Bringing Interpretability to Neural Audio Codecs

Samir Sadok[1,*], *Julien Hauret*[2,3,*], *Éric Bavu*[2]

[1]Inria, Université Grenoble Alpes CNRS, LJK, France
[2]LMSSC, Conservatoire national des arts et métiers, Paris, France
[3]APC, French-German Research Institute of Saint-Louis, France

[*]Equal contribution

samir.sadok@inria.fr, julien.hauret@lecnam.net, eric.bavu@lecnam.net

Abstract

The advent of neural audio codecs has increased in popularity due to their potential for efficiently modeling audio with transformers. Such advanced codecs represent audio from a highly continuous waveform to low-sampled discrete units. In contrast to semantic units, acoustic units may lack interpretability because their training objectives primarily focus on reconstruction performance. This paper proposes a two-step approach to explore the encoding of speech information within the codec tokens. The primary goal of the analysis stage is to gain deeper insight into how speech attributes such as content, identity, and pitch are encoded. The synthesis stage then trains an AnCoGen network for post-hoc explanation of codecs to extract speech attributes from the respective tokens directly.

Index Terms: Interpretability, Neural Audio Codec

1. Introduction

In the era of deep learning, audio representations are usually built using self-supervised learning. However, there is a notable distinction in whether the task is generating or understanding audio. In speech understanding, notable works such as HuBERT [1] and WavLM [2] employ a strategy of leveraging a transformer network's partially masked output for high-level speech comprehension and semantic token generation. Another category of representations has emerged with Soundstream [3] and Encodec [4], with acoustic tokens produced by neural codecs primarily created for audio compression. These acoustic tokens have since become the de facto building blocks of generation tasks, as they retain all the necessary acoustic details for constructing high-quality audio, unlike semantic tokens. The present paper proposes a study of four contemporary codecs capable of compressing and decompressing audio while maintaining a high level of quality. Each of these codecs possesses a distinctive feature :

- DAC [5]: The first codec to exceed Encodec and Soundsteam in terms of quality, facilitated by two pivotal techniques: factorized codes and L2-normalized codes, which enhance codebook usage. The model is trained with adversarial loss and feature matching on discriminator embeddings while learning codebooks like the original VQ-VAE formulation [6]. Additionally, the classic residual vector quantization (RVQ) is employed to derive a two-dimensional discrete representation of the audio data.
- SpeechTokenizer [7] enhances the DAC framework by introducing a hierarchical disentanglement of speech information within its RVQ structure. A semantic teacher, such as HuBERT, enables the first RVQ layer to focus on phonetic content, while subsequent layers capture paralinguistic features such as timbre and prosody. This design ensures strong phonetic discriminability, as reflected in a superior Phone-Normalized Mutual Information compared to systems such as Encodec.
- Mimi [8] also employs distillation techniques, utilising a split RVQ to circumvent semantic leakage at higher RVQ scales. Additionally, it integrates a transformer network into both the encoder and decoder, while maintaining full causality. It has been demonstrated to outperform both SpeechTokenizer and DAC regarding quality and bitrate.
- Bigcodec [9] is distinct from previous codecs in that it employs a single VQ with a higher cardinality and sampling rate, yielding a similar bitrate as Mimi. The quality of Bigcodec is aligned with the standards of Mimi, with a level of preference comparable to that of the reference signal.

The following contributions constitute the core of the present study. The first experiment explores the deterministic mapping between acoustic and semantic tokens. Then, two codecs have been selected for an analysis phase using a pre-trained AnCoGen-Melspectrogram [10] to determine how pitch, speaker identity, and linguistic content are encoded. Finally, two AnCoGen-Codec are trained as a plugin to enable direct prediction of speech attributes from tokens, facilitating both a deeper understanding and manipulation within the token space.

2. Analysis

As illustrated in Figure 1, the analysis stage aims to delineate the coding of linguistic content, speaker identity, and pitch into neural codecs. To this end, we propose a framework in which AnCoGen-Melspectrogram [10] predicts speech attributes (e.g., content and pitch) in parallel with codecs computing acoustic tokens. When the sampling rates of speech attributes and codecs differ, we eliminate some tokens from the longest sequence while maintaining time alignment to align the sequence length. In the following, the cardinality of the tokens of a given codec is denoted by N_{codec}. When a codec involves M RVQ scales, it consequently exhibits $(N_{\text{codec}})^M$ possible code combinations for a given timestep, provided that the same cardinality is maintained across each RVQ scale.

