Toward Automatic Discovery of a Canine Phonetic Alphabet

Theron S. Wang[1] **Xingyuan Li**[2] **Hridayesh Lekhak**[1]
Tuan M. Dang[1] **Mengyue Wu**[2] **Kenny Q. Zhu**[3]
[1,3]University of Texas at Arlington, Arlington, Texas, USA
[2]Shanghai Jiao Tong University, Shanghai, China
[1]{sxw7663,hxl7195,txd0904}@mavs.uta.edu
[2]{xingyuan,mengyuewu}@sjtu.edu.cn [3]kenny.zhu@uta.edu

Abstract

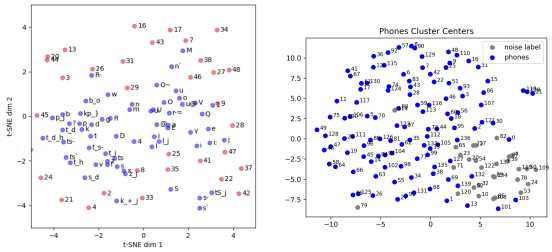
Dogs communicate intelligently through vocalizations, yet the phonetic properties of their communication remain poorly understood. This paper introduces an iterative algorithm inspired by human phonetic discovery, leveraging minimal pairs to identify distinct *canine cognitive vocal units* and construct a comprehensive phonetic alphabet. Additionally, the algorithm derives *canine vocal pattern* that exhibit structured correlations with specific environments and activities, suggesting potential meaningful communicative patterns. Our approach provides a novel framework for analyzing non-human vocalization systems and offers potential applications beyond canines to other animal species, advancing the study of cross-species communication.

1 Introduction

Animals have a variety of ways of communicating information to each other, including vocalizations, behaviors, and smells (Meijer, 2020). Among them, vocalizations are one of the most common communication methods for many animals, such as frogs, birds, primates, and elephants (Schwartzkopff, 1977; Slabbekoorn and Smith, 2002; Zuberbühler, 2001). Scholars from different fields have been interested in studying vocalizations in animal communication. Several studies have demonstrated that vocalizations between animals contain a wealth of information about individuals, behaviors, emotions and contextual environments (Abzaliev et al., 2024; Hantke et al., 2018).

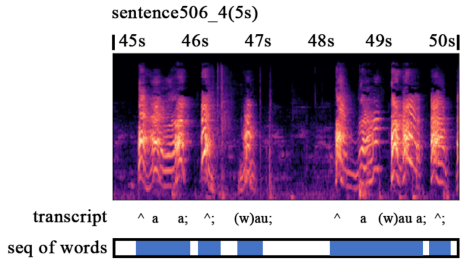
The exploration of animal vocal communication is a complex and challenging task. One important aspect is identifying the basic phonetic units that may exist in animal vocal communication. Although the phonetic units alone are not sufficient to constitute a "language"; they provide the foundation for exploring the grammatical and semantic information embedded in vocal communications. Recent advancements in machine learning have broadened the path of "animal language" research, offering the possibility of a data-driven study that can be conducted at a more granular level. Huang et al. (2023) applied human International Phonetic Alphabet (IPA) directly to transcribe Shiba Inu voices (Figure 1c), which restricts dog sounds to a subset of human IPAs, ignoring the differences between human pronunciations and dog vocals. Hagiwara et al. (2024) proposed Inter-Species Phonetic Alphabet (ISPA), a system designed to transcribe animal sounds into discrete symbols. However, the system was trained on datasets from hundreds of many different species (containing marine mammal, birds, dogs, and even mosquitos) to obtain non-human phones (Figure 1a). This coarse-grained approach overlooks the differences in vocal characteristics across species. Wang et al. (2024) used HuBERT (Hsu et al., 2021) to obtain "phonemes" in Shiba Inu, but many phonemes thus discovered actually refer to noises rather than true dog vocals since there are no mechanisms to completely filter out noise from the input audio (Figure 1b). Sharma et al. (2024) obtained the sperm whale phonetic alphabet by analyzing different combinations of rhythm and tempo in sperm whale communication (Figure 1d). However this method is specialized to sperm whales or other similar marine mammals that produce simple clicking sounds using specialized phonic lips and nasal air sacs.

In this paper, we propose an iterative approach to discover phonetic inventory through simultaneously filtering qualified canine *cognitive vocal units* (akin to human phonemes) and canine *vocal patterns* (akin to human words). We define a canine *vocal unit* (like a phone) as a short, continuous fragment of dog vocalization. A *cognitive vocal unit* is a set-based representation of acoustically similar vocal units that are grouped based on perceptual and contextual similarity, analogous to the relationship between phonemes and phones in hu-

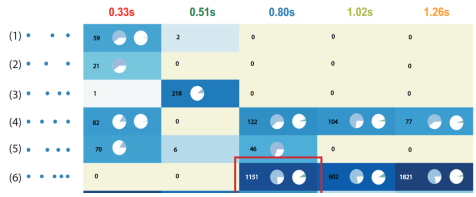


(a) Inter-species phones by Hagiwara et al. (2024).

(b) Canine phones by Wang et al. (2024).



(c) A transcript of Shiba Inu (Huang et al., 2023).



(d) Part of sperm whale phonetic alphabet (Sharma et al., 2024).

Figure 1: Previous attempts of searching for animal phonetic alphabets.

man language. A canine *vocal pattern* is then defined as a sequence of cognitive vocal units that co-occur with, or correspond to, specific environmental or behavioral contexts. While these definitions bear some similarity with human phonemes and words (Twaddell, 1935), they should not be confused or interchangeable. Our starting point is the symbolic transcripts from a large dog vocalization corpus (Wang et al., 2024) which are noisy and inaccurate. The goal of our method is to filter out phonetic symbols that are not dog vocalizations and merge symbols that are similar both acoustically and semantically by using the *minimal pairs principle* in human linguistics.

