Speech Prosody 2026
26-29 May 2026, Philadelphia, Pennsylvania, USA

Is Prosody Encoded in the Neural Audio Codec Mimi?

Nicolas Ballier[1], *Artem Saloev* [1], *Erin Pacquetet*[2], *Taylor Arnold*[3]

[1] ALTAE, Université Paris Cité, F-75013 Paris, France
[2] SCIAM, 10 rue de Penthièvre, F-75008 Paris, France
[3] Linguistics & Cognitive Science, University of Richmond, USA

nicolas.ballier@u-paris.fr, artem.saloev@gmail.com, erin.pacquetet@sciam.fr, tarnold2@richmond.edu

Abstract

Neural audio codecs (NACs) have become a fundamental infrastructure for modern speech-language models, discretizing continuous speech signals into token sequences that enable large language models to process and generate audio. Most research evaluates codec quality at the decoding phase, assessing naturalness of reconstructed speech. However, little research investigates what phonetic information is captured during encoding. In this paper, we investigated the potential correlates of prosody in the NAC Mimi, a codec used by conversational speech agents such as Moshi and Sesame and used in downstream tasks such as speech-to-speech translation (Hibiki). We analyzed the eight codebooks generated by Mimi when encoding incrementally resynthesized English speech (pitch, duration, and glissando value of the last syllable). We investigated the potential prosodic correlates of the codebooks using Canonical Correlation Analysis (CCA) of the different codebooks generated by Mimi, and assessed the correlation between the codebook values and the resynthesized version. We show the overimportance of the first codebook (codebook_0) compared to the others and that it clearly correlates with segmental features; the CCA shows moderate correlation between the prosodic features and the other codebooks used to represent speech.
Index Terms: speech foundation models, neural audio codecs, computational linguistics.

1. Introduction

Neural audio codecs (NACs) have become central in the design of modern speech models, enabling efficient audio tokenization. These models use neural networks to compress speech into discrete code sequences for high-quality reconstruction at low bitrates [1]. Many codecs have been developed for industry needs, especially in conversational agents and specific applications. For instance, NotebookLM is an application from Google that leverages the Soundstream codec to produce podcasts [2]. We can also cite Google Research's AudioLM model [3], an audio generation model, that employs SoundStream's discrete representations as targets during its training process. In comparison, Meta Research's VALL-E utilizes the Encodec codec [4] as an intermediate feature for text-to-speech tasks. Neural models such as Vall-E [5] have been used as language modeling tools for tasks like text-to-speech (TTS) synthesis and speech-to-speech neural translation. Vall-E's architecture demonstrates the potential of neural vocoders for accurate voice synthesis by efficiently encoding speaker identity, prosody, and linguistic content into compressed representations. Finally, [6] introduced VALL-E X, a cross-lingual neural codec language model specifically designed for multilingual speech synthesis. VALL-E X leverages large-scale training data to facilitate cross-lingual transfer, enabling high-quality speech generation across languages with minimal speaker-specific data.

In this paper, we focus on Mimi, a codec designed for the conversational agent Moshi [1], also used by the conversational agent Sesame AI [1] and the speech-to-speech translation system Hibiki [7]. The NAC Mimi uses Residual Vector Quantization (RVQ) to generate hierarchical discrete speech representations [8]. RVQ not only compresses phonetic information, but may also distill semantic and suprasegmental speech features. In the encoding phase, Mimi is reported to encode the semantic representation of speech in the first codebook, using distillation from WaveLM. The other codebooks are specialized for "acoustic representations" [1]. These range from codebook_0 to codebook_7, in a process referred to as semantic "distillation". We aim to discuss this division of labour between the first codebook specialized in segmental representation and the seven other codebooks, apparently the most likely to capture prosodic variability. Our paper aims to verify this, testing the correlations between the 2048 tokens used in the 8 codebooks [1] with phone posteriorgram predictions and three prosodic features that we have systematically modified using the WinPitch software [9]. The features we chose to modify incrementally are duration, pitch, and glissando values. Duration and pitch are core dimensions of prosodic structure, while the glissando threshold was used as a way to triangulate pitch value and duration.

Most codec studies focus on decoding quality, with limited research on encoding behavior or codebook properties. Little research has been carried out on the encoding side, especially when trying to find correlates in the codec codebooks when investigating the properties of the speech tokenizer codebook. Others [10], [11] use the posteriorgram method to try to find correspondence between the probability of detecting a phone and the codebook. To the best of our knowledge, no systematic investigation has been carried out in the prosodic domain.