2.1. Content

Mapping between Hubert and Codec tokens: A crucial preliminary inquiry pertains to the deterministic nature of the mapping of codec tokens on linguistic content, as represented by the $N_{\text{HuBERT}} = 100$ tokens. For this purpose, the Librispeech clean test set [11] was tokenized, and the co-occurrences of acoustic tokens on every RVQ scale and Hubert tokens were observed.

[*]These authors contributed equally.

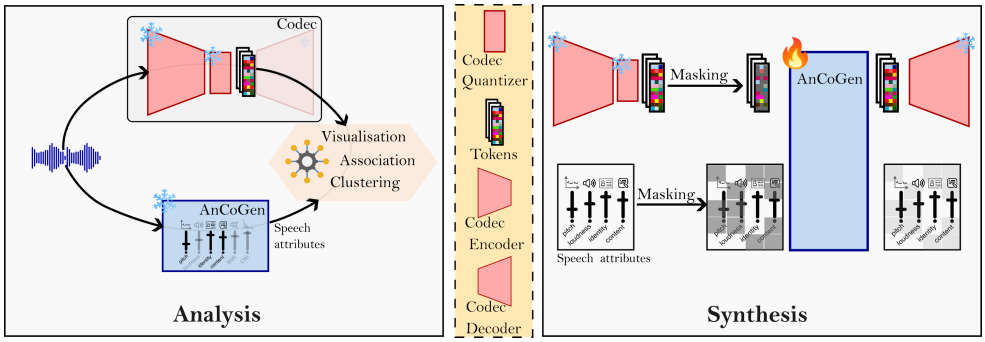


Figure 1: ***Analysis (left):*** *Links between codec tokens and speech attributes are observed by performing two parallel forwards from the waveform.* ***Synthesis (right):*** *Once trained, AnCoGen model directly predicts speech attributes from codec tokens and vice-versa*

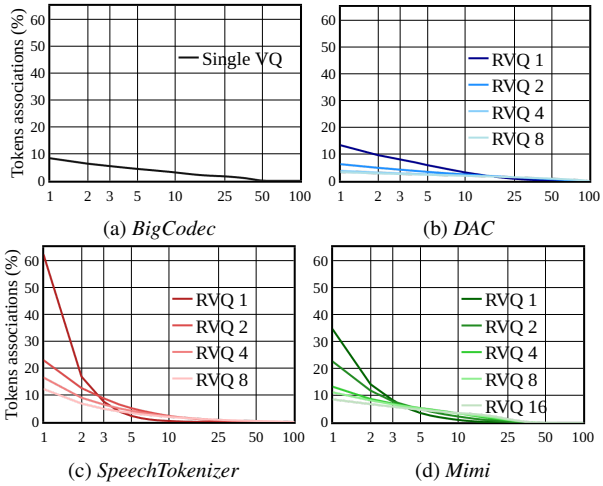


Figure 2: *HuBERT/Codecs tokens associations on LibriSpeech test-clean : top-k HuBERT tokens usage per codec token/scale*

For each scale of each codec, a matrix of size $N_{\text{HuBERT}} \times N_{\text{codec}}$ is thus obtained. From this matrix, HuBERT tokens are ranked for each codec token based on their level of association, from highest to lowest. The results are then averaged for each codec and scale, expressed as percentages, and visualized in Figure 2.