Our contribution can be summarized as follows:

- We create the biggest ever dog vocalization dataset with 152 hours of pure dog barks from 6 different common breeds.
- We propose an iterative canine cognitive vocal unit discovery algorithm that produces a phonetic alphabet of 105 distinct canine cognitive vocal units and 69 canine vocal patterns from 6,285 hours of YouTube videos of six popular dog breeds. [1]
- Our experiments show that 42% of the vocal units we get from canine vocalization datasets are distinct cognitive vocal units, as verified by human evaluation, and 11% of the vocal patterns have a high likelihood of appearing in specific contexts. In addition, the set of vocal patterns discovered closely follows the Zipf law Figure 6.

2 Approach

We employ a data-driven approach to uncover the canine phonetic alphabet using data sourced from YouTube. The pipeline, as outlined in Figure 2, consists of three steps: *preprocessing*, *transcription*, and *vocal units refinement*. Besides canine vocal units and canine vocal patterns, which were defined earlier, we define a *canine vocal segment* (similar to a human sentence) as a continuous dog vocalization bounded by long pauses (> 0.5 sec). The preprocessing step extracts canine vocal segments from the audio track of dog videos. The transcription step converts each canine vocal segment into a sequence of frame vectors and then further into a sequence of symbols which represent canine vocal units through clustering. The vocal unit refinement step refines the vocal units into cognitive vocal units by removing false vocal units and merging similar ones, while producing possible canine vocal patterns at the same time. We discuss these steps in detail next. All terms used in this paper are defined in Table 1.

2.1 Preprocessing

Similar to the work of Huang et al.'s (2023), YouTube videos often contain background music, human speech and other noises than dog vocalizations. Therefore, it is essential to exclude as much noise as possible. First, we employed AudioSep (Liu et al., 2023) to remove extraneous noise from the video data. Then, we utilized the fine-tuned DCASE2023 Challenge Task 4 baseline model to extract dog vocal segments. We replace the default encoder with a pretrained BEATs (Chen et al., 2022) for better sound event detection performance. To further enhance accuracy, we manually labeled over 9,000 seconds of pure dog vocal data

[1] Our data and code is available at: https://github.com/TheronWang/Canine-phonetic

Term	Definition	Similar Concept in Linguistics
Vocal Segment	A continuous dog vocalization bounded by long pauses (> 0.5 sec).	sentence
Vocal Unit	A short, continuous fragment of dog vocalization.	phone
Cognitive Vocal Unit	A set-based representation of perceptually and contextually similar vocal units.	phoneme
Vocal Pattern	A sequence of cognitive vocal units co-occurring with specific contexts.	word

Table 1: Glossary of terms.

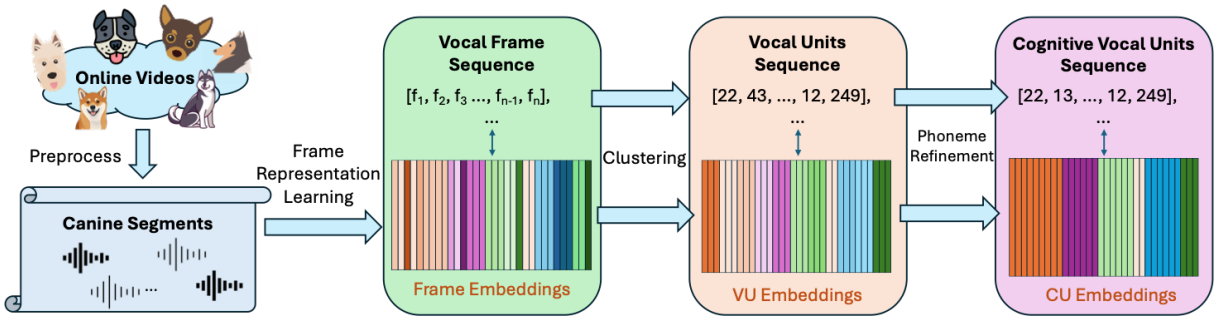


Figure 2: Data processing flow: from dog videos to cognitive vocal units.

for post training the SED model, achieving an F1 score of 0.8556 on the test set. This process enabled us to obtain cleaner and higher-quality dog vocal segments.

2.2 Transcription

We then trained HuBERT (Hsu et al., 2021), a self-supervised audio representation model, on large number of canine segments (Li et al., 2024), and obtain a frame embedding representation for each 20ms audio frame in the canine segments. Next, we cluster the frames into K clusters, where K is determined by the elbow method (Thorndike, 1953). The cluster center, calculated as the average of all frames within each cluster, served as the feature embedding for a unique phone. We assign a unique symbol to each vocal unit and make it an initial "cognitive vocal unit." These initial vocal units are subject to refinement and modification in the next step. At the end of the transcription step, each canine segment is represented by a sequence of vocal unit embeddings and corresponding symbols.

2.3 Canine Vocal Units Refinement

Canine cognitive vocal unit refinement adopts the *minimal pairs method* (Swadesh, 1934) which has been used to discover phonemes in unknown languages. A minimal pair is a pair of words that differ by only a single phoneme, which can change the meaning of the words (Ladefoged, 2006). For example, in English, the words "bat" and "pat" form a minimal pair, where the difference in the initial phoneme /b/ and /p/ alters the word's meaning.

