



# Speech Resynthesis from Discrete Disentangled Self-Supervised Representations

Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia,  
Wei-Ning Hsu, Abdelrahman Mohamed, Emmanuel Dupoux

## ► To cite this version:

Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, et al.. Speech Resynthesis from Discrete Disentangled Self-Supervised Representations. INTERSPEECH 2021 - Annual Conference of the International Speech Communication Association, Aug 2021, Brno, Czech Republic. hal-03329245

**HAL Id: hal-03329245**

**<https://inria.hal.science/hal-03329245>**

Submitted on 11 Oct 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Speech Resynthesis from Discrete Disentangled Self-Supervised Representations

Adam Polyak<sup>1,2\*</sup>, Yossi Adi<sup>1</sup>, Jade Copet<sup>1</sup>, Eugene Kharitonov<sup>1</sup>, Kushal Lakhota<sup>1</sup>, Wei-Ning Hsu<sup>1</sup>, Abdelrahman Mohamed<sup>1</sup>, Emmanuel Dupoux<sup>1,3</sup>

<sup>1</sup>Facebook AI Research, USA

<sup>2</sup>Tel-Aviv University, Israel

<sup>3</sup>EHESS, France

adampolyak@fb.com

## Abstract

We propose using self-supervised discrete representations for the task of speech resynthesis. To generate disentangled representation, we separately extract low-bitrate representations for speech content, prosodic information, and speaker identity. This allows to synthesize speech in a controllable manner. We analyze various state-of-the-art, self-supervised representation learning methods and shed light on the advantages of each method while considering reconstruction quality and disentanglement properties. Specifically, we evaluate the F0 reconstruction, speaker identification performance (for both resynthesis and voice conversion), recordings' intelligibility, and overall quality using subjective human evaluation. Lastly, we demonstrate how these representations can be used for an ultra-lightweight speech codec. Using the obtained representations, we can get to a rate of 365 bits per second while providing better speech quality than the baseline methods. Audio samples can be found under the following link: [speechbot.github.io/resynthesis](https://speechbot.github.io/resynthesis).

**Index Terms:** speech generation, speech resynthesis, self-supervised learning, speech codec.

## 1. Introduction

Learning unsupervised speech representations, both continuous and discrete, has seen a significant leap in performance following the recent success of Self-Supervised Learning (SSL) methods [1, 2, 3, 4]. In the self-supervised setting, unlabeled inputs define an auxiliary task that can generate pseudo-labeled training data. This data can then be used to train a model using supervised techniques. The learned representations are often used for downstream tasks with a minimal amount of supervised data. For example, learning speech recognition with only 10 minutes of transcribed speech and 53K hours of untranscribed speech as in wav2vec 2.0 [3] and HuBERT [4]. The learned self-supervised discrete representations also showed impressive performance on the conditional and unconditional spoken generative language modeling (GSLM) task [5].

Despite its success, most studies on SSL for speech are focused on generating and evaluating the quality of the learned representations in the context of Automatic Speech Recognition (ASR). It remains unclear how suitable these representations are for speech synthesis. Moreover, in the context of expressive and controllable generation, it is unknown to what extent the speaker identity and F0 information are encoded in the learned representations.

\*The contribution of Adam Polyak is part of a Ph.D. thesis research conducted at Tel Aviv University.

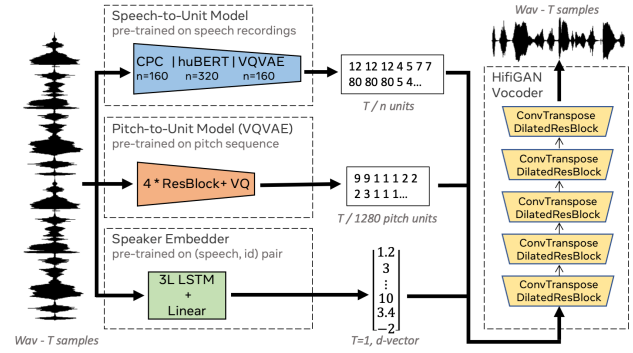


Figure 1: **The overall proposed speech resynthesis architecture.** Three parallel encoders extract discrete representations from the raw input signal. These are then being used as a conditioning to reconstruct the signal using a decoder network.