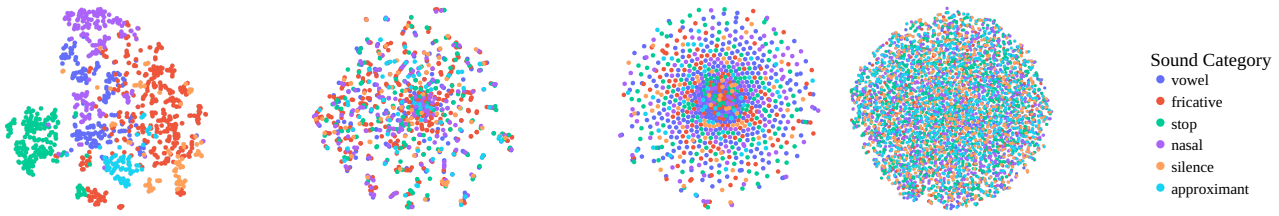
As might be anticipated, a purely deterministic mapping of token (*i.e.*, when a given code is systematically mapped to the same HuBERT token) is improbable due to the receptive fields of networks. Indeed, codes also gain meanings thanks to their surroundings, just as words in a sentence are not always translated to the same word in another language, depending on the context. Nevertheless, analyzing the basic one-to-one correspondence between acoustic and semantic tokens reveals several noteworthy properties. First, the single VQ scale of BigCodec exhibits a mean mapping of its token to its most associated HuBERT token of 8%, with a long-tail distribution that results in a suboptimal mapping. This result can be justified by the fact that the codec employs a single scale and was not trained to align with HuBERT. A subsequent examination of DAC shows that the initial RVQ scale correlates more consistently with HuBERT. As scales increase, the mapping gets less precise, suggesting most linguistic content is encoded in the first scale. The first scale of SpeechTokenizer exhibits a high degree of alignment with HuBERT tokens, which is to be expected given that Zhang *et al.* [7] distilled the HuBERT embeddings into this scale. All residual scales also display good associations, indicating they also contain linguistic content. At last, although Mimi introduced the splitRVQ tokenizer to restrict this linguistic leakage into higher RVQ scales, but the plot demonstrates that the larger RVQ scales still contain a considerable amount of linguistic content. This may be attributed to the extremely low frequency of Mimi, 12.5 Hz, which encompasses more HuBERT tokens.

Embeddings clustering: in this analysis, our focus is narrowed to BigCodec and SpeechTokenizer, which present markedly contrasting association schemes. We posit a strong association between a specific codec token and its predominant HuBERT token. This assumption enables a two-component t-SNE [12] visualization of learned token embeddings in the high-dimensional code space, as in Figure 3 of [13]. The resulting t-SNE visualization is depicted in Figure 3, with the same perplexity applied to all graphs. The sound categories delineated in the dendogram of sound classification are presented in [14]. This analysis enables us to assess the structural organization of the embedding space in relation to the linguistic content.

As illustrated by the Figure 3a plot, there is a clear delineation of clusters, indicative of the precise mapping between the initial scale of the SpeechTokenizer and HuBERT semantic tokens. In contrast, the Figures 3b and 3c demonstrate an absence of this behavior. The second RVQ scale exhibits some clusters, though they do not map to linguistic content. The eighth RVQ scale, on the other hand, appears much more diffuse, lacking any discernible structure. The BigCodec Figure 3d exhibits a comparable behavior to the SpeechTokenizer second RVQ scale. It should be noted, however, that $N_{\text{BigCodec}} > N_{\text{SpeechTokenizer}}$. This results in a greater number of clusters that do not align effectively with semantic content.

2.2. Identity

We also studied SpeechTokenizer and BigCodec with a two-component t-SNE visualization to examine how speaker identity is encoded. For this purpose, a subset of 720 utterances and 24 speakers from the LibriSpeech test clean corpus was tokenized, and the resulting embeddings were averaged on a per-utterance basis over time. A colormap depicting speaker identity was employed to obtain Figure 4. Analysis of this figure reveals that speaker identity information is barely present in the first RVQ scale of SpeechTokenizer, is highly prevalent on the fourth scale, and is still present on the final scale, albeit with more dispersed clusters. For BigCodec, clustering is also observed, although some data points are outside the clusters. Fi-

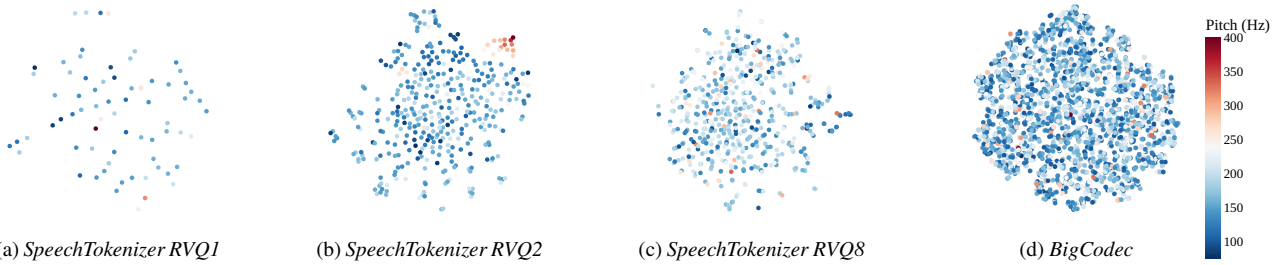


(a) *SpeechTokenizer RVQ1* (b) *SpeechTokenizer RVQ2* (c) *SpeechTokenizer RVQ8* (d) *BigCodec*