The core of our algorithm is to identify minimal pairs among candidate vocal patterns. We define a minimal pair as two vocal patterns that (1) differ by exactly one cognitive vocal unit and (2) are empirically distinguishable (e.g., through their associated contexts or acoustic profiles). Furthermore, we assume that all valid vocal patterns must be composed exclusively of cognitive vocal units, and that a cognitive vocal unit must occur as the minimal contrasting unit in at least one such pair.

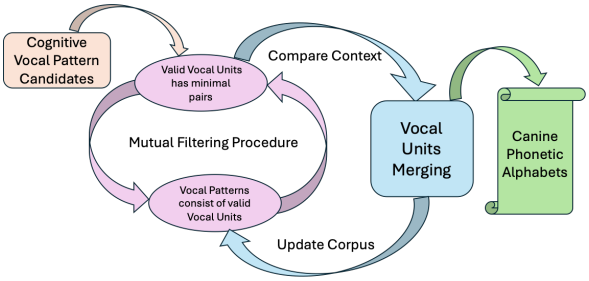


Figure 3: Vocal units refinement algorithm.

Our iterative algorithm proceeds as follows Figure 3: we first generate initial candidates for vocal units and vocal patterns; then, we identify all valid minimal pairs. Since we cannot definitively determine whether two vocal patterns differ semantically, we incorporate a vocal unit merging step: if a pair of cognitive vocal units consistently appears as the contrasting unit in minimal pairs that occur in similar vocal environments, we consider them acoustically and functionally equivalent and merge them. After merging, we retain only those vocal units that serve as contrasting units in at least

one valid minimal pair and filter the vocal patterns to include only those composed entirely of valid vocal units. This refinement process continues until convergence and is shown in Algorithm 1. We then consider the result as the final set of valid cognitive vocal units and vocal patterns, which are subsequently analyzed in downstream experiments.

2.3.1 Vocal Patterns Candidates.

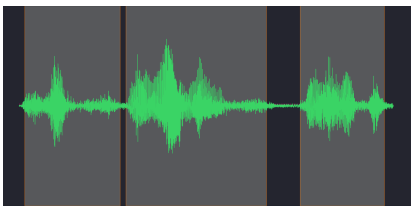


Figure 4: Canine vocal pattern candidate segmentation.

To apply the minimal pair methodology to canine vocal data, we first generate candidate vocal patterns by segmenting contiguous acoustic regions with elevated energy and minimal internal pauses (see Figure 4). Because there is no predefined lexicon or transcription framework for animal vocalizations, these segments serve as proxies for vocal patterns. We then iteratively refine them, updating segment boundaries and underlying vocal unit representations in a data-driven process.

To obtain structurally complete vocal patterns, we first segment continuous canine vocalizations using the Auditok library (Sehili, 2024), which identifies energy-based silences to detect natural pause boundaries. This process yields a sequence of *candidate canine vocal patterns*, each corresponding to a continuous region of elevated acoustic energy.

To accommodate variation in vocal intensity across clips, we adopt a dynamic thresholding strategy. For each utterance, we compute the segmentation threshold as the product of a tunable coefficient ρ and the root mean square (RMS) energy of the full audio segment. We empirically explore a range of ρ values and manually select the one that achieves the most consistent and interpretable segmentation on a development set.

As shown in Figure 4, we present a waveform example with the detected pattern boundaries. The resulting candidate vocal patterns are separated by low-energy intervals determined by the adaptive threshold.

Algorithm 1 Mutual Filtering Algorithm

Input: VP — candidate vocal patterns
Output: VP
Abbreviations:
CU — cognitive vocal units
MP — minimal pairs

```
CU ← EXTRACT_CU(VP)
while HAS_CHANGED(CU) do
    MP ← FIND_MINIMAL_PAIRS(VP)
    CU ← FILTER_CU_FROM_MP(MP)
    VP ← FILTER_VALID_VP(VP, CU)
end while
return VP
```

Algorithm 2 Cognitive Vocal Unit Refinement Algorithm

Input: T — vocal segment transcriptions
Output: CU, VP
Abbreviations:
VP — vocal patterns
CU — cognitive vocal units
JS — contextual JS divergence

```
VP ← SEGMENT_VP(T)
CU ← EXTRACT_CU(VP)
while HAS_CHANGED(CU) do
    VP ← MUTUAL_FILTER(VP)
    JS ← COMPUTE_CONTEXT_JS(VP, T)
    T ← MERGE_CU_AND_UPDATE()
    VP ← SEGMENT_VP(T)
    CU ← EXTRACT_CU(VP)
end while
VP ← MUTUAL_FILTER(VP)
CU ← EXTRACT_CU(VP)
return CU, VP
```

2.3.2 Minimal Vocal Pattern Pairs.

Given the absence of explicit semantic labels for canine vocalizations, we cannot directly determine whether two vocal patterns differ in meaning. Instead, we identify *minimal pair candidates*—pairs of vocal patterns that differ by exactly one cognitive vocal unit. However, due to possible oversegmentation or clustering artifacts, some of these differences may reflect superficial variations of the same vocal pattern rather than distinct forms.

To address this, we evaluate the contextual similarity of each minimal pair candidate by examining the *phonetic environments* in which they occur. Specifically, we extract context distributions based on (1) preceding and following vocal units, (2) co-occurring canine patterns, and (3) local N-gram structures. For each pair, we compute the Jensen–Shannon (JS) divergence (Dagan et al., 1997) between their context distributions. A low JS divergence suggests that the surrounding contexts are highly similar, indicating that the differing vocal units may encode the same functional category.