Traditionally, speech synthesis and text-to-speech models produce Mel-spectrogram autoregressively given textual features as input. Next, a vocoder is applied to reconstruct the phase from the Mel-spectrogram (e.g., Griffin-Lim [6], WaveNet [7], WaveGlow [8], or HiFi-GAN [9]). In this study, we suggest using the learned speech units as an input to a vocoder module with no spectrogram estimation. We additionally augment the learned units with quantized F0 representation and a global speaker embedding. Figure 1 presents the overall proposed method. This allows the evaluation of the learned units with respect to speech content, speaker identity, and F0 information, as well as better control the audio synthesis. We experiment with signal reconstruction, voice conversion, and F0 manipulation using several datasets and encoder models. Finally, equipped with our previous findings, we demonstrate how the learned units can function as an ultra-lightweight speech codec. Following the proposed method, we can reach an encoding rate of 365 bits per second (bps) while being superior to the baseline methods by a significant margin, including lightweight and heavyweight codecs.

**Our Contribution:** (i) We demonstrate the usage of discrete speech units, learned in a self-supervised manner, for high-quality synthesis purposes (no Mel-spectrogram estimation); (ii) We provide an extensive evaluation of the SSL speech units from a synthesis point of view, i.e., signal reconstruction, voice conversion, and F0 manipulation; (iii) We build an ultra-lightweight speech codec from the obtained speech units.

## 2. Related Work

**Unsupervised Speech Representation Learning** Studies on unsupervised speech representation learning can roughly be di-

vided into reconstruction and self-supervised learning methods. Auto-encoding is the common approach for signal reconstruction, where speech is first encoded into a low-dimensional latent representation, and then decoded back to speech. Various constraints can be imposed on the encoded space, such as temporal smoothness [10], discreteness [11], and hierarchy [12].

SSL methods have shown remarkable results for ASR [2, 3], phoneme segmentation [13], and GSLM [5]. The authors in [1, 2] suggested training a convolutional neural network to distinguish true future samples from random distractor samples using a Contrastive Predictive Coding (CPC) loss function. Similar to CPC, the authors in [3] use an encoder and a predictor, which is trained contrastively to distinguish positive and negative samples. However, unlike [2] it discretizes and masks segments of the encoder’s output. In HuBERT [4], the model is trained with a masked prediction task similar to BERT [14] but with masked continuous audio signals.

**Speech Resynthesis** Recent advancements in neural-based vocoders enabled generating natural-sounding speech and music [7, 15, 9]. These are often conditioned on the log Mel-spectrogram for the generation process. Recently, unsupervised learning of low bitrate speech representations was explored under the Zero-Resource Challenge [16, 17]. Vector-Quantized Variational Auto-Encoder (VQ-VAE) [11] employs a learned fixed-sized codebook, decoded by a WaveNet model for speech synthesis. In [18] the authors proposed a VQ-VAE model followed by an FFTNet vocoder model [19]. The authors in [20] suggested using transformer [21] together with a VQ-VAE model for unsupervised unit discovery, and in [22] they combine vector quantization with contrastive predictive coding for acoustic unit discovery. The study by [23] is the closest to our work. In which, the authors suggest training a VQ-VAE with two encoders, one for the waveform and the other for F0. The authors demonstrated how such modeling improves overall generation quality. In contrast, we study SSL-based speech encoders and empirically show these representations are better disentangled, and apply them as an ultra-low bitrate speech codec. Another line of work suggests using intermediate representations obtained from an ASR acoustic model. These representations are being used together with the identity and prosodic information for voice conversion [24, 25, 26]. Unlike all of the above, we suggest synthesizing speech directly from the discrete units. Moreover, the resynthesis process sheds light on the encoded information in each of the evaluated representations.

**Speech Codec** Speech codecs typically employ a carefully hand-engineered pipeline combining an encoder and a decoder which is influenced by speech production physics to remove redundancies in the data and yield a compact bitstream. Low bitrate parametric speech codecs have long been studied [27], but their quality has been severely limited. Despite some advances [28, 29], modeling the excitation signal has remained a challenge. Neural speech codecs have been recently proposed and demonstrated promising results [30, 31]. In [32] an LPC-Net [33] vocoder was conditioned on hand-crafted features and a uniform quantizer. In [34] a WaveNet model was conditioned on discrete units obtained from a VQ-VAE model, while in [35] the Opus codec [36] was fed to WaveNet.

### 3. Method

The proposed architecture is comprised of three pre-trained and fixed encoders, namely: (i) content encoder; (ii) F0 encoder; (iii) speaker identity encoder; and a decoder network. The first two encoders extract discrete representation from the raw au-

dio while the latter extracts a single global representation. The overall architecture is depicted in Figure 1.