Figure 3: *t-SNE visualisation. Color is attributed via codec-to-HuBERT and HuBERT-to-sound mappings.*

(a) *SpeechTokenizer RVQ1* (b) *SpeechTokenizer RVQ4* (c) *SpeechTokenizer RVQ8* (d) *BigCodec*

Figure 4: *t-SNE visualisation of identity. Each dot corresponds to one utterance averaged over time, each color to one speaker.*



(a) *SpeechTokenizer RVQ1* (b) *SpeechTokenizer RVQ2* (c) *SpeechTokenizer RVQ8* (d) *BigCodec*

Figure 5: *t-SNE visualisation of pitch. Each dot corresponds to a token associated at least once with a specific (vowel) HuBERT token. The legend shows the mean pitch of all associations.*

nally, it is notable that clusters that are neighbors on one scale tend to maintain proximity on other scales or codecs.

2.3. Pitch

To analyze pitch structuration in the embedding space, the tokenized Librispeech test-clean dataset was filtered to retain only codec tokens associated with the 39th HuBERT token, which has high occurrence counts. This HuBERT token is identified as a vowel in [14]. Pitch labels were computed as the mean pitch of these associations, with results shown in Figure 5. Pitch standard deviations are not displayed but are significant (approx. 40 Hz). Note that the number of points increases with the RVQ scale because the associations become less precise, as explained in Section 2.1. The analysis of Figure 5 shows that the embeddings lack interpretability in terms of pitch, with the only visible feature being a small cluster of high pitch values on the second RVQ scale of SpeechTokenizer (Figure 5b). This highlights that pitch, being based on human auditory perception rather than the refinement of RVQ scales, is likely the most challenging speech attribute to decode from codec tokens.

2.4. Mutual information

A quantitative analysis was performed to evaluate the mutual information (MI) between SpeechTokenizer RVQ scales and speech attributes using the Contrastive Log-ratio Upper Bound (CLUB) method [15]. The resulting histogram (Figure 6) dis-

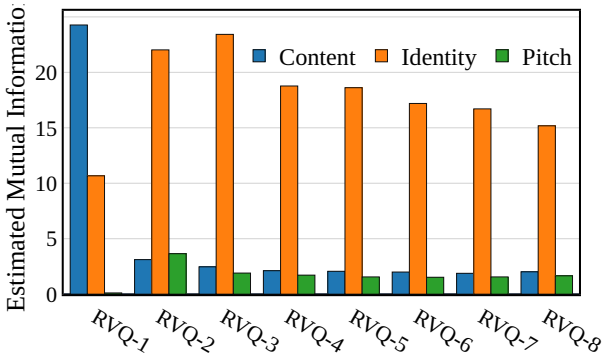


Figure 6: *Mutual information between SpeechTokenizer RVQ scales and speech attributes*

plays non-normalized MI to highlight the relative scarcity of pitch MI compared to identity and content. For all attributes, the MI peaked where clusters were most distinct in the t-SNE visualizations, corroborating our previous observations.

3. Synthesis

The previous section highlighted that key speech features—pitch, identity, and linguistic content—are entangled in the codec representation, limiting the interpretability of encoded speech signals. Only a few codecs stand out as exceptions to this trend [16, 17, 18]. Inspired by [19, 20, 10], we introduce a bidirectional framework to analyze, control, and generate speech audio from codec tokens. This approach maps codec tokens to speech attributes (analysis) and reverses the process to map attributes back to codec tokens (synthesis). As in the previous section, our study focuses on *SpeechTokenizer* and *BigCodec*, chosen for their contrasting design paradigms (see Section 1). All AnCoGen-Codec models are trained on the LibriSpeech-100-clean dataset, with performance evaluations conducted primarily on the LibriSpeech test dataset.

3.1. Method

Our AnCoGen-Codec utilizes two speech signal representations: codec tokens and four high-level attributes capturing linguistic, prosodic, and acoustic features. These attributes include *linguistic content* (from HuBERT encoder outputs), *pitch contour* (extracted using CREPE [21]), *loudness* (computed via root mean square signal), and *speaker identity* (derived from embeddings of the pre-trained ECAPA-TDNN model [22]). For further details, refer to the original AnCoGen paper [10].

