



# Learning Optimal Prosody Embedding Codebook based on F0 and Energy

David Porteš, Aleš Horák

Faculty of Informatics, Masaryk University, Brno, Czech Republic

xportes@fi.muni.cz, hales@fi.muni.cz

## Abstract

Both the Fundamental frequency (F0) and Energy are prominent features of prosody. Together, they have been used across a wide variety of speech-processing tasks. However, there is a lack of freely available pre-trained vector representations of these features. Therefore, in this paper, we provide the research community with high-quality joint embeddings of the frame-level F0 and Energy features, using the VQ-VAE architecture. By converting the F0 and Energy into a single stream of vector embeddings, we make it possible to seamlessly use prosody in modern architectures, such as multimodal LLMs. In order to ensure maximum embedding quality, we conduct a large-scale hyperparameter search, totaling over 150 experiments on the LibriTTS dataset. We outperform previous works on F0 embeddings, reaching FFE error below 1 percent, while simultaneously embedding the additional feature of Energy. We publish our best-performing models on the HuggingFace website.

**Index Terms:** Prosody, VQ-VAE, Fundamental frequency, F0, Energy, Embeddings

## 1. Introduction

The Fundamental frequency (F0) and Energy are two of the most prominent features of prosody. Along with phoneme duration, they have been used as a de-facto default representation of prosody across many different speech processing tasks, such as text-to-speech [1, 2], speaker anonymization [3], speech resynthesis [4] or prosody transfer [5, 6]. However, there is a lack of freely available pre-trained vector representations of these features, which makes it hard to integrate them into state-of-the-art deep learning architectures. Additionally, to the best of our knowledge, frame-level Energy embeddings have not been constructed before.

Therefore, in this paper, we provide the research community with high-quality joint embeddings of the frame-based F0 and Energy features, which can be used in an off-the-shelf manner. By embedding multiple prosodic features into a single series of discrete vectors, we make it possible to seamlessly use prosody in multimodal Large language models based on the Transformer [7] architecture. This could further enhance existing use cases, or potentially unlock new ones, such as generative prosody modeling or even prosody-conditioned text generation, as proposed in [8].

As our model of choice, we make use of the Vector Quantized Variational Autoencoder (VQ-VAE) [9]. The unique properties of the VQ-VAE make it ideal for the task, since it learns a codebook of discrete vectors to represent the input features. This latent representation can easily be extracted and used as embeddings.

To maximize the embedding quality, we present a large-scale hyperparameter study over the VQ-VAE parameters (Number of bins, Embedding vector size), as well as parameters of the training process (Batch size, Learning rate). Additionally, we take special care with respect to the handling of the unvoiced frames of F0. These are frames where F0 is not defined, because the vocal folds are not vibrating, and are typically set to 0 Hz by the F0 extraction algorithms. This introduces many discontinuous jumps into the signal, causing issues for the VQ-VAE model. In order to mitigate these jumps, we either interpolate over the unvoiced regions of the F0, or, to account for use cases where the unvoiced regions need to be preserved, we normalize the voiced F0 regions (see Figure 2). We also explore the combination of these two strategies.

For each strategy, we conduct a separate hyperparameter search using the Optuna framework [10] on the LibriTTS dataset [11], in total comprising over 150 experiments. We outperform the F0 reconstruction accuracy of the F0-only embeddings of [12], while at the same time embedding the entire additional feature of Energy, and we publish our best models for each strategy on the HuggingFace website.<sup>1</sup>

## 2. Related Work

Fundamental frequency and Energy have been used to represent prosody across many different speech-processing tasks, with highly varying representations. As an example, FastSpeech2 [2] uses F0 and Energy of target speech as conditioning input during training, and employs continuous wavelet transform on the pitch to improve pitch prediction. USD-AC [13] uses binned Gaussian-blurred F0 and Energy features for accent conversion. Valle et al. [14] use frame-based continuous F0 and Energy to obtain a highly controllable text-to-speech system, while in Daft-Exprt [5], they predict both frame-level and utterance level Energy and F0 features for the purpose of cross-speaker prosody transfer.