#### 3.1. Encoders

Denote the domain of audio samples by  $\mathcal{X} \subset \mathbb{R}$ . The representation for a raw signal is therefore a sequence of samples  $\mathbf{x} = (x_1, \dots, x_T)$ , where  $x_t \in \mathcal{X}$  for all  $1 \leq t \leq T$ .

**Content Encoder** The input to a content encoder network,  $E_c$ , is a speech utterance,  $\mathbf{x}$ , and the output is a sequence of spectral representations sampled at a low frequency as follows  $E_c(\mathbf{x}) = (\mathbf{v}_1, \dots, \mathbf{v}_{T'})$ . We evaluated three state-of-the-art unsupervised representation learning functions as  $E_c$ . Specifically, we experimented with: (i) CPC [1] which attempts to predict the future states of the encoder based on the past and optimizes a contrastive loss comparing the actual future from that of random sequences; (ii) HuBERT [4] which was trained with a masked prediction task similar to BERT [14] on masked continuous audio signals as inputs. The targets are obtained through clustering of raw speech features or learned features from earlier iterations; (iii) and VQ-VAE [11] which performs similarly to a Variational Auto Encoder [37] where the encoder’s output is discrete rather than continuous.

Since the representations learned by CPC and HuBERT are continuous, a k-means algorithm is applied over the models’ outputs to generate discrete units, denoted as  $\mathbf{z}_c = (z_1, \dots, z_L)$ . Each element  $z_i$  in  $\mathbf{z}_c$  is a positive integer,  $z_i \in \{0, 1, \dots, K\}$  for  $1 \leq i \leq L$ , where  $K$  is the number of discrete units. We did not follow the same approach for VQ-VAE as its representations are already quantized.

**F0 Encoder** To generate low frequency discrete F0 representation,  $\mathbf{z}_{F_0} = (z_1, \dots, z_{L'})$ , a separate encoder,  $E_{F_0}$ , is applied over the F0 extracted from the input signal. Each element in  $\mathbf{z}_{F_0}$  is an integer  $z_s \in \{0, 1, \dots, K'\}$ , where  $K'$  is the encoder dictionary size. The YAAPT [38] algorithm is used to extract the F0 from the input signal,  $\mathbf{x}$ , generating  $\mathbf{p} = (p_1, \dots, p_{T'})$ .

$E_{F_0}$  is trained using the VQ-VAE framework. VQ-VAE employs a convolutional encoder,  $E_{F_0}$ , a bottleneck with a learned codebook  $C = (\mathbf{e}_1, \dots, \mathbf{e}_{K'})$ , where each item in  $C$  is a 128-dimensional vector, and a decoder  $D_{F_0}$ . The encoder extracts a sequence of latent vectors  $E_{F_0}(\mathbf{p}) = (\mathbf{h}_1, \dots, \mathbf{h}_{L'})$  from the raw audio, where  $\mathbf{h}_i \in \mathbb{R}^{128}$ , for all  $1 \leq i \leq L'$ . Then, the bottleneck maps each latent vector to its nearest vector in the codebook  $C$ . The embedded latent vectors are then being fed into the decoder  $D_{F_0}(\mathbf{e}_{z_1}, \dots, \mathbf{e}_{z_{L'}}) = \hat{\mathbf{p}}$  which reconstructs the original F0 signal. Similar to [39], we use Exponential Moving Average updates to learn the codebook and employ random restarts for unused embeddings.

To generate  $\mathbf{z}_{F_0}$ , we use the indices of the mapped latent vectors rather than the vectors.

**Speaker Encoder** Lastly, a speaker encoder,  $E_{spk}$ , is used to extract speaker embedding. A pre-trained speaker verification model similar to the one proposed in [40] is used. Formally,  $E_{spk}$  gets as input the speech utterance,  $\mathbf{x}$ , extracts the Mel-spectrogram and outputs a  $d$ -vector speaker representation, denoted as  $\mathbf{z}_{spk} \in \mathbb{R}^{256}$ . We experimented with learning the speaker embedding via a lookup-table. Although such method performs slightly better, it is limited to speakers seen during training.

#### 3.2. Decoder

A neural vocoder is employed to decode the speech signal from the discrete representation. This study considers the decoder to be a modified version of the HiFi-GAN [9] neural vocoder.