The rest of the paper is organized as follows: Section 2 presents the main studies on Neural Audio Codecs. Section 3 explains how we have analyzed the Mimi codebooks. Section 4 presents our results, Section 5 discusses them and Section 6 suggests future research and concludes.

2. Previous Research

2.1. The multiplication of codecs

Research of phonetic features in codecs is still an emergent field compared to a more established literature on the probing of phonetic representation in speech models like Wav2vec [12] or Whisper [13, 14]. [15] analyzed tokens from four self-supervised learning models (HuBERT, ContentVec, data2vec,

[1] https://www.ai-sesame.com/

10.21437/SpeechProsody.2026-115

emotion2vec) manipulating prosody and speaker identity from the TIMIT corpus [16] to observe their sensitivity to prosody.

Recently, [17] focused on discrete acoustic codec space analysis, still emphasizing speech reconstruction quality. They analyze audio restoration in clean and noisy conditions and compare various codecs. They present their own codec (Language-codec), devised using a Masked Channel Residual Vector Quantization (MCRVQ) mechanism to limit what is learnt on the first three channels (codebooks)[2] and adds a fourth codebook using Short-Time Fourier Transform.

[18] employed a masked encoder [19] to investigate pitch representations in NACs. They visualized encoder internal representations with t-SNE, examining pitch feature encoding and layer organization. For voice conversion and prosody manipulation, [20] developed a prosody-aware audio codec encoder (PACE). The model explicitly conditions encoding on prosodic features, enabling fine control in speech synthesis or modification.

2.2. Probing codec representations

Most neural audio codec analyses focus on decoding [21], [22], using mean opinion score (MOS) to mimic human speech quality judgments. MOS is the standard benchmark for evaluating codec speech reconstruction quality. Human ratings typically range from 1 (bad) to 5 (excellent) and assess reconstructed .wav quality, while automatic speech prediction systems [23] emerge for codec benchmarking. Another decoding quality metric is word error rate, measured from reconstructed signals used in speech recognition [24].

In contrast, encoding-focused studies often align codec codebooks to phoneme-like units via ABX methodology [25], testing minimal pair discrimination. Another approach uses segmental realization with datasets like TIMIT to study encoder and speech decoder performance. For example, posteriogram techniques [26] estimate what was in the signal and how it could correspond to individual tokens of the codebook. This enables calculation of metrics emphasizing the frequency of codebook token usage matched to segmental units. However, systematic analyses relating to acoustic tokens and their corresponding values remain limited. To address this, benchmarks such as AnCoGen [19] have been proposed to provide comprehensive evaluations, incorporating supra-segmental features alongside segmental ones. These frameworks capture and analyze prosodic correlates within encoded speech representations.

A downside is that acoustic property studies often coincide with new codecs (e.g., FlexiCodec [27], Language-Codec [17]), making it harder to maintain a consistent analysis. For example, [28] proposed an audio codec toolkit to facilitate research but also produced their HiFi-Codec. Research on phonetic correlates has mostly focused on segmental features to the detriment of prosodic features, and, to the best of our knowledge, no canonical correlation analysis has been proposed for codecs to investigate prosodic correlates.

2.3. Residual Vector Quantization and Semantic Distillation

Most recent codecs resort to Residual Vector Quantization (RVQ), which is a hierarchical compression technique used by neural audio codecs such as Mimi. RVQ quantizes residual speech signals across multiple sequential layers of codebooks, each layer encoding information not captured by the preceding layers [29]. This layered architecture produces a discrete representation that includes both segmental phonetic details and higher-level acoustic and semantic features. Semantic distillation denotes the hypothesis that the composite representation derived from RVQ codebooks implicitly encodes semantic and suprasegmental speech aspects.

[2]In the literature that "codebook" refers to the successions of tokens produced when encoding the sound file, and to the inventory of the tokens (in our case, from 0 to 2,047) assigned through the RQV process to a corresponding frame of the acoustic space.

3. Material and methods

In our experiments, we aim to identify which aspects of the speech signal are encoded by the Mimi neural audio codec's codebooks. The main research question we aim to answer is: are the speech correlates more segmental or suprasegmental?

3.1. Methods

Mimi is described as producing one "semantic" codebook (codebook_0), while the remainder of codebooks are "acoustic" tokens. Our study evaluates this hypothesis with two experiments, first by correlating codebook embeddings with prosodic features (pitch, duration, glissando) and second with posteriorgram phone predictions. We performed a Canonical Correlation Analysis (CCA) for the prosodic features as well as a classification task to see if we can predict the correlates of speech with the codebook token values.

