GMSLM: Generative Marmoset Spoken Language Modeling

Talia Sternberg[1], Mickey London[2], David Omer[2]*, Yossi Adi[1]*
[1]The School of Computer Science and Engineering
[2]The Edmond and Lily Safra center for Brain Sciences (ELSC)
Hebrew University of Jerusalem, Israel
talia.sternberg@mail.huji.ac.il

Abstract

Marmoset monkeys exhibit complex vocal communication, challenging the view that nonhuman primates' vocal communication is entirely innate, and show similar features of human speech, such as vocal labeling of others and turn-taking. Studying their vocal communication offers a unique opportunity to link it with brain activity—especially given the difficulty of accessing the human brain in speech and language research. Since Marmosets communicate primarily through vocalizations, applying standard LLM approaches is not straightforward. We introduce *Generative Marmoset Spoken Language Modeling* (GMSLM), an optimized spoken language model pipeline for Marmoset vocal communication. We designed a novel zero-shot evaluation metrics using unsupervised in-the-wild data, alongside weakly labeled conversational data, to assess GMSLM and demonstrate its advantage over a basic human-speech-based baseline. GMSLM generated vocalizations closely matched real resynthesized samples acoustically and performed well on downstream tasks. Despite being fully unsupervised, GMSLM effectively distinguish real from artificial conversations and may support further investigations of the neural basis of vocal communication and provides a practical framework linking vocalization and brain activity. We believe GMSLM stands to benefit future work in neuroscience, bioacoustics, and evolutionary biology. Samples are provided under: this link.

1 Introduction

Marmoset monkeys are small primates with surprisingly complex vocal communication. Although it was once assumed that nonhuman primates' vocal communication is entirely innate and inflexible, recent studies show that Marmosets can learn new vocalizations and even label their conspecifics (Oren et al., 2024a). These findings open the door for advanced computational analyses powered by large language models—to systematically investigate and decode Marmoset vocalization patterns, as a novel approach to uncover the cognitive and evolutionary mechanisms behind vocal communication. However, as Marmosets use vocal communication, with no well-defined orthography, it is unclear how recent advancements in LLMs, typically operate over discrete tokens, could be effectively utilized.

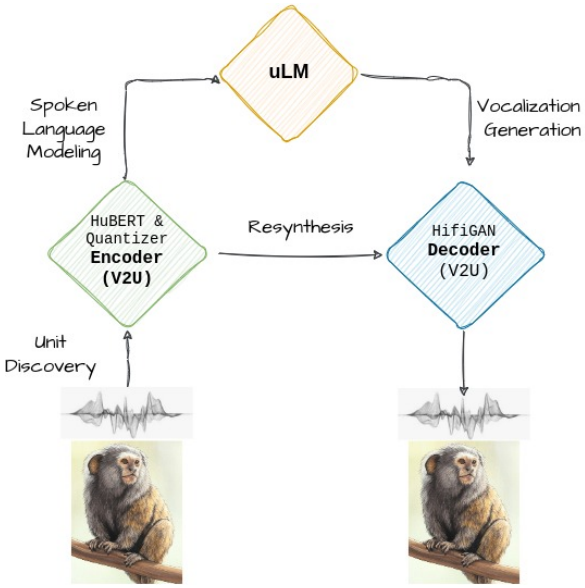


Figure 1: High level description of GMSLM.

In recent years, the field of *Generative Spoken Language Modeling* (GSLM), often known to as textless NLP, has gained traction (Lakhotia et al., 2021; Hassid et al., 2024), demonstrating remarkable performance across various speech and audio tasks, encompassing both generative (Kharitonov et al., 2023; Kreuk et al., 2022) and discriminative approaches (Chang et al., 2024). The objective of GSLM is to develop a spoken language model that operates without text. The typical GSLM framework begins with an unsupervised process

*Equal-contribution last authors

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 20599–20618
November 4-9, 2025

for acoustic units discovery, producing a discrete representation of the audio signal. Then, a language model is applied to these units to estimate the likelihood of the audio signal. Finally, a generative network converts the units back into a time-domain audio signal. Since the GSLM pipeline enables efficient encoding, sequential modeling, and decoding of raw audio signals without relying on textual supervision, it could be ideally suited for modeling Marmosets vocalizations.

In this work, we introduce GMSLM (stands for Generative Marmoset Spoken Language Modeling), a tailored adaptation of the GSLM pipeline for Marmosets-vocalizations, where each modeling component is optimized specifically for Marmoset-vocalization. To evaluate GMSLM, we developed a set of zero-shot evaluation metrics to measure the quality of its sequential modeling and conducted additional evaluations using weakly labeled conversational Marmoset data. A visual description can be seen on Figure 1. We compare the proposed approach to the naive human-speech based baseline, and demonstrate GMSLM significantly enhances system performance. Finally, we assess the quality of the sampled and generated Marmosets-vocalizations and empirically show these are comparable to their resynthesized counterparts. Leveraging manually annotated data (at the speaker level), we empirically demonstrate that, despite being entirely unsupervised, the proposed method assigns higher likelihood to authentic Marmosets conversations compared to unnatural ones. We evaluate GMSLM on downstream supervised tasks, where it performs strongly, confirming its utility beyond the training objective. Ablation studies support the need for our model architecture given the data's complexity. We also use GMSLM to explore marmosets' vocal communication and interpret learned representations. Overall, our work establishes a foundation for future research in animal language and related downstream applications. More broadly, it opens a new direction for unsupervised spoken language modeling in nonhuman species. We hope it inspires further research in this emerging field.

2 Background

2.1 Generative spoken language modeling

The general Generative Spoken Language Modeling (GSLM) pipeline is comprised of three main modules: (i) Speech-to-unit, (ii) Unit language model, and (iii) Unit-to-speech, where each of these modules is trained separately. Speech resynthesis can be achieved while ignoring the language model and directly feeding the quantized units into the unit-to-speech module (Polyak et al., 2021). In the following paragraphs, we give detailed background for the three components mentioned above.

Speech-to-unit module encodes the raw speech signal into a discrete representation. The common approach is first to encode the speech into a continuous representation and then quantize the representation to achieve a sequence of discrete units (Tjandra et al., 2020; Polyak et al., 2021).

Formally, denote the domain of audio samples by $\mathcal{X} \subset \mathbb{R}$. A raw signal is represented by a sequence of samples $x = (x_1, \ldots, x_T)$, where $x_t \in \mathcal{X}$ for all $1 \leq t \leq T$. Consider an encoder, f that gets as input the speech utterance and produces low-frequency features $f(x) = (v_1, \ldots, v_{T'})$. Lakhotia et al. (2021), evaluated several speech encoders, namely, Mel-spectrogram, wav2vec2 (Baevski et al., 2020), and HuBERT (Hsu et al., 2021). A k-means algorithm is applied over the models' outputs to generate discrete units, denoted as $z = (z_1, \ldots, z_{T'})$. Each element z_i in z is a positive integer, $z_i \in \{1, .., K\}$ for $1 \leq i \leq T'$, K is the number of discrete units.

Unit language model is trained on the unit sequences z, learning a distribution over them. This enables unconditional or conditional speech generation without text. Unlike text-based models, unit LMs can capture prosody (Kharitonov et al., 2021), speaker identity (Borsos et al., 2022), or even natural dialogues (Nguyen et al., 2022).

Unit-to-speech module converts the speech discrete units to a raw waveform. Lakhotia et al. (2021) used a Tacotron2.0 (Shen et al., 2018) based model followed by WaveGlow (Prenger et al., 2019) vocoder. Later, Polyak et al. (2021) proposed a more efficient HiFi-GAN-based unit which convert units to speech directly. Further work (Kreuk et al., 2021; Lee et al., 2021) added emotional and duration modeling for tasks like emotion transfer and speech-to-speech translation

2.2 Marmoset vocalization background

Marmoset monkey (Callithrix jacchus) are highly social primates that exhibit a wide repertoire of vocalizations (Bezerra and Souto, 2008). Their vocal production flexibility, including modulation of call features such as duration, intensity, complexity, and timing (Brumm et al., 2004; Eliades

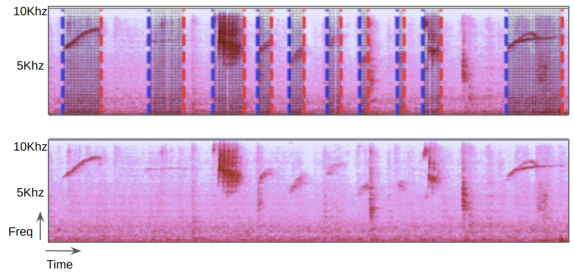


Figure 2: Segmentation example over an 18-second spectrogram from the original dataset. The x-axis represents time, while the y-axis represents frequency. As observed, the call bandwidth exceeds 5 kHz, with well-defined onset (blue lines) and offset (red lines) points computed in the preprocessing phase.

and Wang, 2012; Roy et al., 2011), enable them to encode a wide range of social and emotional information (Seyfarth and Cheney, 2003), including callers' identity, sex and group affiliation,as well as group dialect, and receivers' identity (Norcross and Newman, 1993; Zürcher and Burkart, 2017; Jones et al., 1993).

One particularly well-studied vocalization is the Phee call (Chen et al., 2009), a contact call ranging from 5.5 to 10kHz, which Marmosets is useed to engage in turn-taking dialogues with other conspecifics. Recent findings further suggest that Marmosets use the Phee call to encode the receiver's identity, in a manner analogous to how humans use names to address each other (Oren et al., 2024a). These vocal characteristics resemble human speech properties, including turn-taking in dialogue and vocal labeling of others (Oren et al., 2024a; Osmanski and Wang, 2023). Thus, Marmosets serve as a comparative model for vocal communication, offering insights for linguistics, cognitive science, and evolutionary biology.

