Vocalizations of the Parus minor Bird: Taxonomy and Automatic Classification

Artem Abzaliev*
University of Michigan
Ann Arbor, MI, USA
abzaliev@umich.edu

Katsumi Ibaraki*
University of Michigan
Ann Arbor, MI, USA
ibaraki@umich.edu

Kohei Shibata
Independent Researcher
Kamakura, Japan
ateliermochamura@gmail.com

Rada Mihalcea
University of Michigan
Ann Arbor, MI, USA
mihalcea@umich.edu

Abstract

Previous research has revealed that Japanese tits (*Parus minor*) use synthetic syntax to combine various elements of their vocalizations and derive complex meanings. We collect and annotate a new dataset of vocalizations produced by the *Parus minor* bird and develop a full taxonomy of individual phonemes for this species, a total of 91 phonemes of different granularities. We provide an in-depth overview of the phonemes and explore methods to classify them automatically. Our best performing few-shot model achieves 13.9% multilabel accuracy on the test data.

CCS Concepts

• **Applied computing** → *Life and medical sciences.*

Keywords

Parus minor, animal vocalizations, machine learning

ACM Reference Format:
Artem Abzaliev, Katsumi Ibaraki, Kohei Shibata, and Rada Mihalcea. 2024. Vocalizations of the Parus minor Bird: Taxonomy and Automatic Classification. In *The International Conference on Animal-Computer Interaction (ACI 2024), December 02–05, 2024, Glasgow, United Kingdom.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3702336.3702344

1 Introduction

Recent studies suggest that phonemes, which constitute the fundamental units of sound in human speech, also feature in the vocalizations of non-human species [2, 8, 16, 24]. In particular, an analysis of the calls of the Japanese tit (*Parus minor*) has revealed that Japanese tits form words using various phonemes, similar to that of the phonemes in human languages, and then combine these phonemes to create words to communicate [24]. To our knowledge, the exact number of phonemes and their taxonomy have not been studied before, which we address in this work. We propose a taxonomy of vocalizations by studying spectrograms and experimenting with few-shot machine learning techniques to automatically label the phonemes in the audiostream.

Our contributions are as follows:

- We propose a taxonomy to classify the phonemes in the vocalizations of *Parus minor*, and present detailed analyses of these phonemes. To our knowledge, no taxonomy for *Parus minor* has been proposed before.
- We collect a comprehensive dataset of vocalizations from *Parus minor* and annotate the dataset with phoneme information using this taxonomy.
- We explore machine learning techniques to automatically detect and classify existing audio files. Since there is no prior taxonomy of vocalizations, machine learning algorithms can serve as a soft check that the proposed taxonomy is valid. If the machine learning algorithm can follow our taxonomy and classify the vocalizations on unseen data with some accuracy, it can indicates that our taxonomy is valid.

This work promises to enrich our understanding of animal communication systems and showcases the potential of computational approaches for deciphering the intricate soundscapes of wildlife.

2 Related work

There is a significant amount of research studying building blocks that serve the role of phonemes in animal vocalizations. Sharma et al. [16] studies sequences of clicks called codas in Sperm whales (*Physeter macrocephalus*), and show that codas exhibit contextual and combinatorial structure. Studies with the Gunnison's prairie dog (*Cynomys gunnisoni*) suggest that they are able to encode labels about predator colors and species in their alarm calls [17, 18]. Engesser et al. [8] studies the ability to generate new meaning by rearranging combinations of meaningless sounds. They show that chestnut-crowned babbler (*Pomatostomus ruficeps*) uses the same acoustic elements (phonemes) in different arrangements to create distinct vocalizations.

Some earlier research found that even for birds that have a small repertoire of calls, there are various types of alarm calls [14]. Field studies have shown complex anti-predator communication in multiple species of birds, which have a high degree of variation in frequency, duration, shape, and repetition rate, as well as combining notes or calls into complex sequences [20, 21]. Dutour et al. [6] suggest a conserved perception of call ordering, where typical

[*] Both authors contributed equally to this research

This work is licensed under a Creative Commons Attribution International 4.0 License.

ACI 2024, December 02–05, 2024, Glasgow, United Kingdom
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1175-6/24/12
https://doi.org/10.1145/3702336.3702344

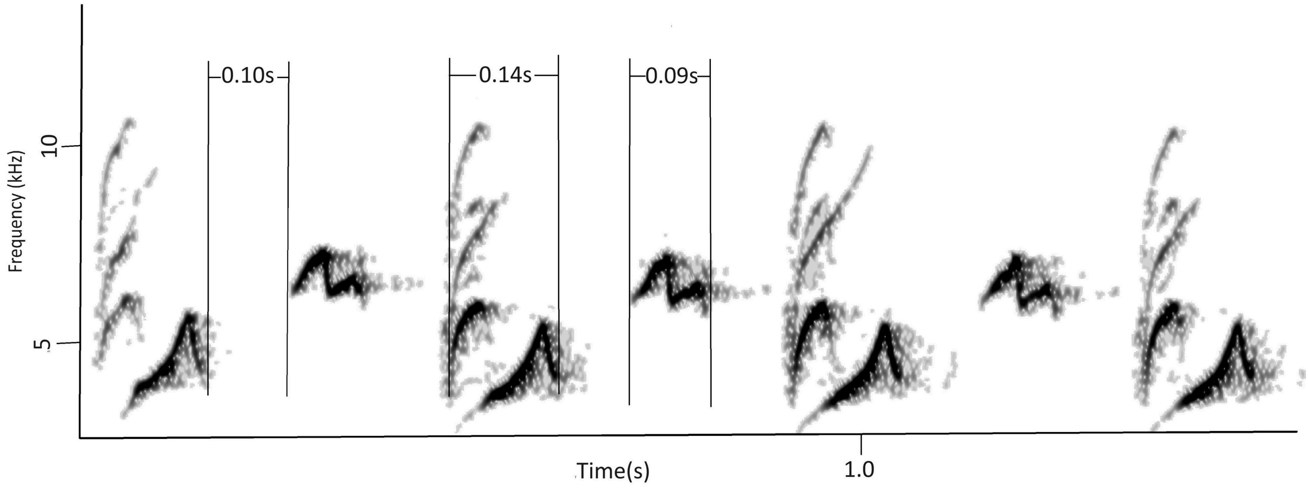

Figure 1: An example spectogram of the audio recording, showing time dimensions of the phonemes and the gap intervals between them.