3.1.1. Canonical Correlation Analysis (CCA)

Canonical Correlation Analysis (CCA) was employed to investigate the relationship between the multidimensional neural codec codebooks and prosodic features derived from resynthesized speech. The goal is to evaluate if prosodic information is encoded in the neural codec's discrete representations. CCA identifies linear combinations of variables from two datasets that are maximally correlated, allowing exploration of shared representations between acoustic features and codec embeddings [30]. Because the codebook tokens were categorical (with token IDs ranging from 1 to 2048), we resorted to kernel CCA using the R [31] package kernel [32]. We defined a Hamming kernel (which counts the proportion of matches across the codebook vectors) for the codebook categorical data, and measured the correlation with the three quantitative prosodic data (pitch, duration, glissando values).

3.1.2. Testing Mimi Representations

Previous work [1], investigating the Mimi neural audio codec's encoder, have primarily employed the XAB perceptual method to analyze its phonetic correlates. We have reasons to believe that this method fails to capture the semantic codebook. Mimi produces a token every 80 milliseconds, which may be about the segment or below the segment for long vowels. We would rather hypothesize that the different token IDs for the semantic codebook representation are more likely to correspond to a hybrid combination of monophones, diphones, triphones, and even quadphones. Our motivation for including these phonetic and suprasegmental correlates in our analysis arises from the hypothesis that Mimi's codebook tokens may also encode connected speech processes. These may include phenomena such as assimilations or persistent repetitive segments, which are characteristic of natural fluent speech.

We analyze the capability of codebooks to capture seg-

mental representations by comparing phone predictions derived from posteriorgrams [26] with the discrete tokens generated by the Mimi codec, and we try to investigate whether the sequence of the eight codebooks can be correlated to the phone prediction obtained by the posteriorgrams (which are the generation of posterior probabilities of the phonetic categories). We aligned the posteriorgram predictions (the phone predicted with the highest probability, one every 20ms) to each time frame corresponding to a Mimi codebook (80ms) for the 6300 files of the TIMIT data. We excluded the NA and [PAD] values, resulting in a dataset of 52,749 Mimi tokens with their four corresponding posteriorgram predictions. The Mimi codec allows for 8, or 32 codebooks. We focused on the 8 version but also computed results for the 32 codebook version for our systematic investigation of the prosodic features and of the segmental features.

3.1.3. Classification Task

We used the Tilburg memory-based learner (TiMBL) that implements the k-nn algorithm for categorical values as predictors [33] to analyze whether we could predict our three prosodic features or the posteriorgram phone labels with the codebook token values. TiMBL operates by storing labeled instances and classifying new data points based on the majority class among their k closest neighbors in the feature space, where closeness is typically measured by a distance metric. In our study, TiMBL is applied to classify tokens generated by the Mimi codec to assess how well discrete codebook embeddings can represent phonetic and prosodic classes. To be more specific, we only analyzed the tokens of the codebooks that corresponded to the synthetically modified syllable. In 75% of the cases (the longest files), the last five tokens corresponded to the final silence and the eight previous tokens matched the representation of the modified syllables. For the prosodic predictions, we used TiMBL with 64 features, corresponding to eight tokens multiplied by eight codebooks.

3.2. Data

For the investigation of segmental correlates, we replicate the codec analysis proposed by the analysis of the SpeechTokenizer codec [24] using the TIMIT dataset. The TIMIT dataset consists of broadband recordings from 630 speakers of eight major American English dialects, each reading ten phonetically rich sentences with time-aligned transcriptions[16]. We probed the posteriorgram correspondence between the phone predicted by posteriorgram generation with a system trained on the TIMIT datasets [26] and the Mimi codebooks that were generated using the TIMIT data.

For the investigation of prosodic correlates, we utilized a synthetic dataset created using the WinPitch software [9]. We controlled for pitch, duration, and glissando for the modification of the last syllable of the sentence "He arrived on time". We systematically modified the pitch (+10Hz), duration (+10ms), and glissando values in regular increments, allowing for the following variations to be observed. We generated the sentences with a value of pitch ranging from 100 to 490 Hz and the duration of the last syllable range from 10 to 390 milliseconds. 1029 variants of prosodic events in total were analyzed and compared to the eight vectors of the corresponding Mimi codebooks.

3.2.1. Artificial generation of speech with codebook token IDs

The Mimi codebook generates sequences of tokens on eight codebooks (or more), which are the representations of the speech signal. We also tested the potential correlates of the Mimi token codebooks by artificially generating codebook vectors with arbitrary token IDs values.