Pairs falling below a predefined similarity thresh-

old (e.g., the 5th percentile of JS divergence across all pairs) are assumed to represent over-clustered variants of the same vocal unit. In such cases, we merge the corresponding vocal units, update the transcription accordingly, and repeat the entire minimal pair mining process. This iterative refinement continues until both the vocal unit inventory and the vocal pattern list converge.

To improve the robustness of the final inventory, we discard rare vocal patterns that occur fewer than K times in the corpus. The complete refinement procedure is detailed in Algorithm 2. Ultimately, this process yields a stable set of cognitive vocal units and vocal patterns, which we refer to as the *canine phonetic alphabet*.

3 Implementation Details

For fine-tuning the BEATs-SED model on canine vocal segment extraction, we implemented a 4-layer RNN classifier and modified the CNN encoder to output 256-dimensional features. The model was trained using a learning rate of 0.001 for 200 epochs on a 9000-second dataset. We used a batch size of 4 and trained the model on two NVIDIA RTX 4090 GPUs, consuming approximately 48 GB of GPU memory over the course of 6 hours.

For self-supervised representation learning, we trained a three-stage HuBERT model on the same dataset. Audio clips shorter than 0.35 seconds were discarded, while those longer than 5 seconds were truncated with a 1-second overlapping window. We used 100 clusters for the first stage, 200 for the second, and selected 250 clusters for the third stage using the elbow method. The learning rate was set to 0.0001. After the second stage, we extracted features from the 11th transformer layer to train a K-Means model for clustering. The first and second stages of HuBERT training took 20 and 18 hours, respectively, using two NVIDIA RTX 4090 GPUs with 48 GB total memory and a batch size of 4.

For vocal pattern segmentation, we empirically evaluated several energy thresholding parameters and selected $\rho = 3$ as the coefficient that yielded the most coherent and consistent segmentation across samples, as judged by human inspection of waveform boundaries and perceptual continuity in the resulting vocal patterns.

We further experimented with the minimum occurrence frequency $K \in \{1, 3, 5\}$ for filtering rare vocal patterns, and tested Jensen–Shannon (JS) divergence thresholds at the 5th, 10th, and 15th percentiles of the pairwise JS divergence distribution computed across all minimal pair candidates. We selected $K = 5$ and the 5th percentile threshold for merging vocal units, as this combination produced the most reliable segmentations—defined as those that align well with perceived acoustic consistency and contextual distinctiveness in minimal pair analysis. Examples of segmented vocal patterns were visually inspected to ensure that merged units occurred in comparable environments and resulted in fewer fragmented or overly granular units.

4 Evaluation

We apply our approach to a total of 6,235 hours of dog videos. After preprocessing, we extracted 306,233 canine vocal segments, yielding approximately 152 hours of pure dog vocalizations.[2] Our transcription pipeline produced 250 initial canine vocal units, which were subsequently refined to 105 cognitive vocal units and 69 cognitive vocal patterns. Table 2 summarizes the source dataset, and selected cognitive vocal units and vocal patterns are presented in Table 3.

Breed	**Videos (hrs)**	**Vocalizations (hrs)**
Shiba Inu	3,542	52
Chihuahua	332	6
Husky	851	25
Pitbull	349	11
Labrador	829	50
German Shepherd	332	8
Total	**6,235**	**152**

Table 2: Total durations of dog videos and extracted dog vocalizations in our dataset.

Type	**Examples**
Cognitive Vocal Units	0, 2, 3, 4, 7, 8, 9, 10, 11, ... 55, 57, 58, 59, 61, 62, 63, ... 121, 122, 124, 125, 127, 128, ...
Vocal Patterns	“0 95 0”, “95 211 72”, “95 180”, “69 62”, “121 95”, ...

Table 3: Selected cognitive vocal units and vocal patterns.

We evaluate the final canine phonetic alphabet in two dimensions: the quality of the *cognitive vocal units* and the utility of the resulting *vocal patterns*. For cognitive vocal units, we conduct both acoustic and semantic evaluations. While vocal patterns are not the primary output of our pipeline, we examine

[2]The dataset is publicly available on our GitHub page.

their potential contextual significance to support the use of minimal pair analysis.

4.1 Cognitive Vocal Unit Evaluation

We first perform acoustic evaluations to test whether our approach effectively removes noise and produces consistent, distinguishable vocal units. We then examine the semantic utility of the cognitive vocal units by applying them in downstream tasks.

4.1.1 Acoustic Evaluation: Noise Removal

From the 250 initial vocal units generated during transcription, we randomly sampled 10 instances for each vocal unit for human inspection. Manual labeling identified 52 of the total 250 as noise and the remaining 198 as legitimate dog vocalizations. After refinement, the final inventory included 105 cognitive vocal units, of which only 22 were labeled as noise. The remaining 26 noisy units were eliminated by the refinement process, and an additional 4 were removed during vocal pattern segmentation. This demonstrates the effectiveness of our iterative refinement in denoising the initial transcription.

4.1.2 Acoustic Evaluation: ABX Test

To assess the acoustic consistency and discriminability of the cognitive vocal units, we conducted an **ABX Test** (Munson and Gardner, 1950), a standard psychoacoustic task used to determine phonetic units. In this test, two reference vocal units, A and B, are presented, followed by a third unit X, which matches either A or B. The goal is to determine which of the two reference units is acoustically closer to X.

Sample Set	Group 1	Group 2	Agreement
Initial Vocal Units	78.00%	84.00%	79.6%
Final Cognitive Units	**82.00%**	**86.00%**	79.5%

Table 4: ABX test results showing accuracy from two independent rater groups and inter-group agreement.