note orderings for alarm calls are shared across species. James et al. [11] studies the Bengalese finch (*Lonchura striata domestica*) from the perspective of acquisition, where they show the evidence of vocal sequence learning between tutors and pupils. Other work have been focusing on Carolina chickadees, *Poecile carolinensis*, identifying structure in their calls with combinations of six notes [9].

Suzuki et al. [24] have explored compositional syntax in the Japanese tit, with experimental evidence for the use of different notes, orderings, and combinations in their vocal repertoire. Additional work show that the Japanese tit have an ordering rule; they are able to discriminate between different orderings of calls and extract a compound meaning when the sequences follow a specific ordering [25]. Zhang et al. [30] show that different populations of Japanese tit use mostly similar notes, but also that each population has some unique note types. Furthermore, referentiality in the Japanese tit has also been explored, where the receiver of a call becomes visually perceptive to an object resembling a predator when hearing an alarm call [22, 23].

3 Data collection

Over a span of sixteen months, Kohei Shibata – an amateur ornithologist and one of the co-authors of this work – has observed the behaviors of Japanese tits (*Parus minor*) within the northern precincts of Kamakura City, situated in Kanagawa Prefecture, Japan. The observations started in winter when the birds were active in flocks, continued through the courtship season, and followed the formation of pairs as they nested, laid eggs, incubated, hatched, raised their young, fledged, left the nest, and resumed flock behavior. The study involved auditory and visual documentation of the subjects, accomplished through the utilization of a high-fidelity sound recorder and a video camera.

The recording was done with a sampling frequency of 44.1kHz and 32-bit floating point format. A total of 2,527 audio files were recorded and subsequently annotated based on the spectral analysis. Classification of these recordings was primarily achieved through visual inspection of spectrograms, with instances requiring nuanced discernment resolved through auditory discrimination, or sometimes visual information. A spectrogram is a visual representation of the spectrum of frequencies of a signal over time. Commonly, the horizontal axis represents time, the vertical axis represents frequency, and a third dimension (usually color or brightness) represents amplitude. Parameters utilized for classification include temporal attributes such as sound duration, alongside spectral characteristics such as middle, lowest, and highest frequencies, as well as the frequency of peak volume.

To add more context to the auditory analysis, visual footage captured by a Panasonic Model No. AG-AC90 camera was used. The sound recorder was a TASCAM Linear PCM Recorder DR-07MK II, and the software used for acoustic analysis was Audacity 3.2.4 and Sonic Visualiser 4.5.1. Table 1 shows the summary of the dataset collected.

3.1 Phoneme Analyses

3.1.1 Phoneme Identification and Naming. The audio recordings were transformed into spectrograms, which we utilize for the identification of similar continuous segments exhibiting variations in frequency or volume. These segments typically range in duration between 0.1 and 0.6 seconds and are designated as phonemes. Each recording comprises a sequence of these phonemes, interspersed with approximately 0.1-second gaps. Figure 1 shows an example of the annotation process.

Phonemes are distinguishable based on the acoustic characteristics of the spectrogram and the frequency of the sound, with

Table 1: Statistics of observations

Number of observations	2527
Number of phonemes	91
Shortest phoneme length	0.10 s
Longest phoneme length	0.64 s
Average phoneme length	0.12 s
Minimum peak frequency	3.29 kHz
Maximum peak frequency	9.98 kHz
Average peak frequency	6.31 kHz
Number of words (phoneme sequences)	171

each isolated phoneme assigned a unique identification symbol according to specified naming conventions. We note that this study excludes phonemes present in the calls of chicks and young birds due to their distinct length and waveform, despite their relevance to the learning process.

The basic form of each phoneme is arbitrarily given an English alphabetical name such as /*P*/ or /*AN*/. If a number is added, the number indicates the main frequency band of the phoneme, for example, /P_6/ is a sound around 6 kHz. Additionally, we add a second letter to provide additional granularity to the phonemes. These lower case letters are not exclusive, and some phonemes contain up to three lower case letters. The lowercase letters after the primary letter(s) are as follows:

- u = up, frequency rising
- d = down, frequency falling
- f = flat, frequency constant
- n = noisy, multiple sounds together
- v = variation,
- a = attached, different sound attached
- o = overlap, multiple frequencies
- s = short length
- m = middle/medium length
- l = long length

Primary and secondary letters result in a total of 91 phonemes. We note that the number of phonemes may change based on additional observations and classification. Each phoneme can be divided into the following nine groups based on its spectrogram shapes. Representative examples of each group are shown in Figure 2.