The HiFi-GAN architecture is comprised of a generator,  $G$ ,

Table 1: *Speech resynthesis results. MOS results are reported with 95% confidence interval. Except MOS for all other evaluation metrics the lower is the better. We did not report EER for LJ since it is a single speaker dataset.*

Dataset	Method	Content		F0		Speaker	Overall Quality
		PER ↓	WER ↓	VDE ↓	FFE ↓	EER ↓	MOS ↑
LJ	GT	6.93	5.60	–	–	–	4.33±0.20
	CPC	9.66	8.51	13.48	15.19	–	3.31±0.33
	HuBERT	<b>9.52</b>	<b>6.96</b>	13.09	15.00	–	<b>3.66±0.33</b>
	VQ-VAE	12.77	8.85	<b>7.19</b>	<b>8.54</b>	–	<b>3.66±0.31</b>
VCTK	GT	17.16	4.32	–	–	3.25	4.08±0.66
	CPC	23.01	14.49	10.56	11.13	<b>4.25</b>	3.33±0.61
	HuBERT	<b>19.66</b>	<b>11.44</b>	9.77	10.43	5.79	<b>3.41±0.66</b>
	VQ-VAE	31.97	19.80	<b>5.20</b>	<b>5.59</b>	4.28	3.39±0.58

and a set of discriminators,  $D$ . The generator is built from a set of look-up tables (LUT) that embed the discrete representation and a series of blocks composed of transposed convolution and a residual block with dilated layers. The transposed convolutions upsample the encoded representation to match the input sample rate, while the dilated layers increase the receptive field.

As an input, the generator receives the encoded representation ( $\mathbf{z}_c, \mathbf{z}_{F_0}, \mathbf{z}_{spk}$ ). The discrete content sequence,  $\mathbf{z}_c$ , and the discrete pitch sequence,  $\mathbf{z}_{F_0}$ , are converted to a continuous representation via  $LUT_c$  and  $LUT_{F_0}$  accordingly. The sequences are up-sampled and concatenated together. The speaker embedding,  $\mathbf{z}_{spk}$ , is concatenated to each frame in the up-sampled sequence.

The discriminator comprises two networks, a Multi-Period Discriminator (MPD) and a Multi-Scale Discriminator (MSD). The MPD consists of multiple sub-discriminators operating on equally spaced samples from the input signal. The period sub-discriminators differ from each other based on the space between the samples. Similar to [9], the MPD employs a total of five-period discriminators with a period hops of [2, 3, 5, 7, 11]. Multi-scale discriminator (MSD) employs multiple sub-discriminators operating at different scales of the input signal. Specifically, we use three scales: the original input scale,  $\times 2$  downsampled scale, and  $\times 4$  downsampled scale. Overall each sub-discriminator  $D_j$  is tasked with minimizing the following,

$$L_{adv}(D_j, G) = \sum_{\mathbf{x}} \|1 - D_j(\hat{\mathbf{x}})\|_2^2, \quad (1)$$

$$L_D(D_j, G) = \sum_{\mathbf{x}} [\|1 - D_j(\mathbf{x})\|_2^2 + \|D_j(\hat{\mathbf{x}})\|_2^2],$$

where  $\hat{\mathbf{x}} = G(LUT_c(\mathbf{z}_c), LUT_{F_0}(\mathbf{z}_{F_0}), \mathbf{z}_{spk})$ , is the resynthesized signal from the encoded representation.

Additionally, two terms are added to the loss function. The first one is a reconstruction term computed between the Mel-spectrogram of the input signal and the generated signal,

$$L_{recon}(G) = \sum_{\mathbf{x}} \|\phi(\mathbf{x}) - \phi(\hat{\mathbf{x}})\|_1, \quad (2)$$

where  $\phi$  is a spectral operator computing Mel-spectrogram. The second term is a feature-matching loss [41] which measures the distance between discriminator activations of the real signal and those of the resynthesized signal,

$$L_{fm}(D_j, G) = \sum_{\mathbf{x}} \sum_{i=1}^R \frac{1}{M_i} \|\psi_i(\mathbf{x}) - \psi_i(\hat{\mathbf{x}})\|_1, \quad (3)$$

Table 2: *Voice conversion & F0 manipulation results. MOS results are reported with 95% confidence interval. VDE, and FFE are reported for F0 manipulation while PER, WER, EER, and MOS are reported for voice conversion. Notice, for VDE, and FFE higher is the better since F0 was flattened.*