3 Approach

Recall that the proposed approach, (GMSLM) builds on the GSLM pipeline, which was originally designed for human speech. In this section, we outline the modifications and adaptations made to better suit Marmoset vocalizations. We begin by detailing the data collection and preprocessing phase, followed by an explanation of the adjusted system components. Finally, we describe the system evaluation process.

3.1 Data

Dataset collection. We utilize two distinct Marmoset-vocalization datasets. The first, referred to as COLONYDB, consists of unsupervised raw Marmoset vocalization recordings and is used for training GMSLM. The second, called PHEEDB, is a smaller, weakly supervised dataset primarily used for model evaluation.

COLONYDB comprises continuous 24/7 audio recordings of spontaneous, multi-speaker Marmoset vocalizations in a colony room. Monkeys, housed in family cages without visual contact but with vocal communication, produced diverse vocal exchanges. The dataset features a rich and diverse range of vocalizations exchanged naturally between different monkeys. Recordings were collected using an omnidirectional microphone positioned at the center of the colony room.

PHEEDB was collected in a controlled setup where monkey pairs, separated by a visual barrier, spontaneously engaged in *Phee* call dialogues. Unlike colony recordings, these sessions captured structured turn-taking with a single vocalization type, and each caller's identity was fully labeled.

Pre-processing. We apply a multi-stage pipeline to segment Marmoset calls while minimizing environmental noise. A high-pass filter at 5 kHz removes low-frequency noise, followed by spectrogram-based noise filtering and duration-based segmentation. This approach ensures robust call detection, achieving a precision of 0.975 and recall of 0.78. For a detailed explanation of the segmentation algorithm, see Appendix A.1. Fig 3 provides statistics on the pre-processed data, while a visual representation of the spectrogram is shown in Fig 2.

3.2 GMSLM

Vocalization-to-Unit (V2U). We adapt the HuBERT model, pretrained on LibriSpeech (Panayotov et al., 2015), to tokenize Marmoset vocalizations. HuBERT learns speech representations via a masked prediction task on continuous audio features. To obtain discrete units, we apply k-means clustering on layer activations. Our adaptation follows a two-stage process: (i) training k-means on intermediate-layer features from HuBERT, then (ii) fine-tuning HuBERT using the new quantizer. Final units are extracted from the first layer of the fine-tuned model trained on COLONYDB.

unit-LM (uLM). We use the vanilla Transformer architecture within a conventional language modeling framework. Unlike previous speech-related studies that found performance improvements by removing sequential unit repetitions (Kharitonov et al., 2021), we discovered that unit length conveys

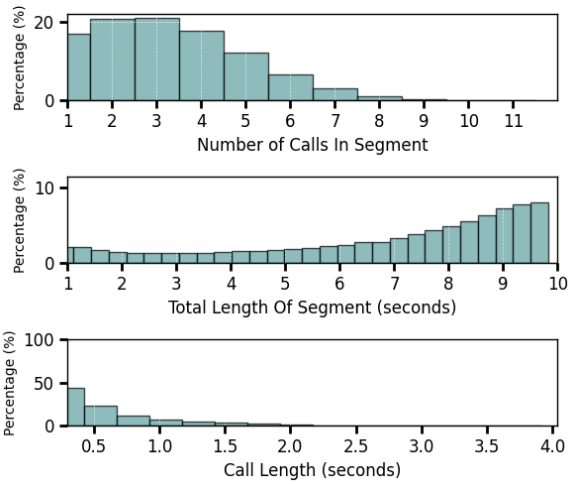


Figure 3: Statistics of the dataset after pre-processing. On average, each segment contains three calls, with an approximately equal distribution ranging from one to five calls. The total segment length is predominantly greater than 7 seconds, while the mean call duration is approximately 0.8 seconds.

important information. As a result, we preserve repeated units rather than collapsing them.

Unit-to-Vocalization(U2V). To reconstruct Marmoset vocalizations from the generated unit sequences, we utilize a unit vocoder, following the approach of Polyak et al. (2021). Since COLONYDB lacks speaker identity annotations, we exclude speaker embeddings and F0 information, relying only on vocalization tokens, obtained by V2U.

3.3 uLM evaluation

Drawing inspiration from established evaluation frameworks for text and speech-based large language models (LLMs) (Lakhotia et al., 2021), we develop a set of zero-shot evaluation tasks.

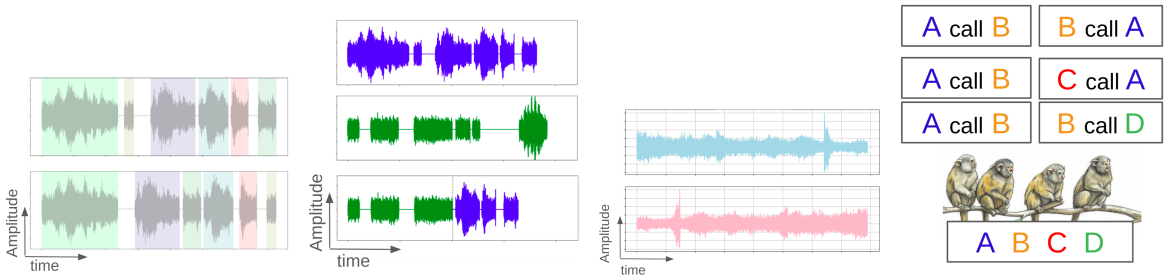
We follow evaluation methods from previous research on lexical and syntactic modeling in SpeechLMs, specifically sWUGGY and sBLIMP (Nguyen et al., 2020). In sWUGGY, the model receives pairs of utterances—one with a real word and the other with a phonotactically plausible non-word—and is assessed based on its ability to assign a higher probability to the real word. Similarly, sBLIMP evaluates the model using paired speech segments, where one contains a grammatically correct sentence and the other an ungrammatical one.

To adapt these methods to our setting, we create pairs of Marmoset speech (positive samples) and pseudo-Marmoset speech (distractor samples) by systematically modifying natural Marmoset call sequences. Specifically, we generate distractor samples using four approaches: (i) *Shuffle* (Figure 4a), where we randomly reorder the sequence of Marmoset calls within the same segment; (ii) *Reversal* (Figure 4c), where we reverse the entire audio segment; and (iii) *Concat* (Figure 4b), where we combine the first half of one audio segment (A) with the second half of another segment (B). Specifically, we take two distinct call segments, each containing six calls, split them at their midpoints, and merge the first half of one segment with the second half of another. (iv) *Phee eval* (Figure 4d), where PHEEDB is divided into call-and-response pairs, with the call originating from one Marmoset and the response from another. To generate pseudo call-response pairs, we keep the original call unchanged but modify the response. This is done by either changing the responder's identity (CallerChange) or altering the intended recipient (ReciverChange), i.e., who is calling versus whom the call is directed to. See Figure 4 for a visual representation of all four evaluation tasks.

Finally, we evaluate the uLM's capability to assign a higher probability to the real Marmoset segment over its pseudo-Marmoset counterpart. Higher accuracy suggests that the model effectively captures meaningful structures in the unit sequences, allowing it to differentiate authentic sequences from artificially modified ones. It is important to note that while the first three evaluation tasks are artificially constructed, they offer a quick and efficient method for assessing the uLM's ability to model the likelihood of Marmoset vocalizations. The final evaluation task (Phee eval) is particularly noteworthy and reliable because (i) it is based on manually labeled data, and (ii) previous research indicates that the likelihood of a pseudo pair naturally occurring in the training data is extremely low (Oren et al., 2024a).

4 Experimental Setup

Dataset. COLONYDB was then divided into training (80%), validation (10%), and test (10%) sets. Each segment was 10 seconds long and contained as many Marmoset calls as possible. Overall, the dataset includes $216K$ samples. Dataset statistics are detailed in Table 6. For PHEEDB, a total of 56 call-response pairs were identified, in which one Marmoset calls another, and the recipient responds, with gaps of up to 10 seconds between calls. Using these pairs, we generated approximately 600 augmented sequences to create the evaluation task,



(a) Shuffle (b) Concat (c) Reverse (d) Phee Evals

Figure 4: Our uLM evaluation tasks. The original samples are positioned above, while the pseudo-vocalizations are placed below.(d) The second row represents the CallerChange task and below the ReceiverChange task.

	Shuffle	Concat	Reversal	PPL
S-Hu	67.75	78.96	83.29	2.02
S-mHu	71.57	70.31	87.33	2
GMSLM	**84.84**	**79.94**	**90.45**	**1.78**

Table 1: Shuffle, Reversal and Concat tasks performance across different models. All baseline models were trained on units extracted from the 9th layer.

	ReceiverChange	CallerChange
S-Hu	50.76	59.67
S-mHu	58.77	59.99
GMSLM	**65.26**	**71.51**

Table 2: Phee evaluation metrics performance across different models.GMSLM significantly outperforms both baselines.

as outlined in Section 3.3. Full details about the dataset can be found on Section A.2.

Training configuration. All training configurations including hyper-parameters, compute resources, implementation details, etc. can be found on Section A.3 in the Appendix.

5 Results

5.1 Main results

We start by evaluating GMSLM using the Shuffle, Concat, and Reversal tasks, along with reporting uLM Perplexity (PPL). We compare the performance of the proposed method against uLMs trained on COLONYDB using vocalization units from three V2U quantizers: (i) S-Hu, where both HuBERT and k-means were trained on speech-only data; (ii) S-mHu, where HuBERT was trained on speech, but k-means was trained on COLONYDB; and (iii) the proposed method where both HuBERT and k-means were trained entirely on COLONYDB.

Results are presented in Table 1. We observe consistent improvements across all evaluation tasks as training progresses, indicating that gradually adapting the GSLM pipeline to Marmoset vocalizations enhances its ability to capture meaningful patterns and more effectively model data likelihood.