4. Results

4.1. Mixed results

For our CCA analysis, about 25% of shared variance ($0.498^2 = 0.248$) is captured by the first canonical dimension. We did not resort to Wilks test as it would be just an approximation since we are using a kernel method (kernel-based similarity, not raw numeric data). We then ran a classification task to realize that the eight codebooks were rather poor predictors of pitch (and subsequently, glissando) (see Table 1) but better predictors of the phone labels predicted by the posteriorgram method. For the phone predictions, the tokens have been realigned to the corresponding time stamps, so that eight tokens from eight codebooks are used as predictors.

Metric	**8 codebooks**
Duration accuracy	92.3%
Glissando accuracy	8.1%
Pitch accuracy	15.5%
Phone accuracy	68.32%

Table 1: *TiMBL prediction performance (with K=5) for the 8 tokens of the 8 codebooks for our three prosodic variables and phone predictions with the aligned token for the TIMIT data*

We evaluated how Mimi encodes three prosodic parameters manipulated in the synthesized corpus: duration, pitch, and glissando. For each parameter, TiMBL was used with Leave-One-Out testing, $k = 5$, Overlap metric and GainRatio weighting.

4.1.1. Prosodic variables

Duration classification reached 92.3% accuracy. This indicates that Mimi's tokenization preserves duration noticeably well. GainRatio showed that the most predictive features were concentrated in the codebooks_1-3. Pitch classification yielded 15.5% accuracy. Although much lower than the duration, the performance is above chance (given the number of classes to predict), demonstrating that pitch information is partially recoverable from the tokens. The highest GainRatio values again came from codebooks_0–3.

Glissando classification reached 8.1% accuracy. Despite the fine-grained target space, glissando direction and slope were recoverable above chance. As with pitch, the strongest predictors were found in codebooks_1-3. The last two tokens in each codebook were consistently the most predictive across all prosodic variables.

4.2. Number of Codebooks

We tested two types of implementations, the 8-codebook and the 32-codebook, the hypothesis being that with the 8-codebook version we would have a better way to predict the correlates of the system and that the 8-codebook version was more likely to predict the prosodic features that we tested. The first 8 codebooks were identical in the 8 and the 32 version.

4.3. Semantic Leakage

Using the information gain ratio as a criterion in the classification task using TiMBL, we observed that the feature importance of the different codebooks varied systematically for all

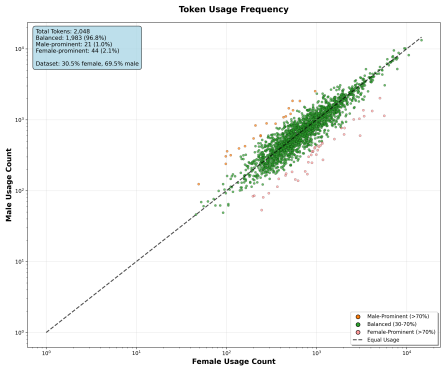

Figure 1: *Gender prominence in the Mimi tokens (TIMIT data)*

the predictions. An important observation, which perhaps mirrors RQV is the number of unique tokens grows with each new codebook. The number of unique tokens in codebook_0 is potentially consistent with allophone representations of the syllable "time", but tokens in the codebook_0 fail to predict the phone labels as accurately as the 8 codebooks taken together, so that some form of the semantic leakage might be at work over the other seven codebooks (see Table 2).

Codebook ID	Count
codebook_0	27
codebook_1	163
codebook_2	409
codebook_3	461
codebook_4	577
codebook_5	649
codebook_6	749
codebook_7	823

Table 2: *Number of unique token IDs per codebook*

4.4. Segmental Correlates

Partially replicating the SpeechTokenizer experiments matching codebook IDs with posteriorgrams, we managed to match codebook IDs to posteriogram predictions. We surmise that the representation in tokens, which spans over 80 milliseconds may correspond to quadphones, triphones, diphones, and single phone representations for the 2048 token IDs of codebook_0. On the TIMIT dataset, for the 80 ms time frames corresponding to the 52,749 tokens produced in codebook_0, we observed 33,540 realisations of identical posteriogram predictions (possibly phones), 18,795 realisations of two different posteriogram predictions (possibly diphones), 413 realisations of three different posteriogram predictions (possibly triphones) and one candidate for quadphone (the plausible frequent phone sequence /ɪnðə/, corresponding to tokenID 1047).

