MYOCODEC: A STREAMING NEURAL CODEC FOR ELECTROMYOGRAPHY

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural codecs encode continuous signals into compact sequences of discrete tokens, providing an interface for efficient transmission, storage, and token-based sequence modeling. This paradigm has been widely adopted in modern speech and audio frameworks; however, the biosignal domain still lacks a neural codec designed specifically for low-bitrate streaming and generalization across diverse downstream tasks. **We present MYOCODEC, a streaming neural codec designed for electromyography (EMG)**. Inspired by recent neural audio codecs, MYOCODEC combines causal Transformers with residual vector quantization to encode continuous EMG signals into different levels of EMG representations spanning from continuous latent features to discrete tokens operating at $50\,\mathrm{Hz}$. Trained on twelve public EMG datasets, MYOCODEC achieves favorable performance in both intrinsic codec quality and representative downstream tasks, including typing (*emg2qwerty*), hand-pose (*emg2pose*), speech decoding (*emg2speech*), and speech-to-EMG synthesis (*speech2emg*). Across these tasks, MYOCODEC exhibits strong performance against prior models while providing a compact and causal EMG representation. During streaming inference, it requires compute time of only $0.482\,\mathrm{ms}$ for each $20\,\mathrm{ms}$ frame, enabling real-time streaming. Also, the discrete token representation provided by MYOCODEC has the potential to support integration into language-model based approaches, creating a path toward LLM-based interactive systems, where tokenized EMG representations are directly processed into such language or speech models. Code and model weights will be released publicly upon publication[1].

1 INTRODUCTION

Foundation models have reshaped various domains, such as text, speech, audio, and biosignal, by replacing task-specific front ends with generalizable representations learned from large-scale unlabeled data. Rather than training a separate encoder for every application, a single pretrained model can provide a shareable representation for various tasks, including recognition, generation, and understanding across a broad range of downstream tasks.

In the biosignal domain, LaBraM learns general-purpose representations for EEG (Jiang et al., 2024), while the generic neuromotor interface of Kaifosh et al. (2025) demonstrates that EMG decoding can generalize across individuals without per-user calibration. However, such approaches in the EMG domain still lack a shared, task-agnostic representation for diverse downstream tasks. To serve as a practical interface in the EMG domain, such a representation must not only transfer across tasks but also be compact, streamable, and lightweight.

In the speech and audio domain, these requirements have been met through the neural codec approach. Neural audio codecs such as SoundStream (Zeghidour et al., 2021), EnCodec (Défossez et al., 2023), DAC (Kumar et al., 2023), Mimi (Défossez et al., 2024), and JHCodec (Lee et al., 2026b) combine an encoder, residual vector quantization (RVQ), and a decoder trained with reconstruction and adversarial objectives to encode continuous waveforms into compact discrete tokens. The resulting representation can serve several purposes, such as reducing transmission bandwidth,

[1]Reviewer version: `https://anonymous.4open.science/r/myocodec-release-8FF9`; the model checkpoint is in the supplementary material.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

limiting storage and caching costs, and providing tokens that can be modeled directly by modern token-based models. These advantages position neural codecs as general-purpose front ends for speech and audio systems.

BioCodec (Avramidis et al., 2025) is one of the first to extend this neural-codec paradigm to electrophysiology signals. It adapts the EnCodec (Défossez et al., 2023) architecture using causal convolutions and RVQ, producing discrete representations that generalize across various downstream tasks, such as clinical, sleep, speech, and motor-imagery tasks. BioCodec establishes the central premise that the neural-codec paradigm can successfully be applied to biosignals.

However, BioCodec has two main limitations. First, it focuses primarily on electroencephalography (EEG), with EMG treated only as a secondary modality. Second, despite its causal architecture, its experimental design and implementation lack a frame-wise streaming aspect. These limitations lead to an open question for on-device EMG frameworks: whether a codec designed with the temporal and deployment constraints of EMG can produce low-bitrate discrete representations in real time. For more details of related works, refer to Appendix A.

Bridging this gap, **we present MYOCODEC, a streaming neural codec designed for EMG signals**. MYOCODEC consists of an encoder, a residual vector quantizer, and a decoder. The encoder and decoder are based on causal Transformers to support streamability. MYOCODEC takes $2\,\mathrm{kHz}$ EMG signals as input and encodes them into different levels of EMG representations spanning from continuous encoder features to discrete token streams operating at $50\,\mathrm{Hz}$ and $2400\,\mathrm{bits/s}$ per channel. MYOCODEC is trained in two phases. First, it is trained with signal reconstruction objectives and then with a gradually introduced adversarial objective. It is trained on twelve EMG datasets that are publicly available. On a single `NVIDIA RTX PRO 6000 Blackwell` GPU, the inference time takes $0.48\,\mathrm{ms}$ per frame, achieving a real-time throughput of $41\times$.

We probe MYOCODEC in two aspects: intrinsic codec quality and generalizability to downstream tasks. Experimental results show that MYOCODEC outperforms existing EMG representation models in signal reconstruction. MYOCODEC also provides useful EMG representations generalizing to various downstream tasks, including typing (*emg2qwerty* (Sivakumar et al., 2024)), hand-pose detection (*emg2pose* (Salter et al., 2024)), speech decoding (*emg2speech* (Gaddy & Klein, 2021)), and EMG synthesis (*speech2emg* (Scheck & Schultz, 2023)). Our experimental results demonstrate that MYOCODEC outperforms or closely matches existing frameworks, even in a streaming setting, showing its generalization to various applications.

Beyond its current performance, MYOCODEC provides a future potential integration between muscle activity signals and modern token-based sequence modeling approaches. By encoding continuous EMG into low-bitrate discrete tokens, it enables training objectives widely used in modern language modeling approaches, such as categorical next-token prediction. Neural audio codecs have similarly enabled Transformer-based modeling of speech and audio in systems such as AudioLM, VALL-E, and Moshi (Borsos et al., 2023; Chen et al., 2025; Défossez et al., 2024), leading to substantial improvement in performance. Similarly, MYOCODEC tokens could further be incorporated into language or speech-language models, enabling unified systems that map muscle activity directly to language, speech, or actions. Such systems could reduce reliance on cascaded pipelines, in which latency and errors accumulate across separately trained stages.

Other potential applications include silent-speech communication, hands-free text entry, gesture-based interaction, prosthetic control, and other assistive or wearable interfaces. Also, the ability of MYOCODEC to generate EMG signals could support synthetic biosignal augmentation, simulation of task-conditioned muscle activity, and personalized neuromuscular digital twins. Moreover, the MYOCODEC's lightweight model size, streamability, and fast inference speed also make on-device processing plausible, reducing communication bandwidth and limiting the transmission of sensitive raw biosignals.

Our main contributions include:

- We introduce MYOCODEC, a streaming neural codec for EMG, operating at $50\,\mathrm{Hz}$ and $2400\,\mathrm{bits/s}$ per channel.
- We demonstrate its strong intrinsic codec quality and transferability across diverse downstream tasks, establishing MYOCODEC as a generalizable EMG representation.
- With lightweight model size and fast streaming inference, MYOCODEC provides a practical framework for low-latency wearable systems and future on-device deployment.

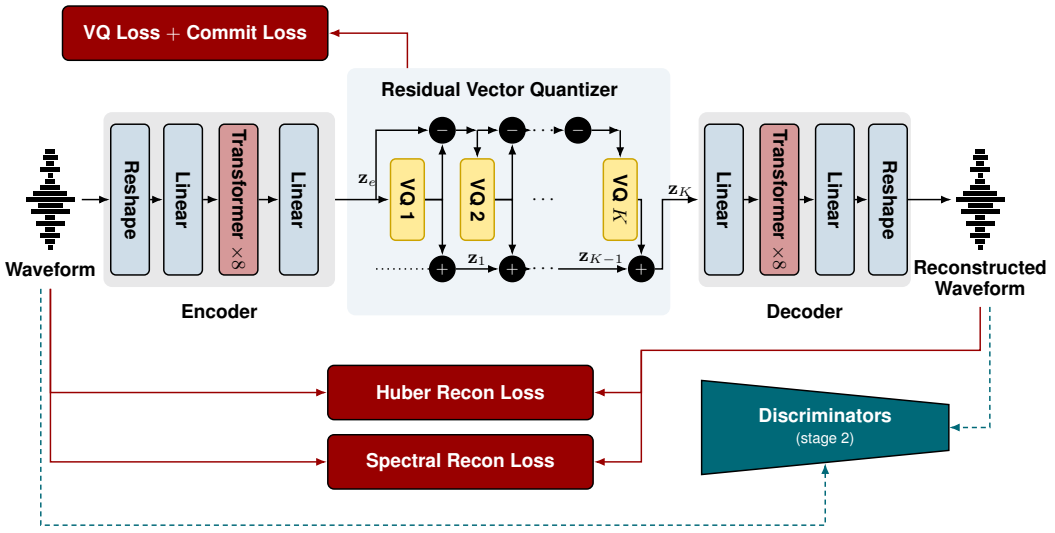


Figure 1: **MYOCODEC architecture.** A causal Transformer encoder–decoder with residual vector quantization (RVQ) processes each EMG channel independently at 50 Hz. The discriminators are used only during the second phase of training.