(1) **Group A**: No change in frequency between the onset and end:
$/P_3a/, /P_4/, /P_4l/, /P_5o/, /Eh/, /P_6/, /P_67s/, /P_67/, /P_68/, /P_7s/, /P_7f/, /P_7u/, /P_8/, /P_8l/, /P_8m/, /P_8mo/, /P_8n/, /P_8so/, /P_8s/, /P_9s/, /P_9la/, /P_10s/$

(2) **Group B**: A rise and then a fall in frequency, with other sounds added to the first and second halves:
$/Au/, /Auh/, /Aua/, /ADD/, /N/, /P_5no/, /AB/, /AFF/$

(3) **Group C**: A rise in frequency for the base tone:
$/Fi/, /Hb/, /P_5u/, /P_8uo/, /P_7uo/, /J/, /Jp/, /Hi/, /P_7a/, /P_6v/, /P_8suv/, /P_8su/, /P_8ua/, /P_9u/$

(4) **Group D**: A fall in frequency for the base tone:
$/L/$, $/P_4do/, /Ad/, /P_8d/, /Y/$

(5) **Group E**: A short rise in frequency:
$/Ah/, /Us/, /P_7/, /P_7d/$

(6) **Group F**: Spectrogram shape resembles the letter N:
$/Af/, /Cy/, /AJ/, /D/, /Q/, /P_6m/, /P_6mv/, /AI/, /Sv/, /AN/$

(7) **Group G**: Spectrogram shape resembles the letter N but also contains another sound near the top:
$/T/, /Gy/, /I/, /Fu/, /S/, /AL/, /X/, /Oh/, /Oi/, /P_7uso/, /He/$

(8) **Group H**: Wave-shaped, repeat finely in high and low tones:
$/Ws/, /Wd/, /W/, /Wl/, /V/, /Vs/$

(9) **Group I**: Unique shape (leftover category, ones not able to be grouped with others):
$/AC/$, $/AE/, /AG/, /AK/, /Ck/, /Fo/, /AM/, /AA/, /G/, /Lv/, /M/$

3.1.2 Word Identification and Writing. The vocalizations of Japanese tits exhibit phoneme groups comprising consecutive phoneme clusters, with inter-phoneme intervals typically less than 0.05 seconds. Each phoneme cluster, arranged in a fixed sequence, is regarded as a single word. Additionally, calls consisting of a single phoneme, such as "chi" or "ji," traditionally categorized as calls, are also classified as words. When two words are concatenated without intervening gaps, forming a continuous sequence, they are counted as a single word. For example, if there are phonemes /A/, /B/, /C/, /D/, /E/, /F/ forming the words [/A//B//C/] and [/D//E//F/], the combined call of [/A//B//C//D//E//F/] is also counted as one word. However, even when employed in similar contexts, calls like [/A//B//C/][/D//E//F/] and [/A//C/][/D//E//F/] (where /B/ is absent) are treated as distinct words, acknowledging potential errors. Instances such as [/F//F//F/][/A//B/] and [/F//F/][/A//B/] are likewise regarded as separate words. Furthermore, words featuring numerous consecutive identical phonemes, such as [/A//B/] followed by eight consecutive [/F/] and [/A//B/] followed by five consecutive [/F/], were deemed identical.

During the observation period, a total of 171 distinguishable words were identified, with a comprehensive listing provided in Appendix A. These words are categorized based on formal distinctions rather than semantic or functional considerations. Among the most straightforward to comprehend are the termed "chirping/twitter/singing" words, characterized by continuous repetition of one or several phonemes for approximately 10 seconds, typically interspersed with brief pauses of a few seconds. The spectrum of identified words ranges from repetitions of single phonemes to sequences of three phonemes. These repetitions are denoted as follows when occurring for four or more sets: a sequence of a single phoneme [/A/*n], a sequence of two phonemes [(/A//B/)*n], and a sequence of three phonemes [(/A//B//C/)*n]. Instances where multiple phoneme groups, such as /A//B//C/, are reiterated, are treated as quasi-words and enclosed within parentheses. Notably, besides being utilized independently, words were at times employed repetitively with varying time intervals between them, fluctuating between approximately 1-2 seconds and 3-5 seconds.

3.2 Observations and Analyses

Bird calls serve various functions, including territorial marking and mate attraction, although distinguishing between communicative calls remains challenging [5, 14]. Our audio analyses demonstrate that Japanese tit vocalizations comprise a finite number of

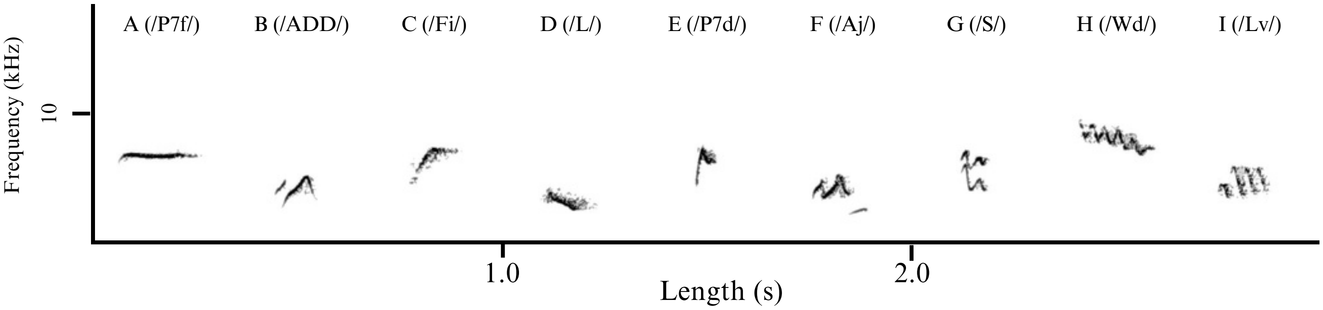

Figure 2: A comparison of the nine phoneme groups. Group A has a flat frequency, while group B has a rise and fall; group C has a rise and group D has a fall. Group E contains a short rise, group F resembles the letter *N*, and group G also resembles an *N* but has an additional sound. Group H is wave-shaped, and group I is the leftover category for phonemes that doesn't fit the other groups.

phonemes, akin to the 50 sounds in the Japanese alphabet or the 44 sounds corresponding to the 26 letters in the English alphabet. Each word is constructed through a combination of these phonemes, forming the basis for constructing phrases and sentences. Despite their audible similarity, Japanese tit chirps and calls exhibit distinct spectrographic profiles. Furthermore, phoneme classification confirms the distinctiveness of these vocalizations, suggesting the existence of multiple versions within commonly used onomatopoeic expressions like "tsutsupi" and "jajaja."