Dataset	Method	Voice Conversion				F0 Manipulation	
		PER ↓	WER ↓	EER ↓	MOS ↑	VDE ↑	FFE ↑
VCTK	GT	17.16	4.32	3.25	4.11±0.29	–	–
LJ	CPC	22.22	16.11	0.46	3.57±0.15	<b>46.68</b>	<b>48.71</b>
	HuBERT	<b>19.09</b>	<b>12.23</b>	<b>0.31</b>	<b>3.71±0.24</b>	39.20	48.42
	VQ-VAE	40.88	36.96	9.65	2.90±0.17	10.54	12.08
VCTK	CPC	23.58	15.98	<b>4.83</b>	3.42 ± 0.24	<b>25.29</b>	<b>26.97</b>
	HuBERT	<b>20.85</b>	<b>12.72</b>	6.01	<b>3.58 ± 0.28</b>	23.46	26.67
	VQ-VAE	36.88	29.44	11.56	3.08 ± 0.34	7.03	7.80

where  $\psi_i$  is an operator which extracts the activations of the discriminator  $i$ -th layer,  $M_i$  is the number of features in layer  $i$ , and  $R$  is the total number of layers in  $D_j$ .

The final loss with respect to the sub-discriminators composing discriminator  $D$  and generator  $G$  is:

$$L_G^{multi}(D, G) = \sum_{j=1}^J [L_{adv}(G, D_j) + \lambda_{fm} L_{fm}(G, D_j)],$$

$$+ \lambda_r L_{recon}(G),$$

$$L_D^{multi}(D, G) = \sum_{j=1}^J L_D(G, D_j), \quad (4)$$

where we set  $\lambda_{fm} = 2$  and  $\lambda_r = 45$ .

## 4. Results

Our results cover three different settings: (i) speech reconstruction experiments; (ii) speaker conversion and F0 manipulation; (iii) bitrate analysis with subjective tests for speech codec evaluation. We employ two datasets: LJ [42] single speaker dataset and VCTK [43] multi-speaker dataset. All datasets were resampled to a 16kHz sample rate.

**Implementation Details** We follow the same setup as in [5]. For CPC, we used the model from [44], which was trained on a “clean” 6k hour sub-sample of the LibriLight dataset [45, 44]. We extract a downsampled representation from an intermediate layer with a 256-dimensional embedding and a hop size of 160 audio samples. For HuBERT we used a BASE 12 transformer-layer model trained for two iterations [4] on 960 hours of LibriSpeech corpus [46]. This model downsamples the raw audio  $\times 320$  into a sequence of 768-dimensional vectors. Similarly to [5], activations were extracted from the sixth layer.

For CPC and HuBERT, the k-means algorithm is trained on LibriSpeech clean-100h [46] dataset to convert continuous frames to discrete codes. We quantize both learned representations with  $K = 100$  centroids. Leading to a bitrate of 700bps for CPC and 350bps for HuBERT.

Similarly to CPC models, we trained the VQ-VAE content encoder model on the “clean” 6K hours subset from the LibriLight dataset. We use an encoder operating on the raw signal to extract discrete units, similar to [39]. In addition, “random restarts” were performed when the mean usage of a codebook vector fell below a predetermined threshold. Finally, we used HiFiGAN (architecture and objective) as the decoder instead of a simple convolutional decoder, as it improved the overall audio quality. This model encodes the raw audio into a sequence of discrete tokens from 256 possible tokens [34] with a hop size of 160 raw audio samples. The VQ-VAE discrete code operates

at a bitrate of 800bps. We additionally experimented with 100 discrete units for VQ-VAE, however results were the best for 256. This finding is consistent with [34].

The speaker verification network uses the architecture proposed in [40]. It was trained on the VoxCeleb2 [47] dataset, achieving a 7.4% Equal Error Rate (EER) for speaker verification on the test split of the VoxCeleb1 [48] dataset.

Only a single F0 representation is considered across all evaluated models, trained on the VCTK dataset. The F0 is extracted from the raw audio using a window size of 20ms and a 5ms hop. As a result, the F0 sequence is sampled at 200Hz. The quantization described at Sec. 3, is applied using an F0 codebook of  $K' = 20$  tokens and an encoder that downsamples the signal by  $\times 16$ . Hence, the discrete F0 representation is sampled at 12.5Hz, leading to a bitrate of 65bps. The final bitrate of the evaluated codecs is the sum of the pitch code bitrate with the content code bitrate.