5. Discussion

5.1. Gender bias or specialization?

The 2,048 tokens used in each codebook may represent distinct acoustic features for each codebook. When looking at codebook specific tokens, the one with the most exclusivity per gender is by far codebook_1. For instance, no female has token_ID 6 in codebook_1, which makes it gender exclusive for this codebook,

Codebook	Total Tokens	Exclusive Tokens	ratio
1	858	265	30.90%
2	1,641	362	22.10%
0	1,821	165	9.10%
3	1,857	176	9.50%
4	1,919	142	7.40%
5	1,916	128	6.70%
6	1,933	145	7.50%
7	1,943	111	5.70%

Table 3: *Statistics of codebook tokens and exclusivity rate.*

but we see token_ID 6 in female data elsewhere. It is also interesting to note that codebook_1 has far fewer unique tokens than other codebooks. When looking at unique token (independent from the codebook they are from), a few tokens show a gender bias (none are exclusive, though).

5.2. Ablation experiments and synthetic codebook values

To test the meaning of the first codebook, we performed a perceptual ablation experiment replacing the first codebook token values by arbitrary numbers between 1 and 100 and the sound file was audibly very similar to the original sound file, so that we may conclude that the first codebook is surprisingly not as important as the combination of all the codebooks taken together. When multiple 8 codebooks were simultaneously modified with synthetic or cycled values, the speech degradation was noticeable, characterized by noise and unintelligible words. This suggests that the speech decoder relies heavily on joint coherence across multiple codebooks to maintain naturalness and clarity.

5.3. Specialization or Distributed semantics?

An alternative experiment is the resynthesis of the codebooks with the assumption that if the codebooks have specific meanings or prosodic values associated to the tokens, we would be expected to observe variability in frequency, pitch or glissando investigating the output from codebooks synthetically generated using artificial numbers from zero to 2,048. The sequences are somehow arbitrary. We suspect that there is a form of solidarity for the eight values of the codebook, so that we also tested on a minimal number of token units that we found in the data the change that we observed modifying these token ideas.

6. Conclusion

In this paper, we have shown that the division of labour for the Mimi codec between the semantic codebook and the acoustic codebooks might not be as clear-cut as initially posited [1]. The so-called 'acoustic codebooks' (from codebook_1 to codebook_7) correlate strongly with duration, but not with glissando and pitch values. It might be the case that tokens are good predictors of duration because, mechanically, as duration increases, tokens are produced every 80 ms. However, clearer correlations can be established with the codebook_0 values with phone, diphone, triphone and quadphone representations than suggested in the ABX discrimination method [34, 1]. Future neural audio codecs may benefit from a more specialized allocation of representational capacity, where different vectors are optimized for distinct prosodic dimensions. This could pave the way for better prosodic controllability in synthesis and robustness in speech models.

7. References

[1] 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," Kyutai, Tech. Rep., September 2024. [Online]. Available: http://kyutai.org/Moshi.pdf

[2] 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.

[3] Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, D. Roblek, O. Teboul, D. Grangier, M. Tagliasacchi *et al.*, "Audiolm: A language modeling approach to audio generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2523–2533, 2023.

[4] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *Trans. Mach. Learn. Res.*, 2023.

[5] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li *et al.*, "Neural codec language models are zero-shot text to speech synthesizers," *arXiv preprint arXiv:2301.02111*, 2023.

[6] Z. Zhang, L. Zhou, C. Wang, S. Chen, Y. Wu, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li *et al.*, "Speak foreign languages with your own voice: Cross-lingual neural codec language modeling," *arXiv preprint arXiv:2303.03926*, 2023.

[7] T. Labiausse, L. Mazaré, E. Grave, A. Défossez, and N. Zeghidour, "High-fidelity simultaneous speech-to-speech translation," in *Forty-second International Conference on Machine Learning*, 2025. [Online]. Available: https://openreview.net/forum?id=fgjN8B6xVX

[8] J. Zhao, X. Wang, and Y. Wang, "Prosody-adaptable audio codecs for zero-shot voice conversion via in-context learning," *arXiv preprint arXiv:2505.15402*, 2025.

[9] P. Martin, "Multi methods pitch tracking," in *Proc. Speech-Prosody 2012*, 2012, pp. 47–50.

[10] Y. Cao, S. Liu, X. Wu, S. Kang, P. Liu, Z. Wu, X. Liu, D. Su, D. Yu, and H. Meng, "Code-switched speech synthesis using bilingual phonetic posteriorgram with only monolingual corpora," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7619–7623.