Instances occur where two phonemes, employed in sounds lasting approximately 10 seconds, seldomly serve the function of as a single word. Consequently, these instances resemble humming tunes, lacking direct meaning like single words but possibly conveying territorial declarations. Moreover, these "songs" are often a medley of several different tunes simultaneously. Successive, spaced vocalizations of words, unlike purely melodic renditions, appear to convey meaning akin to songs with lyrics, particularly observed in male nest-guarding tits. Furthermore, these songs are not improvised but adhere to set "tunes," indicating that all tits share a common repertoire of a few songs.

Interestingly, these “songs” are often performed in the form of several different songs at once. Successive, spaced vocalizations of words that could be used alone (unlike the song-only ones) often have meaning, similar to songs with lyrics. This phenomenon was particularly evident in male nest-guarding tits. Moreover, these songs are not improvised spontaneously but adhere to predetermined "tunes", suggesting that all tits share a common repertoire of a few songs. On the other hand, it was possible to deduce from the context that words used alone or in combination with several words are used as meaningful words to signify concepts such as danger or food.

4 Machine Learning based Phoneme Classification

The process of discovering and classifying phonemes or vocalizations even for a single species is laborious and requires significant of effort from domain experts. However, as recent advances in Natural Language Processing have shown [4], a large volume of data might be the key to successfully modeling human language. If this assumption holds for animal communication studies, we might need a large amount of data to make significant progress. Given the large number of species in the world and limited number of domain experts, we might never have enough effort to label the data to advance our understanding of animal communication. Instead we attempt to use advances in audio processing and machine learning to alleviate this issue. We formulate the problem as follows: given very few (2-5 examples) human-labelled annotations, can machine learning models determine what leads to specific phoneme discovery, i.e., frequency, spectrogram shape, etc, and infer the phonemes in the unlabelled sequences?

We experiment with two approaches:

- Spectogram-based audio classification. The underlying audio signal is represented as an image by the spectrogram, and an image classification model (efficientnet_b0 in our case) is used to make a prediction.
- Audio conversion to discrete tokens. We use an EnCodec model [7], which is a neural audio codec, to transform raw audio waves into a sequence of discrete tokens. The discreteness of the underlying sequence has some desirable properties for interpretability, allowing us to use methods from Natural Language Processing to understand which tokens contributed to the prediction of specific phonemes.

To evaluate these approaches, we hold out a small sample of audio files, without removing any background noise and without any editing. These files are usually longer than in the training data and the pauses are longer in between the words/phonemes. Since our problem is multilabel classification, we report multilabel precision, accuracy, recall and F1, as defined in [19]. Table 4 shows the results for both approaches.

4.1 Spectrogram Based Phoneme Classification

Similar to how phonemes were analyzed in section 3, we experiment with using a pre-trained computer vision model on mel-spectrograms of vocalizations. The sample rate of raw audios is 32kHz. We construct melspectograms using PyTorch for each five second interval of the audio, with the following parameters: 128

mel filterbanks, 2048 for size of FFT and window length, 512 for hop length, 20 Hz as minimum frequency. We use a variant of F0 Normalizer-Free ResNet [1] [3, 26] pre-trained on ImageNet as a backbone. Since the amount of data is very scarce, we apply data augmentation techniques: we use SpecAugment [15] and Mixup [27, 29]. For SpecAugment, both the frequency and time mask is applied with a probability of 30% with mask_max_length=10 and mask_max_masks=3. We use Binary Cross Entropy multi-label loss since each audio file might include several phonemes. We train for 50 epochs with AdamW optimizer and learning rate 1e-3. During the inference, we slice the audio into 5-second intervals and predict a single phoneme for each 5-second interval. The final prediction for a single vocalization clip includes all the unique phonemes from those 5-second intervals. The average number of predicted phonemes per file is 3.03. Table 4 shows the precision, recall, F1 score and accuracy for this approach. Given the scarcity of data per single phoneme, we consider these results being rather significant. With more data and human-in-the-loop annotation scheme similar to [12], our proposed approach can be used to accelerate human annotations.

4.2 Audio Conversion to Discrete Tokens for Phoneme Classification

Our second approach consists of converting the audio signal into discrete tokens. We use EnCodec, which is an encoder-decoder architecture with residual vector quantizer [28] as a bottleneck to compress the audio, i.e. at 48 kHz the model outputs 150 tokens per second instead of 48000 per second. We hypothesize that compressed audio tokens can be correlated with the phonemes we described in section 3, and thus allow us to learn the correspondence between them. I.e. some sort of simple rules: "if token 123 is observed in the discretized audio, it indicates that the phoneme F is present". To verify this hypothesis, we trained an EnCodec model from scratch exclusively on *Parus minor* audio vocalizations. While the EnCodec's training data includes bird vocalizations from the Audioset (74.6 hours in total) [10], the total fraction of bird vocalizations is less than 0.5% of the total duration of the training data (17537 hours), with the majority of the data sources being human speech, music, and general audio. There is also no splitting on bird species, only one general category "birds".