To evaluate both the initial and refined cognitive vocal units, we generated 500 ABX samples from the HuBERT-derived initial units and 210 samples from the final refined set (number of samples depends on the number of vocal units). These sample sizes were selected to ensure representative coverage across the unit inventories. To reduce evaluation bias, each ABX test was assessed independently by two groups of graduate student annotators, all of whom were familiar with dog behavior and vocalization. Each annotator evaluated approximately 100 samples with the same instructions: listen to three segments and indicate whether X more closely resembled A or B.

As shown in Table 4, the inter-group agreement reached approximately 80%, validating the reliability of the test setup. Both groups also achieved higher accuracy when evaluating the final cognitive vocal units compared to the initial set. These results suggest that the refined units are not only acoustically distinct but also more consistently perceived—supporting the effectiveness of our refinement algorithm.

4.1.3 Contextual Evaluation: Downstream Tasks

In the absence of a gold standard for canine phonetic structure, evaluating feature quality through downstream tasks is a well-established practice in representation learning (Chung et al., 2020). Such evaluations indirectly demonstrate the semantic utility and discriminative power of the learned cognitive vocal units. In this section, we employ two classification tasks to assess the effectiveness of our features.

The first task is **dog bark type classification**. We use bark-type-labeled clips from AudioSet (Gemmeke et al., 2017), which defines six distinct bark categories with corresponding time stamps (see Appendix C). No additional annotation was required, as bark type labels were pre-annotated in the dataset.

The second task is **dog bark context classification**. We randomly sample 1,000 bark-containing video clips from AudioSet and annotate each clip along three dimensions: *ear movement*, and *tail movement*. These dimensions reflect contextual and behavioral signals associated with the dog's vocalization. The full set of annotation categories is shown in Table 5.

Semantic Type	Possible Values
Surroundings	Well-known Human; Stranger; Items; Other Sounds
Ear Movement	Standing; Laying
Tail Movement	Wagging; Stationary

Table 5: Semantic categories used in downstream task 2.

Five graduate students participated in the annotation process. Prior to annotation, they were provided with written instructions and labeled examples illustrating each semantic category. Since

cognizing ear/tail movement and environmental context does not require domain-specific expertise, the task was designed to be accessible and repeatable by non-experts. Annotators viewed 10-second clips and labeled all three semantic fields. If uncertain, they marked the instance as "N/A." These ambiguous labels were excluded from downstream evaluation.

We trained standard classifiers—including Logistic Regression (LR), Random Forest (RF), and XGBoost (XGB)—on both tasks. To ensure comparability, we extracted four types of features for each vocal segment: 1. 13-dimensional MFCCs, 2. 768-dimensional frame-level embeddings, 3. 768-dimensional vocal unit embeddings, 4. 768-dimensional cognitive vocal unit embeddings.

All features were temporally average-pooled before classification.

Feature Type	LR	RF	XGB
MFCC	0.186	0.207	0.242
Frame Embedding	**0.365**	0.255	0.346
Vocal Unit Embedding	0.351	0.220	0.327
Cognitive Vocal Unit Embedding	0.354	**0.281**	**0.353**

Table 6: Macro F1 scores for dog bark type classification.

Results: Bark Type Classification Table 6 shows that cognitive vocal unit embeddings outperformed other features on RF and XGB, suggesting higher semantic alignment. However, frame embeddings achieved slightly better performance on LR, indicating that low-level acoustic features still play a role in bark type discrimination.

Results: Bark Context Classification As shown in Table 7, cognitive vocal unit embeddings achieve the best overall performance in most cases, particularly for tail and ear movement prediction. This indicates that our final symbolic representations capture context-sensitive features beyond raw acoustics. We also observe a general trend: as features become more abstract and symbolic—progressing from MFCCs to cognitive vocal units—the classification performance improves, reinforcing the hypothesis that discrete vocal representations offer stronger semantic grounding.

4.2 Canine Vocal Pattern Evaluation

To explore whether the discovered vocal patterns (analogous to words) carry potential semantic value, we analyze the *environment* and *activity* associated with the dog when each vocal pattern is uttered. Specifically, we investigate whether particular vocal patterns consistently co-occur with specific contexts, which would suggest possible referential or functional roles.

4.2.1 Setup

We use the Janus-Pro-7B vision-language model (Chen et al., 2025) to describe the dog's environment and behavior. For each vocal pattern occurrence, we extract three video frames sampled from a window spanning three seconds before to three seconds after the utterance. Descriptions are generated using a consistent prompt (see Appendix B).

The generated textual descriptions are then mapped into 10 predefined categories each for environment and activity, as listed in Table 8.

To quantify correlations, we compute the *relative frequency* (RF) of each category for every vocal pattern:

$$\mathrm{RF}(w, c) = \frac{C(c, w)}{C(w)} - \frac{\sum_{w'} C(c, w')}{\sum_{w'} C(w')}, \qquad (1)$$

where $C(c, w)$ is the count of vocal pattern w appearing in category c, and $C(w)$ is the total number of times w occurs. The first term represents the category frequency specific to vocal pattern w, and the second term is the global category frequency across all vocal patterns. A positive RF score indicates overrepresentation of a category in a given vocal pattern, suggesting a possible contextual association.

4.2.2 Results

Figure 5 shows an example for the vocal pattern `0_95_0`. This pattern is negatively correlated with the "Indoor" environment and passive activities like "Observing & Watching," suggesting it rarely appears in calm, indoor contexts. Conversely, it shows positive correlations with active behaviors such as "Movement" and "Exploring & Investigating," hinting that it may functionally correspond to more dynamic, outdoor settings.