During training, the input speech signal is converted into codec and speech attribute token sequences, which are partially masked following a predefined strategy. AnCoGen, leveraging an encoder-decoder Transformer, embeds visible tokens into learned vectors processed by the encoder using multi-head self-attention. Mask tokens are added, and the sequence is fed to the decoder which predicts masked token indices thanks to several linear layers for multi-scale codecs. Training optimizes the cross-entropy loss between predicted and ground-truth indices. *At inference*, AnCoGen supports both speech analysis and generation. For analysis, the codec token sequence of an input speech signal is fed to the encoder with all speech attribute tokens masked, allowing the decoder to predict the corresponding speech attributes. For generation, the encoder processes the speech attribute token sequence while all codec tokens are masked. The decoder predicts the codec tokens, which are then used to reconstruct the audio signal. Finally, the frozen codec decoder converts the reconstructed codec tokens into an audio waveform.

3.2. Results

This section discusses the results presented in Table 1 for three tasks: pitch and content estimation (analysis task), resynthesis task, and voice conversion (control task).

Speech analysis: This experiment evaluates AnCoGen's ability to estimate speech attributes, such as pitch and content, from codec token representations. Pitch estimation is evaluated on the PTDB dataset [23] with ground-truth pitch values, using the Average Absolute Error (AAE) metric. While content estimation is performed on the LibriSpeech test set, measuring accuracy across 100 classes defined by HuBERT. Both BigCodec and SpeechTokenizer tokens in AnCoGen yield strong pitch estimation results, with SpeechTokenizer showing slightly better accuracy on average. For content analysis, AnCoGen achieves its highest performance with SpeechTokenizer tokens, reaching 82% accuracy. This outcome aligns with expectations, as SpeechTokenizer incorporates HuBERT's semantic information to regularize the first RVQ scale.

Speech resynthesis: In this experiment, we use AnCoGen to map back and forth between codec tokens and speech attributes, evaluating potential information loss. We use two non-intrusive audio quality metrics: DNSMOS Background Noise Quality (BAK) [24] and Noresqa MOS (N-MOS) [25, 26], along with two intrusive metrics: Short-Time Objective Intelligibility (STOI) [27] and speechBertScore [28]. As a reference, we compare against the codec decoder output without AnCoGen to isolate potential artifacts. The three AnCoGen models—Melspectrogram, BigCodec, and SpeechTokenizer—show good synthesis quality across the four speech attributes. AnCoGen-BigCodec achieves the highest N-MOS score (4.39) but demonstrates less accurate content reconstruction due to BigCodec's content analysis errors. AnCoGen-SpeechTokenizer offers good listening quality but introduces frame artifacts, as seen in BAK scores, due to parallel pattern scale prediction [29]. However, SpeechTokenizer excels in content reconstruction, reflecting the strong linguistic interpretability of its first RVQ scale.

Speech control: In this task, we investigate voice conversion using AnCoGen through the *analysis-control-generation* framework. During the control step, the speaker identity of the source signal is replaced with that of the target while preserving all other source attributes. Speech signals from 10 randomly selected identities in the LibriSpeech test set are used for evaluation. Speaker similarity is measured using cosine similarity (COS) computed with Resemblyzer embeddings [30], while audio quality is assessed using the N-MOS metric. All three approaches achieve good voice conversion quality, with AnCoGen-SpeechTokenizer delivering the highest target voice similarity, evidenced by a COS score of 0.74.

4. Conclusion

This work studies two prominent neural audio codecs, assessing their ability to represent content, pitch, and speaker identity. Our visual and quantitative analysis reveals that these speech attributes are entangled within the codecs' quantized latent spaces, limiting interpretability - even for codecs designed with disentanglement in mind. We also introduce a framework inspired by AnCoGen, which bridges these codecs to improve interpretability. This bidirectional approach links codec tokens to key speech attributes (e.g., pitch, content, loudness, identity), enabling their extraction for analysis and speech generation.

5. Acknowledgements

This work was granted access to the HPC/AI resources of [CINES / IDRIS / TGCC] under the grant 2022-AD011013469 awarded by GENCI and partially funded by the French National Research Agency under the ANR Grant No. ANR-20-THIA-0002.

6. References

[1] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.

		Analysis		Resynthesis				Control	
Task		Pitch	Content	Non-intrusive		Intrusive		Voice Conversion	
Metrics	Bitrate (kbps)	AAE ↓	Acc. ↑	N-MOS ↑	BAK ↑	STOI ↑	BertScore ↑	N-MOS ↑	COS ↑
AnCoGen - *Melspectrogram* [10]	51.2	**4.60**	<u>81.20</u>	4.26	**4.16**	**0.79**	<u>0.87</u>	4.24	<u>0.72</u>
AnCoGen - *BigCodec*	1.04	7.04	73.58	**4.39**	<u>4.07</u>	0.72	0.82	**4.37**	0.68
AnCoGen - *SpeechTokenizer*	4.00	<u>6.48</u>	**82.00**	<u>4.34</u>	3.91	<u>0.75</u>	**0.88**	<u>4.26</u>	**0.74**

Table 1: *AnCoGen results (best score in each column is in bold, second best score is underlined).*

[2] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[3] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "Soundstream: An end-to-end neural audio codec," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 495–507, 2021.