Next, we evaluate GMSLM using the Phee eval evaluation task, as described in Section 3.3, on PHEEDB. Similar to the previous evaluations, we

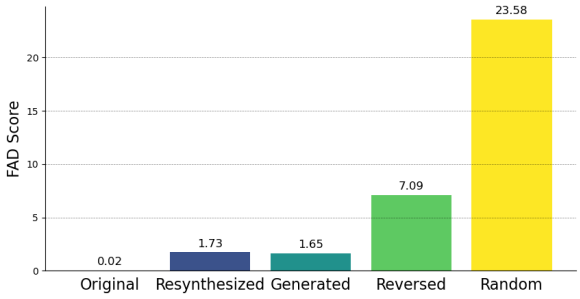


Figure 5: FAD scores for different audio manipulations (lower is better). The x-axis shows subsets compared to an original reference set. See Appendix A.4 for details.

compare GMSLM's performance against S-Hu and S-mHu. The results, presented in Table 2, show that S-Hu, trained exclusively on human speech, performs at near-random levels ($\sim$ 50%) on the ReceiverChange task. Adapting the k-means module to model Marmoset vocalizations (S-mHu) improves performance, achieving $\sim$ 59% on both ReceiverChange and CallerChange. GMSLM significantly outperforms both baselines considering both ReceiverChange and CallerChange tasks.

Lastly, we evaluate the quality of our generated Marmoset vocalizations using the Fréchet Audio Distance (FAD) (Kilgour et al., 2019), computed with a VGGish encoder (Hershey et al., 2017). FAD is a perceptual similarity metric that measures the distributional distance between real and generated audio embeddings, lower FAD scores indicate

Task	Recall	Precision	F1
Vocalization Type	91.96	90.14	90.72
Speaker Identity	90.25	90.03	90.12

Table 3: Classification performance scores. Results report the average across four different random seeds, std is smaller than 0.001 in all cases.

greater similarity to real vocalizations. As shown in Figure 5, both the generated and resynthesized vocalizations exhibited closely match, low FAD scores, significantly lower than those of reversed and random audio samples. These findings underscore the naturalness of our generated vocalizations in comparison to real data. We provide examples of resynthesized and generated Marmoset vocalizations visually in Figure 12, and audio in link.

5.2 Model evaluation

To further assess the quality and generalizability of the learned representations, we conducted supervised classification experiments using the final uLM outputs from our GmSLM model. Specifically, we trained simple classifiers on the open-source Marmoset vocalization dataset, INFANTMARMOSETSVOX (Sarkar and Magimai-Doss, 2023), which contains 11 vocalization types recorded from 10 individual marmosets (licensed under Creative Commons Attribution 4.0). We evaluated two classification tasks: predicting the vocalization type and identifying the individual speaker. As shown in Table 3, the GmSLM representations achieved high performance across both tasks, with an F1 score of 90.72 for vocalization type classification and 90.12 for speaker identity classification. These results demonstrate that the model captures rich, discriminative information about both vocal content and speaker identity. Importantly, they highlight the model's ability to generalize beyond its training objective, and suggesting that the model generalizes well across different marmoset setups, supporting its applicability in broader downstream tasks that may serve as baselines. Further implementation details regarding the classifiers are provided in Appendix A.5 and Figure 7 and Table 7 provide comparison of different feature representations on the same classification tasks. GMSLM consistently outperforms all baselines.

Dedup Units	Clusters	Shuffle	Concat	Reversal
V	100	62.64	77.29	75.55
V	50	64.87	77.68	75.59
X	**50**	**67.75**	**78.96**	**83.29**

Table 4: Number of clusters and duplications analysis.

5.3 Ablation study

We conduct an ablation study to examine the impact of (i) the number of clusters and unit duplication (ii) the choice of HuBERT layer used for feature extraction when training the uLM and (iii) the impact of using SSL features and a Transformer uLM versus simpler alternatives in the GmSLM model.

The effect of number of clusters and unit duplications. Table 4 presents the results for the Shuffle, Concat, and Reversal evaluation tasks. We assess the proposed method using both 50 and 100 units, with and without unit deduplication. When comparing the number of units, results indicate that 50 clusters perform best on the Shuffle evaluation task and yield comparable performance on the other evaluations. As a smaller vocabulary size results in shorter sequences and more efficient uLM training, we prioritize a vocabulary size of 50. When examining unit deduplication, results show that preserving unit repetitions improves performance across all tasks, suggesting that unit length is beneficial in Marmoset vocalization modeling. This contrasts with spoken language, where repetitions hurt performance (Lakhotia et al., 2021). We hypothesize this stems from structural differences and fewer vocalizations in marmosets. This aligns with Section 5.4, where low unit purity suggests meaning arises from context rather than individual units.

The effect of layer selection. Next, we analyze the impact of selecting different layers for quantization and uLM training. We extract and quantize features from layers $1, 3, 6$, and 9, training a separate uLM for each. To ensure the results are consistent across different encoders, we repeat this process using three HuBERT models, each differing in the layer used for teacher supervision. For readability, we report the average results and standard deviations for the Shuffle, Concat, and Reversal evaluation tasks, as shown in Figure 6. The findings indicate that uLM performance declines as deeper layers are used. We hypothesize that this occurs because Marmoset vocalizations are more acoustic than semantic (contextualized), making it easier for the uLM to capture fine-grained acoustic variations in

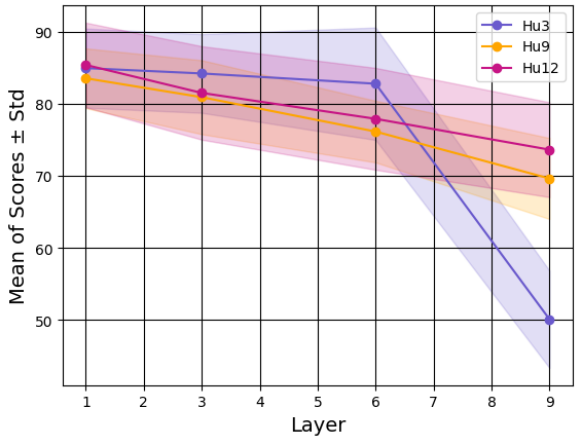


Figure 6: Mean and standard deviation results of uLM trained on tokens derived from various HuBERT models $Hu < n >$, where $n \in \{3, 9, 12\}$, indicating different layer supervision). The x-axis represents layers used for discretized token creation.

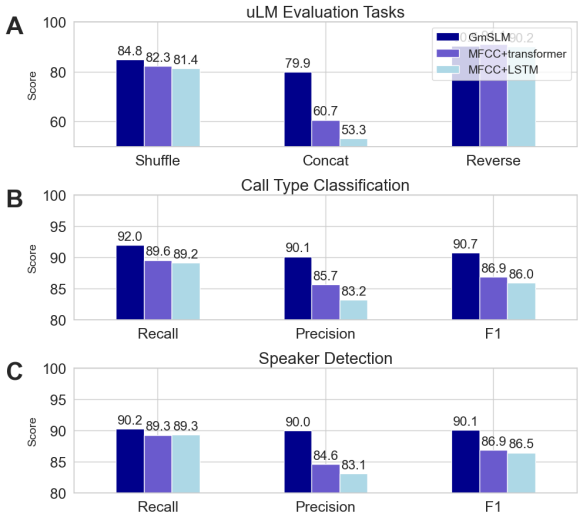


Figure 7: GmSLM vs. Simpler Models

earlier layers rather than in deeper layers.

The effect of model complexity. To evaluate whether GMSLM is overly complex for modeling Marmoset vocalizations, we compare it against simpler alternatives varying both input features and uLM architecture: (i) MFCC v2U + Transformer uLM and (ii) MFCC v2U + LSTM uLM. Results (Figure 7A) show that GMSLM outperforms both, particularly on the Concat task requiring cross-call context. Additionally, we experimented with a custom filterbank tailored to the energy profile of Marmoset vocalizations. This alternative achieved performance comparable to the MFCC-based setup (Appendix A.7). Classifier evaluations (trained as described in Section 5.2) further confirm that GMSLM representations significantly surpass those from the simpler models, underscoring the effectiveness of SSL features and Transformer modeling for this data (Figure 7B,C). These results suggest that SSL-based representations provide richer features, particularly for capturing contextual aspects of vocalizations and for classification performance, and that the advantage of HuBERT shallower layers does not indicate a lack of semantic properties (contextualized information) but rather an effective balance between acoustic and semantic information.(Full training details for the LSTM model are provided in Appendix A.6).

5.4 Analysis

A central question in Marmoset vocalization research is the role of context in communication (Eliades and Miller, 2017). To explore this, we systematically limited the model's attention to a fixed context length by masking all tokens beyond a threshold (set to $-\infty$), effectively removing them from the softmax.We tested context spans of 500, 400, 300, 200, and 50 tokens, with and without preserving the first token or first five tokens—known to influence attention (Sun et al., 2024; Oren et al., 2024b). Table 5 reports results for the Shuffle, Concat, and Reversal tasks.

Without preserving early tokens, performance matched the full-context setup only at 500 tokens ($\sim$ 10 seconds). When keeping the first token, similar performance persisted down to 300 tokens, suggesting that essential contextual information in Marmoset vocalizations spans $\sim$ 6 seconds.

To assess the role of context alongside unit repetitions, we examined how uLM units—derived from k-means clustering of HuBERT features—align with annotated call types. On a balanced INFANT-MARMOSETSVOX subset, we found low vocalization (0.18) and unit purity (0.26), much lower than in speech LMs. This suggests that unit meaning relies on context and span length rather than being inherently discrete, hinting at an n-gram–like distribution (Section 5.2). Full distributions are in Figures 8, 9, and 10. Attention analysis (Figure 11) revealed no consistent alignment with call labels.

