# Exploring VQ-VAE with Prosody Parameters for Speaker Anonymization

Sotheara Leang, Anderson Augusma, Eric Castelli, Frédérique Letué, Sethserey Sam, Dominique Vaufreydaz

# EXPLORING VQ-VAE WITH PROSODY PARAMETERS FOR SPEAKER ANONYMIZATION

AUTHOR VERSION

**Sotheara Leang[1,2], Anderson Augusma[1,3], Éric Castelli[1], Frédérique Letué[3], Sethserey Sam[2], Dominique Vaufreydaz[1]**

[1] Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
[2] Institute of Digital Research and Innovation, CADT, Phnom Penh, Cambodia
[3] Univ. Grenoble Alpes, CNRS, LJK, 38000 Grenoble, France

## ABSTRACT

Human speech conveys prosody, linguistic content, and speaker identity. This article investigates a novel speaker anonymization approach using an end-to-end network based on a Vector-Quantized Variational Auto-Encoder (VQ-VAE) to deal with these speech components. This approach is designed to disentangle these components to specifically target and modify the speaker identity while preserving the linguistic and emotionalcontent. To do so, three separate branches compute embeddings for content, prosody, and speaker identity respectively. During synthesis, taking these embeddings, the decoder of the proposed architecture is conditioned on both speaker and prosody information, allowing for capturing more nuanced emotional states and precise adjustments to speaker identification. Findings indicate that this method outperforms most baseline techniques in preserving emotional information. However, it exhibits more limited performance on other voice privacy tasks, emphasizing the need for further improvements.

***K*eywords**: speech anonymization, speech synthesis, vector-quantized variation auto-encoder, emotional state.

## 1 Introduction

Preserving privacy has become a key concern in artificial intelligence research, especially due to the widespread use of deep learning architectures that rely heavily on large datasets, often containing personal information. In contemporary applications, audio speech serves various purposes, including voice recognition systems for security and accessibility, virtual assistants for personalized user interactions, customer service automation for efficient query handling, and emotion analysis for enhancing user experience. Each of these applications underscores the need for robust privacy-preserving techniques to protect individuals' sensitive information while leveraging the power of AI. However, the inherent risk of speaker identification presents a significant threat to personal privacy. In response to this problem, the Voice Privacy Challenge 2024 [18] aims to tackle the critical task of anonymizing speech while preserving pertinent information, notably the emotional state of the speaker. Drawing on insights from previous challenges [3, 8, 13, 14, 16, 18], This article depicts a novel approach using discrete representation by a vector-quantized neural network for the speaker anonymization. The approach builds upon the foundations laid by existing research, leveraging advancements in vector quantization and neural network techniques to achieve effective speaker anonymization while preserving the emotional nuances conveyed in the speech. Organizers proposed several baselines for the Voice Privacy Challenge. The proposed architecture employs a similar approach as baselines B1, B5, and B6 [18]. In the proposed approach, the fundamental frequency (F0) and x-vectors are extracted for the purpose of anonymization, similar to the methods used in B1. Specifically, we focus on modifying the F0 component and integrating Vector-Quantized Variational Auto-Encoder (VQ-VAE). This strategy aligns with the techniques employed in B5 and B6 [18], aiming to leverage VQ-VAE's capabilities to enhance the effectiveness of the anonymization process while preserving essential speech characteristics. The proposed architecture aims to strike a balance between two essential objectives: anonymizing speaker information to protect privacy and retaining the emotional context of speech. By employing vector-quantized neural networks, we introduce a robust framework capable of achieving both objectives simultaneously. Throughout this paper, comprehensive explanations and analysis of the approach are provided, showcasing its efficacy in anonymizing speaker identity while faithfully preserving the emotional content of the speech data.

In the subsequent sections, the technical details of the proposed architecture are delved into, with its design principles, implementation strategies, and experimental results being elucidated. The effectiveness of the approach in achieving the goals set by the Voice Privacy Challenge 2024 is demonstrated through rigorous evaluation and comparison with existing methods.

