# Exploring Vector-quantized Variational Auto-Encoder with Prosody Parameters for Speaker Anonymization

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

[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 linguistic content, prosody, and speaker-specific attributes. Our approach to speaker anonymization employs an end-to-end network using a Vector Quantization Variational Auto-Encoder (VQ-VAE) to separate these components, explicitly targeting speaker information. We condition the decoder with both speaker and prosody features to enhance performance. This method allows us to precisely modify characteristics key to speaker identity while preserving linguistic and emotional content. The findings indicate that the proposed method yields superior results and effectively preserves emotional information in all test sets, although privacy enhancement still requires further improvement.

**Index Terms**: speech recognition, human-computer interaction, computational paralinguistics

## 1. Proposed Method

Our approach leverages vector quantization alongside a variational auto-encoder to effectively separate speaker and content information. To enhance the model to focus on content, we condition the decoder with not only the speaker information (x-vector) but also with prosody features, including the fundamental frequency (F0) and the energy of the spectrum. Integrating prosody features improves the network's ability to preserve emotional information during speech synthesis. The detailed architecture of the proposed model is depicted in 1, and the description of each module is provided in Table 2. While the proposed network shares similarities with Baselines 5 and 6 of the challenge, our method uses the vector quantization in conjunction with a variational auto-encoder, which has proven more effective at disentangling content information [1, 2].

### 1.1. Content Module

The content module comprises an encoder module followed by a vector quantization module. The encoder features two front-end convolution blocks, each with a kernel size of 3, a stride of 1, and 768 channels. This setup is followed by a downsampling convolution block that employs a kernel size of 4 with a stride of 2, which compresses the signal 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 depicted in Figure 1. The encoder processes an 80 mel-spectrogram as input and generates a 256-dimensional output representation. This network bears similarities to the approach described in [1].

Lastly, the vector quantization module features a large codebook containing 1,024 codes, each with 256 dimensions. This design enables the network to learn more complex representations from the data.

### 1.2. 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 was extracted from the audio waveform using pYAAPT [1], which is given by the current challenge guidelines.

In addition, the f0 and the spectral energy were normalized according to the procedures outlined in Baseline 3 [3]. As depicted in Figure 1, the normalized f0 and energy are subsequently fed into a Bi-directional Gated Recurrent Unit (Bi-GRU) network with a hidden state dimensionality of 128, which allows for a robust temporal analysis of the prosody information.

### 1.3. Anonymization Module

Our speaker anonymization process closely follows that of Baseline 1. However, the ECAPA-TDNN [4] was used to compute the x-vector, which is known for its effectiveness in capturing robust speaker characteristics. The extracted x-vector was replaced with a pseudo-x-vector computed by averaging the x-vectors that are most distant from the original, selected from a specially constructed speaker pool. This pool was created using the mean x-vector computed for each speaker during training, ensuring a comprehensive representation of diverse speaker traits.

We are also investigating the effects of changing the fundamental frequency (F0) during the anonymization process. Modifying the F0 can prevent the disclosure of identifiable speaker information. Firstly, we suggest randomly adjusting the F0 by 20%, which is a method similar to the one used in Baseline 3 [3]. Secondly, we propose normalizing the F0 using the mean of the most dissimilar speakers from the speaker pool.

### 1.4. Decoder Module

The HiFiGAN vocoder [5] was used as the decoder to synthesize speech. The embedding from the content module was upsampled back 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, while the pseudo x-vector offers overarching speaker characteristics