6 Related work

Textless NLP was introduced by Lakhotia et al. (2021), showing how raw speech can be used to build GSLM systems. Kharitonov et al. (2021) extended this with multi-stream speech language models (SLMs) that combine pseudo-text with

Context Length	First Tokens	Shuffle	Concat	Reversal
-	-	83.15	79.6	93.44
500	0	83.15	79.54	93.44
	1	83.15	79.54	93.44
	5	83.15	79.54	93.42
400	0	81.54	79.42	85.72
	1	82.79	79.42	93.42
	5	**83.04**	**79.54**	**93.42**
300	0	75.14	79.01	73.5
	1	81.54	78.68	93.18
	5	82.33	78.85	93.28
200	0	62.71	75.76	67.71
	1	79.19	78.23	92.69
	5	80.4	78.19	93.1
50	0	30.74	32.49	35.81
	1	68.95	70.14	84.94
	5	71.9	70.69	88.23

Table 5: The effect of context length on GMSLM language modeling performance, considering the first token, the first five tokens, or no tokens at all.

prosodic features, opening new directions in spoken language modeling. Later works improved performance by initializing SLMs from text LMs (Hassid et al., 2024), framing tasks like emotion (Kreuk et al., 2021) and speaking style conversion (Maimon and Adi, 2023) as discrete translations, and jointly modeling dialogue (Nguyen et al., 2022). Borsos et al. (2022) proposed a cascade of LMs for semantic and acoustic tokens, enabling high-quality, speaker-consistent speech generation (Wang et al., 2023; Kharitonov et al., 2023). Semantic units from these models correlate with phonemes (Sicherman and Adi, 2023), while others adapted this approach for speech translation (Popuri et al., 2022; Peng et al., 2024). Additional advances include augmenting text LMs with speech for QA (Nachmani et al., 2024), state-space SLMs for long-context modeling (Park et al., 2024), and preference-tuned SLMs using LLM feedback (Lin et al., 2024; Rafailov et al., 2024).

Animal language. Traditional analysis of marmoset vocalizations (e.g., caller identity, call type) relied on ML models trained on acoustic features (Turesson et al., 2016; Verma et al., 2017). More broadly, ML, and particularly deep learning—has been used across various species for vocalization analysis, considering birds (Kahl et al., 2021; Ghani et al., 2023), dogs (Huang et al., 2023), mice (Coffey et al., 2019), call classification (Zhang et al., 2018), and nonhuman primates (Pellegrini, 2021; Oikarinen et al., 2019), including marmosets (Uesaka et al., 2023).

Recent advances in self-supervised learning (SSL), originally developed for human speech, have enabled large-scale bioacoustic learning from unlabeled data, supporting tasks like birdsong detection (Saeed et al., 2021), event detection (Bermant et al., 2022), and multi-species classification (Hagiwara, 2023). SSL has been used for phonetic and lexical discovery in dogs (Wang et al., 2024; Abzaliev et al., 2024), caller/call-type classification in marmosets (Sarkar and Magimai-Doss, 2023), and gibbon identity recognition (Cauzinille et al., 2024). These representations have also been leveraged to infer animal emotions and health from vocalizations (Manikandan and Neethirajan, 2024). Parallel to our work, Kobayashi et al. (2025) proposed FinchGPT—a Transformer-based model trained on birdsong transcripts—demonstrating its capacity to capture long-range dependencies in syllable sequences, consistent with our findings.

7 Conclusion

We introduce **Generative Marmoset Spoken Language Modeling** (GMSLM), a novel pipeline for modeling, generating, and evaluating Marmoset vocalizations without labeled supervision. Our approach bridges nonhuman primates vocal communication and modern generative language models. Through zero-shot evaluations, GMSLM effectively captures structural properties of Marmoset vocalizations, distinguishing authentic from artificial utterances. It outperforms human-speech-only baselines and achieves high performance on downstream tasks.

Layer selection and model complexity analyses show that SSL-based representations provide richer features, with the first layer performing best—likely reflecting an optimal balance between acoustic and semantic information in Marmoset vocalizations. Contextual analysis indicates key communicative information spans ≈six seconds.

Our work introduces a new direction for unsupervised spoken language modeling beyond human speech. We position GMSLM as a foundational tool for studying vocal communication in species with limited annotations, fostering interdisciplinary research at the intersection of computational linguistics, bioacoustics, and neuroscience.

Acknowledgments

This research was funded by the Center for Interdisciplinary Data Science Research (CIDR) grant number 3035000503, Israeli Science Foundation (ISF) grant numbers 2049/22 and 1331/23, and The Gatsby Charitable Foundation.

Ethical Statement

Since the proposed approach involves encoding, modeling, and decoding Marmosets vocalizations, careful attention must be given to data collection. All data collection procedures were conducted with the necessary ethical committee approvals, ensuring that the well-being of the Marmosets was maintained to the highest standards.

References

Artem Abzaliev, Humberto Pérez Espinosa, and Rada Mihalcea. 2024. Towards dog bark decoding: Leveraging human speech processing for automated bark classification. *Preprint*, arXiv:2404.18739.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *NeurIPS*.

P. C. Bermant, L. Brickson, and A. J. Titus. 2022. Bioacoustic event detection with self-supervised contrastive learning. *bioRxiv*.

Bruna Martins Bezerra and Antonio Souto. 2008. Structure and Usage of the Vocal Repertoire of Callithrix jacchus. *International Journal of Primatology*, 29(3):671.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. 2022. Audiolm: a language modeling approach to audio generation. *arXiv preprint arXiv:2209.03143*.

Henrik Brumm, Katrin Voss, Ireen K"ollmer, and Dietmar Todt. 2004. Acoustic communication in noise: regulation of call characteristics in a new world monkey. *Journal of Experimental Biology*, 207(3):443–448.

Jules Cauzinille, Benoit Favre, Ricard Marxer, Dena Clink, Abdul Hamid Ahmad, and Arnaud Rey. 2024. Investigating self-supervised speech models' ability to classify animal vocalizations: The case of gibbon's vocal identity. In *Proceedings of Interspeech 2024*.

Kai-Wei Chang, Haibin Wu, Yu-Kai Wang, Yuan-Kuei Wu, Hua Shen, Wei-Cheng Tseng, Iu-thing Kang, Shang-Wen Li, and Hung-yi Lee. 2024. Speechprompt: Prompting speech language models for speech processing tasks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

H.-C. Chen, G. Kaplan, and L.j. Rogers. 2009. Contact calls of common marmosets (Callithrix jacchus): influence of age of caller on antiphonal calling and other vocal responses. *American Journal of Primatology*, 71(2):165–170.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. 2022. Beats: Audio pre-training with acoustic tokenizers. *Preprint*, arXiv:2212.09058.

Kevin R Coffey, Ruby E Marx, and John F Neumaier. 2019. Deepsqueak: a deep learning-based system for detection and analysis of ultrasonic vocalizations. *Neuropsychopharmacology*, 44(5):859–868.

Steven J Eliades and Cory T Miller. 2017. Marmoset vocal communication: behavior and neurobiology. *Developmental neurobiology*, 77(3):286–299.

Steven J Eliades and Xiaoqin Wang. 2012. Neural correlates of the lombard effect in primate auditory cortex. *Journal of Neuroscience*, 32(31):10737–10748.

Burooj Ghani, Tom Denton, Stefan Kahl, and Holger Klinck. 2023. Global birdsong embeddings enable superior transfer learning for bioacoustic classification. *Scientific Reports*, 13(1):22876.

Masato Hagiwara. 2023. Aves: Animal vocalization encoder based on self-supervision. In *Proceedings of ICASSP*, pages 1–5.

Michael Hassid, Tal Remez, Tu Anh Nguyen, Itai Gat, Alexis Conneau, Felix Kreuk, Jade Copet, Alexandre Defossez, Gabriel Synnaeve, Emmanuel Dupoux, et al. 2024. Textually pretrained speech language models. *Advances in Neural Information Processing Systems*, 36.

Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Jieyi Huang, Chunhao Zhang, Mengyue Wu, and Kenny Q. Zhu. 2023. Transcribing vocal communications of domestic shiba inu dogs. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13819–13832, Toronto, Canada. Association for Computational Linguistics.

Bidda S Jones, Duncan HR Harris, and Clive K Catchpole. 1993. The stability of the vocal signature in phee calls of the common marmoset, callithrix jacchus. *American Journal of Primatology*, 31(1):67–75.

Stefan Kahl, Connor M. Wood, Maximilian Eibl, and Holger Klinck. 2021. BirdNET: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, 61:101236.

Eugene Kharitonov, Jade Copet, Kushal Lakhotia, Tu Anh Nguyen, Paden Tomasello, Ann Lee, Ali Elkahky, Wei-Ning Hsu, Abdelrahman Mohamed, Emmanuel Dupoux, et al. 2022. textless-lib: a library for textless spoken language processing. *arXiv preprint arXiv:2202.07359*.

Eugene Kharitonov, Ann Lee, Adam Polyak, Yossi Adi, Jade Copet, Kushal Lakhotia, Tu-Anh Nguyen, Morgane Rivière, Abdelrahman Mohamed, Emmanuel Dupoux, et al. 2021. Text-free prosody-aware generative spoken language modeling. *arXiv preprint arXiv:2109.03264*.

Eugene Kharitonov et al. 2023. Speak, read and prompt: High-fidelity text-to-speech with minimal supervision. *Transactions of the Association for Computational Linguistics*, 11:1703–1718.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. 2019. Fréchet audio distance: A metric for evaluating music enhancement algorithms. *Preprint*, arXiv:1812.08466.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kosei Kobayashi, Kosuke Matsuzaki, Masaya Taniguchi, Keisuke Sakaguchi, Kentaro Inui, and Kentaro Abe. 2025. Finchgpt: a transformer based language model for birdsong analysis. *Preprint*, arXiv:2502.00344.

Jungil Kong, Jaehyeon Kim, and Juhee Bae. 2020. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. In *Proc. NeurIPS*.