2 METHOD

2.1 ARCHITECTURE

We base our architecture on the recent neural audio codec JHCodec (Lee et al., 2026b) and adopt the training objectives from BioCodec (Avramidis et al., 2025). As illustrated in Figure 1, our model follows the standard encoder–quantizer–decoder structure in neural codec approaches. To support online inference, the entire architecture is causal. For more details on the architecture, refer to Appendix C.

Encoder. The encoder takes continuous EMG signals at 2 kHz as input. Following BioCodec, each channel is treated independently. Each frame consists of patchified 40 samples, resulting in a frame rate of 50 Hz. A linear layer projects each frame to match the model dimension. The projected sequence is then processed by an eight-layer Transformer with eight attention heads. We use the rotary positional embeddings (Su et al., 2024) and apply causal self-attention with a sliding window of 64 frames per layer. During streaming inference, each layer maintains a 64-frame key–value cache.

Quantizer. The encoder output is projected to 32 dimensions and quantized using a residual vector quantizer with 6 stages, each stage containing a codebook of 256 entries. A linear projection follows to map the quantized representation back to the Transformer dimension before the decoder. To prevent codebook collapse, entries whose exponentially averaged usage falls below a fixed threshold are reinitialized from active encoder outputs every 250 optimization steps. This procedure maintains full codebook utilization and prevents a failure mode in which the vector-quantization loss approaches zero while reconstruction quality deteriorates.

Decoder. The decoder mirrors the encoder, using an eight-layer causal Transformer with the same configuration. A final linear projection maps each decoder output back to a frame of 40 samples, resulting in reconstructed EMG signals at 2 kHz. The encoder and decoder contain approximately 6.3 million parameters each, for a total model size of 12.72 million parameters.

Discriminators. In the second phase of training, we apply adversarial training using two discriminators: a multi-period discriminator and a multi-scale STFT discriminator.

Table 1: **Intrinsic codec quality.** Reconstruction quality is evaluated using SI-SDR and SNR (dB), waveform Pearson correlation r_{wav}, and envelope Pearson correlation r_{env}. N_q is the number of codebooks used. MYOCODEC outperforms TinyMyo and BioCodec in signal reconstruction by using only two codebooks ($800\,\text{bits/s}$).

Model	Params	Hz	N_q	bitrate/ch	reconstruction ↑				codebook	
					SI-SDR	SNR	r_{wav}	r_{env}	util.	perp.
TinyMyo	3.6 M	100.0	—	≤614k†	0.46	1.70	0.735	0.445	—	—
BioCodec	12.3 M	27.8	6	1333	−1.51	2.20	0.643	0.907	1.000	203.4
MYOCODEC (Ours)	12.7 M	50.0	2	800	3.62	4.99	0.832	0.933	1.000	172.1
			3	1200	5.89	6.75	0.881	0.949	1.000	177.8
			6	2400	**9.80**	**9.98**	**0.933**	**0.967**	1.000	186.0

† Rate assumes `float32`.

Table 2: Per-codebook utilization and perplexity for MYOCODEC. All six codebooks are fully utilized, with higher perplexity in later RVQ stages.

	CB 1	CB 2	CB 3	CB 4	CB 5	CB 6	mean
Utilization	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Perplexity	154.9	189.2	189.2	192.6	194.1	196.0	186.0

2.2 TRAINING CURRICULUM

We train MYOCODEC in two phases. Phase 1 jointly optimizes signal reconstruction and vector-quantization objectives without adversarial training. To further refine reconstruction fidelity, Phase 2 adds adversarial and feature-matching objectives while retaining the Phase 1 objectives. More details are in Appendix C.

Phase 1: Reconstruction and quantization training. In Phase 1, we focus on joint training of vector quantization and signal reconstruction, encouraging stable, well-utilized codebooks. Following BioCodec, we use multiple training objectives including time-domain reconstruction $\mathcal{L}_{\text{Huber}}$, multi-scale spectral reconstruction $\mathcal{L}_{\text{spec}}$, and vector-quantization $\mathcal{L}_{\text{VQ}}$, and commit $\mathcal{L}_{\text{commit}}$ losses:

$$\mathcal{L}_{\text{Phase1}} = \mathcal{L}_{\text{Huber}}(x, \hat{x}) + \mathcal{L}_{\text{VQ}} + \lambda_c \mathcal{L}_{\text{commit}} + \frac{\lambda_s}{3} \sum_{n \in \{64,128,256\}} \mathcal{L}_{\text{spec}}^{(n)}(x, \hat{x}), \tag{1}$$

where x and $\hat{x}$ denote the input and reconstructed signals. We set $\lambda_s = 1$ and $\lambda_c = 0.25$. For an STFT $S^{(n)}$ with FFT size n, the spectral loss $\mathcal{L}_{\text{spec}}^{(n)}$ is defined as:

$$\mathcal{L}_{\text{spec}}^{(n)} = \big\| \log |S_x^{(n)}| - \log |S_{\hat{x}}^{(n)}| \big\|_1 + \frac{\big\| |S_x^{(n)}| - |S_{\hat{x}}^{(n)}| \big\|_{\mathcal{F}}}{\big\| |S_x^{(n)}| \big\|_{\mathcal{F}}} + 0.1 \big\| \angle S_x^{(n)} - \angle S_{\hat{x}}^{(n)} \big\|_1, \tag{2}$$

where $\| \cdot \|_{\mathcal{F}}$ indicates the Frobenius norm. Each component captures log-magnitude error, spectral convergence, and phase error, respectively. We train Phase 1 for 100k steps using the Muon optimizer (Jordan et al., 2024) with a batch size of 1024.

Phase 2: Adversarial training. In Phase 2 training, we introduce gradual adversarial training to improve the fine-grained quality of the reconstructed signals while retaining the Phase 1 objectives:

$$\mathcal{L}_{\text{Phase2}} = \mathcal{L}_{\text{Phase1}} + \lambda_{\text{adv}} \mathcal{L}_{\text{adv}} + \lambda_{\text{FM}} \mathcal{L}_{\text{FM}}. \tag{3}$$

where $\mathcal{L}_{\text{adv}}$ and $\mathcal{L}_{\text{FM}}$ denote the adversarial and feature-matching loss, respectively. Their weights are increased linearly from zero to $\lambda_{\text{adv}} = 0.1$ and $\lambda_{\text{FM}} = 2.0$ over the first 50k steps of Phase 2, providing a stable transition to adversarial training. We train Phase 2 for an additional 100k steps with a batch size of 192.