[1]pYAAPT:http://bjbschmitt.github.io/AMFM_decompy/pYAAPT.html

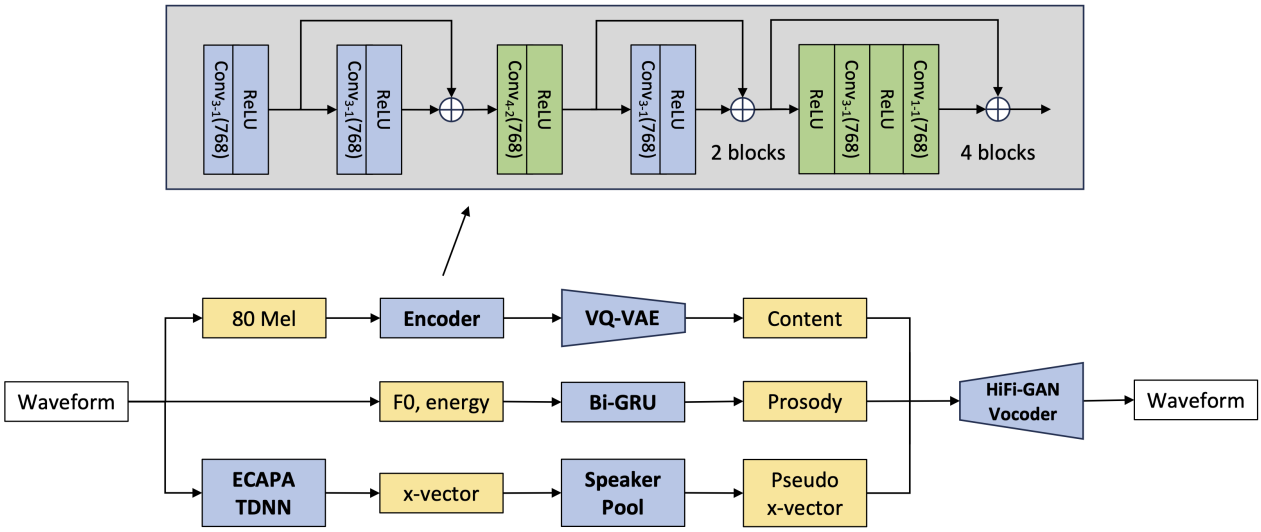


Figure 1: *The proposed architecture: The top figure shows the encoder of the content module, while the bottom figure depicts the anonymization system, including the content, prosody, anonymization, and decoder modules. The system takes as input an 80 mel-spectrogram, F0, energy, and x-vector. The x-vector, modified with or without normalized F0, is fed to the network to produce anonymized speech.*

as global conditioning

The upsampling procedure employs factors of 10, 4, and 4, totaling 160, to accommodate a sampling rate of 16kHz. Each stage employs kernels sized 20, 8, and 8, respectively. This structured upsampling effectively reconstructs the waveform, guaranteeing that the synthesized speech corresponds to the original sample rate. This multi-stage technique fills the gap between the low-dimensional embedded space and the high-resolution audio required for speech synthesis.

# 2. Experiment

## 2.1. Datasets

All datasets used in the experiments adhered to the guidelines defined by the challenge. Our training data consists of subsets from LibriSpeech [6] and additional data from CREMA-D [7] to enhance the model's ability to recognize and synthesize emotions. 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 [8]. These datasets were specifically chosen to evaluate performance of the model across multiple tasks, including automatic speaker verification, speech recognition, and speaker emotion recognition.

Table 1: *Statistical information about training datasets.*

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

## 2.2. Evaluation Metrics

This challenge evaluated the anonymization system using three objective metrics. The Equal Error Rate (EER) is the privacy metric, and two utility metrics, Word Error Rate (WER) for Automatic Speech Recognition (ASR) and Unweighted Average Recall (UAR) for Speech Emotion Recognition (SER), were used. The EER and WER were used to evaluate the system on Librispeech, while UAR was used on IEMOCAP test sets.

## 2.3. Experimental Setup

Two types of discriminators were used during the training process: the Multi-Period Discriminator (MPD) and the Multi-Scale Discriminator (MSD). Our MPD and MSD closely follow the implementation described in [9]. 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-spectrograms, fundamental frequency (F0), and energy, were computed using a window length of 25 milliseconds and a hop length of 10 milliseconds. We used an FFT size of 1024 to generate the spectrogram. The training was conducted over 150 epochs with a batch size of 128. We utilized the AdamW optimizer with $\beta_1 = 0.8$, $\beta_2 = 0.99$. The learning rate started at an initial value of $2 \times 10^{-4}$ and was gradually decreased by a factor of 0.999 following each epoch. This configuration is consistent with the approach used in HiFiGAN [5].

# 3. Results and Conclusion

The performance of our three systems is outlined in Tables 3, 4 and 5. Our system v1 features no F0 normalization, system v2 applies a random 20% scaling to F0, and system v3 normalize

F0 using the mean values obtained from the most distant speakers within speaker pool.

Our systems achieved a lower EER than most baselines. This indicates a somewhat reduced effectiveness in terms of privacy, which may derive from the traditional method used to compute the pseudo-x-vector. However, system v2 is the best and is greater than the Orig. configuration, Baselines 1 and 2. Additionally, it recorded a significant UAR in speaker emotion recognition across test sets, ranking second only to B4. This emphasizes the method's ability to retain considerable information pertinent to emotional characteristics.

Furthermore, systems v2 yielded superior results in ASR than B2 and B6. The performance enhancements were particularly notable for system v1, which demonstrated a lower WER than B4. This suggests that although scaling the F0 by 20% might lead to some loss of content information, the overall emotional expression is primarily maintained, indicating robustness in capturing emotional nuances.

Despite its lower performance in SER compared to system v1 and v2, system v3 outperformed many baselines. Nevertheless, it registered the highest WER across all test sets, indicating that normalizing F0 based on the mean values of the most distant speakers adversely impact crucial content information within the speech. This normalization process distorts essential speech characteristics, compromising the speech output's clarity and intelligibility.