To assess broader trends, we evaluated all 69 high-frequency vocal patterns. We found that the majority exhibited strong positive or negative correlation with one to two semantic categories—an observation that supports their potential referential consistency. Selected examples are shown in Table 9. A comprehensive analysis of all patterns is included in Appendix D.

Feature Type	Surroundings			Ear Movement			Tail Movement		
	LR	RF	XGB	LR	RF	XGB	LR	RF	XGB
MFCC	0.253	0.251	0.241	0.421	0.444	0.491	0.498	0.501	0.509
Frame Embedding	0.363	0.256	0.259	**0.544**	0.466	0.464	0.655	0.589	0.580
Vocal Unit Embedding	0.306	0.244	**0.328**	0.448	0.461	0.464	0.610	0.610	0.584
Cognitive Vocal Unit Embedding	**0.373**	**0.353**	0.321	0.455	**0.467**	**0.515**	**0.656**	**0.616**	**0.591**

Table 7: Macro F1 scores for dog bark context classification across three dimensions.

Environment	Activity
Indoor	Passive Action
Walls & Windows	Movement
Furniture	Observing & Watching
Outdoor	Exploring & Investigating
Plants	Interaction with Items
Play Area	Feeding
Vehicles	Drinking
Water	Interaction with Animals
Pet Space	Water-Related
Crowd	Interaction with Humans

Table 8: Environment and activity categories extracted from Janus-Pro-7B outputs.

Vocal Pattern	Inferred Contextual Description
0 95 0	Outdoor, near furniture; active movement; not observing.
95 211 72	Outdoor, alert; interacting with surroundings.
15 138 55	Indoor, passive; minor movement, some watching.
95 180	Green outdoor space; light water-related activity.
69 62	Residential interior; feeding and light interaction.
121 95	Open area; observing surroundings; engaging with animals.

Table 9: Example contextual descriptions for selected vocal patterns.

Although only 69 vocal patterns show strong contextual associations, our overall transcription contains over 1,000 vocal patterns. This raises questions about the robustness of our minimal pair refinement method. However, the existence of even a moderate-sized subset of patterns with stable contextual correlates provides empirical support for the utility of the minimal pair paradigm in discovering cognitive vocal units. Future work can focus on improving vocal pattern frequency thresholds and refining the category mapping process to enhance semantic precision.

5 Related Work

In the following, we discuss some previous work on discrete representation of structures in animal communication, and phonetic discovery in unknown human languages.

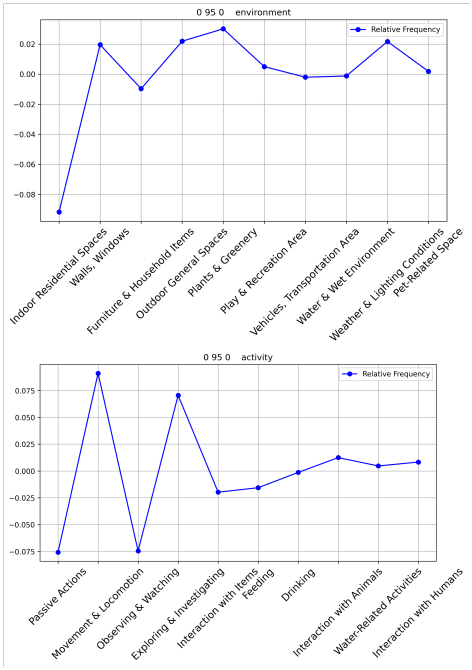


Figure 5: Relative frequency of environmental and activity categories for vocal pattern `0_95_0`.

5.1 Discrete Representation of Animal Vocalization

Besides canines, previous research has shown that many animal species communicate with discrete patterns (Kershenbaum et al., 2014; Cartmill, 2023). Ficken et al. (1994)'s work illustrates that Mexican chickadees have simple units of sound and can form a call system with a variety of meanings by combining different units.

Research on chimpanzees reveals that they combine these vocal units into longer, structured sequences to create a versatile vocal communication system (Girard-Buttoz et al., 2022). Japanese tits produce distinct alarm calls for different threats and combine these calls into structured sequences, illustrating the complexity of their communication systems (Suzuki, 2021). Similarly, bottlenose dolphins employ unique signature whistles, characterized by specific frequency modulation patterns, to broadcast their identity (Janik et al., 2013). Additionally, Egyptian fruit bats use their vocalizations

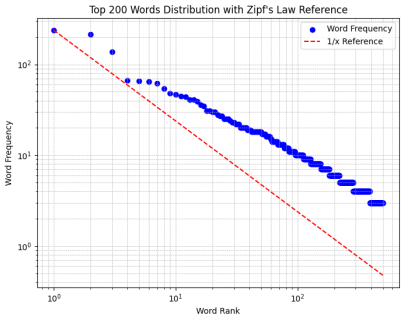


Figure 6: Rank distribution of top 200 vocal patterns vs. Zipf's Law.

to convey detailed information about the emitter, context, and addressee (Prat et al., 2016). These examples highlight that structures similar to linguistic phones and their combinations are present in animal communication, reflecting the complexity and sophistication of these systems across different species.

5.2 Phoneme Discovery in Human Languages

Since there are no written letters associated with canine language, it is reminiscent of the documentation of unwritten human languages when we're trying to do the phonemic analysis of canine language. Generally, there are two types of language documentation, which are extremely challenging. One is considered as the documentation of extinct languages without any speech resources, while the other is the documentation of unwritten languages with very few language consultants. Our phonemic analysis of canine language is inspired by the documentation of unwritten human languages.