Felix Kreuk, Adam Polyak, Jade Copet, Eugene Kharitonov, Tu-Anh Nguyen, Morgane Rivière, Wei-Ning Hsu, Abdelrahman Mohamed, Emmanuel Dupoux, and Yossi Adi. 2021. Textless speech emotion conversion using decomposed and discrete representations. *arXiv preprint arXiv:2111.07402*.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. 2022. Audiogen: Textually guided audio generation. *arXiv preprint arXiv:2209.15352*.

Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021. On Generative Spoken Language Modeling from Raw Audio. *TACL*.

Ann Lee, Peng-Jen Chen, Changhan Wang, Jiatao Gu, Xutai Ma, Adam Polyak, Yossi Adi, Qing He, Yun Tang, Juan Pino, et al. 2021. Direct speech-to-speech translation with discrete units. *arXiv preprint arXiv:2107.05604*.

Guan-Ting Lin, Prashanth Gurunath Shivakumar, Aditya Gourav, Yile Gu, Ankur Gandhe, Hung-yi Lee, and Ivan Bulyko. 2024. Align-slm: Textless spoken language models with reinforcement learning from ai feedback. *arXiv preprint arXiv:2411.01834*.

Gallil Maimon and Yossi Adi. 2023. Speaking style conversion in the waveform domain using discrete self-supervised units. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8048–8061.

Venkatraman Manikandan and Suresh Neethirajan. 2024. Decoding poultry vocalizations: Natural language processing and transformer models for semantic and emotional analysis. *arXiv preprint arXiv:2412.16182*.

Eliya Nachmani et al. 2024. Spoken question answering and speech continuation using spectrogram-powered llm. In *The Twelfth International Conference on Learning Representations*.

Tu Anh Nguyen, Emmanuel Dupoux, and Ewan Dunbar. 2020. Investigating the learning of phonological representations in neural speech models. In *Proceedings of Interspeech*.

Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoit Sagot, Abdelrahman Mohamed, et al. 2022. Generative spoken dialogue language modeling. *arXiv preprint arXiv:2203.16502*.

Tu Anh Nguyen, Benjamin Muller, Bokai Yu, Marta R. Costa-jussa, Maha Elbayad, Sravya Popuri, Christophe Ropers, Paul-Ambroise Duquenne, Robin Algayres, Ruslan Mavlyutov, Itai Gat, Mary Williamson, Gabriel Synnaeve, Juan Pino, Benoit Sagot, and Emmanuel Dupoux. 2024. Spirit lm: Interleaved spoken and written language model. *Preprint*, arXiv:2402.05755.

JL Norcross and John D Newman. 1993. Context and gender-specific differences in the acoustic structure of common marmoset (callithrix jacchus) phee calls. *American journal of primatology*, 30(1):37–54.

Juha Oikarinen, Outi Tervo, and Risto Uusitalo. 2019. Automated identification of primate calls with deep learning. In *Proceedings of IEEE International Conference on Machine Learning for Wildlife Conservation*.

Guy Oren, Aner Shapira, Reuven Lifshitz, Ehud Vinepinsky, Roni Cohen, Tomer Fried, Guy P. Hadad, and David Omer. 2024a. Vocal labeling of others by nonhuman primates. *Science*, 385(6712):996–1003.

Matanel Oren, Michael Hassid, Nir Yarden, Yossi Adi, and Roy Schwartz. 2024b. Transformers are multi-state rnns. *arXiv preprint arXiv:2401.06104*.

Michael S Osmanski and Xiaoqin Wang. 2023. Perceptual specializations for processing species-specific vocalizations in the common marmoset (callithrix jacchus). *Proceedings of the National Academy of Sciences*, 120(24):e2221756120.

M Ott. 2019. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Se Jin Park, Julian Salazar, Aren Jansen, Keisuke Kinoshita, Yong Man Ro, and RJ Skerry-Ryan. 2024. Long-form speech generation with spoken language models. *arXiv preprint arXiv:2412.18603*.

Thomas Pellegrini. 2021. Primate species classification using deep neural networks. In *Proceedings of the International Conference on Bioacoustics*.

Yifan Peng et al. 2024. Mslm-s2st: A multitask speech language model for textless speech-to-speech translation with speaker style preservation. *arXiv preprint arXiv:2403.12408*.

Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021. Speech Resynthesis from Discrete Disentangled Self-Supervised Representations. In *Proc. Interspeech 2021*.

Sravya Popuri, Peng-Jen Chen, Changhan Wang, Juan Pino, Yossi Adi, Jiatao Gu, Wei-Ning Hsu, and Ann Lee. 2022. Enhanced direct speech-to-speech translation using self-supervised pre-training and data augmentation. *arXiv preprint arXiv:2204.02967*.

Ryan Prenger, Rafael Valle, and Bryan Catanzaro. 2019. Waveglow: A flow-based generative network for speech synthesis. In *ICASSP*.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Sabyasachi Roy, Cory T Miller, Dane Gottsch, and Xiaoqin Wang. 2011. Vocal control by the common marmoset in the presence of interfering noise. *Journal of Experimental Biology*, 214(21):3619–3629.

A. Saeed, D. Grangier, and N. Zeghidour. 2021. Contrastive learning of general-purpose audio representations. In *Proceedings of ICASSP*, pages 3875–3879.

E. Sarkar and M. Magimai-Doss. 2023. Can self-supervised neural representations pre-trained on human speech distinguish animal callers? In *Proceedings of Interspeech*, pages 1189–1193.

Robert M Seyfarth and Dorothy L Cheney. 2003. Signalers and receivers in animal communication. *Annual review of psychology*, 54(1):145–173.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *ICASSP*.

Amitay Sicherman and Yossi Adi. 2023. Analysing discrete self supervised speech representation for spoken language modeling. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. 2024. Massive activations in large language models. In *First Conference on Language Modeling (COLM)*.

Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. 2020. Transformer vq-vae for unsupervised unit discovery and speech synthesis: Zerospeech 2020 challenge. In *Interspeech*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

H. K. Turesson, S. Ribeiro, D. R. Pereira, J. P. Papa, and V. H. C. de Albuquerque. 2016. Machine learning algorithms for automatic classification of marmoset vocalizations. *PLOS ONE*, 11:1–14.

Minato Uesaka, Hideto Kawauchi, Kouei Yamaoka, Yukoh Wakabayashi, Yuma Kinoshita, Nobutaka Ono, Jun Noguchi, Satoshi Watanabe, Noritaka Ichinohe, Seico Benner, and Hidenori Yamasue. 2023. Automatic call classification of autism model marmosets by deep learning and analysis of their vocal development. In *Proceedings of the 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1650–1655. IEEE.

Sakshi Verma, K L Prateek, Karthik Pandia, Nauman Dawalatabad, Rogier Landman, Jitendra Sharma, Mriganka Sur, and Hema A Murthy. 2017. Discovering language in marmoset vocalization. In *Proceedings of Interspeech 2017*, pages 2093–2097. International Speech Communication Association.

Chengyi Wang et al. 2023. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.

Theron S. Wang, Xingyuan Li, Chunhao Zhang, Mengyue Wu, and Kenny Q. Zhu. 2024. Phonetic and lexical discovery of canine vocalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

Y. Zhang, J. Huang, N. Gong, Z. Ling, and Y. Hu. 2018. Automatic detection and classification of marmoset vocalizations using deep and recurrent neural networks. *The Journal of the Acoustical Society of America*, 144:478–487.

Yvonne Zürcher and Judith M Burkart. 2017. Evidence for dialects in three captive populations of common marmosets (callithrix jacchus). *International Journal of Primatology*, 38:780–793.

A Appendix

A.1 Pre-processing Details

We develop a multi-stage pipeline aimed at precisely segmenting Marmoset calls while effectively reducing environmental noise contamination.

We applied a high-pass filter at 5 kHz, as Marmoset calls primarily occur above this frequency, while lower frequencies mostly consist of noise or human speech. A spectrogram was generated using a hop length of 512 and a window size of 2048, allowing for frame-wise filtering in the frequency domain. Individual frames were classified as noise based on the following algorithm. We begin by filtering time-frequency bins with energy below a predefined threshold. Next, individual frames are identified as noise candidates based on their filtered time-frequency variance and density. These noise candidates are then aggregated over time, and segments are classified as noise if their duration is either less than 0.5 seconds or greater than 2 seconds. A similar analysis is then performed on the remaining audio, keeping only segments whose durations fall within the typical marmoset call range of $[0.25, 4]$ seconds. Finally, call boundaries are determined by the first and last frames of each detected segment. This method results in a robust call segmentation while reducing false detections, achieving a precision of 0.975 and a recall of 0.78 (computed on a held-out manually labeled set).

A.2 Dataset

Audio segments from the preprocessing stage were downsampled to 16 kHz, ensuring the necessary frequency range for detecting Marmoset vocalizations while preserving call integrity. The dataset was then divided into training (80%), validation (10%), and test (10%) sets. Each segment was 10 seconds long and contained as many Marmoset calls as possible. The average call length was 0.8 seconds, with approximately 3.3 calls per segment. In total, the dataset comprises $\sim$ 360 hours of audio, with approximately 40% consisting of detected Marmoset calls and the remaining 60% comprising inter-call gaps, which may still include background noise. Overall, the dataset includes $216K$ samples. Dataset statistics are detailed in Table 6.

For PHEEDB, the dataset consists of three one-hour-long audio recordings at 16kHz, capturing interactions among four different Marmoset. A total of 56 call-response pairs were identified, in which one Marmoset calls another, and the recipient responds, with gaps of up to 10 seconds between calls. Using these pairs, we generated approximately 600 augmented sequences to create the evaluation task, as outlined in Section 3.3.

	Train	Test	Validation	Total
N. Samples	173,328	21,666	21,666	216,660
Total Hours	292	36	37	365
Call Hours	125	15	15	155

Table 6: Dataset statistics after preprocessing and filtering. The dataset includes detected marmoset calls and inter-call gaps.