However, none of the commonly used representations of these two features can be used seamlessly in modern deep-learning architectures. To address this gap, we continue in the line of work started by Polyak et al. [4], who used the VQ-VAE model to generate embeddings of F0 for the purpose of speech synthesis from disentangled features. This was followed by [12], who conducted a full grid search over the parameters of the VQ-VAE and improved the F0 reconstruction error. We expand upon this line of work by jointly embedding the Fundamental frequency and Energy features into a single embedding, thus obtaining prosodic embeddings that can be easily used across a wide range of speech-processing tasks. Addition-

<sup>1</sup>[https://huggingface.co/MU-NLPC/F0\\_Energy\\_joint\\_VQVAE\\_embeddings](https://huggingface.co/MU-NLPC/F0_Energy_joint_VQVAE_embeddings)

ally, we make use of the Optuna framework [10], which allows us to efficiently sweep over a wider range of parameters.

### 3. Methods

For our study, we make use of the Vector Quantized Variational Autoencoder (VQ-VAE) model, whose implementation was adapted from Polyak et al. [4]. The VQ-VAE model, as all autoencoders, works by learning to reconstruct the input signal from a limited-size bottleneck layer representation. However, it differs from the vanilla autoencoder in that it restricts the allowed bottleneck activations to a set of discrete vectors from a fixed-size, learnable codebook (see Figure 1).

In greater detail, the input data

$$(x_1, x_2, x_3, \dots, x_n), (w_1, w_2, w_3, \dots, w_n) @ 200 \text{ Hz},$$

with  $x_n$  and  $w_n$  being the F0 and Energy values, respectively, are fed into the encoder. By passing the input through successive 1D convolutional layers, the encoder down-samples the data by a factor of 16, obtaining the latent representation

$$(z_1, z_2, z_3, \dots, z_m) @ 12.5 \text{ Hz}$$

$$z_1, z_2, \dots, z_m \in \mathbb{R}^{emb.size}.$$

Then, nearest neighbor-based vector quantization is applied, collapsing each vector into its closest vector in the codebook.<sup>2</sup>

$$(z_{q1}, z_{q2}, z_{q3}, \dots, z_{qm}) @ 12.5 \text{ Hz}$$

$$z_{q1}, z_{q2}, \dots, z_{qm} \in \mathbb{R}^{emb.size}$$

$$\forall x, z_{qx} \in \text{codebook}.$$

The reconstruction is then generated by an inverse process in the decoder, obtaining

$$(\hat{x}_1, \hat{x}_2, \hat{x}_3, \dots, \hat{x}_n), (\hat{w}_1, \hat{w}_2, \hat{w}_3, \dots, \hat{w}_n) @ 200 \text{ Hz}.$$

The VQ-VAE is trained with the standard Mean Square Error (MSE) reconstruction loss. The Fundamental frequency and Energy each use a separate channel in the convolutional layers.

#### 3.1. Data and Preprocessing

We conduct our experiments on the English multispeaker dataset LibriTTS, using the train-clean-100 (33,236 files), dev-clean (5,736 files), and test-clean (4,837 files) splits. We use audio files sampled at 16 kHz. The F0 is extracted using the YAAPT algorithm [15], using hop size 5 ms and frame size 20 ms, resulting in the F0 being sampled at 200 Hz. The unvoiced frames are set to 0 Hz. For Energy<sup>3</sup>, the PRAAT [16] package was used, with a sampling rate identical to F0 at 200 Hz. To mitigate the discontinuous jumps caused by unvoiced regions of F0, we investigate three different strategies for their handling.

- F0 interpolation (**FI**)

We interpolate over the unvoiced regions of F0 using the built-in interpolation algorithm of the YAAPT package, so that the interpolated F0 does not descend down to and up from zero, but instead travels smoothly from the last voiced frame to the first voiced frame of the next voiced region. Note that the reconstruction quality is measured against the

<sup>2</sup>These quantized vectors are used as the embedding.

<sup>3</sup>We use the intensity feature from PRAAT, which is expressed in dB, instead of raw Energy, for better interpretability.