**Evaluation Metrics** We consider both subjective and objective evaluation metrics. For subjective tests, we report the Mean Opinion Scores (MOS). In which human evaluators rate the naturalness of audio samples on a scale of 1–5. Each experiment, included 50 randomly selected samples rated by 30 raters. For objective evaluation, we consider: (i) Equal Error Rate (EER) as an automatic speaker verification metric obtained using a pre-trained speaker verification network. We report EER between test utterances and enrolled speakers; (ii) Voicing Decision Error (VDE) [49], which measures the portion of frames with voicing decision error; (iii) F0 Frame Error (FFE) [50], measures the percentage of frames that contain a deviation of more than 20% in pitch value or have a voicing decision error; (iv) Word Error Rate (WER) and Phoneme Error Rate (PER), proxy metrics to the intelligibility of the generated audio. We used a pre-trained ASR network [3] on both reconstructed and converted samples to calculate both metrics.

**Reconstruction & Conversion** We start by reporting the reconstruction performance. Results are summarized in Table 1. When considering the intelligibility of the reconstructed signal HuBERT reaches the lowest PER and WER scores across all models, where both CPC and HuBERT are superior to VQ-VAE. However, when considering F0 reconstruction VQ-VAE outperforms both HuBERT and CPC by a significant margin. This results are somewhat intuitive, bearing in mind VQ-VAE objective is to fully reconstruct the input signal. In terms of subjective evaluation, all models reach similar MOS scores, with one exception of CPC on LJ.

To better evaluate the disentanglement properties of each method with respect to speaker identity and F0, we conducted an additional set of experiments aiming at speaker conversion and F0 manipulation. For voice conversion, we converted each test utterance into five random target speakers. Next, we employed a speaker verification network, which extracts  $d$ -vector representation to evaluate speaker-converted utterances’ similarity to real speaker utterances (low error-rate indicates good conversion), providing measurement to the speaker identity’s disentanglement from the evaluated coding method. The error-rate is reported between converted test utterances and enrolled speakers. For the LJ speech single speaker dataset, we converted samples from the VCTK dataset to the single speaker and enrolled all VCTK speakers together with the single speaker. Results are summarized in Table 2 (left). Unlike resynthesis results, on voice conversion CPC and HuBERT outperform VQ-VAE on both LJ and VCTK datasets, indicating VQ-VAE contains more information about the speaker in the encoded units, hence producing more artifacts. Notice, this also affects WER,

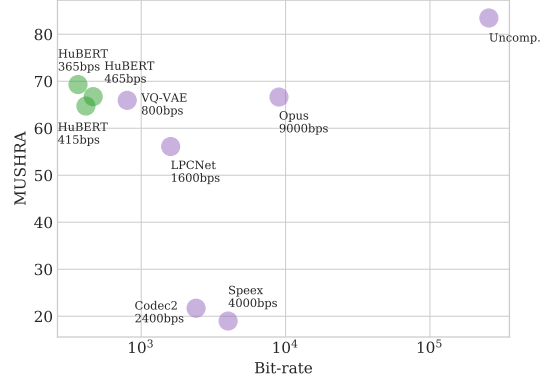


Figure 2: *MUSHRA subjective quality results as a function of bitrate per second. Purple dots denote the baseline methods, and green dots the proposed SSL based method.*

PER, and the overall subjective quality (MOS).

Next, to evaluate the presence of F0 in the discrete units, we flattened the F0 units before synthesizing the signal and calculated VDE and FFE with respect to the original F0 values. F0 flattening was done by setting the speakers’ mean F0 value across all voiced frames. In this experiment, we expected units that contain F0 information to be better at F0 reconstruction over disentangled units. Results are summarized in Table 2 (right). Notice VQ-VAE can still reconstruct the F0 almost at the same level as when using the original F0 as conditioning (5.2 vs 7.03, and 5.59 vs 7.8), in contrast to CPC and HuBERT.

**Speech Codec** Our final experiment evaluates the obtained speech units as a low bitrate speech codec. We use a subjective MUSHRA-type listening test [51] to measure the perceived quality of the proposed speech codec with regard to its bitrate constraints. In MUSHRA evaluations, listeners are presented with a labeled uncompressed signal for reference, a set of test samples to rate, a copy of the uncompressed reference, and a low-quality anchor. Listeners are asked to rate each test utterance and the copy of the uncompressed reference with respect to the labeled reference in a scale of 1-100.

The experiment is performed on the VCTK dataset [43]. For evaluation, we used 20 utterances from 5 speakers. The set of speakers in the test data is disjoint with those in the training data. For this experiment, HuBERT models with 50, 100, and 200 units were trained as described in Sec. 4. For comparison, we included other speech codecs in our evaluation: Opus [36] wideband at 9 kbps VBR, Codec2 [52] at 2.4 kbps and LPCNet [32] operating at 1.6 kbps. The LPCNet model was trained from scratch on the VCTK dataset following the experimental setup in [32]. The VQ-VAE model employs the HiFiGAN decoder trained on the LibriLight dataset to match the amount of data reported in [34]. We compressed the anchor sample with Speex [53] at 4 kbps as a low anchor. Fig. 2 depicts the results. HuBERT with 50 units reaches the best MUSHRA score while its bitrate is only 365bps, which is significantly lower than the baseline methods.