We train the EnCodec model on a subset of all unlabeled audioclips, total of 1125 audio recordings, 15.22 hours. While the size of the dataset is much smaller than the original EnCodec model was trained on, it exclusively consists of *Parus minor* vocalizations, which we hypothesize to be beneficial. We train a causal model for 20 epochs, with all the default parameters as in the EnCodec base: SeaNet as an encoder/decoder, hidden dimension 128, two LSTM layers. The model achieves a 3.052 Signal-to-Noise ratio [13] on validation data. We also manually inspected a set of reconstructed samples and we were unable to distinguish between the original and reconstructed samples, which indicates the model was able to reconstruct the signal specific for *Parus minor*. We note that the reconstruction happens from the discrete tokens alone, thus these discrete tokens should contain all the information that is present in raw audio. We can analyze the discrete tokens to see which information they contain.

Table 2: Example of top 10 mutual information scores (normalized) for phoneme *E*, tokens are from the first quantizer.

Token	MI score
184	0.5337
844	0.4669
311	0.3734
727	0.3664
309	0.3107
44	0.3003
512	0.3003
319	0.3003
612	0.2864
167	0.2656

Mutual Information. Up to this point, the model was trained without any information about the phonemes, in a fully unsupervised fashion. At this step, we would like to analyze how the tokens from the model can be used to predict the phonemes. We fully tokenize our dataset, i.e. each audio becomes a sequence of discrete tokens. We study how tokens/bigrams/trigrams can be used to predict phonemes in the audio.

To do this, we calculate the adjusted mutual information scores between the occurrence of a token/bigram from the first quantizer and the occurrence of a phoneme in the audio recording. I.e. given a tokenized audio sequence [1, 2, 3, 4, 5] and a phoneme sequence (/A, /B), for each pair [(1, /A), (1, /B) ... (5, /B)] we calculate a mutual information score. This score represents how large the information gained for predicting a phoneme from knowing that the audio token was present in the audio. If the mutual information score is 1, it means this particular token is only present for this particular phoneme and is never associated with other phonemes. More broadly, the closer the score is to 1, the stronger the association between a token and a phoneme.

An example of the mutual information scores for a single phoneme is shown in Table 2.

We repeated the process with bigrams, trigrams, and four-grams for the tokens, to see if a certain combination of tokens was also unique to a set of phonemes. We analyze bigrams and trigrams for sequences, as the length of two tokens combined was closest to the average phoneme lengths. From the mutual information scores for sequences using bigrams and trigrams of tokens, we extract a list of the most-used phonemes with the highest normalized mutual information score of 1. In theory, this means that when a certain bigram or trigram of tokens appears in a file, a certain phoneme is present.

Next, using the same EnCodec model, given an unseen data stream, we predict if certain phonemes appear in the audio sequence based on the token sequences generated by the model. If the bigram is present in the tokenized audio sequence, we predict this specific phoneme. The resulting accuracy of this approach is shown in table 4. The resulting performance is not statistically significant from random, thus we did not experiment further with this approach.

[1] https://huggingface.co/timm/eca_nfnet_l0

Table 3: Statistics of the train and test datasets

	Train	Test
# files	180	79
duration, #mins	13.77	75.4
# unique labels	76	39
# labels	407	154
Avg. # phonemes per file	2.26	1.94

Table 4: Test set accuracy for two different approaches for the automatic classification of *Parus minor* phonemes.

Approach	Precision	Recall	F1	Accuracy
Spectogram + NFNet	21.77	19.00	17.88	13.89
EnCodec + Mutual information	2.63	1.79	2.02	1.2

We believe one potential reason for the failure is that we only use the first quantizer.

5 Conclusion and Future Work

In this work, we explored the individual phonemes produced by *Parus minor* and the compositionality of these phonemes. We proposed a taxonomy for the phonemes that incorporates the granularity of the spectral signals. We showed that these phonemes can be combined into words of various lengths. We also presented our results when automatically classifying individual phonemes with two different approaches, achieving 21.7% accuracy on the test data.

Our current work focuses on the phonetic side of the vocalization without any semantics. The meaning and grounding of each word will be the subject of future work, as they also require the video recordings of the vocalizations. All the observations were conducted in one area, the northern part of Kamakura City, and previous research has shown that there are some differences in the vocalizations depending on the area. Hence, comparing the vocalizations of the tits to the other regions seeme to be an important direction. Also, observing changes in language as the chicks learn and acquire their vocabulary might bring interesting research directions.

The notion that tits, like humans, can generate words by combining phonemes and construct sentences by assembling these words implies the potential for them to innovate new vocabulary to adapt to their surroundings or circumstances. This possibility opens up interesting research directions.

It would be also interesting to see how the model pre-trained on human speech performs on bird vocalizations since there is some evidence that it helps for other animal vocalization tasks [1]. We plan to investigate this in future work.

6 Ethics Statement

All the data was collected observationally, and no birds were harmed or had their behavior adjusted in collecting this dataset.

References

[1] Artem Abzaliev, Humberto Perez-Espinosa, and Rada Mihalcea. 2024. Towards Dog Bark Decoding: Leveraging Human Speech Processing for Automated Bark Classification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, Torino, Italia, 16480–16486. https://aclanthology.org/2024.lrec-main.1432

[2] Daniel L Bowling and W Tecumseh Fitch. 2015. Do animal communication systems have phonemes? *Trends in Cognitive Sciences* 19, 10 (2015), 555–557.