Phonemic analysis is a fundamental part of the description and documentation of a language, which is primarily concerned with identifying the contrastive sounds (Kempton and Moore, 2014). Since the process of a phonemic analysis involves looking for evidence of contrast between every possible pair of sounds, which is often very time-consuming, we designed an automated and reiterative algorithm to utilize one of the most effective methods: minimal pairs. Minimal pairs are two different words that differ in exactly one sound in the same location and are considered the only method to conclusively establish contrast between sounds (Hayes, 2011).

Previously, to find the putative minimal pairs (with noisy data) in a "word list" of unwritten and undocumented Kua-nsi language data, Kempton and Moore (2014) used a program called Minpair written in C. The minimal pair algorithms they applied had surprisingly poor performance based on the ROC-AUC evaluation measure. Therefore, a much more innovative method of minimal pair algorithm is highly needed to expedite the procedures related to the phonemic analysis of any unwritten language, including the canine language in our current study.

6 Conclusion

We presented an iterative framework for discovering an alphabet of canine cognitive vocal units and their corresponding vocal patterns from raw dog vocalizations. The method is general and robust, with potential applicability to vocal communication analysis in other animal species. To ensure statistical reliability, we aggregated data from six popular dog breeds, as individual breed recordings were insufficient for training a stable HuBERT model or deriving reliable statistics. In future work, we aim to enhance the preprocessing pipeline to extract higher-quality, breed-specific vocalizations. Furthermore, improving the algorithm's ability to automatically filter out noise vocal units remains an important challenge, particularly given the prevalence of environmental interference in animal recordings.

Limitations

This work introduces an algorithm to discover a phoneme alphabet for canine vocalizations. Verification of these phonemes is currently done by averaging their embeddings within transcripts, which does not fully capture potential linguistic structures. A more robust approach could involve treating the transcript as a sequence of discrete tokens and applying a transformer-based model directly to it.

Ethical Considerations

This work does not involve live animals and uses public-domain data only. The released dataset consists of processed vocal segments, transcripts, and embeddings (no original videos) for research purposes. Researchers must sign an agreement to ensure it is used solely for research, mitigating any privacy concerns.

Acknowledgment

This work is partially supported by NSF Award No. 2349713.

References

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

Erica A Cartmill. 2023. Overcoming bias in the comparison of human language and animal communication. Proceedings of the National Academy of Sciences, 120(47):e2218799120.

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

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025. Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811.

Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2020. Rethinking embedding coupling in pre-trained language models. arXiv preprint arXiv:2010.12821.

Ido Dagan, Lillian Lee, and Fernando Pereira. 1997. Similarity-based methods for word sense disambiguation. In 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics, pages 56–63, Madrid, Spain. Association for Computational Linguistics.

Millicent Sigler Ficken, Elizabeth D Hailman, and Jack P Hailman. 1994. The chick-a-dee call system of the mexican chickadee. The Condor, 96(1):70–82.

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), pages 776–780. IEEE.

Cédric Girard-Buttoz, Emiliano Zaccarella, Tatiana Bortolato, Angela D Friederici, Roman M Wittig, and Catherine Crockford. 2022. Chimpanzees produce diverse vocal sequences with ordered and recombinatorial properties. Communications Biology, 5(1):410.

Masato Hagiwara, Marius Miron, and Jen-Yu Liu. 2024. ISPA: Inter-species phonetic alphabet for transcribing animal sounds. arXiv preprint arXiv:2402.03269.

Simone Hantke, Nicholas Cummins, and Bjorn Schuller. 2018. What is my dog trying to tell me? the automatic recognition of the context and perceived emotion of dog barks. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5134–5138. IEEE.

Bruce Hayes. 2011. Introductory phonology. John Wiley & Sons.

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, 29:3451–3460.

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

Vincent M Janik, Stephanie L King, Laela S Sayigh, and Randall S Wells. 2013. Identifying signature whistles from recordings of groups of unrestrained bottlenose dolphins (tursiops truncatus). Marine Mammal Science, 29(1):109–122.

Timothy Kempton and Roger K Moore. 2014. Discovering the phoneme inventory of an unwritten language: A machine-assisted approach. Speech Communication, 56:152–166.

Arik Kershenbaum, Ann E Bowles, Todd M Freeberg, Dezhe Z Jin, Adriano R Lameira, and Kirsten Bohn. 2014. Animal vocal sequences: not the markov chains we thought they were. Proceedings of the Royal Society B: Biological Sciences, 281(1792):20141370.

Peter Ladefoged. 2006. A course in phonetics. Thomson Wadsworth.

Xingyuan Li, Sinong Wang, Zeyu Xie, Mengyue Wu, and Kenny Q Zhu. 2024. Phonetic and lexical discovery of a canine language using HuBERT. arXiv preprint arXiv:2402.15985.

Xubo Liu, Qiuqiang Kong, Yan Zhao, Haohe Liu, Yi Yuan, Yuzhuo Liu, Rui Xia, Yuxuan Wang, Mark D Plumbley, and Wenwu Wang. 2023. Separate anything you describe. arXiv preprint arXiv:2308.05037.

Eva Meijer. 2020. Animal languages. MIT Press.

WA Munson and Mark B Gardner. 1950. Standardizing auditory tests. The Journal of the Acoustical Society of America, 22(5_Supplement):675–675.

Yosef Prat, Mor Taub, and Yossi Yovel. 2016. Everyday bat vocalizations contain information about emitter, addressee, context, and behavior. Scientific Reports, 6(1):1–10.

Johann Schwartzkopff. 1977. Auditory communication in lower animals: Role of auditory physiology. Annual review of psychology.

Amehdi Sehili. 2024. Auditok: Audio activity detection and audio segmentation library. https://github.com/amsehili/auditok. Accessed: 2025-02-10.