## 2 Related Work

Prior research has explored various approaches to speaker anonymization and speech synthesis, employing a range of methodologies including deep learning and statistical techniques. Many studies have utilized prosodic features such as fundamental frequency (F0) and energy, as well as speaker embedding like x-vector proposed by Snyder et al. [17] for speaker information manipulation.

Wawalim et al. [12] developed an anonymization system based on F0 analysis and modified x-vectors, using Singular Value Modification and statistical regression models to enhance speaker privacy and alter speaker identifiable characteristics. Their approach underscores the importance of sophisticated statistical methods in handling speaker-related data transformations for privacy-sensitive applications. Gaznepoglu et al. [9] investigate F0 trajectory correction with a DNN where F0 trajectory is predicted in a logarithmic scale with a global mean-variance normalization. Champion et al. [4] apply a modification of F0 using a linear transformation based on the mean and standard deviation of log-scaled F0. In their work, the linear transformation is performed only on voiced frames. Meyer et al. [13] explored using random offsets for prosody cloning to maintain the naturalness and variability of speech characteristics essential for preserving speaker identity nuances. Their approach complements existing methods by focusing on the nuanced manipulation of prosodic elements to ensure that synthesized speech retains authenticity and remains intelligible.

This research investigates the use a vector-quantized variational auto-encoder (VQ-VAE) combined with prosody information to improve the disentanglement of content and speaker information [5, 6, 19] during speaker anonymization. Integrating prosody parameters enhances emotional expression and improves the fidelity of synthesized speech, aligning with recent advancements in speech synthesis [11].

## 3 Proposed Method

The proposed network employs a vector-quantized variational auto-encoder to separate speaker and content information. In addition to speaker information (x-vector), the decoder is conditioned on prosody information learned from the fundamental frequency (F0) and the energy of the spectrum. This conditioning enhances the model to focus more on the content of the speech. The detailed architecture of the proposed model is depicted in Figure 1.

### 3.1 Content Module

The content module comprises an encoder followed by vector quantization. The encoder includes two front-end convolution blocks, each with a kernel size of 3, a stride of 1, and 768 channels. This is followed by a downsampling convolution block with a kernel size of 4 and a stride of 2, which reduces the temporal resolution of the input feature from 100Hz to 50Hz. The sequence continues with two additional residual convolution blocks that mirror the configuration of the front-end blocks and concludes with four residual blocks, as shown in Figure 1. This network is based on the encoder of [6] except for the last four residual blocks, composed of ReLU activation functions and the convolutions. The encoder processes 80 mel-spectrograms as input and produces a 256-dimensional output representation. This output is then projected into discrete codes using a vector quantization module, which employs a codebook with 1024 codes, each 256-dimensional.

### 3.2 Prosody Module

To enhance the ability to capture the subtle nuances of intonation and emotional expression in speech, we propose incorporating two pivotal parameters: the fundamental frequency (F0), and the energy of the spectrum. These parameters are essential in enriching the prosody information supplied to the decoder, significantly improving the accuracy of speech reconstruction and elevating the efficacy of emotion detection in our model. The fundamental frequency (F0) was extracted from the audio waveform using pYAAPT[1], following the challenge guidelines. The F0 was normalized using a logarithmic scale, while the energy was normalized using the mean. The F0 and energy are then fed into a Bi-directional Gated Recurrent Unit (Bi-GRU) network with a hidden state dimensionality of 128, enabling robust temporal analysis of the prosody information.

### 3.3 Anonymization Module

The speaker anonymization process closely follows that of Baseline 1. However, the pre-trained ECAPA-TDNN [7] was used to compute the x-vector, known for its effectiveness in capturing robust speaker characteristics. The original x-vector was replaced with a pseudo-x-vector, obtained by averaging 100 x-vectors randomly selected from the 200 most distant x-vectors based on Euclidean distance. These distant x-vectors were chosen from the speaker pool, created with the mean x-vector of the 1417 training speakers, ensuring a diverse representation of speaker traits.