A.3 Training Configuration

For all HuBERT configurations, we used the `hubert-base` model with 12 Transformer layers. This model encodes raw audio into 768-dimensional frame representations, generating one frame every 20ms. The masking span is set to $l = 10$, with $p = 8\%$ of encoder output frames masked. Optimization is performed using Adam (Kingma and Ba, 2014) with $\beta = (0.9, 0.98)$, applying linear warm-up of the learning rate to 0.0035 over the first 3% of steps, followed by a linear decay to zero. Training is conducted for $150K$ steps on four $24GB$ GPUs, with a batch size of up to 32 seconds of audio.

Clustering was conducted using the MiniBatchK-Means function from scikit-learn, utilizing 50% of the training and validation data, with a mini-batch size of $10K$ frames and k-means++ initialization with 20 random starts. All experiments are implemented using the *fairseq* library (Ott, 2019). Inference for the first stage, leveraging the pre-trained speech model, was performed using *textlesslib* (Kharitonov et al., 2022).

For the uLM, we use a vanilla Transformer model as implemented in *fairseq* (Ott, 2019). Specifically, we adopt the `transformer_lm_big` architecture, which consists of 12 layers, 16 attention heads, an embedding size of 1024, a feed-forward network (FFN) size of 4096, and a dropout probability of 0.1. The model is trained as a causal language model (LM) on sequences of pseudotext units, with each sample containing up to $3,072$ units. The model is trained on 4 GPUs for $50K$ steps with a batch size of up to 700 samples. We use the `inverse_sqrt` learning rate scheduler with $4k$ warm-up steps, starting with an initial learning rate of $1e{-}7$ and reaching a peak learning rate of

$5e-45$. Optimization is performed using Adam with $\beta = (0.9, 0.98)$. For generation, we apply a beam-search sampling, with beam-size of 5 and a temperature of 1.5.

For U2V, we use a modified version of the HiFi-GAN neural vocoder (Kong et al., 2020), following the adaptation by Polyak et al. (2021) for unit-to-waveform conversion. The unit vocoder is trained on three $24GB$ GPUs for $400K$ steps.

A.4 FAD groups

The FAD scores across different audio comparisons (the lower the better). Different types of audio subsets, were compared to the same Original group (A) , together with modified versions of another subset of the original data (B), with no intersection between them. Resynthesized refers to reconstructed versions of B, Generated includes vocalizations generated using first 3 seconds prompts from B, Reversed contains temporally inverted sequences of B, and Random corresponds to Gaussian noise. All groups consist of 2K examples, each lasting more than 5 seconds.

A.5 Classifier Training Configuration

To evaluate the generalization capability of the model, we trained lightweight classifiers on the INFANTMARMOSETSVOX dataset for both vocalization type and caller identity prediction (see Section 5.2). Below, we provide implementation details and training setup for these classifiers. We randomly split the labeled dataset into 90% training and 10% validation sets, ensuring the splits were balanced with respect to the class labels (either call type or caller identity). Each classifier was trained for 20 epochs using a single GPU with $24GB$ memory. The models were optimized using the Adam optimizer combined with a polynomial learning rate scheduler. The classifier takes as input pooled statistical features derived from the given representations outputs — specifically, the mean and variance computed over the temporal dimension of the representations, which are concatenated to form the input vector. This input is passed through three fully connected layers of decreasing size, each followed by layer normalization and ReLU activation functions. The final layer is a linear projection that outputs the logits for classification.

A.6 LSTM training configuration

For the lstm, we used a model as implemented in *fairseq* (Ott, 2019). Specifically, we adopt the `lstm_lm` architecture, consists of a single-layer unidirectional LSTM decoder with 512-dimensional embeddings and hidden states, followed by a linear projection to the token vocabulary and a dropout probability of 0.2 similar to the ulm transormer training, the model is trained as a causal language model (LM) on sequences of pseudotext units, with each sample containing up to $3,072$ units. The model is trained on 4 GPUs for $50K$ steps with a batch size of up to 700 samples. We use the `inverse_sqrt` learning rate scheduler with $4k$ warm-up steps, starting with an initial learning rate of $1e{-}7$ and reaching a peak learning rate of $5e-45$. Optimization is performed using Adam with $\beta = (0.9, 0.98)$. For generation, we apply a beam-search sampling, with beam-size of 5 and a temperature of 1.5.

Model	Recall	Precision	F1
GmSLM	**91.96**	**90.14**	**90.72**
Marmoset HuBERT	85.66	84.15	84.89
Speech HuBERT	88.88	83.60	86.15

Table 7: Comparison of performance on the vocalization type classification task using different representations. GmSLM uses the final uLM layer, while HuBERT and Speech HuBERT use representations from the 9th layer.

A.7 Marmoset Filterbank

Marmoset vocalizations exhibit concentrated energy in the 5–8 kHz range. Although MFCCs were originally designed for speech signals, they have become a standard representation for general audio tasks and are widely used across various fields. However, due to the acoustic mismatch between human speech and Marmoset vocalizations, we explored a feature representation better aligned with the Marmoset specific characteristics. To this end, we constructed a custom filterbank focused on the 5–8,kHz band. This filterbank maintains the same dimensionality as the MFCCs, replaces the Mel scale with a linear frequency scale, and was used to train a uLM from scratch.

As shown in Table 8, the 5–8 kHz filterbank performs similarly to MFCCs, with a slight improvement on the Shuffle task. Nonetheless, GmSLM, which leverages HuBERT representations, achieves the best overall results.

A.8 More Advanced Modeling Components

Our pipeline currently relies on relatively dated components, including HuBERT and legacy

Model	Shuffle	Concat	Reversal
5–8 kHz Filterbank	84.85	59.52	90.01
MFCC + uLM	82.30	60.70	89.90
GmSLM (ours)	**84.84**	**79.94**	**90.45**

Table 8: Comparison of different input feature representations for uLM training. The 5–8 kHz filterbank is designed to match the dominant frequency range of marmoset vocalizations.

vocoders. In recent years, the field has seen significant advancements in open-source speech and language models. For example, Whisper (Radford et al., 2022) and BEATs Chen et al. (2022) demonstrate strong performance in speech understanding, while modern pre-trained language models like LLaMA and Qwen offer improved capabilities over earlier LLMs (Qwen et al., 2025; Touvron et al., 2023). However, applying these advanced tools in our setting poses specific challenges. Whisper is a fully supervised ASR model, which assumes the availability of written language—unsuitable in the case of Marmosets, where no such orthographic form exists. Similarly, BEATs is primarily optimized for audio classification, limiting its effectiveness in generative or sequence modeling tasks. By contrast, self-supervised learning (SSL) models, such as HuBERT, are better suited for modeling spoken language without written form. This approach aligns with standard practices in prior work on speech language models (Lakhotia et al., 2021; Hassid et al., 2024; Nguyen et al., 2024). We additionally explored the use of more recent language models by replacing the uLM with a pre-trained Qwen2.5-0.5B model, fine-tuned using the TWIST (Hassid et al., 2024) initialization strategy. For a controlled comparison, we trained two variants with the same architecture: (1) **GSLM**, where the Qwen model is trained from scratch, and (2) **TWIST**, where the Qwen model is initialized from pre-trained weights and fine-tuned on speech representations. Performance results are summarized in Table 9. Both approaches achieve comparable results across all evaluation tasks. Finally, while it is true that a more advanced vocoder could improve the quality of audio output, our goal is not to maximize audio fidelity but to analyze the linguistic structure of uLM predictions. Thus, we leave the development of a more sophisticated vocoder to future work.

Model	Shuffle	Concat	Reversal
TWIST	84.88	68.21	88.78
GSLM	85.90	69.23	89.90

Table 9: Evaluation results for Qwen-based uLMs trained from scratch (GSLM) and fine-tuned (TWIST).

A.9 Limitations

While GmSLM introduces effective evaluation proxies, a key limitation remains the overall system evaluation. Assessing such a complex model is particularly challenging in a fully unsupervised setting. Future research should explore more fine-grained, task-specific evaluation methods. Additionally, since our approach was primarily tested under a single recording condition, evaluating its performance in out-of-domain scenarios remains an important direction for future work.

A.10 AI Tools Usage

AI tools have been used to assist in fixing grammar mistakes and sentence paraphrasing. Additionally, AI tools have been partially used to enhance code implementations. However, the authors carefully reviewed all content, ensuring these tools were only used as supportive aids and in responsible manner.

A.11 Additional Results

See units distribution for different vocalization types and vice versa in Figures 8-10. Attention maps visualization in Figure 11 and spectrogram visualizations in Figure 12. Audio samples can be found under this link.