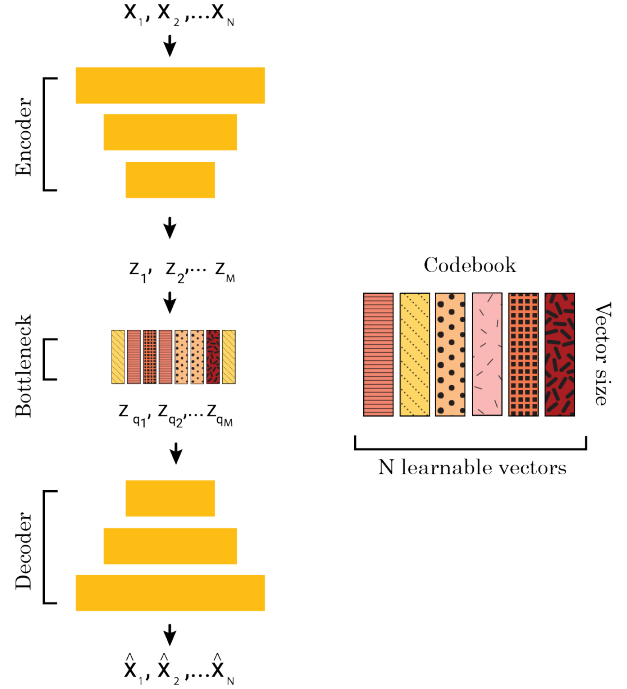


Figure 1: Diagram of the VQ-VAE architecture.

interpolated ground truth F0. This strategy is useful in case speaker statistics for normalization cannot be obtained and unvoiced regions do not need to be preserved, or for use cases like linguistic interpretation of prosody.

- Voiced F0 normalization + Energy normalization + voicedness mask (**FN+EN+VM**)

We normalize the voiced frames of the F0 with respect to the speaker mean and standard deviation, while leaving the unvoiced frames at 0 Hz. Energy is normalized with respect to speaker mean and standard deviation as well. We make use of the voicedness mask in order to recover the unvoiced regions of the F0. This serves to distinguish between unvoiced frames and frames with F0 close to the mean (See Figure 2). The voicedness mask is added using a third channel. The normalization of both F0 and Energy is reversed before calculating the final reconstruction metrics, and the F0 regions with voicedness mask reconstruction  $< 0.5$  are set to 0 Hz. This strategy is useful in case the unvoiced regions should be preserved.

- F0 and Energy normalization + F0 interpolation (**FN+EN+FI**)

This setting is a combination of the above two strategies. The normalization of both F0 and Energy is reversed before calculating the final reconstruction metrics. Note that the voicedness mask is not needed, since the unvoiced regions are interpolated over. The reconstruction quality is measured against the interpolated ground truth F0.

#### 3.2. Training

We conduct training on one-second-long segments (192 frames) of the input F0, which are dynamically sampled during the training process so that, at each epoch, a different segment is used. We exclude audio files shorter than 1 second from the training.

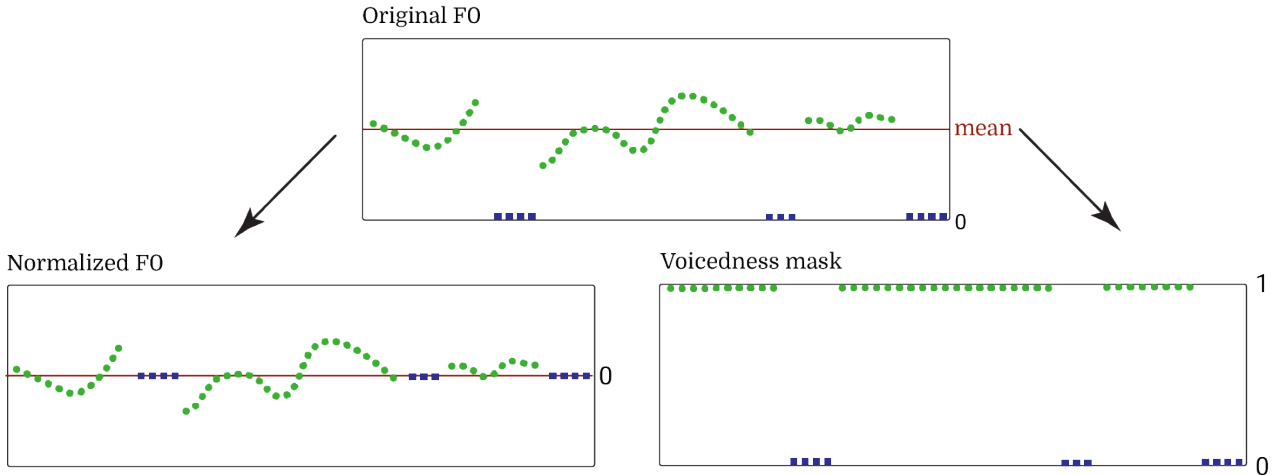


Figure 2: Diagram of the voicedness mask feature used to distinguish voiced frames (green circles) from unvoiced frames (blue squares) when normalization is used without interpolation. Note that without the mask, the frames with F0 very close to the mean are indistinguishable from the unvoiced frames.