The effects of the fundamental frequency (F0) were investigated during the anonymization. Modifying the F0 can prevent the disclosure of identifiable speaker information. Firstly, we propose randomly adjusting the utterance F0 uniformly and independently between 0.8 and 1.2 to shift individual prosodic patterns while preserving the overall prosody of the utterance. Secondly, we propose normalizing the F0 using the mean of the most 100 dissimilar speakers obtained when computing the pseudo-x-vector.

### 3.4 Decoder Module

The HiFiGAN vocoder [11] is used as the decoder to synthesize the speech. The embedding from the content module was upsampled to 100Hz, concatenated with the embedding from the prosody module, and fed into the decoder along with the pseudo-x-vector. In this setup, the prosody embedding provides nuanced, time-varying information as local conditioning,

[1] `http://bjbschmitt.github.io/AMFM_decompy/pYAAPT.html` (last seen 08/2024).

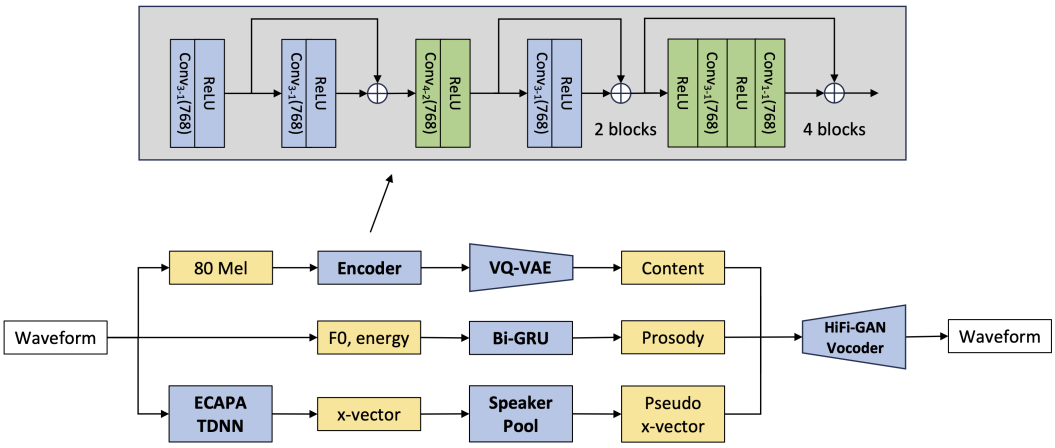


***Figure 1:*** *The proposed architecture: The top figure shows the encoder of the content module, while the bottom figure depicts the anonymization system, including the content, prosody, anonymization, and decoder modules. The system takes as input 80 mel-spectrogram, F0, energy, and x-vector. The pseudo-x-vector, with content and prosody embedding, is fed to the network to produce anonymized speech.*

while the pseudo-x-vector offers overarching speaker characteristics as global conditioning. The embeddings were upsampled with factors of 10, 4, and 4, totaling 160, to effectively reconstruct the waveform, guaranteeing that the synthesized speech corresponds to the original sample rate. Each upsampling stage employs kernels sized 20, 8, and 8, respectively.

## 4 Experiments

This work investigates the vector quantized-variational autoencoder (VQ-VAE) alongside fundamental frequency (F0) in three scenarios. System 1 employs the standard normalized F0 using a logarithmic transformation. System 2 applies a random scaling factor between 0.8 and 1.2 to F0. System 3 normalizes F0 using the mean value of the speakers with the most variation within the speaker pool. The distant speakers were obtained through the same process during the computation of the pseudo-x-vector. It is important to note that all experiments and evaluations follow the challenge guidelines [18]. The proposed system was evaluated against six baselines (B1-6) as specified in the challenge.

### 4.1 Datasets

All datasets used in the experiments complied with the challenge guidelines. The training data includes subsets from LibriSpeech [15] and additional data from CREMA-D [2] aiming to enhance the ability of the model to recognize and synthesize emotions. The speakers from the LibriSpeech were utilized to create a speaker pool for anonymization. Table 1 provides detailed statistical information on the composition and distribution of the training data. The development and test sets included subsets of both LibriSpeech and IEMOCAP [1]. These datasets were explicitly chosen to evaluate the performance across multiple tasks, including Automatic Speaker Verification (ASV), Automatic Speech Recognition (ASR), and Speaker Emotion Recognition (SER).