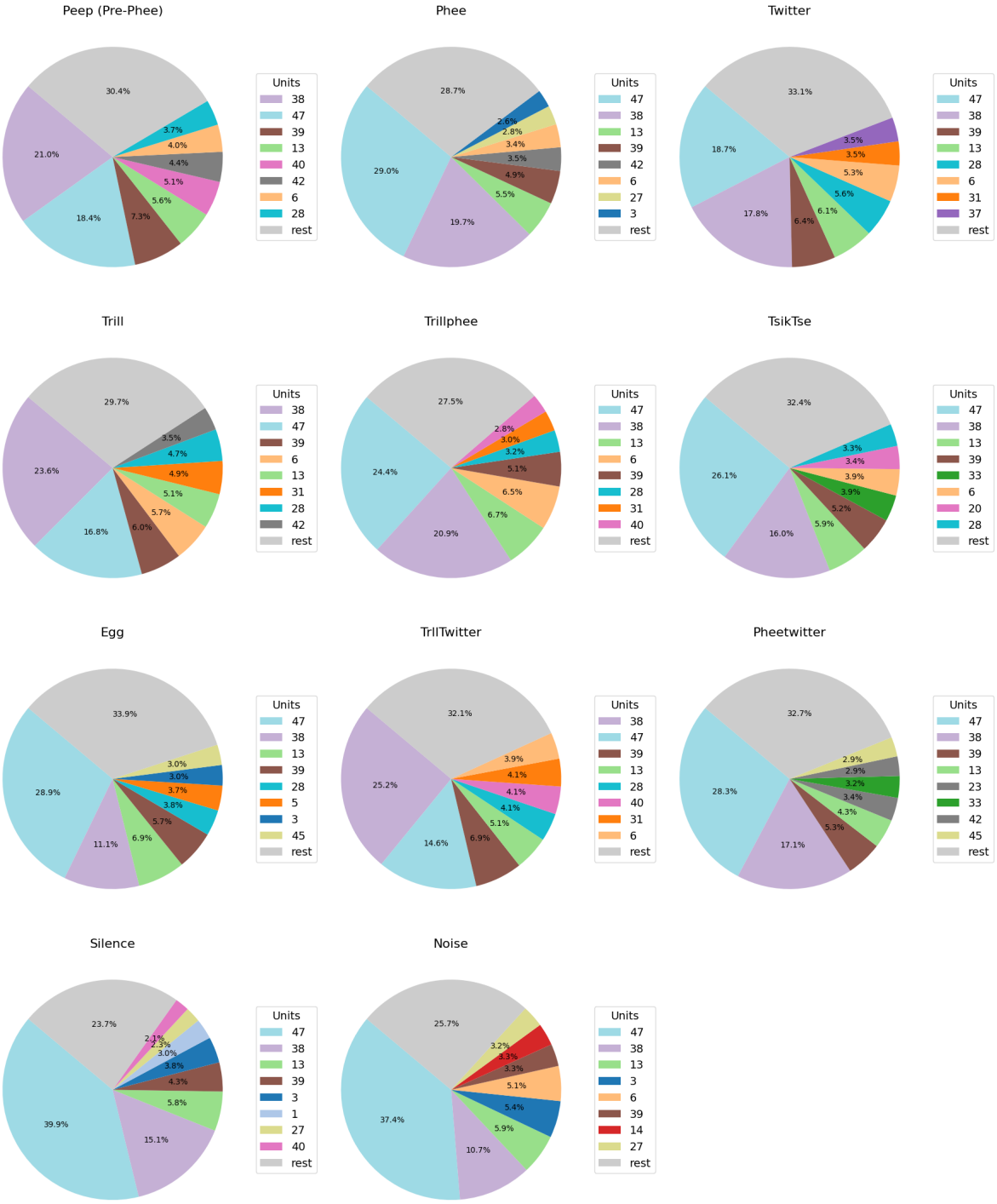


Figure 8: Units distribution for different call-types given from INFANTMARMOSETSVOX . Analysis is based on a balanced subset from the dataset, where each call-type occur equaly often and was chosen randomly. The call type names (Peep, Phee etc.) represents known vocal patterns in the marmoset vocalizations.

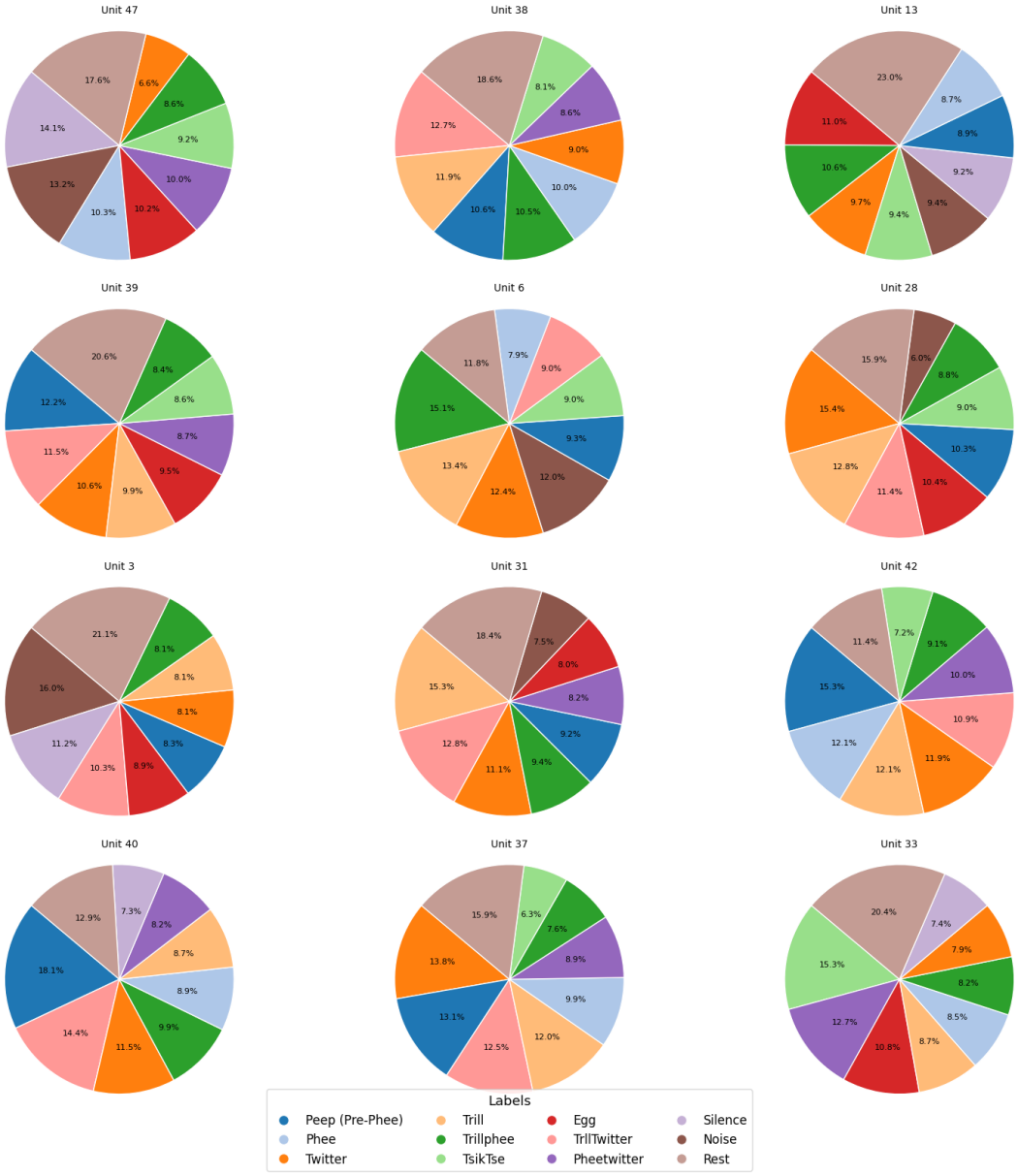


Figure 9: Vocalization Type Labels distribution for the most frequent HuBERT units. The analysis is based on a balanced subset of INFANTMARMOSETSVOX dataset, where each call-type occur equaly often and was chosen randomly. The call type names (Peep, Phee etc.) represents known vocal patterns in the marmoset vocalizations. HuBERT units are given from the 9th layer of a the marmoset-HuBERT model used for GmSLM