## 5. Conclusion

We applied self-supervised discrete representations for the task of speech resynthesis. Furthermore, we demonstrated the efficiency of disentangled representations for signal reconstruction, voice conversion, and F0 manipulation. Our evaluations shed light on the properties encoded by each method in the context of speech synthesis. Finally, we adapt the HuBERT speech representation as an ultra lightweight speech codec, providing superior subjective results than the baselines with lower bitrate.

## 6. References

- [1] A. van den Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” *arXiv preprint arXiv:1807.03748*, 2018.
- [2] S. Schneider, A. Baevski, R. Collobert, and M. Auli, “wav2vec: Unsupervised Pre-Training for Speech Recognition,” in *INTER-SPEECH*, 2019.
- [3] A. Baevski *et al.*, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” in *ICLR*, 2020.
- [4] W.-N. Hsu *et al.*, “Hubert: How much can a bad teacher benefit ASR pre-training?” in *NeurIPS Workshop on Self-Supervised Learning for Speech and Audio Processing Workshop*, 2020.
- [5] K. Lakhota *et al.*, “Generative spoken language modeling from raw audio,” *arXiv preprint arXiv:2102.01192*, 2021.
- [6] D. Griffin and J. Lim, “Signal estimation from modified short-time fourier transform,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.
- [7] A. v. d. Oord *et al.*, “Wavenet: A generative model for raw audio,” *arXiv preprint arXiv:1609.03499*, 2016.
- [8] R. Prenger, R. Valle, and B. Catanzaro, “Waveglow: A flow-based generative network for speech synthesis,” *ICASSP*, 2019.
- [9] J. Kong *et al.*, “Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis,” in *NeurIPS*, 2020.
- [10] J. Ebbers *et al.*, “Hidden markov model variational autoencoder for acoustic unit discovery,” in *INTERSPEECH 2017*, 2017.
- [11] A. van den Oord, O. Vinyals *et al.*, “Neural discrete representation learning,” in *NeurIPS*, 2017.
- [12] W.-N. Hsu, Y. Zhang, and J. Glass, “Unsupervised learning of disentangled and interpretable representations from sequential data,” in *Advances in Neural Information Processing Systems*, 2017.
- [13] F. Kreuk, J. Keshet, and Y. Adi, “Self-supervised contrastive learning for unsupervised phoneme segmentation,” *arXiv preprint arXiv:2007.13465*, 2020.
- [14] J. Devlin *et al.*, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *NAACL*, 2019.
- [15] R. Prenger, R. Valle, and B. Catanzaro, “Waveglow: A flow-based generative network for speech synthesis,” in *ICASSP*, 2019.
- [16] E. Dunbar *et al.*, “The Zero Resource Speech Challenge 2019: TTS Without T,” in *Proc. Interspeech 2019*, 2019, pp. 1088–1092.
- [17] —, “The Zero Resource Speech Challenge 2020: Discovering Discrete Subword and Word Units,” in *Proc. Interspeech 2020*, 2020, pp. 4831–4835.
- [18] R. Eloff *et al.*, “Unsupervised Acoustic Unit Discovery for Speech Synthesis Using Discrete Latent-Variable Neural Networks,” in *INTERSPEECH*, 2019.
- [19] Z. Jin, A. Finkelstein, G. J. Mysore, and J. Lu, “Fftnet: A real-time speaker-dependent neural vocoder,” in *ICASSP*, 2018.
- [20] A. Tjandra, S. Sakti, and S. Nakamura, “Transformer VQ-VAE for Unsupervised Unit Discovery and Speech Synthesis: ZeroSpeech 2020 Challenge,” in *INTERSPEECH*, 2020.
- [21] A. Vaswani *et al.*, “Attention is all you need,” in *NeurIPS*, 2017.
- [22] B. van Niekerk, L. Nortje, and H. Kamper, “Vector-Quantized Neural Networks for Acoustic Unit Discovery in the ZeroSpeech 2020 Challenge,” in *INTERSPEECH*, 2020.
- [23] Y. Zhao *et al.*, “Improved prosody from learned f0 codebook representations for vq-vae speech waveform reconstruction,” *arXiv preprint arXiv:2005.07884*, 2020.
- [24] A. Polyak, L. Wolf, and Y. Taigman, “TTS Skins: Speaker Conversion via ASR,” in *INTERSPEECH*, 2020.
