenge: Results and findings," *Computer Speech & Language*, vol. 74, p. 101362, Jul. 2022.