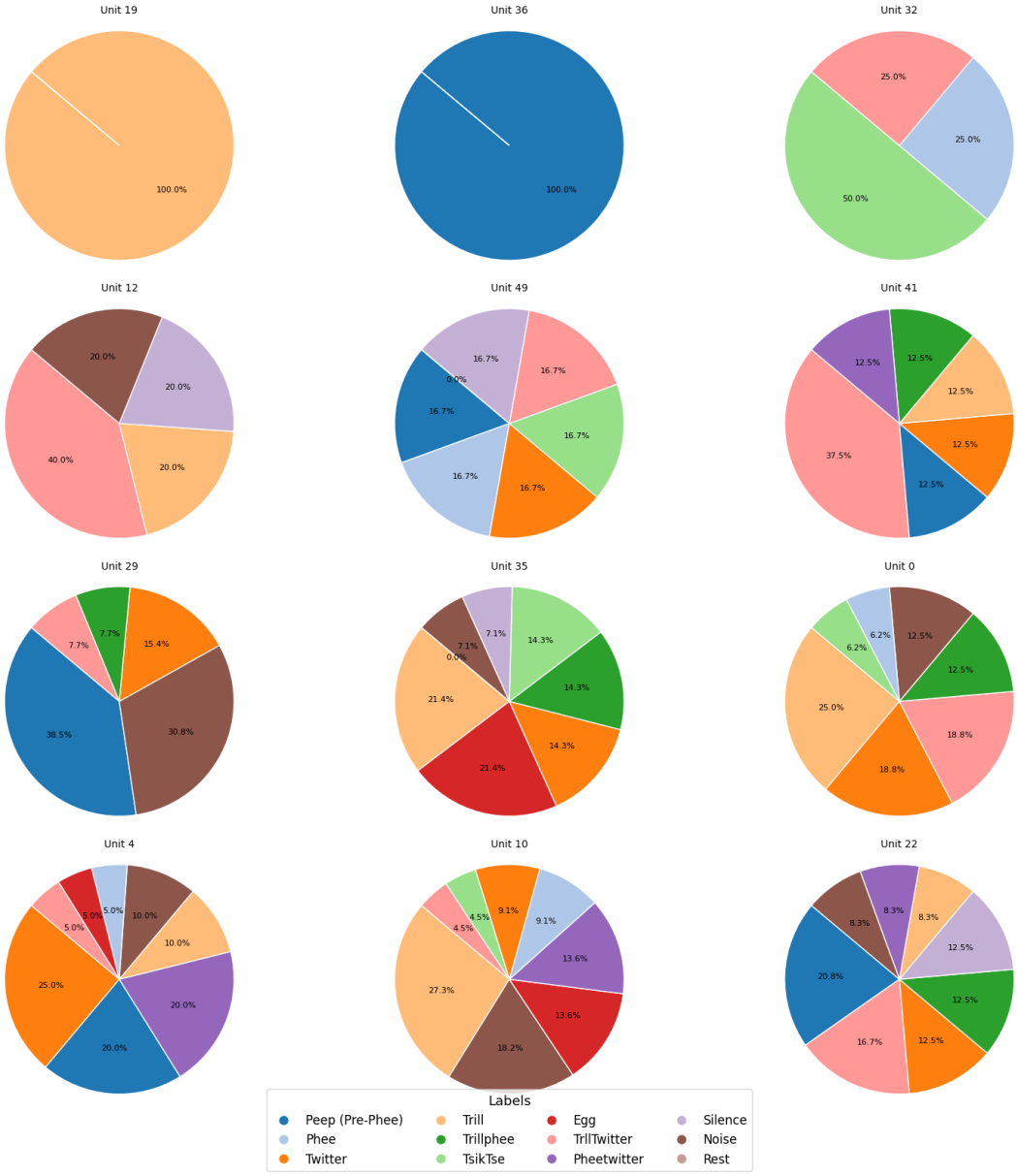


Figure 10: Vocalization Type Labels distribution for the less frequent HuBERT units. The analysis is based on a balanced subset of INFANTMARMOSETSVOX dataset, where each call-type occur equaly often and was chosen randomly. The call type names (Peep, Phee etc.) represents known vocal patterns in the marmoset vocalizations. HuBERT units are given from the 9th layer of a the marmoset-HuBERT model used for GmSLM

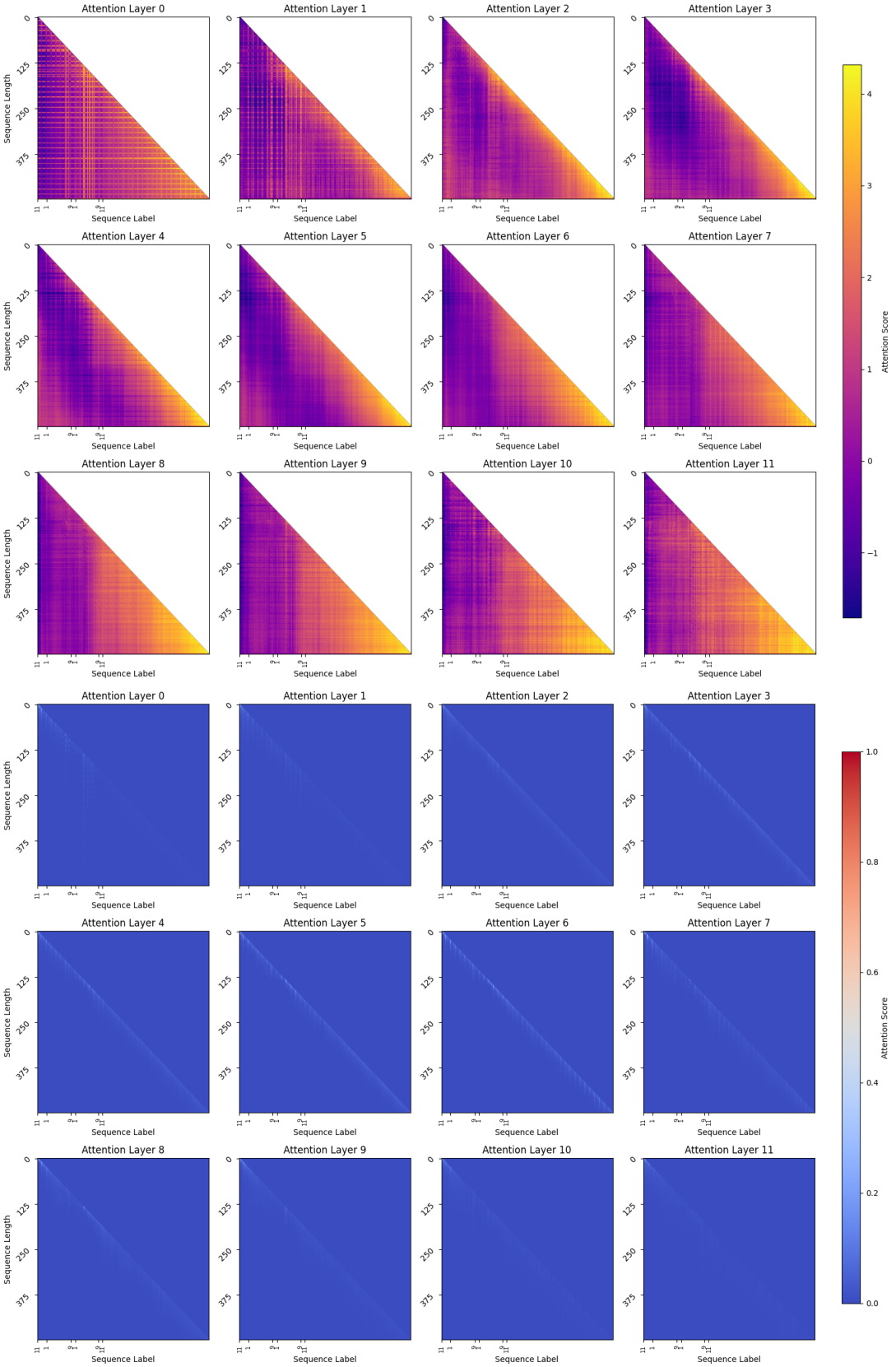


Figure 11: Attention map averaged across different heads of a sample from INFANTMARMOSETSVOX before and after applying softmax. the x-axis represents the start time of call labels where '11' represents silence, '1' is a Phee call, '9' is a 'Pheetwitter' call. We did not observe a clear relationship between attention patterns and the asigned labels.

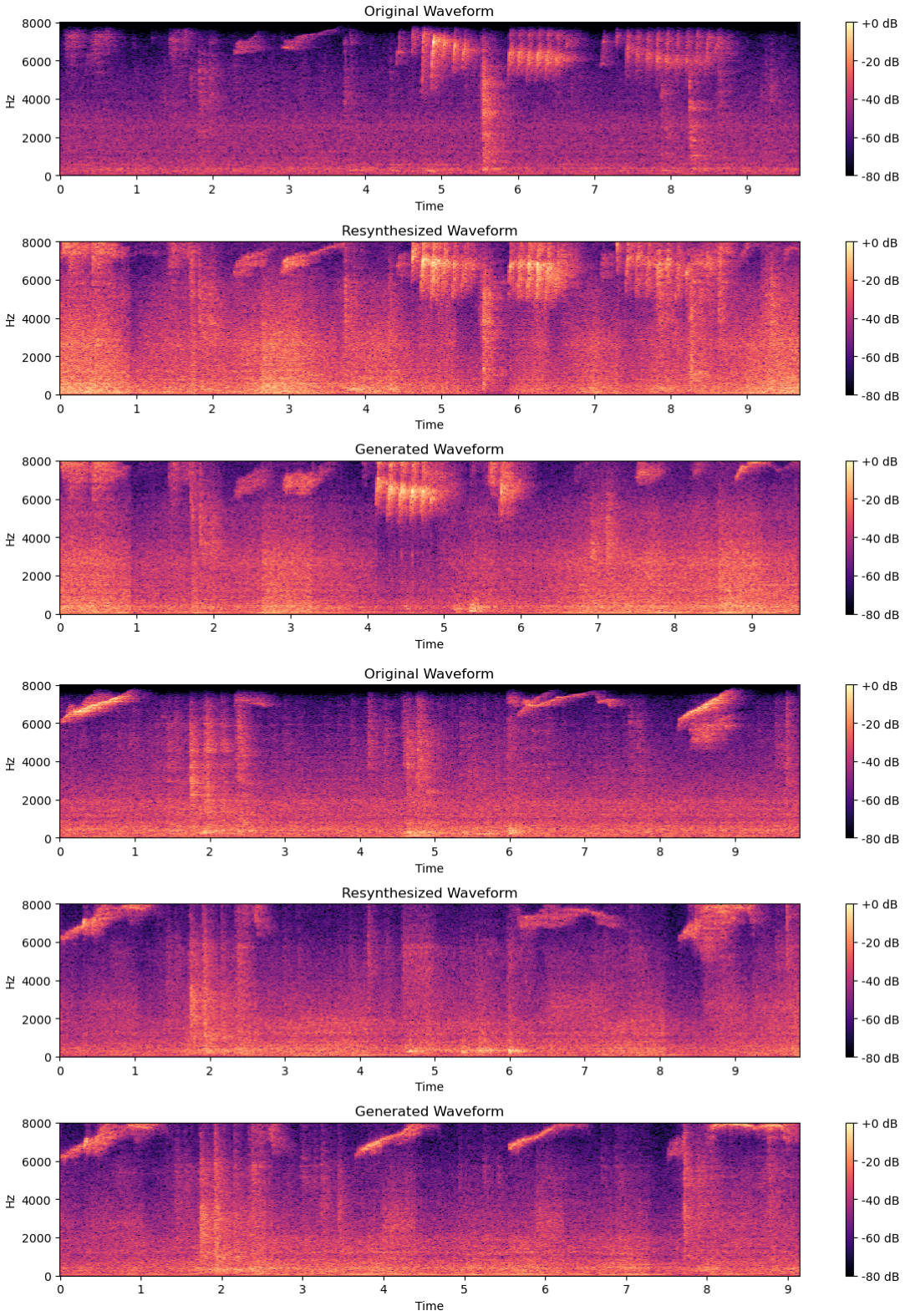


Figure 12: Spectrograms of an original Marmoset vocalization, its resynthesized version, and a generated version produced using GMSLM, conditioned on the first 3 seconds as a prompt.