[3] Andrew Brock, Soham De, Samuel L. Smith, and Karen Simonyan. 2021. High-Performance Large-Scale Image Recognition Without Normalization. arXiv:2102.06171 [cs.CV]

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[5] Clive K Catchpole. 1987. Bird song, sexual selection and female choice. *Trends in Ecology & Evolution* 2, 4 (1987), 94–97.

[6] Mylène Dutour, Toshitaka N Suzuki, and David Wheatcroft. 2020. Great tit responses to the calls of an unfamiliar species suggest conserved perception of call ordering. *Behavioral Ecology and Sociobiology* 74 (2020), 1–9.

[7] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2022. High Fidelity Neural Audio Compression. arXiv:2210.13438 [eess.AS]

[8] Sabrina Engesser, Jodie MS Crane, James L Savage, Andrew F Russell, and Simon W Townsend. 2015. Experimental evidence for phonemic contrasts in a nonhuman vocal system. *PLoS biology* 13, 6 (2015), e1002171.

[9] Todd Freeberg, Jeffrey Lucas, and Indrikis Krams. 2012. The Complex Call of the Carolina Chickadee What can the chick-a-dee call teach us about communication and language? *American Scientist* 100 (04 2012), 398–407. https://doi.org/10.1511/2012.98.398

[10] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 776–780.

[11] Logan S James, Herie Sun, Kazuhiro Wada, and Jon T Sakata. 2020. Statistical learning for vocal sequence acquisition in a songbird. *Scientific reports* 10, 1 (2020), 2248.

[12] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment Anything. arXiv:2304.02643 [cs.CV] https://arxiv.org/abs/2304.02643

[13] Yi Luo and Nima Mesgarani. 2019. Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27, 8 (Aug. 2019), 1256–1266. https://doi.org/10.1109/taslp.2019.2915167

[14] Peter R Marler and Hans Slabbekoorn. 2004. *Nature's music: the science of birdsong.* Elsevier.

[15] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Interspeech 2019*. ISCA. https://doi.org/10.21437/interspeech.2019-2680

[16] Pratyusha Sharma, Shane Gero, Roger Payne, David F Gruber, Daniela Rus, Antonio Torralba, and Jacob Andreas. 2024. Contextual and combinatorial structure in sperm whale vocalisations. *Nature Communications* 15, 1 (2024), 3617.

[17] CN Slobodchikoff and J Placer. 2006. Acoustic structures in the alarm calls of Gunnison's prairie dogs. *The Journal of the Acoustical Society of America* 119, 5 (2006), 3153–3160.

[18] Con N Slobodchikoff, Andrea Paseka, and Jennifer L Verdolin. 2009. Prairie dog alarm calls encode labels about predator colors. *Animal cognition* 12 (2009), 435–439.

[19] Mohammad S. Sorower. 2010. A Literature Survey on Algorithms for Multi-label Learning. https://api.semanticscholar.org/CorpusID:13222909

[20] Toshitaka N. Suzuki. 2014. Communication about predator type by a bird using discrete, graded and combinatorial variation in alarm calls. *Animal Behaviour* 87 (2014), 59–65. https://doi.org/10.1016/j.anbehav.2013.10.009

[21] Toshitaka N Suzuki. 2016. Semantic communication in birds: evidence from field research over the past two decades. *Ecological Research* 31 (2016), 307–319.

[22] Toshitaka N Suzuki. 2019. Imagery in wild birds: retrieval of visual information from referential alarm calls. *Learning & Behavior* 47 (2019), 111–114.

[23] Toshitaka N Suzuki. 2021. Animal linguistics: exploring referentiality and compositionality in bird calls. *Ecological Research* 36, 2 (2021), 221–231.

[24] Toshitaka N Suzuki, David Wheatcroft, and Michael Griesser. 2016. Experimental evidence for compositional syntax in bird calls. *Nature communications* 7, 1 (2016), 10986.

[25] Toshitaka N Suzuki, David Wheatcroft, and Michael Griesser. 2017. Wild birds use an ordering rule to decode novel call sequences. *Current Biology* 27, 15 (2017), 2331–2336.

[26] Ross Wightman, Hugo Touvron, and Hervé Jégou. 2021. ResNet strikes back: An improved training procedure in timm. arXiv:2110.00476 [cs.CV]

[27] Kele Xu, Dawei Feng, Haibo Mi, Boqing Zhu, Dezhi Wang, Lilun Zhang, Hengxing Cai, and Shuwen Liu. 2018. Mixup-Based Acoustic Scene Classification Using Multi-Channel Convolutional Neural Network. arXiv:1805.07319 [cs.CV]

[28] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2021. SoundStream: An End-to-End Neural Audio Codec. arXiv:2107.03312 [cs.SD]

[29] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond Empirical Risk Minimization. arXiv:1710.09412 [cs.LG]

[30] Li Zhang, Jiangping Yu, Chao Shen, Dake Yin, Longru Jin, Wei Liang, and Haitao Wang. 2022. Geographic variation in note types of alarm calls in Japanese tits (Parus minor). *Animals* 12, 18 (2022), 2342.

A Word List

Below is a list of words (phoneme sequences) observed more than three times.