***Table 1:*** *Statistical information about training data.*

| Corpus | Subset | Hour | Speaker |
|---|---|---|---|
| LibriSpeech | train-clean-100 | 100.6 | 251 |
| | train-other-500 | 496.7 | 1,166 |
| CREMA-D | | 5.2 | 91 |

### 4.2 Evaluation Metrics

The anonymization systems were evaluated using three objective metrics. The Equal Error Rate (EER) served as the privacy metric, while the Word Error Rate (WER) and Unweighted Average Recall (UAR) were used as utility metrics for ASR and SER, respectively. The EER and WER were employed to assess the systems on Librispeech, and the UAR was used for the IEMOCAP test sets.

### 4.3 Experimental Setup

Two types of discriminators were used during the training process: the Multi-Period Discriminator (MPD) and the Multi-Scale Discriminator (MSD). These discriminators deal with multi-resolution concerning the temporal and local features, allowing the model to capture fine-grained speech details. The MPD and MSD follow the implementation proposed in [10]. The MPD was specifically simplified by targeting periods with factors of 3, 5, and 7. This modification was aimed at reducing the complexity of the discriminator, while ensuring that the model remains robust yet computationally feasible, aligning with the objectives of producing realistic and natural-sounding synthetic speech.

All the input features, including the 80 mel-spectrogram, fundamental frequency (F0), and energy, were computed using a window length of 25 milliseconds and a hop length of 10 milliseconds. The FFT size of 1024 was used to generate the spectrogram. The training was conducted over 150 epochs with a batch size of 128. The AdamW optimizer was utilized with $\beta_1 = 0.8$, $\beta_2 = 0.99$. The learning rate started at an initial value of $2 \times 10^{-4}$ and gradually decreased by 0.999 factor following each epoch. This configuration is consistent with the approach used in HiFiGAN [11].

## 5 Results and Discussions

The performance of the three systems, as detailed in Tables 2, 3, 4, and 5, indicates lower EERs than most baselines. This suggests a somewhat reduced effectiveness in terms of privacy, which may be due to the traditional approach used to compute the pseudo-x-vector and the disentanglement challenges associated with vector quantization when the codebook size is relatively large. However, system 2 outperforms the original configuration and baselines 1 and 2. Furthermore, it achieved a better UAR in speaker emotion recognition across test sets, ranking second after baseline B2. This highlights that the method is able to retain substantial information related to emotional information.

Additionally, system 2 yielded superior results in ASR than baselines B2 and B6. The performance enhancements were particularly notable for system 1, demonstrating a lower WER than B4. This suggests that although scaling the F0 with random factor between 0.8 and 1.2 might lead to some content information loss, the overall emotional expression is primarily maintained. This indicates robustness in capturing emotional nuances.

Despite its lower performance in SER compared to systems 1 and 2, system 3 outperformed some baselines. Nevertheless, it reported the highest WER across all test sets, indicating that normalizing F0 based on the mean values of the most distant speakers adversely impacts crucial content information within the speech. This normalization process distorts essential speech characteristics, compromising speech clarity and intelligibility.

## 6 Conclusion

This work examined a vector-quantized variational auto-encoder with prosody parameters, including the fundamental frequency (F0) and the spectrum's energy for speaker anonymization. The findings demonstrate that employing a vector quantization on the variational auto-encoder to disentangle content and speaker identity involves loss of information. However, that loss does not significantly reduce the efficacy of voice conversion processes. Moreover, the introduced method outperforms most of the baselines in terms of emotion recognition. This underlines the advantages of integrating discrete representations with the prosody parameters. However, the systems perform less significantly than most baselines for speaker anonymization.