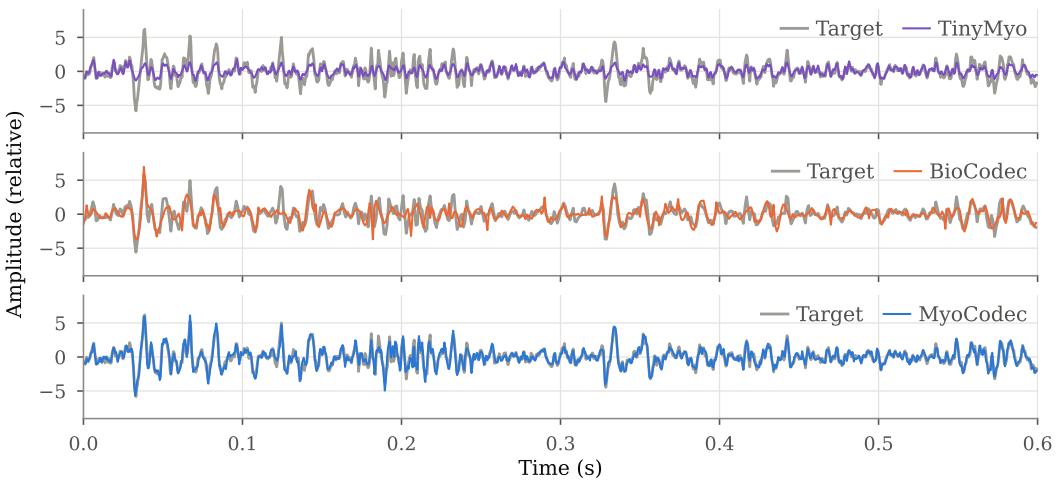


Figure 2: An example of EMG signals reconstructed by each model: TinyMyo (top), BioCodec (middle), and MYOCODEC (bottom). MYOCODEC shows the best reconstruction quality, more properly capturing signal peaks missed by the other models.

2.3 TRAINING DATA AND PREPROCESSING

We train MYOCODEC on twelve publicly available EMG datasets. They contain approximately 24k channel-hours of data from 250k recordings. We use the official training, validation, and test set partitions of each dataset when available and construct speaker- and/or session-disjoint random splits otherwise. For datasets subsequently used in downstream evaluation, codec pretraining used only the official training partition. As the datasets differ in acquisition protocols and release formats, we follow the official preprocessing pipeline published with each corpus. We apply across-channel normalization, preserving relative amplitudes across electrodes while bringing signals from different datasets onto a comparable scale range. For more details on the datasets, refer to Appendix E.

2.4 DOWNSTREAM TASKS

To assess the generalizability of MYOCODEC across various applications, we use MYOCODEC as the EMG featurizer for four representative downstream tasks. In *emg2qwerty*, the task is to predict typed characters from wrist EMG signals, and *emg2pose* estimates hand poses. In *emg2speech*, the aim is to decode speech waveforms from orofacial EMG signals, whereas *speech2emg* addresses the reverse mapping, reconstructing EMG signals from speech waveforms.

For each task, we replace the original EMG representation module with MYOCODEC while preserving the remainder of the architecture, apart from minor task-specific modifications. We compare against TinyMyo (Fasulo et al., 2026), BioCodec (Avramidis et al., 2025), and the corresponding task-specific EMG encoder or decoder from the original work. Depending on the task, we use either the quantized codec output or continuous pre-quantization features, exploring the utility of different levels in the MYOCODEC representation hierarchy. All downstream tasks are trained causally unless stated otherwise. For controlled comparison, some task-specific data augmentation or perturbation is omitted for some downstream tasks. Full implementation details are provided in Appendix D.

2.5 EVALUATION

Our evaluation on MYOCODEC focuses on two aspects: intrinsic codec quality and application to downstream tasks.

Intrinsic codec quality. We assess reconstruction fidelity using scale-invariant signal-to-distortion ratio (SI-SDR) (Roux et al., 2019), signal-to-noise ratio (SNR), waveform-level Pearson correlation, and envelope correlation. We additionally evaluate the quality of the quantized representation using the utilization rate and perplexity of each codebook.

Table 3: **emg2qwerty.** Character error rate (CER) is reported with and without language-model beam search, across *seen* and *unseen* subjects. The best and second-best are **bolded** and underlined, respectively.

EMG Representation	Hz	bitrate/ch	CER (%) ↓			
			greedy		LM beam	
			seen	unseen	seen	unseen
Log-spectrogram (Sivakumar et al., 2024)	125.0	≤132k†	24.02	52.98	16.86	48.57
TinyMyo (Fasulo et al., 2026)	100.0	≤614k†	21.16	**50.02**	14.48	**44.55**
BioCodec (Avramidis et al., 2025)	27.8	1.3k	18.52	51.87	13.07	47.78
MyoCodec (Ours)	50.0	2.4k	**18.43**	50.47	**12.74**	45.77

† Rate assumes `float32`.

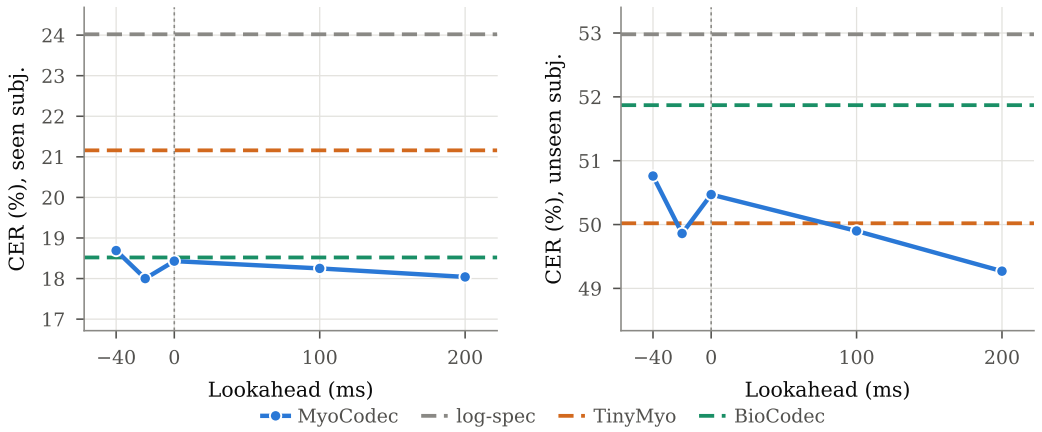


Figure 3: **Effect of lookahead on emg2qwerty.** Decoding performance improves at a negative lookahead of −20 ms, decreases slightly at zero lookahead, and then gradually improves with increasing positive lookahead.

Downstream tasks. To assess generalizability, we evaluate performance on each downstream task using its commonly used evaluation metrics. For *emg2qwerty*, we report character error rate (CER). For *emg2pose*, we report mean joint-angle error with mean landmark and fingertip distances. For *emg2speech*, we report character, word, and phoneme error rates (CER, WER, and PER), phone feature error rate (PFER), and UTMOS (Saeki et al., 2022). PFER measures phonetic errors while accounting for similarity in phonological features, whereas UTMOS provides an automatic estimate of perceived speech quality. For *speech2emg*, we report envelope correlation of synthesized EMG signals. Complete evaluation protocols and implementation details are provided in Appendix D.

3 Results and Discussion

3.1 Intrinsic Codec Quality

We evaluate the intrinsic quality of MyoCodec along two dimensions: reconstruction fidelity and codebook usage. These analyses assess whether the codec preserves relevant signal information while learning a well-utilized discrete representation.

Signal reconstruction. As shown in Table 1, MyoCodec achieves the best reconstruction performance across all reported metrics. MyoCodec already outperforms both BioCodec and TinyMyo on all four metrics when only two codebooks (800 bits/s) are used. The full codebook ablation results are shown in Figure 5 in Appendix B. As depicted in Figure 2, MyoCodec reconstructs the target EMG waveform better than TinyMyo and BioCodec, including transient peaks that the other models often miss.

Table 4: **emg2pose.** Mean joint-angle error (degrees) and mean landmark and fingertip distances (mm).

EMG encoder	joint-angle error (°)↓		land. (mm)↓	fingertip (mm)↓
	val	test		
Log-spectrogram (Salter et al., 2024)	13.13	15.23	20.76	35.07
TinyMyo (Fasulo et al., 2026)	13.15	14.83	20.13	33.97
BioCodec (Avramidis et al., 2025)	13.11	14.75	20.12	34.00
MYOCODEC (Ours)	**12.75**	**14.24**	**19.04**	**32.02**

Table 5: **emg2speech results.** We report character, word, and phoneme error rates (CER, WER, and PER), phone feature error rate (PFER), and UTMOS. All of the downstream models are causally trained with $100\,\mathrm{ms}$ of lookahead.

EMG Encoder	CER↓	WER↓	PER↓	PFER↓	UTMOS↑
Ground truth	1.43	3.76	7.13	4.19	3.41
Gaddy & Klein (Gaddy & Klein, 2021)	22.95	38.49	31.62	14.98	1.90
TinyMyo (Fasulo et al., 2026)	16.39	27.77	26.47	13.29	1.91
BioCodec[†]	16.56	28.33	24.63	12.59	1.93
MYOCODEC (Ours)	**12.15**	**21.12**	**21.61**	**11.54**	**2.00**

[†] We used a retrained version of BioCodec, as the preprocessing used for its released checkpoint removes the frequency range crucial for speech-related information in EMG signals.