In this manner, we train for 700 epochs or until the experiment is pruned by the optimizer. We then calculate the final validation metrics using the checkpoint<sup>4</sup> with the lowest validation loss.

The hyperparameter search is conducted over the batch size, learning rate, number of vectors (bins) in the codebook, and embedding vector size parameters, and is guided by the Optuna optimizer with pruning. We use the TPESampler along with the Hyperband pruner, with the maximum resource set to the maximum number of epochs (700) and a reduction factor of 7. We set the minimum resource to 1 epoch. The hyperparameter ranges we search over are summarized in table 1. For each strategy, we conduct over 50 experiments, and we evaluate the best-performing model for each strategy on the test set. Each of our experiments was run on a 10GB VRAM fraction of the NVIDIA A100 using NVIDIA Multi-Instance GPU (MIG), taking on average 4 hours (if not pruned).

Table 1: Parameter ranges searched

Parameter	Values
N bins	20, 40, 80, 160, 320
Emb size	32, 64, 128, 256, 512, 1024
Batch size	32, 64, 128, 256
Learning rate	0.001, 0.0005, 0.0001

#### 4. Experiments and Evaluation

For evaluation, we make use of the standard Voicing Decision Error (VDE) [17] and F0 Frame Error (FFE) [18] metrics for the Fundamental frequency. The VDE is defined as the percentage of frames that contain a voicing decision error

$$\text{VDE}(i) = \begin{cases} 1, & \text{if } \text{voiced}(\hat{f}_0(i)) \neq \text{voiced}(f_0(i)) \\ 0, & \text{otherwise} \end{cases}$$

$$\text{VDE} = \frac{1}{N} \sum_{i=1}^N \text{VDE}(i)$$

<sup>4</sup>Checkpoints are taken each 20000 steps.

where  $f_0$  is the original F0,  $\hat{f}_0$  is the F0 reconstruction and  $i$  is the frame index. A voicing decision error occurs when a frame is reconstructed as unvoiced when the original is voiced, or vice-versa. We do not report VDE when interpolation is used, as all unvoiced regions are interpolated over.

The FFE measures the overall accuracy of the reconstruction, measured as the portion of frames that differ by more than X percent from the true value or contain a VDE error. We report FFE using X = 20 and X = 10 percentage thresholds.

$$\text{FFE}(i, X) = \begin{cases} 1, & \text{if } \left| \frac{\hat{f}_0(i) - f_0(i)}{f_0(i)} \right| * 100\% > X \\ 0, & \text{otherwise} \end{cases}$$

$$\text{FFE}(X) = \frac{1}{N} \sum_{i=1}^N \text{FFE}(i, X) \vee \text{VDE}(i)$$

For Energy, we use the RMSE between the reconstruction and the original frames. All metrics are calculated per audio file, and we report their average.

#### 5. Results and Discussion

The five top-performing experiments for each strategy are showcased in tables 2-4, sorted by FFE. Since the results follow clear stratification depending on the number of bins, with a larger number of bins leading to better reconstruction accuracy, we also showcase the best-performing model for each number of bins setting<sup>5</sup>. In tables 5 and 6, we show the results for the lowest FFE model for each strategy on the test set, and in Figure 3, we show a sample test set reconstruction. The full results are available on the model’s HuggingFace website.

Our results outperform [12] both on interpolated (0.49% vs 1.60% FFE<sub>20</sub>) and non-interpolated data (3.30% vs 4.29% FFE<sub>20</sub>), while simultaneously embedding the additional feature of Energy. This is likely due to our search sweeping across larger parameter space. The performance variation

<sup>5</sup>Note that, due to the hyperparameter optimizer, this is best thought of as the upper bound on the performance drop magnitude, as fewer configurations were searched for the smaller numbers of bins.

Table 2: Dev set results for the F0 interpolation strategy (FI).

Bins	Emb	Batch	LR	VDE%	FFE <sub>20</sub> %	RMSE
320	512	256	0.0005	-	0.38	5.26
320	1024	256	0.001	-	0.40	4.57
160	512	128	0.001	-	0.41	4.80
320	512	128	0.0005	-	0.42	4.77
320	32	256	0.001	-	0.43	4.53
80	64	256	0.0005	-	0.59	5.79
40	128	64	0.0001	-	1.19	7.88
20	512	32	0.0001	-	1.63	8.85

Table 3: Dev set results for the F0 and Energy normalization + F0 interpolation strategy (FN+EN+FI).