Further investigation into the trade-offs associated with different codebook sizes could enhance disentanglement and improve anonymization. Reducing the quantizer's focus on speaker information would also be advantageous. Implementing a speaker classifier with gradient reversal as an auxiliary network on the content encoder's output could effectively penalize speaker-related content. Lastly, exploring advanced methods for generating pseudo-x-vectors could further refine anonymization. These combined efforts will contribute to more effective and nuanced anonymization strategies.

***Table 2:*** *The Equal Error Rates (EER, %) on LibriSpeech-dev achieved for male (M) and female (F) by the baselines (B1-6) and original (Orig.) data vs. the proposed systems*

| **Models** | **F ↑** | **M↑** | **Avg↑** |
|---|---|---|---|
| Orig. | 10.51 | 0.93 | 5.72 |
| B1 | 10.94 (9) | 7.45 (5) | 9.20 (7) |
| B2 | 12.91 (8) | 2.05 (9) | 7.48 (9) |
| B3 | 28.43 (3) | 22.04 (3) | 25.24 (3) |
| B4 | 34.37 (2) | 31.06 (2) | 32.71 (2) |
| **B5** | **35.82 (1)** | **32.92 (1)** | **34.37 (1)** |
| B6 | 25.14 (4) | 20.96 (4) | 23.05 (4) |
| System 1 | 16.47 (6) | 2.79 (7) | 9.63 (6) |
| System 2 | 17.91 (5) | 2.32 (8) | 10.11 (5) |
| System 3 | 14.05 (7) | 3.09 (6) | 8.57 (8) |

***Table 3:*** *The Equal Error Rates (EER, %) on LibriSpeech-test achieved for male (M) and female (F) by the baselines (B1-6) and original (Orig.) data vs. the proposed systems[2].*

| **Models** | **F ↑** | **M↑** | **Avg↑** |
|---|---|---|---|
| Orig. | 8.76 | 0.42 | 4.59 |
| B1 | 7.47 (8) | 4.68 (5) | 6.07 (6) |
| B2 | 7.48 (7) | 1.56 (8) | 4.52 (8) |
| B3 | 27.92 (3) | 26.72 (3) | 27.32 (3) |
| B4 | 29.37 (2) | 31.16 (2) | 30.26 (2) |
| **B5** | **33.95 (1)** | **34.73 (1)** | **34.34 (1)** |
| B6 | 21.15 (4) | 21.14 (4) | 21.14 (4) |
| System 1 | 8.76 (6) | 2.67 (6) | 5.72 (7) |
| System 2 | 11.31 (5) | 2.67 (6) | 6.99 (5) |
| System 3 | 6.38 (9) | 2.00 (7) | 4.19 (9) |

## Acknowledgments

This research was partially supported by the PERSYVAL Labex (ANR-11-LABX-0025), by the TALISMAN project (ANR-22-CE38-0007), by a French Government Scholarship

[1] In all tables, for each metric, the best system is in bold, the best proposal is underlined and rank of all systems are provided.

***Table 4:*** *Word Error Rates (WER, %) achieved by the baselines (B1-6) and original (Orig.) data vs. the proposed systems[2].*

| Models | LibriSpeech-dev ↓ | LibriSpeech-test ↓ |
|---|---|---|
| Orig. | 1.80 | 1.85 |
| **B1** | **3.07 (1)** | **2.91 (1)** |
| B2 | 10.44 (8) | 9.95 (8) |
| B3 | 4.29 (2) | 4.35 (2) |
| B4 | 6.15 (5) | 5.90 (6) |
| B5 | 4.73 (3) | 4.37 (3) |
| B6 | 9.69 (7) | 9.09 (7) |
| System 1 | 6.13 (4) | 5.27 (4) |
| System 2 | 6.59 (6) | 5.39 (5) |
| System 3 | 13.65 (9) | 11.04 (9) |