Codebook usage. As shown in Table 2, MYOCODEC achieves full utilization of all six codebooks, indicating no representational collapse. Perplexity increases across successive RVQ stages, consistent with a coarse-to-fine decomposition in which the first stage captures the dominant signal structure and later stages encode progressively finer residual details.

3.2 DOWNSTREAM TASKS

emg2qwerty. As shown in Table 3, MYOCODEC achieves the best performance on seen subjects and second-best on unseen subjects, while operating at only $2400\,\mathrm{bits/s}$ per channel. We also explore how different temporal lookaheads influence the decoding performance (Figure 3). Decoding performance improves slightly at $-20\,\mathrm{ms}$, worsens at zero lookahead, and then gradually improves as additional future context is provided. One possible explanation is that muscle activation precedes the registered keystroke, whereas mechanical vibrations at key contact may introduce artifacts into the EMG signal; further analysis is required for more clarification, though. Note that, for controlled comparison, we omit the EMG augmentation methods used in the original *emg2qwerty* framework (Sivakumar et al., 2024), including SpecAugment (Park et al., 2019) and temporal jittering.

emg2pose. As shown in Table 4, MYOCODEC achieves the best performance across all three metrics: joint-angle error, landmark distance, and fingertip distance. All models use the same LSTM-based pose decoder from Salter et al. (2024) and only the input EMG representation differs.

emg2speech. As shown in Table 5, MYOCODEC outperforms all baselines across every evaluated metric, including CER, WER, PER, PFER, and UTMOS. Each downstream model is causally trained with $100\,\mathrm{ms}$ lookahead, where the prediction target is JHCodec speech features (Lee et al., 2026b). We use a retrained version of BioCodec specifically for this task, as its publicly released model weights were trained on band-pass-filtered EMG that excludes frequency content important for speech-related muscle activity. As shown in Figure 4, the best performance is observed when lookahead between $100\,\mathrm{ms}$–$140\,\mathrm{ms}$ is applied. Note that, to isolate the effect of the EMG representation, we apply no EMG augmentation or perturbation, including temporal jittering. Speech samples are available here.[2]

[2] https://myocodec.github.io/myocodec-release/samples/

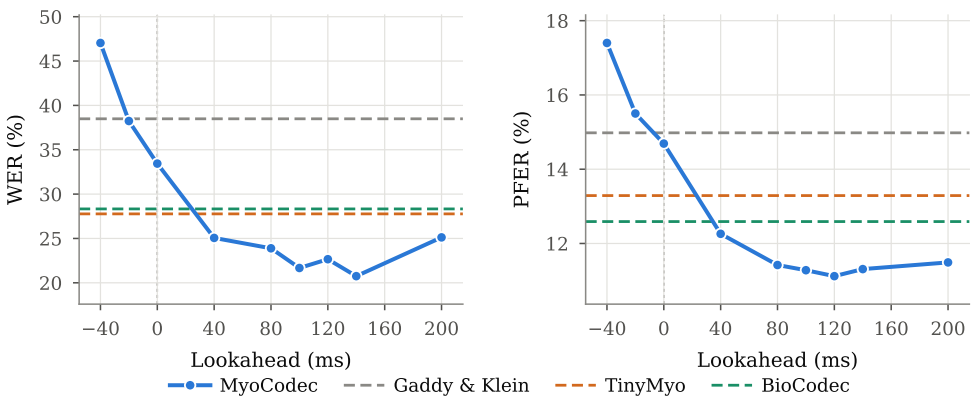


Figure 4: **emg2speech with different lookahead.** The best performance is achieved when lookahead between 100 ms–140 ms is applied.

Table 6: **speech2emg.** Envelope correlation coefficients (Env. CC) with 95% confidence intervals are reported using 20 ms and 50 ms smoothing windows. We compare offline and causal systems with different lookahead. The best and second-best are **bolded** and underlined, respectively.

		Env. CC ↑	
System	lookahead	50 ms window	20 ms window
STE-GAN (Scheck & Schultz, 2023)	∞ (offline)	**0.705** ± 0.014	**0.578** ± 0.012
TinyMyo (Fasulo et al., 2026)	∞ (offline)	0.577 ± 0.011	0.474 ± 0.010
BioCodec (Avramidis et al., 2025)	0 ms	0.511 ± 0.010	0.434 ± 0.008
	144 ms	0.600 ± 0.009	0.522 ± 0.008
MyoCodec (Ours)	0 ms	0.587 ± 0.010	0.503 ± 0.009
	120 ms	<u>0.655</u> ± 0.010	**0.578** ± 0.009

speech2emg. Table 6 compares MyoCodec with baseline systems under causal, limited-lookahead, and offline inference. Following Scheck & Schultz (2023), we use envelope correlation coefficient (Env. CC) as a proxy metric to assess the quality of generated EMG signals, evaluated with 20 ms and 50 ms smoothing windows. Non-causally trained STE-GAN (Scheck & Schultz, 2023) achieves the best performance, while causally trained MyoCodec closely matches 120 ms of lookahead. Note that STE-GAN and TinyMyo are trained non-causally, while BioCodec and MyoCodec are trained causally.

Streaming inference. We measure inference speed on a single `RTX PRO 6000 Blackwell` GPU. Table 7 compares inference speed for each codec on two bases: per-second chunks and per-frame streaming. On a per-second basis, MyoCodec requires 1.474 ms for encoding and decoding, compared with 2.032 ms for BioCodec. For frame-level streaming inference, MyoCodec maintains a state cache and processes each 20 ms frame using compiled `CUDA Graphs`. Encoding and decoding require 0.253 ms and 0.229 ms, respectively, for a total of 0.482 ms per frame, corresponding to 41× real-time throughput. On a single CPU thread (Intel Xeon Platinum 8559C), streaming inference compute times are 3.27 ms for encoding and 2.91 ms for decoding, 6.18 ms in total, corresponding to a real-time throughput of 3.24×.

3.3 Potential Future Applications

MyoCodec's compact size, causal operation, and discrete output at low bitrate make it promising for on-device streaming. Processing EMG input directly on a wearable could enable low-latency interaction, reduce transmission bandwidth, and limit the need to transmit raw biosignals. As EMG signals constitute sensitive physiological data, local processing could also improve privacy by keep-

Table 7: **Inference speed.** All measurements are obtained on a single `NVIDIA RTX PRO 6000 Blackwell` GPU. Per-second speed is measured using 5 s input windows and normalized by signal duration, while MYOCODEC's per-frame speed is measured using its compiled `CUDA Graph` implementation.

		Compute time (ms) ↓			Real-time throughput (×) ↑		
System	basis	Encoder	Decoder	Total	Encoder	Decoder	Total
MYOCODEC (Ours)	per second	0.782	0.692	1.474	1279	1445	678
	per frame	0.253	0.229	0.482	79	87	41
BioCodec	per second	1.091	0.941	2.032	917	1062	492
	per frame	N/A[†]					

[†] The official BioCodec release provides neither a state cache nor a frame-wise streaming interface.

ing raw signals on the device. These properties could broaden the practical and secure use of wearable EMG devices in assistive communication, gesture and prosthetic control, and hands-free input.

Beyond on-device deployment, encoding continuous EMG signals into low-bitrate discrete tokens provides a potential direct integration to modern language and speech-language models. As modern speech and language modeling approaches mostly operate through next-token prediction over discrete representations, MYOCODEC makes EMG signals more compatible with similar architectures and training objectives. This creates a path toward interactive multimodal systems that condition directly on muscle activity to generate text, speech, or actions without intermediate steps, potentially reducing the latency, information loss, and error propagation associated with cascaded pipelines.

The high reconstruction fidelity achieved by MYOCODEC's decoder suggests a complementary application in synthetic EMG generation. Generative models operating on MYOCODEC tokens could synthesize EMG signals from speech, motion, or task descriptors, providing synthetic training data when collecting large-scale labeled EMG is costly or impractical. Such signals could augment data-scarce training sets and support neuromuscular digital-human models that simulate task- and subject-dependent muscle activity. Realizing these applications, however, will require further validation of physiological realism, cross-channel consistency, and downstream utility.