(1) $[(/AA//Au/) * n]$
(2) $[(/AB//Af/) * n]$
(3) $[(/AC//Cy/) * n]$
(4) $[(/Ad//P6m/) * n]$
(5) $[(/ADD//AE/) * n]$
(6) $[(/AFF//AG/) * n]$
(7) $[(/AI//AJ/) * n]$
(8) $[(/Au//P7u/) * n]$
(9) $[(/Auh//AI/) * n]$
(10) $[(/Auh//P7u/) * n]$
(11) $[(/Cy//AB/) * n]$
(12) $[(/J/*n)(/G/*n)]$
(13) $[(/P67//Cy/) * n]$
(14) $[(/P7f//M/) * n]$
(15) $[(/P8//Hi//S//Hi/)/Cy/(/G/*n)]$
(16) $[(/P8//Hi//S//Hi/)/Cy//G//G/]$
(17) $[(/P8//Hi//S//Hi/)/Cy/]$
(18) $[(/P8//Hi//S//Hi/)/P8//P8//L/]$
(19) $[(/Q//Af/) * n]$
(20) $[(/Q//Aua/) * n]$
(21) $[(/V//P67s/) * n]$
(22) $[(/V//V//Hb/) * n] <<<<<< [(/Vd//Vd//Hb/) * n]$
(23) $[(/W/*n)(/Wd/*n)(/Ws/*n)]$
(24) $[(/W/*n)(/Ws/*n)]$
(25) $[(/Wl/*n)(/W/*n)/Wd//Ws//W//W/]$
(26) $[/AK/*n]$
(27) $[/Ck//P7/(/J/*n)]$
(28) $[/Ck//P7/]$
(29) $[/D//P7u/]$
(30) $[/Fi//Fi//L/(/G/*n)]$
(31) $[/Fi//Fi//L/]$
(32) $[/Fi//Fu//Gy/]$
(33) $[/Fi//Fu//Hu//Fi/]$
(34) $[/Fi//Fu//Hu/]$
(35) $[/Fi//Fu//L/]$
(36) $[/Fi//Hu/]$
(37) $[/Fi//L/]$
(38) $[/Fi//P6m//Gy/]$
(39) $[/Fo//Fo/]$
(40) $[/G/*n]$
(41) $[/Hi//AN//Hi//AN/]$
(42) $[/Hi//Hi//L/]$
(43) $[/Hi//Hi//N/]$
(44) $[/J/(/G/*n)]$
(45) $[/J/(/L/*n)]$
(46) $[/J/*n]$
(47) $[/J//J//Af/]$
(48) $[/J//J//J//L//L//L/]$
(49) $[/J//J//N/]$
(50) $[/J//J/P4o/]$
(51) $[/J//L//J//J/]$
(52) $[/J//L/]$
(53) $[/J//T/(/G/*n)]$
(54) $[/Jp/*n]$
(55) $[/Js//Ls/]$
(56) $[/L///P7/s//L/]$
(57) $[/L//He//He//Eh//Eh/(/J/*n)]$
(58) $[/L//He//He//Eh//Eh/]$
(59) $[/L//P8//L//P8//L/]$
(60) $[/P10s//P9s//P9s/]$
(61) $[/P4///P8///P4/]$
(62) $[/P4/]$
(63) $[/P5n//P8n//P5n/]$
(64) $[/P68//Gy/]$
(65) $[/P68//P7//S//P7/]$
(66) $[/P68//P7f//P7f/]$
(67) $[/P6v/]$
(68) $[/P7//AI//AL/(/G/*n)]$
(69) $[/P7//AI//AL/]$
(70) $[/P7//P6mv//AL/]$
(71) $[/P7//P6v/]$
(72) $[/P7//P7/(/P8//S//T//P8/)]$
(73) $[/P7//P7//N/]$
(74) $[/P7//S//AL/]$
(75) $[/P7//S//T/(/G/*n)]$
(76) $[/P7//T//G//G//G/]$
(77) $[/P7//X//Gy/]$
(78) $[/P7d//P5no/]$
(79) $[/P7d/]$
(80) $[/P7f//AL/]$
(81) $[/P7f//Gy/]$
(82) $[/P7f//Hi//I//Hi//P3a/]$
(83) $[/P7f//Hi//S//Hi//P3a/]$
(84) $[/P7f//Lv//L//L//L/]$
(85) $[/P7f//P7f//P7f/]$
(86) $[/P7f//P7f/]$
(87) $[/P7s/]$
(88) $[/P7u//D//P8/]$
(89) $[/P7u//Hb//P7u/]$
(90) $[/P7u//L//L///L/]$
(91) $[/P7u//P5//Hb//P7u/]$
(92) $[/P7u//P6u//Oh//Oh//Oh//Oh/]$
(93) $[/P7u//P7/(/Oh/*n)]$
(94) $[/P7u//P7u//Ad/]$
(95) $[/P7u//P7u//L/]$
(96) $[/P7u//P7u//Lo//Lo//Lo/]$
(97) $[/P7u//P7u//P7u/]$
(98) $[/P7u//P7u/]$
(99) $[/P7u/]$
(100) $[/P7uo//AM///P7uo/]$
(101) $[/P7uo//Hb//P7uo/]$