***Table 5:*** *Unweighted Average Recall (UAR, %) achieved by the baselines (B1-6) and original (Orig.) data vs. the proposed systems[2].*

| Models | IEMOCAP-dev ↑ | IEMOCAP-test ↑ |
|---|---|---|
| Orig. | 69.08 | 71.06 |
| B1 | 42.71 (4) | 42.78 (4) |
| **B2** | **55.61 (1)** | **53.49 (1)** |
| B3 | 38.09 (7) | 37.57 (7) |
| B4 | 41.97 (6) | 42.78 (4) |
| B5 | 38.08 (8) | 38.17 (5) |
| B6 | 36.39 (9) | 36.13 (8) |
| System 1 | 45.45 (3) | 44.23 (3) |
| System 2 | 45.56 (2) | 44.85 (2) |
| System 3 | 42.28 (5) | 38.06 (6) |

(BGF), and was granted access to the HPC resources of IDRIS under the allocation 2023-AD010614233 made by GENCI. This research was made possible by the collaboration between M-PSI team ↗ at Grenoble Informatics Laboratory (LIG) and Cambodia Academy of Digital Technology (CADT).

## References

[1] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359, 2008.

[2] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014.

[3] Pierre Champion. *Anonymizing Speech: Evaluating and Designing Speaker Anonymization Techniques*. PhD thesis, Université de Lorraine, 2023.

[4] Pierre Champion, Denis Jouvet, and Anthony Larcher. A Study of F0 Modification for X-Vector Based Speech Pseudo-Anonymization Across Gender. Research report, INRIA Nancy, équipe Multispeech, November 2020.

[5] Pierre Champion, Denis Jouvet, and Anthony Larcher. Are disentangled representations all you need to build speaker anonymization systems? In *INTERSPEECH 2022 - Human and Humanizing Speech Technology*, incheon, South Korea, September 2022.

[6] Jan Chorowski, Ron J Weiss, Samy Bengio, and Aäron Van Den Oord. Unsupervised speech representation learning using wavenet autoencoders. *IEEE/ACM transactions on audio, speech, and language processing*, 27(12):2041–2053, 2019.

[7] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. ECAPA-TDNN: Emphasized Channel Attention, propagation and aggregation in TDNN based speaker verification. In *Interspeech 2020*, pages 3830–3834, 2020.

[8] Fuming Fang, Xin Wang, Junichi Yamagishi, Isao Echizen, Massimiliano Todisco, Nicholas Evans, and Jean-François Bonastre. Speaker anonymization using x-vector and neural waveform models. In *10th ISCA Workshop on Speech Synthesis*, Vienna (Austria), September 2019.

[9] Ünal Ege Gaznepoglu, Anna Leschanowsky, and Nils Peters. Voiceprivacy 2022 system description: speaker anonymization with feature-matched f0 trajectories. *arXiv preprint arXiv:2210.17338*, 2022.

[10] Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, pages 5530–5540. PMLR, 2021.

[11] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33:17022–17033, 2020.

[12] Candy Olivia Mawalim, Kasorn Galajit, Jessada Karnjana, and Masashi Unoki. X-vector singular value modification and statistical-based decomposition with ensemble regression modeling for speaker anonymization system. In *Interspeech*, pages 1703–1707, 2020.

[13] Sarina Meyer, Florian Lux, Julia Koch, Pavel Denisov, Pascal Tilli, and Ngoc Thang 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)*, pages 1–5. IEEE, 2023.

[14] Michele Panariello, Francesco Nespoli, Massimiliano Todisco, and Nicholas Evans. Speaker anonymization using neural audio codec language models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4725–4729. IEEE, 2024.

[15] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE, 2015.

[16] Jose Patino, Natalia Tomashenko, Massimiliano Todisco, Andreas Nautsch, and Nicholas Evans. Speaker Anonymisation Using the McAdams Coefficient. In *Interspeech 2021*, pages 1099–1103, Brno, Czech Republic, August 2021. ISCA.

[17] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5329–5333. IEEE, 2018.

[18] Natalia Tomashenko, Xiaoxiao Miao, Pierre Champion, Sarina Meyer, Xin Wang, Emmanuel Vincent, Michele Panariello, Nicholas Evans, Junichi Yamagishi, and Massimiliano Todisco. The voice privacy 2024 challenge evaluation plan. In *4th Symposium on Security and Privacy in Speech Communication 2024*, 2024.

[19] Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.