3.4 LIMITATIONS

The publicly available EMG datasets are highly imbalanced to a couple of main sources, *emg2qwerty* (Sivakumar et al., 2024) and *emg2pose* (Salter et al., 2024), accounting for 71% of the collected data, whereas speech-related EMG data contribute only 1.9% (Appendix E). Such concentration could potentially limit uniform generalization across the broader range of EMG applications, sensor configurations, and recording conditions.

Our downstream comparisons also prioritize experimental control over task-specific optimization. For each downstream task, we change the EMG representation primarily while retaining the rest of the architecture. However, most downstream task models were originally designed for their original continuous EMG feature inputs and may not fully exploit discrete codec tokens. The reported results should therefore be interpreted as measuring the MYOCODEC's plug-in compatibility, rather than the maximum performance attainable with architectures designed specifically for such EMG representation.

4 CONCLUSION

We introduce MYOCODEC, a streaming neural codec that encodes EMG signals at 50 Hz and 2400 bits/s using causal Transformer blocks and residual vector quantization. Trained on twelve public corpora, MYOCODEC achieves the strongest reconstruction performance among the compared models. Its EMG representations can generalize across multiple downstream tasks, such as *emg2qwerty*, *emg2pose*, *emg2speech*, and *speech2emg*. This provides a practical foundation for on-device wearable systems and, more broadly, for direct integration of muscle activity with modern language and speech-language models.

Reproducibility Statement

To support reproducibility, we will publicly release the complete codebase and model weights upon publication.

Ethics statement

This work uses only publicly released EMG datasets, collected and distributed by their original authors under their own consent and IRB arrangements. No additional human-subject data were collected.

AI use statement

We used AI for literature summarization, experimental implementation assistance, and refining writing. The authors reviewed AI-assisted code implementation and verified its outputs. The research ideas, experimental design, interpretation of results, and scientific claims are from the authors. The authors take full responsibility for the final content of this work.

References

Christoph Amma, Thomas Krings, Jonas Böer, and Tanja Schultz. Advancing muscle-computer interfaces with high-density electromyography. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pp. 929–938, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450331456. doi: 10.1145/2702123.2702501. URL https://doi.org/10.1145/2702123.2702501.

Manfredo Atzori, Arjan Gijsberts, Claudio Castellini, Barbara Caputo, Anne-Gabrielle Mittaz Hager, Simone Elsig, Giorgio Giatsidis, Franco Bassetto, and Henning Müller. Electromyography data for non-invasive naturally-controlled robotic hand prostheses. *Scientific Data*, 1:140053, 2014. doi: 10.1038/sdata.2014.53.

Kleanthis Avramidis, Tiantian Feng, Woojae Jeong, Jihwan Lee, Wenhui Cui, Richard M Leahy, and Shrikanth Narayanan. Neural codecs as biosignal tokenizers. *arXiv preprint arXiv:2510.09095*, 2025.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Marco E. Benalcázar, Lorena Barona, Leonardo Valdivieso, Xavier Aguas, and Jonathan Zea. EMG-EPN-612 dataset, 2020.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. Audiolm: A language modeling approach to audio generation. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 31:2523–2533, June 2023. ISSN 2329-9290. doi: 10.1109/TASLP.2023.3288409. URL https://doi.org/10.1109/TASLP.2023.3288409.

Sanyuan Chen, Chengyi Wang, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. Neural codec language models are zero-shot text to speech synthesizers. *IEEE Transactions on Audio, Speech and Language Processing*, 33:705–718, 2025. doi: 10.1109/TASLPRO.2025.3530270.

Matteo Cognolato, Arjan Gijsberts, Valentina Gregori, Gianluca Saetta, Katia Giacomino, Anne-Gabrielle Mittaz Hager, Andrea Gigli, Diego Faccio, Cesare Tiengo, Franco Bassetto, Barbara Caputo, Peter Brugger, Manfredo Atzori, and Henning Müller. Gaze, visual, myoelectric, and inertial data of grasps for intelligent prosthetics. *Scientific Data*, 7:43, 2020. doi: 10.1038/s41597-020-0380-3.

Tri Dao, Daniel Y Fu, Stefano Ermon, Atri Rudra, and Christopher Re. Flashattention: Fast and memory-efficient exact attention with IO-awareness. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=H4DqfPSibmx.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=ivCd8z8zR2. Featured Certification, Reproducibility Certification.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: A speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*, 2024.

Matteo Fasulo, Giusy Spacone, Thorir Mar Ingolfsson, Yawei Li, Luca Benini, and Andrea Cossettini. Tinymyo: a tiny foundation model for flexible emg signal processing at the edge, 2026. URL https://arxiv.org/abs/2512.15729.

David Gaddy and Dan Klein. Digital voicing of silent speech. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5521–5530, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.445. URL https://aclanthology.org/2020.emnlp-main.445/.

David Gaddy and Dan Klein. An improved model for voicing silent speech. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 175–181, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.23. URL https://aclanthology.org/2021.acl-short.23/.

Weidong Geng, Yu Du, Wenguang Jin, Wentao Wei, Yu Hu, and Jiajun Li. Gesture recognition by instantaneous surface EMG images. *Scientific Reports*, 6:36571, 2016. doi: 10.1038/srep36571.

Harshavardhana T Gowda, Daniel C Comstock, and Lee M. Miller. emg2speech: synthesizing speech from electromyography using self-supervised speech models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16490–16507, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.750. URL https://aclanthology.org/2026.acl-long.750/.

Injune Hwang, Jaejun Lee, and Kyogu Lee. Emg-to-speech with fewer channels. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7261–7265, 2026. doi: 10.1109/ICASSP55912.2026.11465132.

Matthias Janke and Lorenz Diener. Emg-to-speech: Direct generation of speech from facial electromyographic signals. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25 (12):2375–2385, 2017. doi: 10.1109/TASLP.2017.2738568.

Weibang Jiang, Liming Zhao, and Bao liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=QzTpTRVtrP.

Xinyu Jiang, Xiangyu Liu, Jiahao Fan, Xinming Ye, Chenyun Dai, Edward A. Clancy, Metin Akay, and Wei Chen. Open access dataset, toolbox and benchmark processing results of high-density surface electromyogram recordings. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:1035–1046, 2021. doi: 10.1109/TNSRE.2021.3082551.

Keller Jordan, Yuchen Jin, Vlado Boza, You Jiacheng, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL https://kellerjordan.github.io/posts/muon/.

Piotr Kaczmarek, Tomasz Mańkowski, and Jakub Tomczyński. putEMG—a surface electromyography hand gesture recognition dataset. *Sensors*, 19(16):3548, 2019. doi: 10.3390/s19163548.

Patrick Kaifosh, Thomas R. Reardon, and CTRL-labs at Reality Labs. A generic non-invasive neuromotor interface for human-computer interaction. *Nature*, 2025. doi: 10.1038/s41586-025-09255-w.

Agamemnon Krasoulis, Iris Kyranou, Mustapha Suphi Erden, Kianoush Nazarpour, and Sethu Vijayakumar. Improved prosthetic hand control with concurrent use of myoelectric and inertial measurements. *Journal of NeuroEngineering and Rehabilitation*, 14(71), 2017.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved RVQGAN. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=qjnl1QUnFA.

Jihwan Lee, Kevin Huang, Kleanthis Avramidis, Simon Pistrosch, Monica Gonzalez-Machorro, Yoonjeong Lee, Björn W. Schuller, Louis Goldstein, and Shrikanth Narayanan. Articulatory Feature Prediction from Surface EMG during Speech Production. In *Interspeech 2025*, pp. 320–324, 2025. doi: 10.21437/Interspeech.2025-565.

Jihwan Lee, Parsa Razmara, Kevin Huang, Sean Foley, Aditya Kommineni, Haley Hsu, Woojae Jeong, Prakash Kumar, Xuan Shi, Yoonjeong Lee, Tiantian Feng, Takfarinas Medani, Ye Tian, Sudarsana Reddy Kadiri, Krishna S. Nayak, Dani Byrd, Louis Goldstein, Richard M. Leahy, and Shrikanth Narayanan. An approach to simultaneous acquisition of real-time mri video, eeg, and surface emg for articulatory, brain, and muscle activity during speech production. In *Interspeech 2026*, 2026a. URL https://arxiv.org/abs/2603.04840.