[4] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *arXiv preprint arXiv:2210.13438*, 2022.

[5] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-fidelity audio compression with improved rvqgan," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

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

[7] X. Zhang, D. Zhang, S. Li, Y. Zhou, and X. Qiu, "Speechtokenizer: Unified speech tokenizer for speech large language models," *arXiv preprint arXiv:2308.16692*, 2023.

[8] A. Défossez, L. Mazaré, M. Orsini, A. Royer, P. Pérez, H. Jégou, E. Grave, and N. Zeghidour, "Moshi: a speech-text foundation model for real-time dialogue," *arXiv preprint arXiv:2410.00037*, 2024.

[9] D. Xin, X. Tan, S. Takamichi, and H. Saruwatari, "Bigcodec: Pushing the limits of low-bitrate neural speech codec," *arXiv preprint arXiv:2409.05377*, 2024.

[10] S. Sadok, S. Leglaive, L. Girin, G. Richard, and X. Alameda-Pineda, "Ancogen: Analysis, control and generation of speech with a masked autoencoder," *arXiv preprint arXiv:2501.05332*, 2025.

[11] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[12] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[13] M. Poli, T. Schatz, E. Dupoux, and M. Lavechin, "Modeling the initial state of early phonetic learning in infants," 2024.

[14] B. van Niekerk, M.-A. Carbonneau, and H. Kamper, "Rhythm modeling for voice conversion," *IEEE Signal Processing Letters*, 2023.

[15] P. Cheng, W. Hao, S. Dai, J. Liu, Z. Gan, and L. Carin, "Club: A contrastive log-ratio upper bound of mutual information," in *International conference on machine learning*. PMLR, 2020, pp. 1779–1788.

[16] X. Bie, X. Liu, and G. Richard, "Learning source disentanglement in neural audio codec," *arXiv preprint arXiv:2409.11228*, 2024.

[17] Y. Zheng, W. Tu, Y. Kang, J. Chen, Y. Zhang, L. Xiao, Y. Yang, and L. Ma, "Freecodec: A disentangled neural speech codec with fewer tokens," *arXiv preprint arXiv:2412.01053*, 2024.

[18] Z. Ju, Y. Wang, K. Shen, X. Tan, D. Xin, D. Yang, Y. Liu, Y. Leng, K. Song, S. Tang *et al.*, "Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models," *arXiv preprint arXiv:2403.03100*, 2024.

[19] 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," *arXiv preprint arXiv:2104.00355*, 2021.

[20] N. Gengembre, O. Le Blouch, and C. Gendrot, "Disentangling prosody and timbre embeddings via voice conversion," in *Proc. Interspeech 2024*, 2024, pp. 2765–2769.

[21] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "Crepe: A convolutional representation for pitch estimation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 161–165.

[22] 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.

[23] G. Pirker, M. Wohlmayr, S. Petrik, and F. Pernkopf, "A pitch tracking corpus with evaluation on multipitch tracking scenario." in *Interspeech*, 2011, pp. 1509–1512.

[24] C. K. Reddy, V. Gopal, and R. Cutler, "Dnsmos p. 835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 886–890.

[25] P. Manocha, B. Xu, and A. Kumar, "Noresqa: A framework for speech quality assessment using non-matching references," *Advances in neural information processing systems*, vol. 34, pp. 22 363–22 378, 2021.

[26] A. Kumar, K. Tan, Z. Ni, P. Manocha, X. Zhang, E. Henderson, and B. Xu, "Torchaudio-squim: Reference-less speech quality and intelligibility measures in torchaudio," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[27] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2010, pp. 4214–4217.

[28] T. Saeki, S. Maiti, S. Takamichi, S. Watanabe, and H. Saruwatari, "Speechbertscore: Reference-aware automatic evaluation of speech generation leveraging nlp evaluation metrics," *arXiv preprint arXiv:2401.16812*, 2024.

[29] J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Défossez, "Simple and controllable music generation," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[30] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.