(102) $[/P7uo//Hi//S//Hi//P3a/]$

(103) $[/P7uo//P7uo/]$

(104) $[/P8//P8/(/P8//Fi//Lv//L/)(/P8//Fi//Lv//L/)]$

(105) $[/P8/(/P8//Fi//Lv//L/)(/P8//Fi//Lv//L/)/P8/]$

(106) $[/P8/(/P8//Fi//Lv//L/)(/P8//Fi//Lv//L/)]$

(107) $[/P8/(/P8//Fi//Lv//L/)]$

(108) $[/P8/(/S//T//P8/)(/S//T//P8/)(/S//T//P8/)]$

(109) $[/P8/(/S//T//P8/)(/S//T//P8/)]$

(110) $[/P8/(/S//T//P8/)]$

(111) $[/P8//Ah//Ah/]$

(112) $[/P8//Fu//L/]$

(113) $[/P8//Hb//P8/]$

(114) $[/P8//Hi//I//Hi//P3a/]$

(115) $[/P8//Hi//S//Hi/]$

(116) $[/P8//Hi//S/]$

(117) $[/P8//I//L//P8/]$

(118) $[/P8//L/]$

(119) $[/P8//Oi//Oi/]$

(120) $[/P8//P5o/]$

(121) $[/P8//P7//AL//L/]$

(122) $[/P8//P8/(/P8//Fi//Lv//L/)(/P8//Fi//Lv//L/)]$

(123) $[/P8//P8//L//P7/]$

(124) $[/P8//P8//P7//Lv//L/]$

(125) $[/P8//P8//P8/(/S//T//P8/)(/S//T//P8/)(/S//T//P8/)]$

(126) $[/P8//P8//P8/(/S//T//P8/)]$

(127) $[/P8//P8//P8//P8//P8/(/S//T//P8/)(/S//T//P8/)]$

(128) $[/P8//P8s//Us/]$

(129) $[/P8d/]$

(130) $[/P8l/*n]$

(131) $[/P8l//Fu//L/]$

(132) $[/P8l//P8l//P8l/]$

(133) $[/P8l//P8l/]$

(134) $[/P8l/]$

(135) $[/P8m//Gy//He/]$

(136) $[/P8m//Hi//Gy/]$

(137) $[/P8m//L/]$

(138) $[/P8m//P7uso//Gy/]$

(139) $[/P8m//P8m//L/]$

(140) $[/P8m//P8m//P8m/]$

(141) $[/P8m//S//AL//P8m/]$

(142) $[/P8mo//P7a//P7a/]$

(143) $[/P8n/]$

(144) $[/P8so//Ah//Ah/]$

(145) $[/P8so/]$

(146) $[/P8su//D//P8l/]$

(147) $[/P8su//D//P8m/]$

(148) $[/P8su//L//L//L/]$

(149) $[/P8su//P7/(/Oh/*n)]$

(150) $[/P8su//P8l//L//P8l//P7]$

(151) $[/P8su//P8su//L/(/G/*n)]$

(152) $[/P8su//P8su//L/]$

(153) $[/P8suv//L//P8suv//P8suv/]$

(154) $[/P8u//Y/]$

(155) $[/P8ua//Ah/]$

(156) $[/P8ua//L//L/]$

(157) $[/P8uo//I//P8uo/]$

(158) $[/P9la/*n]$

(159) $[/P9s//P9so/]$

(160) $[/P9u/]$

(161) $[/Q//Af/]$

(162) $[/S//T//P7f/]$

(163) $[/S//V/]$

(164) $[/V/*n]$

(165) $[/V/]$

(166) $[/Vs/*n]$

(167) $[/Wd/]$

(168) $[/Wl//Wm//Wds/]$

(169) $[/Wl//Ws//W//Wd/]$

(170) $[/Wl/]$

(171) $[/Ws/*n]$

B Bigram Mutual Information

Table 2 is for single tokens, while the table below contains the mutual information scores for bigrams of tokens. The sharp drop in mutual information scores suggest that for the first 41 bigrams, these are completely unique for the phoneme, but the rest are observed elsewhere and does not provide much information in classification. This also shows that between phonemes there may be more similarities than what we were able to distinguish manually, and further investigation is needed in matching which n-gram of tokens appear in which set of phonemes.

Table 5: **Example of top 75 mutual information scores (normalized) for phoneme *E*, tokens are from the first quantizer.**

Tokens	MI score
(70, 554)	1.000
(794, 788)	1.000
(554, 794)	1.000
(419, 317)	1.000
(997, 478)	1.000
(153, 83)	1.000
(972, 997)	1.000
(380, 55)	1.000
(423, 419)	1.000
(693, 380)	1.000
(680, 55)	1.000
(757, 423)	1.000
(384, 944)	1.000
(613, 956)	1.000
(235, 997)	1.000
(944, 577)	1.000
(554, 979)	1.000
(554, 607)	1.000
(814, 325)	1.000
(577, 235)	1.000
(833, 567)	1.000
(65, 70)	1.000
(55, 384)	1.000
(350, 833)	1.000
(317, 132)	1.000
(843, 109)	1.000
(27, 554)	1.000
(132, 975)	1.000
(956, 838)	1.000
(83, 972)	1.000
(838, 814)	1.000
(975, 554)	1.000
(979, 757)	1.000
(464, 731)	1.000
(997, 350)	1.000
(577, 65)	1.000
(788, 843)	1.000
(109, 464)	1.000
(478, 680)	1.000
(731, 717)	1.000
(607, 153)	1.000
(325, 27)	0.543
(55, 577)	0.543
(567, 693)	0.543
(823, 604)	0.031
(766, 766)	0.029
(685, 766)	0.029
(604, 685)	0.029
(685, 685)	0.027
(604, 604)	0.022
(766, 817)	0.019
(766, 465)	0.016
(685, 604)	0.016
(766, 685)	0.013
(604, 766)	0.012
(685, 465)	0.012
(823, 823)	0.011

Continued on next page

Table 5: **Example of top 75 mutual information scores (normalized) for phoneme *E*, tokens are from the first quantizer.** (Continued)

Continued on next page

Table 5: **Example of top 75 mutual information scores (normalized) for phoneme *E*, tokens are from the first quantizer.** (Continued)

(465, 604)	0.011
(465, 685)	0.011
(465, 465)	0.010
(585, 604)	0.010
(677, 823)	0.009
(465, 766)	0.009
(677, 677)	0.009
(604, 465)	0.009
(733, 604)	0.008
(966, 766)	0.008
(604, 733)	0.008
(733, 733)	0.008
(823, 685)	0.008
(823, 585)	0.008
(766, 604)	0.008
(256, 604)	0.008
(643, 656)	0.008
(766, 966)	0.008