Junhyeok Lee, Xiluo He, Jihwan Lee, Helin Wang, Shrikanth Narayanan, Thomas Thebaud, Laureano Moro-Velazquez, Jesús Villalba, and Najim Dehak. Reconstruct! don't encode: Self-supervised representation reconstruction loss for high-intelligibility and low-latency streaming neural audio codec. *arXiv preprint arXiv:2603.05887*, 2026b.

Geoffrey S. Meltzner, James T. Heaton, Yunbin Deng, Gianluca De Luca, Serge H. Roy, and Joshua C. Kline. Development of sEMG sensors and algorithms for silent speech recognition. *Journal of Neural Engineering*, 15(4):046031, 2018.

Francesca Palermo, Matteo Cognolato, Arjan Gijsberts, Henning Müller, Barbara Caputo, and Manfredo Atzori. Repeatability of grasp recognition for robotic hand prosthesis control based on semg data. In *2017 International Conference on Rehabilitation Robotics (ICORR)*, pp. 1154–1159, 2017. doi: 10.1109/ICORR.2017.8009405.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Interspeech 2019*, pp. 2613–2617, 2019. doi: 10.21437/Interspeech.2019-2680.

Simon Pistrosch, Kleanthis Avramidis, Zhao Ren, Tiantian Feng, Jihwan Lee, Monica Gonzalez-Machorro, Anton Batliner, Tanja Schultz, Shrikanth Narayanan, and Björn W. Schuller. Affect decoding in phonated and silent speech production from surface emg, 2026. URL https://arxiv.org/abs/2603.11715.

Stefano Pizzolato, Luca Tagliapietra, Matteo Cognolato, Monica Reggiani, Henning Müller, and Manfredo Atzori. Comparison of six electromyography acquisition setups on hand movement classification tasks. *PLOS ONE*, 12(10):e0186132, 2017.

Ashirbad Pradhan, Jiayuan He, and Ning Jiang. Multi-day dataset of forearm and wrist electromyogram for hand gesture recognition and biometrics. *Scientific Data*, 9:733, 2022. doi: 10.1038/s41597-022-01836-y.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R. Hershey. Sdr – half-baked or well done? In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 626–630, 2019. doi: 10.1109/ICASSP.2019.8683855.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022. In *Interspeech 2022*, pp. 4521–4525, 2022. doi: 10.21437/Interspeech.2022-439.

Sasha Salter, Richard Warren, Collin Schlager, Adrian Spurr, Shangchen Han, Rohin Bhasin, Yujun Cai, Peter Walkington, Anuoluwapo Bolarinwa, Robert Wang, Nathan Danielson, Josh Merel, Eftychios Pnevmatikakis, and Jesse Marshall. emg2pose: A large and diverse benchmark for surface electromyographic hand pose estimation. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 55703–55728. Curran Associates, Inc., 2024. doi: 10.52202/079017-1770. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/64f8884f6ba3d9ace5bb647c4f917896-Paper-Datasets_and_Benchmarks_Track.pdf.

Kevin Scheck and Tanja Schultz. STE-GAN: Speech-to-Electromyography Signal Conversion using Generative Adversarial Networks. In *Interspeech 2023*, pp. 1174–1178, 2023. doi: 10.21437/Interspeech.2023-174.

Tanja Schultz and Michael Wand. Modeling coarticulation in EMG-based continuous speech recognition. *Speech Communication*, 52(4):341–353, 2010.

Viswanath Sivakumar, Jeffrey Seely, Alan Du, Sean R Bittner, Adam Berenzweig, Anuoluwapo Bolarinwa, Alexandre Gramfort, and Michael I Mandel. emg2qwerty: A large dataset with baselines for touch typing using surface electromyography. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 91373–91389. Curran Associates, Inc., 2024. doi: 10.52202/079017-2899. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/a64d53074d011e49af1dfc72c332fe4b-Paper-Datasets_and_Benchmarks_Track.pdf.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomput.*, 568(C), February 2024. ISSN 0925-2312. doi: 10.1016/j.neucom.2023.127063. URL https://doi.org/10.1016/j.neucom.2023.127063.

Haibin Wu, Naoyuki Kanda, Sefik Emre Eskimez, and Jinyu Li. TS3-Codec: Transformer-Based Simple Streaming Single Codec. In *Interspeech 2025*, pp. 604–608, 2025. doi: 10.21437/Interspeech.2025-921.

Qiantong Xu, Alexei Baevski, and Michael Auli. Simple and Effective Zero-shot Cross-lingual Phoneme Recognition. In *Interspeech 2022*, pp. 2113–2117, 2022. doi: 10.21437/Interspeech.2022-60.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. SoundStream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2021.

A Related Work

Biosignal foundation models. A parallel line of work pretrains large encoders on physiological signals and transfers their continuous representations to downstream tasks. LaBraM (Jiang et al., 2024), for example, pretrains a masked-prediction Transformer on thousands of hours of EEG data drawn from around dozens of different datasets. It uses a vector-quantized spectral tokenizer to construct pretraining targets, transferring the resulting encoder to abnormality detection, event classification, and emotion recognition. For EMG, Kaifosh et al. (2025) train a unified wrist-worn decoder across thousands of participants and demonstrate generalization to previously unseen users without

person-specific calibration. This provides strong evidence that pretrained EMG representations can transfer across individuals. TinyMyo (Fasulo et al., 2026) is a compact EMG foundation model built for edge deployment. It embeds non-overlapping 20-sample patches of each channel, giving a frame rate of $100\,\mathrm{Hz}$, and carries a learned bank of 16 channel embeddings so that a single model serves different electrode layouts. At roughly $3.6\,\mathrm{M}$ parameters, it is one of the smallest of these encoders, outputting 192-dimensional continuous embeddings. We use TinyMyo as a baseline representation on all four downstream tasks.

Neural audio codecs. Discrete neural audio codecs typically combine an encoder, RVQ, and a decoder trained with reconstruction and adversarial objectives. SoundStream (Zeghidour et al., 2021) introduced the RVQ-GAN formulation, EnCodec (Défossez et al., 2023) incorporated multi-scale spectral discrimination and streaming inference, and DAC (Kumar et al., 2023) improved codebook utilization through factorized and L_2-normalized codes. Mimi (Défossez et al., 2024) further distilled phonetic information into the first RVQ stage, producing tokens useful for language modeling as well as waveform reconstruction. TS3-Codec (Wu et al., 2025) adopts a fully causal Transformer decoder for neural audio codecs. JHCodec (Lee et al., 2026b) extends this framework with a self-supervised representation reconstruction loss to enable stable streaming without lookahead.

BioCodec. BioCodec (Avramidis et al., 2025) is a general-purpose neural codec for biosignal tokenization and therefore serves as our primary codec baseline. It follows an EnCodec-style convolutional architecture comprising a causal SEANet encoder and a six-stage RVQ with 256 entries per stage. In its EMG configuration, BioCodec operates at $27.78\,\mathrm{Hz}$ and $1333\,\mathrm{bits/s}$ per channel. The EMG model was pretrained on *emg2qwerty* (Sivakumar et al., 2024) and evaluated primarily as a tokenizer for downstream gesture classification.

emg2qwerty. The task is to decode wrist-worn EMG into the keystrokes a person types (Sivakumar et al., 2024). Its published front end is a log-spectrogram followed by a rotation-invariant band MLP and a time-depth-separable convolutional encoder trained with CTC, and its headline result is that a generic model trained across participants transfers to held-out ones.

emg2pose. The target is to decode the joint angles of the hand, with motion capture providing the reference. Salter et al. (2024) pair wrist EMG with an optical hand-tracking rig over 193 participants and release both the corpus and a recurrent decoder.

emg2speech. Articulatory EMG recorded from the face and neck carries enough information to decode speech, including when the speech is silently mouthed. Schultz & Wand (2010) established continuous recognition from this signal and showed that modeling coarticulation is what makes it tractable, and Janke & Diener (2017) went further by generating a speech waveform directly from the EMG rather than passing through text. Gaddy & Klein (2020; 2021) released a single-speaker *emg2speech* corpus with both vocalized and silent speech. Meltzner et al. (2018) moved the setting toward practical use, developing EMG sensor arrays and recognition algorithms for silent speech with a vocabulary large enough to be useful rather than demonstrative. Gowda et al. (2026) extend the direction to a multi-speaker corpus, including a participant with ALS. A related line explores what information is encoded in the signals. Lee et al. (2025) predict articulatory features directly from surface EMG during speech production, which is used as the intermediate representation for speech synthesis. Taking that further, Lee et al. (2026a) explore acquiring real-time MRI, EEG, and EMG simultaneously during speech, which places EMG explicitly between the neural command and the articulator motion it produces. Pistrosch et al. (2026) decode affect from surface EMG in both phonated and silent production. Hwang et al. (2026) study how far the electrode count can be reduced, finding that the best subsets are those whose channels are complementary rather than individually strongest.

speech2emg. This is the reverse task of *emg2speech*, namely synthesizing EMG from speech input. Scheck & Schultz (2023) introduce STE-GAN, an adversarial model that generates articulatory EMG conditioned on speech features.