[11] C. Churchwell, M. Morrison, and B. Pardo, "High-fidelity neural phonetic posteriorgrams," in *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 2024, pp. 823–827.

[12] P. C. English, J. Kelleher, and J. Carson-Berndsen, "Domain-informed probing of wav2vec 2.0 embeddings for phonetic features," in *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, 2022, pp. 83–91.

[13] N. Ballier, L. Burin, B. Namdarzadeh, S. B. Ng, R. Wright, and J.-B. Yunès, "Probing Whisper predictions for French, English and Persian transcriptions," in *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, 2024, pp. 129–138.

[14] S. Liang, N. Ballier, G.-A. Levow, and R. Wright, "Beyond wer: Probing whisper's sub-token decoder across diverse language resource levels," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025, pp. 31 225–31 235.

[15] K. Onda, S. Fukayama, D. Saito, and N. Minematsu, "Benchmarking prosody encoding in discrete speech tokens," *arXiv preprint arXiv:2508.11224*, 2025.

[16] J. S. Garofolo, L. F. Lamel, W. M. Fisher, D. S. Pallett, N. L. Dahlgren, V. Zue, and J. G. Fiscus, "Timit acoustic-phonetic continuous speech corpus," *NASA STI/Recon technical report n*, 1993.

[17] S. Ji, M. Fang, J. Zuo, Z. Jiang, D. Wang, H. Wang, H. Huang, and Z. Zhao, "Language-codec: bridging discrete codec representations and speech language models," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025, pp. 13 332–13 345.

[18] S. Sadok, J. Hauret, and Éric Bavu, "Bringing Interpretability to Neural Audio Codecs," in *Interspeech 2025*, 2025, pp. 5023–5027.

[19] S. Sadok, S. Leglaive, L. Girin, G. Richard, and X. Alameda-Pineda, "Ancogen: Analysis, control and generation of speech with a masked autoencoder," *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025.

[20] J. Zhao, X. Wang, and Y. Wang, "Prosody-Adaptable Audio Codecs for Zero-Shot Voice Conversion via In-Context Learning," in *Interspeech 2025*, 2025, pp. 4893–4897.

[21] T. Muller, S. Ragot, L. Gros, P. Philippe, and P. Scalart, "Speech quality evaluation of neural audio codecs," in *Interspeech 2024*, 2024, pp. 1760–1764.

[22] W. Mack, N. Topaloglu, L. Lechler, I. Balić, A. Craciun, M. Yesilbursa, and K. Wojcicki, "Assessing speech quality metrics for evaluation of neural audio codecs under clean speech conditions," *arXiv preprint arXiv:2509.24457*, 2025.

[23] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022," vol. 2022, pp. 4521–4525, 2022.

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

[25] T. Schatz, V. Peddinti, F. Bach, A. Jansen, H. Hermansky, and E. Dupoux, "Evaluating speech features with the minimal-pair ABX task: analysis of the classical mfc/plp pipeline," pp. 1781–1785, 2013.

[26] C. Churchwell, M. Morrison, and B. Pardo, "High-fidelity neural phonetic posteriorgrams," in *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*. IEEE, 2024, pp. 823–827.

[27] J. Li, Y. Qian, Y. Hu, L. Zhang, X. Wang, H. Lu, M. Thakker, J. Li, S. Zhao, and Z. Wu, "Flexicodec: A dynamic neural audio codec for low frame rates," *arXiv preprint arXiv:2510.00981*, 2025.

[28] D. Yang, S. Liu, R. Huang, J. Tian, C. Weng, and Y. Zou, "Hifi-codec: Group-residual vector quantization for high fidelity audio codec," *CoRR*, 2023.

[29] 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, p. 1505, 2022.

[30] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[31] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2024. [Online]. Available: https://www.R-project.org/

[32] A. Karatzoglou, A. Smola, K. Hornik, and A. Zeileis, "kernlab – an S4 package for kernel methods in R," *Journal of Statistical Software*, vol. 11, no. 9, pp. 1–20, 2004.

[33] W. Daelemans, J. Zavrel, K. van der Sloot, and A. van den Bosch, "TiMBL: Tilburg memory based learner, version 6.0, reference guide," ILK Research Group, Tilburg University, Tilburg, The Netherlands, Tech. Rep., 2004. [Online]. Available: https://ilk.uvt.nl/timbl/

[34] T. Schatz, "ABX-discriminability measures and applications," Ph.D. dissertation, Université Paris 6 (UPMC), 2016.