In summary, these results suggest that while using vector quantization to separate content from speaker identity leads to some information loss, it does not drastically affect voice conversion effectiveness. More importantly, the proposed method has proven highly effective in SER, mainly system v2, compared to most baselines, validating the benefits of integrating discrete representations with prosody information. This integration enhances the model's capability to recognize and replicate emotional states, making it a valuable approach for applications where emotional accuracy is critical.

# 4. References


[1] J. Chorowski, R. J. Weiss, S. Bengio, and A. Van Den Oord, "Unsupervised speech representation learning using wavenet autoencoders," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 12, pp. 2041–2053, 2019.

[2] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[3] 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.

[4] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.

[5] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in neural information processing systems*, vol. 33, pp. 17 022–17 033, 2020.

[6] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[7] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

[8] 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, pp. 335–359, 2008.

[9] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5530–5540.

Table 2: *Module Descriptions and Outputs: The model was trained end-to-end using the proposed training data.*

| # | Module | Description | Output features |
|---|---|---|---|
| 1 | ECAPA-TDNN | Pretrained model<br>Input: Waveform (16KHz) | X-vector$^{192}$ |
| 2 | Encoder | Input: 80 mel-spectrogram<br>5 convolutions blocks<br>4 residual blocks | Embedding$^{256}$ |
| 3 | VQ-VAE | Vector Quantizer<br>Input: Embedding$^{256}$<br>Codebook: 1024 codes, each with 256 dimensions | Codes$^{256}$ |
| 4 | Bi-GRU | Prosody Encoder<br>Input: F0 + Energy<br>Hidden: 128 | Prosody$^{256}$ |
| 5 | Speaker Pool | X-vector Pool<br>The mean x-vector of the speakers from the training set | Pseudo-x-vector$^{192}$ |
| 6 | HiFi-GAN Vocoder | Input: Codes$^{256}$ + Prosody$^{256}$ (local conditioning)<br>Global conditioning: Pseudo-x-vector$^{192}$ | Waveform (16KHz) |

Table 3: *The Equal Error Rates (EER, %) achieved by the baselines and original (Orig.) data vs. the proposed method.*

| | LibriSpeech-dev | | | LibriSpeech-test | | |
|---|---|---|---|---|---|---|
| **Models** | **Female** | **Male** | **Average** | **Female** | **Male** | **Average** |
| Orig. | 10.51 | 00.93 | 05.72 | 08.76 | 00.42 | 04.59 |
| B1 | 10.94 | 07.45 | 09.20 | 07.47 | 04.68 | 06.07 |
| B2 | 12.91 | 02.05 | 07.48 | 07.48 | 01.56 | 04.52 |
| B3 | 28.43 | 22.04 | 25.24 | 27.92 | 26.72 | 27.32 |
| B4 | 34.37 | 31.06 | 32.71 | 29.37 | 31.16 | 30.26 |
| B5 | **35.82** | **32.92** | **34.37** | **33.95** | **34.73** | **34.34** |
| B6 | 25.14 | 20.96 | 23.05 | 21.15 | 21.14 | 21.14 |
| Ours v1 | *16.47* | ***02.79*** | *09.63* | *08.76* | *02.67* | *05.72* |
| Ours v2 | ***17.91*** | *02.32* | ***10.11*** | ***11.31*** | ***02.67*** | ***06.99*** |
| Ours v3 | *14.05* | *03.09* | *08.57* | *06.38* | *02.00* | *04.19* |

Table 4: *Word Error Rates (WER, %) achieved by the baselines and original (Orig.) data vs. the proposed method.*

| Models | LibriSpeech-dev | LibriSpeech-test |
|---|---|---|
| Orig. | **01.80** | **01.85** |
| B1 | 03.07 | 02.91 |
| B2 | 10.44 | 09.95 |
| B3 | 04.29 | 04.35 |
| B4 | 06.15 | 05.90 |
| B5 | 04.73 | 04.37 |
| B6 | 09.69 | 09.09 |
| Ours v1 | ***06.13*** | ***05.27*** |
| Ours v2 | *06.59* | *05.39* |
| Ours v3 | *13.65* | *11.04* |

Table 5: *Unweighted average recall (UAR, %) achieved by the baselines and original (Orig.) data vs. the proposed method.*

| Models | IEMOCAP-dev | IEMOCAP-test |
|---|---|---|
| Orig. | **69.08** | **71.06** |
| B1 | 42.71 | 42.78 |
| B2 | 55.61 | 53.49 |
| B3 | 38.09 | 37.57 |
| B4 | 41.97 | 42.78 |
| B5 | 38.08 | 38.17 |
| B6 | 36.39 | 36.13 |
| Ours v1 | *45.45* | *44.23* |
| Ours v2 | ***45.56*** | ***44.85*** |
| Ours v3 | *42.28* | *38.06* |