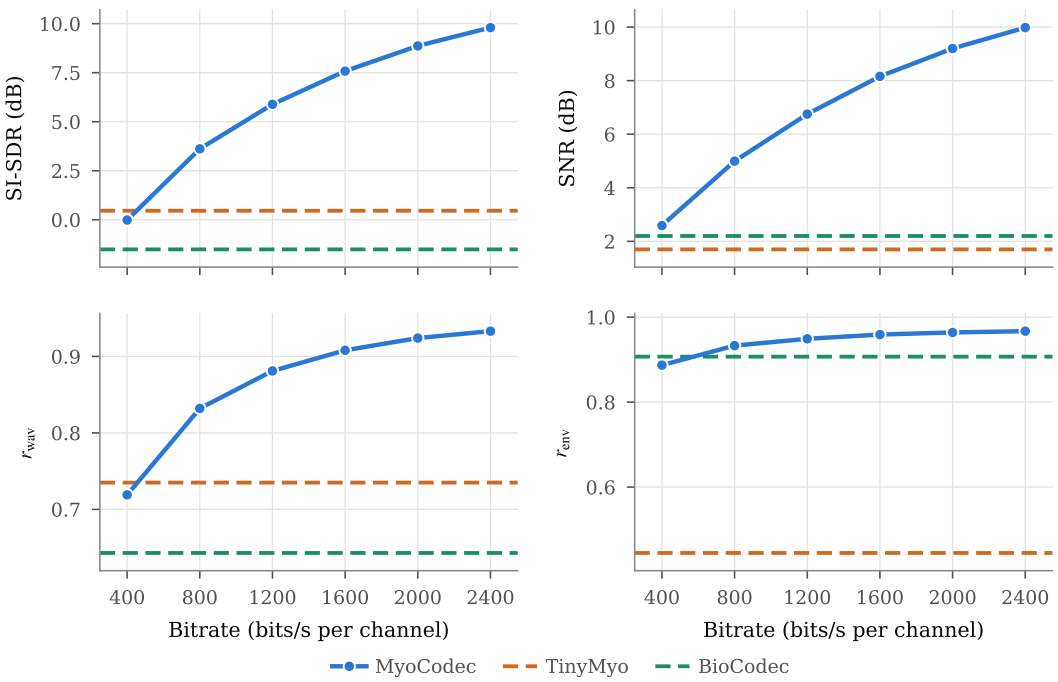


Figure 5: **Reconstruction quality across various bitrates.** MYOCODEC outperforms TinyMyo and BioCodec on reconstruction quality using only two codebooks.

B BITRATE COMPARISON

As the quantizer is residual, it allows to use only the first N_q codebooks, reducing the bitrate in increments of $400\,\text{bits/s}$. Figure 5 shows the bitrate–quality trade-off for $N_q \in \{1, \ldots, 6\}$. Using only two codebooks ($800\,\text{bits/s}$, approximately 60% of BioCodec's $1333\,\text{bits/s}$), MYOCODEC outperforms BioCodec across all four reconstruction metrics. With only one codebook ($400\,\text{bits/s}$), MYOCODEC continues to lead in SI-SDR, SNR, and r_{wav}, while trailing in r_{env}.

C DETAILED IMPLEMENTATION OF MYOCODEC

C.1 ARCHITECTURE

Encoder. All EMG inputs are sampled at $2\,\text{kHz}$ and patchified into non-overlapping frames of $F = 40$ samples, corresponding to a frame duration of $20\,\text{ms}$ and a token rate of $50\,\text{Hz}$. We treat each channel independently, as the datasets use different electrode configurations, supporting arbitrary electrode configurations. Each frame is projected linearly from 40 samples to a model width of $d = 256$. The encoder comprises eight pre-norm Transformer layers, each with eight attention heads of dimension 32, an MLP hidden dimension of 1024, and no dropout. We use rotary positional embeddings (Su et al., 2024) and FlashAttention (Dao et al., 2022) for all attention operations. Self-attention is causal and restricted to a 64-frame sliding window. During streaming inference, each layer retains a 64-frame key–value cache, keeping the per-frame computation independent of sequence length. The encoder size is approximately $6.3\,\text{M}$ parameters.

Quantizer. The encoder output is projected from 256 to 32 dimensions before quantization. We use a residual vector quantizer with $N_q = 6$ stages, each containing $K = 256$ entries of dimension 32. A learned projection maps the quantized representation back to 256 dimensions before fed into the decoder. Codebook entries are initialized from $\mathcal{N}(0, 1)$, giving an expected norm of approximately $\sqrt{32}$. Assignment counts are tracked separately for each entry using an exponential moving average with decay 0.99. Every 250 optimization steps, entries whose averaged count falls below 0.9 are reinitialized from encoder outputs sampled from the current batch, and their count is reset

to 1.8. This mechanism provides a grace period after reinitialization, maintains full utilization of all six codebooks, and helps prevent codebook collapse.

Decoder. The decoder mirrors the encoder, using eight pre-norm causal Transformer layers with the same width, attention configuration, MLP dimension, rotary embeddings, and 64-frame sliding window. It takes the dequantized 256-dimensional sequence as input and applies a final linear projection to reconstruct $F = 40$ waveform samples at each frame. During streaming, the decoder maintains the same per-layer key–value cache as the encoder. It contains approximately $6.3\,\mathrm{M}$ parameters; including the quantizer and projection layers, the complete codec contains $12.72\,\mathrm{M}$ parameters.

Discriminators. Adversarial training uses two complementary discriminator branches. The multi-period discriminator reshapes the waveform using periods $\{2, 3, 5, 7, 11\}$, with 8 base filters increasing to at most 128. The multi-scale STFT discriminator uses FFT and window sizes $\{64, 128, 256, 512\}$, hop lengths $\{17, 31, 67, 131\}$, and 16 base filters increasing to at most 128. The discriminators are used only during Phase 2 and are discarded at inference.

C.2 Training Objectives

Phase 1. Phase 1 optimizes the codec objective in Equation 1, which combines a time-domain Huber loss, multi-scale spectral reconstruction, vector-quantization loss, and commitment loss. The spectral objective is averaged over FFT sizes $\{64, 128, 256\}$, corresponding to $32\,\mathrm{ms}$ to $128\,\mathrm{ms}$ at $2\,\mathrm{kHz}$. Each scale uses a Hann window with hop size $n/4$ and combines log-magnitude error, spectral convergence, and phase error, with the phase term weighted by 0.1 as defined in Equation 2. Magnitudes are floored at 10^{-7} before taking the logarithm, and phase is compared using principal-value angles without unwrapping. Phase 1 runs for 100k steps on the Muon optimizer (Jordan et al., 2024) with a batch size 1024 and no discriminator.

Huber loss. The time-domain reconstruction term is the mean Huber penalty between the target x and reconstruction $\hat{x}$:

$$\mathcal{L}_{\mathrm{Huber}}(x, \hat{x}) = \frac{1}{T} \sum_{t=1}^{T} h_\beta(\hat{x}_t - x_t), \qquad h_\beta(d) = \begin{cases} \frac{1}{2} d^2, & |d| < \beta, \\ \beta \left(|d| - \frac{1}{2}\beta \right), & |d| \geq \beta. \end{cases}$$

where $\beta = 1$. The quadratic region provides smooth gradients for small reconstruction errors, while the linear region limits the influence of isolated high-amplitude bursts and motion artifacts.