- [25] A. Polyak *et al.*, “Unsupervised Cross-Domain Singing Voice Conversion,” in *INTERSPEECH*, 2020.
- [26] —, “High fidelity speech regeneration with application to speech enhancement,” *ICASSP*, 2021.
- [27] B. S. Atal and S. L. Hanauer, “Speech analysis and synthesis by linear prediction of the speech wave,” *The journal of the acoustical society of America*, vol. 50, no. 2B, pp. 637–655, 1971.
- [28] D. Griffin and J. Lim, “A new model-based speech analysis/synthesis system,” in *ICASSP*, 1985.
- [29] A. McCree, K. Truong, E. B. George, T. P. Barnwell, and V. Viswanathan, “A 2.4 kbit/s melp coder candidate for the new us federal standard,” in *ICASSP*, 1996.
- [30] W. B. Kleijn, F. S. Lim, A. Luebs, J. Skoglund, F. Stimberg, Q. Wang, and T. C. Walters, “Wavenet based low rate speech coding,” in *ICASSP*, 2018.
- [31] F. S. Lim *et al.*, “Robust low rate speech coding based on cloned networks and wavenet,” in *ICASSP*, 2020.
- [32] J.-M. Valin and J. Skoglund, “A real-time wideband neural vocoder at 1.6 kb/s using lpcnet,” *arXiv preprint arXiv:1903.12087*, 2019.
- [33] —, “Lpcnet: Improving neural speech synthesis through linear prediction,” in *ICASSP*, 2019.
- [34] C. Gârbaea *et al.*, “Low bit-rate speech coding with vq-vae and a wavenet decoder,” in *ICASSP*, 2019.
- [35] J. Skoglund and J.-M. Valin, “Improving opus low bit rate quality with neural speech synthesis,” *arXiv preprint arXiv:1905.04628*, 2019.
- [36] J.-M. Valin, K. Vos, and T. Terriberry, “Definition of the opus audio codec,” *IETF, September*, 2012.
- [37] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *ICLR*, 2014.
- [38] K. Kasi and S. A. Zahorian, “Yet another algorithm for pitch tracking,” *ICASSP*, 2002.
- [39] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, “Jukebox: A generative model for music,” *arXiv preprint arXiv:2005.00341*, 2020.
- [40] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, “End-to-end text-dependent speaker verification,” in *ICASSP*, 2016.
- [41] A. B. L. Larsen *et al.*, “Autoencoding beyond pixels using a learned similarity metric,” in *ICML*, 2016.
- [42] K. Ito and L. Johnson, “The lj speech dataset,” <https://keithito.com/LJ-Speech-Dataset/>, 2017.
- [43] C. Veaux *et al.*, “CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit,” 2017.
- [44] M. Rivière and E. Dupoux, “Towards unsupervised learning of speech features in the wild,” in *SLT 2020: IEEE Spoken Language Technology Workshop*, 2020.
- [45] J. Kahn *et al.*, “Libri-light: A benchmark for asr with limited or no supervision,” in *ICASSP*, 2020.
- [46] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: an asr corpus based on public domain audio books,” in *ICASSP*, 2015.
- [47] J. S. Chung, A. Nagrani, and A. Zisserman, “Voxceleb2: Deep speaker recognition,” *INTERSPEECH*, 2018.
- [48] A. Nagrani, J. S. Chung, and A. Zisserman, “Voxceleb: a large-scale speaker identification dataset,” *INTERSPEECH*, 2017.
- [49] T. Nakatani *et al.*, “A method for fundamental frequency estimation and voicing decision: Application to infant utterances recorded in real acoustical environments,” *Speech Communication*, 2008.
- [50] W. Chu and A. Alwan, “Reducing f0 frame error of f0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend,” *ICASSP*, 2009.
- [51] B. Series, “Method for the subjective assessment of intermediate quality level of audio systems,” *International Telecommunication Union Radiocommunication Assembly*, 2014.
- [52] D. Rowe, “Codec 2-open source speech coding at 2400 bits/s and below,” in *TAPR and ARRL 30th Digital Communications Conference*, 2011, pp. 80–84.
- [53] J.-M. Valin, “Speex: A free codec for free speech,” *arXiv preprint arXiv:1602.08668*, 2016.