Bins	Emb	Batch	LR	VDE%	FFE <sub>20</sub> %	RMSE
320	128	128	0.001	-	0.71	2.73
320	64	128	0.0005	-	0.74	2.57
320	128	128	0.0005	-	0.75	2.76
320	256	128	0.001	-	0.76	2.71
320	128	128	0.001	-	0.76	2.68
160	512	256	0.001	-	0.81	2.90
80	32	256	0.001	-	0.98	3.42
40	256	64	0.001	-	1.45	3.70
20	32	256	0.001	-	1.99	4.45

Table 4: Dev set results for the Voiced F0 normalization + Energy normalization + voicedness mask strategy (FN+EN+VM).

Bins	Emb	Batch	LR	VDE%	FFE <sub>20</sub> %	RMSE
320	64	128	0.0005	1.92	3.02	2.85
320	512	128	0.0005	2.13	3.26	3.00
320	64	64	0.0005	2.17	3.29	3.03
160	256	256	0.001	2.27	3.44	3.40
160	32	256	0.0005	2.62	3.82	3.43
80	512	256	0.001	3.14	4.47	3.51
40	64	256	0.0005	4.03	5.60	4.00
20	1024	256	0.001	5.97	7.89	4.90

Table 5: Test set results for strategies with interpolation. Compared with [12], denoted by \*.

Strategy	VDE%	FFE <sub>20</sub> %	FFE <sub>10</sub> %	RMSE
FI	-	0.49	3.83	6.33
FN+EN+FI	-	0.78	3.54	3.00
Reference*	-	1.60	-	-

Table 6: Test set results for strategies without interpolation. Compared with [12], denoted by \*.

Strategy	VDE%	FFE <sub>20</sub> %	FFE <sub>10</sub> %	RMSE
FN+EN+VM	2.07	3.30	6.70	3.00
No strategy	1.17	5.73	17.41	16.30
Reference*	2.47	4.29	-	-

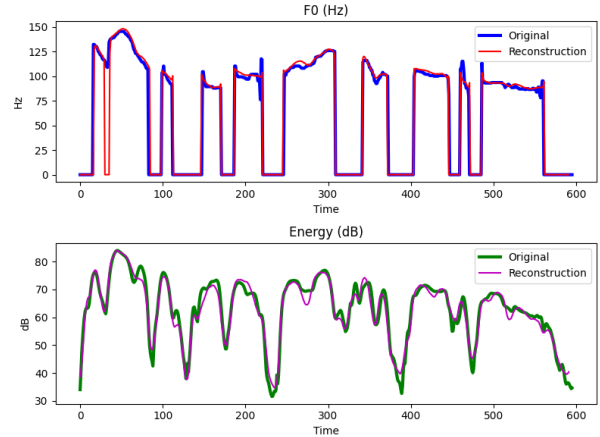


Figure 3: Sample test set reconstruction using the FN+EN+VM strategy.

across strategies follows a similar pattern to [12], with interpolated F0 being easier to reconstruct. The strategies using normalization, however, outperform the unnormalized case (FI) on the reconstruction of Energy (3 dB vs 6.3 dB RMSE). It seems the models prefer to minimize F0 loss, since, without normalization, F0 has larger values than Energy.

Our results suggest that the number of bins parameter is the main limiting factor of reconstruction accuracy. On the other hand, the embedding vector size parameter was surprisingly insignificant, with smaller sizes often outperforming larger ones, other parameters being equal. With regards to the training parameters, we obtain the best results with the batch size setting of 128 or 256, with a learning rate of either 0.001 or 0.0005. The 0.0001 setting underperforms the other learning rate values.

## 6. Conclusion

In this paper, we have presented a thorough study on generating high-quality, joint vector embeddings of the Fundamental frequency and Energy features. In total, we conducted over 150 experiments, using three different strategies for handling unvoiced frames of the F0. We reach a very low F0 reconstruction error on the test set both on interpolated (below 1% FFE<sub>20</sub>) and non-interpolated (3.30% FFE<sub>20</sub>) data, outperforming previously published works on F0 embedding generation, while also embedding Energy with reconstruction RMSE at 3-6.3 dB. Our best models are freely available on the Hugging-Face website, along with supporting code to make their use as easy as possible. We believe our models will prove valuable to the speech-processing community, allowing prosody to be straightforwardly used in modern deep-learning architectures. As future work, we plan to use these embeddings for experiments regarding prosody-conditioned text generation.