Vector-quantization losses. Let $r^{(1)} = z$ denote the projected encoder output, $e^{(i)}$ the nearest entry selected from codebook i, and $\mathrm{sg}[\cdot]$ the stop-gradient operator. Successive RVQ stages operate on the residual

$$r^{(i+1)} = r^{(i)} - \mathrm{sg}\left[e^{(i)}\right].$$

The vector-quantization and commitment losses are summed over all $N_q = 6$ stages:

$$\mathcal{L}_{\mathrm{VQ}} = \sum_{i=1}^{N_q} \left\| \mathrm{sg}\left[r^{(i)}\right] - e^{(i)} \right\|^2, \qquad \mathcal{L}_{\mathrm{commit}} = \sum_{i=1}^{N_q} \left\| r^{(i)} - \mathrm{sg}\left[e^{(i)}\right] \right\|^2.$$

Each squared norm is averaged over the 32 code dimensions and the batch and time axes. The two losses differ in their gradient paths: $\mathcal{L}_{\mathrm{VQ}}$ updates the selected codebook entries, whereas $\mathcal{L}_{\mathrm{commit}}$ updates the encoder and discourages its outputs from drifting away from the codebook. We assign the commitment term a weight of 0.25, compared with unit weight for the VQ loss.

Let $q = \sum_{i=1}^{N_q} e^{(i)}$ denote the quantized representation. As nearest-neighbor assignment is non-differentiable, the decoder receives the straight-through estimate

$$z + \mathrm{sg}[q - z],$$

which equals q during the forward pass but has an identity gradient with respect to z. Detaching each selected entry in the residual recursion also prevents later RVQ stages from propagating gradients through earlier codebooks.

Phase 2. Phase 2 retains the complete Phase 1 objective and adds the adversarial and feature-matching terms in Equation 3. Their target weights are

$$\lambda_{\text{adv}} = 0.1, \qquad \lambda_{\text{FM}} = 2.0.$$

Both weights are increased linearly from zero over the first $50\,000$ steps of Phase 2, providing a gradual transition to adversarial training. The discriminators are optimized with AdamW using a learning rate of 10^{-4}, weight decay of 10^{-4}, and gradient clipping at 1.0. As the discriminators dominate memory usage, the batch size is reduced from 1024 to 192. Phase 2 runs for an additional $100\,000$ steps, giving $200\,000$ optimization steps in total.

D Downstream implementation

emg2qwerty. The downstream model only differs in inputted EMG features. For MyoCodec and BioCodec, the downstream model receives the quantized RVQ output, while TinyMyo takes continuous encoder features from its encoder. We retain the multi-band rotation-invariant MLP and CTC output layer from Sivakumar et al. (2024). The MLP has one hidden layer of width 384 and produces a 768-dimensional representation, followed by a `Linear(768, 99)` classifier. The temporal encoder is a four-layer causal Transformer with input and output projections of $768 \to 256$ and $256 \to 768$, respectively. Each layer uses four attention heads of dimension 64, an MLP hidden dimension of 1536, and dropout 0.1. Beam-search decoding follows Sivakumar et al. (2024), using a beam width of 50, a language-model weight of 2.0, an insertion bonus of 2.0, and a 6-gram character language model trained on WikiText-103. To isolate the effects of the input representation and temporal lookahead, we omit EMG augmentation. In particular, SpecAugment (Park et al., 2019) is not applicable to the non-spectrogram MyoCodec representation, and temporal jittering is excluded because it would confound the lookahead analysis.

emg2pose. The codec replaces the `TdsNetwork` featurizer (Salter et al., 2024), a causal strided convolutional network that maps 16-channel EMG sampled at $2\,\text{kHz}$ to 64-dimensional features at $50\,\text{Hz}$ with an overall stride of 40. For all three front ends the downstream model receives the encoder output, the hidden state before quantization. Multi-head attention mixes information across the 16 electrodes within each frame without introducing additional lookahead, and a two-layer projection with hidden width 256 maps the result to a 128-dimensional feature vector. All subsequent components are retained from the published *emg2pose* model (Salter et al., 2024), including the velocity-integrating LSTM decoder, whose 148-dimensional input concatenates the projected EMG features with the 20 joint angles from the previous predicted state.

emg2speech. The three pretrained front ends are frozen and provide continuous features. The prediction target is the continuous 1024-dimensional pre-quantization latent of JHCodec. For the three pretrained front ends, a three-layer bidirectional channel Transformer mixes the eight electrodes within each frame at width 256 using four attention heads and dropout 0.2, followed by a linear projection to 512 dimensions. A causal temporal Transformer then models the sequence using eight layers, four 128-dimensional attention heads, an MLP hidden dimension of 2048, dropout 0.2, rotary embeddings, and a 128-frame sliding window. The output head consists of $512 \to 1024$, SiLU, and $1024 \to 1024$ layers, and an auxiliary $512 \to 48$ phoneme classifier contributes to the loss with weight 0.5. The JHCodec latent is predicted using an ℓ_1 loss. All components from the temporal Transformer onward are shared across the four systems. Training uses the vocalized set of the Gaddy and Klein corpus. Note that the Gaddy and Klein encoder is bidirectional, serving as an offline reference.

speech2emg. *speech2emg* predicts EMG from speech. The input is JHCodec's 1024-dimensional pre-quantization speech representation at $50\,\text{Hz}$. The continuous prediction target depends on the EMG encoder: MyoCodec and BioCodec use the bottleneck projection, and TinyMyo uses its encoder output. The predictor comprises a six-layer causal temporal Transformer (width 512, eight heads) and a two-layer bidirectional channel Transformer over the eight electrodes (width 256, four heads), followed by a linear regression head. Dropout is 0.1. The objective combines mean-squared error with STE-GAN's multi-time-domain loss (Scheck & Schultz, 2023).

Evaluation metrics. We report a separate set of evaluation metrics per considered task:

- *emg2qwerty.* We report CER, which is measured using greedy decoding and the beam-search decoder with a 6-gram character language model, identical to Sivakumar et al. (2024).
- *emg2pose.* We report the mean absolute error of the 20 joint angles in degrees. We additionally report Euclidean landmark and fingertip distances in millimetres after applying forward kinematics with the published hand model. Landmark distance is averaged over all articulated landmarks, and fingertip distance over the five fingertips.
- *emg2speech.* We report CER, WER, PER, and PFER, and UTMOS (Saeki et al., 2022). CER and WER are measured using Whisper large-v3 (Radford et al., 2023). PER and PFER are measured by a wav2vec 2.0 phoneme recognizer (Baevski et al., 2020; Xu et al., 2022). PFER assigns partial substitution costs according to shared articulatory features, while UTMOS provides a proxy to perceptual quality.
- *speech2emg.* We report the envelope correlation coefficient (Env. CC). Following STEGAN (Scheck & Schultz, 2023), envelopes are obtained by rectifying each signal and applying a moving-average filter.

E PRETRAINING CORPUS

Table 8: The twelve public EMG corpora. Channel-hours count only the channels and recordings retained after filtering.

Corpus	ch.	subj.	ch.-hours	content
emg2qwerty (Sivakumar et al., 2024)	32	108	11,074	typing
emg2pose (Salter et al., 2024)	16	193	5,925	hand pose
Hyser (Jiang et al., 2021)	256	20	2,593	finger force
putEMG (Kaczmarek et al., 2019)	24	44	1,045	hand gestures
EMG-EPN-612 (Benalcázar et al., 2020)	8	612	1,014	hand gestures
MeganePro (Cognolato et al., 2020)	12	45	610	grasping
GRABMyo (Pradhan et al., 2022)	28	43	597	hand gestures
Ninapro*	12–14	82	503	hand movements
emg2speech (ALS) (Gowda et al., 2026)	31	2	304	speech (incl. ALS)
CSL-HDEMG (Amma et al., 2015)	192	5	232	finger gestures
emg2speech (Gaddy) (Gaddy & Klein, 2020)	8	1	156	speech
CapgMyo DB-a (Geng et al., 2016)	128	18	51	hand gestures
Total			**24,104**	

* Only DB2 (Atzori et al., 2014), DB4 (Pizzolato et al., 2017), DB6 (Palermo et al., 2017) and DB7 (Krasoulis et al., 2017) are used.