## 7. Acknowledgements

This work has been partly supported by the Ministry of Education, Youth and Sports of the Czech Republic within the LINDAT-CLARIAH-CZ project LM2023062. This publication was written with the support of the Specific University Research provided by the Ministry of Education, Youth and Sports of the Czech Republic. Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

## 8. References

- [1] N.-Q. Wu, Z.-C. Liu, and Z.-H. Ling, "Discourse-Level Prosody Modeling with a Variational Autoencoder for Non-Autoregressive Expressive Speech Synthesis," in *ICASSP 2022 - IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2022, pp. 7592–7596, iSSN: 2379-190X.
- [2] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *International Conference on Learning Representations*, 2021.
- [3] S. Leang, A. Augusma, D. Vaufraydaz, E. Castelli, S. Sam, and F. Letué, "Exploring VQ-VAE with Prosody Parameters for Speaker Anonymization," in *4th Symposium on Security and Privacy in Speech Communication*. ISCA, Sep. 2024, pp. 127–131.
- [4] A. Polyak, Y. Adi, J. Copet, E. Kharitonov, K. Lakhotia, W.-N. Hsu, A. Mohamed, and E. Dupoux, "Speech Resynthesis from Discrete Disentangled Self-Supervised Representations," in *Interspeech 2021*. ISCA, Aug. 2021, pp. 3615–3619.
- [5] J. Zaïdi, H. Seuté, B. Van Niekerk, and M.-A. Carbonneau, "Daft-Exprt: Cross-Speaker Prosody Transfer on Any Text for Expressive Speech Synthesis," in *Interspeech 2022*. ISCA, Sep. 2022, pp. 4591–4595.
- [6] A. T. Sigurgeirsson and S. King, "Do Prosody Transfer Models Transfer Prosody?" in *ICASSP 2023 - IEEE International Conference on Acoustics, Speech and Signal Processing*, Jun. 2023, pp. 1–5, iSSN: 2379-190X.
- [7] A. Vaswani *et al.*, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.
- [8] D. Porteš, "Towards Using Speech Melody to Guide Large Language Models," in *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2023*, 2023, pp. 133–141.
- [9] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., Dec. 2017, pp. 6309–6318.
- [10] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD '19. New York, NY, USA: Association for Computing Machinery, Jul. 2019, pp. 2623–2631.
- [11] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech," in *Interspeech 2019*. ISCA, Sep. 2019, pp. 1526–1530.
- [12] D. Porteš and A. Horák, "Generating High-Quality F0 Embeddings Using the Vector-Quantized Variational Autoencoder," in *Text, Speech, and Dialogue*, E. Nöth, A. Horák, and P. Sojka, Eds. Cham: Springer Nature Switzerland, 2024, pp. 139–148.
- [13] J.-H. Huang, W.-T. Lee, and C.-H. Wu, "USD-AC: Unsupervised Speech Disentanglement for Accent Conversion," in *Interspeech 2024*. ISCA, Sep. 2024, pp. 4388–4392.
- [14] R. Valle, J. F. Santos, K. J. Shih, R. Badlani, and B. Catanzaro, "High-Acoustic Fidelity Text To Speech Synthesis With Fine-Grained Control Of Speech Attributes," in *ICASSP 2023 - IEEE International Conference on Acoustics, Speech and Signal Processing*, Jun. 2023, pp. 1–5, iSSN: 2379-190X.
- [15] K. Kasi and S. A. Zahorian, "Yet Another Algorithm for Pitch Tracking," in *IEEE International Conference on Acoustics Speech and Signal Processing*. Orlando, FL, USA: IEEE, May 2002, pp. 1–361–1–364.
- [16] P. Boersma and V. Van Heuven, "Speak and unspeak with praat," *Glott International*, vol. 5, no. 9/10, pp. 341–347, 2001.
- [17] T. Nakatani, S. Amano, T. Irino, K. Ishizuka, and T. Kondo, "A method for fundamental frequency estimation and voicing decision: Application to infant utterances recorded in real acoustical environments," *Speech Communication*, vol. 50, no. 3, pp. 203–214, Mar. 2008.
- [18] W. Chu and A. Alwan, "Reducing F0 Frame Error of F0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, Apr. 2009, pp. 3969–3972, iSSN: 2379-190X.