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):3617.

Hans Slabbekoorn and Thomas B Smith. 2002. Bird song, ecology and speciation. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 357(1420):493–503.

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

Morris Swadesh. 1934. The phonemic principle. Language, 10(2):117–129.

Robert L Thorndike. 1953. Who belongs in the family? Psychometrika, 18(4):267–276.

W Freeman Twaddell. 1935. On defining the phoneme. Language, 11(1):5–62.

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, pages 13972–13983. Association for Computational Linguistics.

Klaus Zuberbühler. 2001. Predator-specific alarm calls in campbell's monkeys, cercopithecus campbelli. Behavioral Ecology and Sociobiology, 50:414–422.

A The Implementation Details of Extracting Sentences and HuBERT Pretraining

For fine-tuning the DCASE2023 challenge task 4 baseline model, we set up a 4-layer RNN network and modified the CNN to 256 dimensions. The learning rate was set to 0.001 and trained for 200 epochs.

For HuBERT training, we removed data smaller than 0.35 seconds and cropped audio larger than 5 seconds by a maximum length of 5 seconds with a one second overlap. We used 100 clusters at the first stage and 200 clusters at the second stage. The learning rate was set to 0.0001. Then, we used features from the 11th transformer layer of the second-stage HuBERT model to train a K-Means model.

B Multi Modal Large Language Model Prompt

Below are the prompts used to generate responses from the LLM:

```
[Instruction-based Prompt]
Describe the environment and the activity
 of the dog in the image using a list of words,
  follow the format:

{"environment": [description1,
 description2, ...],
"activity": [description1,
description2, ...]}

[Response Example]
{"environment":
["living room", "rug", "chairs",
 "dog crate", "kitchen", "blinds"],
"activity":
 ["dog jumping",
  "dog jumping over person"]}
```

C Dog Bark Types

Bark Type	Description
Bark	Principal communication sound produced by dogs.
Yip	Sharp, high-pitched bark, often from small breeds.
Howl	Long, plaintive vocalization produced by canids.
Bow-wow	Tonal vocalization, less abrupt than a classic bark.
Growl	Low-pitched, guttural warning or aggressive signal.
Whimper	Subdued vocalization expressing fear, pain, or submission.

Table 10: Dog bark types used in downstream task 1.

D Canine Word Meaning

Table 11 gives the descriptive meaning of each of 69 discovered canine vocal patterns. We feed the relative frequencies of each word into the GPT-4o model to generate a summary of each word's meaning, and we validated that all summaries corresponded to their respective frequencies.

Table 11: Environmental and Activity-Based Descriptions

Canine Word	Duration Mean (s)	Duration Std-Dev (s)	Description
0	0.12	0.05	Resting on furniture, passive but slightly observant.
12	0.14	0.07	Watching outdoors, focused and still.
17	0.17	0.07	Indoors near walls, watching and drinking.
19	0.15	0.09	Calm indoors, occasional drinking and social.
32	0.17	0.08	Observing indoors, little movement, slightly curious.
72	0.10	0.01	Active near water, engaged in movement and drinking.
88	0.11	0.02	Outdoors, passive but engaged in drinking and watching.
95	0.06	0.00	Outdoor general space, mild movement and feeding.
103	0.18	0.15	Passive in pet-related space, light water interaction.
121	0.16	0.09	Observing outdoors near beach, engaging in interactions.
122	0.13	0.05	In pet spaces, watching and slightly moving.
151	0.13	0.05	Active in pet spaces, interacting and feeding.
165	0.12	0.04	Near walls, engaged in feeding and mild movement.
243	0.13	0.05	Observing in pet-related areas, some drinking and feeding.
0 95	0.07	0.03	Slightly passive indoors, little interaction.
19 95	0.09	0.05	Play area focus, occasional feeding and watching.
3 95	0.07	0.03	Indoors, near furniture, engaged in feeding.
141 95	0.06	0.00	Outdoors near the beach, interacting with animals.
233 95	0.08	0.04	Near water, slight movement, observing surroundings.
64 95	0.07	0.03	Indoor space near windows, exploring surroundings.
171 95	0.07	0.03	Indoor, near walls and windows, observing and feeding.
95 0	0.08	0.04	Indoor, household items and furniture, watching surroundings.
56 95	0.07	0.03	Indoor, near furniture, engaged in interactions.
199 95	0.06	0.01	Outdoor, general spaces, slightly interactive.
21 95	0.08	0.03	Outdoor, near beach, engaging with surroundings.
32 95 32	0.23	0.13	Indoor, near windows, exploring surroundings.
159 95	0.07	0.02	Exploring near furniture, some interaction with humans.
131 95	0.06	0.01	Indoor, curious about surroundings, occasional movement.
211 72	0.12	0.03	Active near water, engaging in movement.
114 95	0.07	0.03	Indoor, social behavior with some feeding and drinking.
95 211 72	0.10	0.01	Outdoor, watching and interacting with surroundings.
95 72	0.11	0.03	Water-related spaces, engaging with other animals.
102 95	0.06	0.02	Indoor, engaging with surroundings, mild exploration.
195 95	0.07	0.02	Observing indoors, occasional feeding and drinking.
127 3	0.15	0.07	Indoor play area, interacting with items.
212 95	0.09	0.05	Passive indoors, light interaction with surroundings.
32 95	0.11	0.08	Near windows, slight movement, observing environment.
73 95	0.08	0.05	Indoor play area, occasional drinking and interaction.
32 12	0.15	0.07	Observing outdoor spaces, some interaction with humans.
231 95	0.08	0.05	Outdoor environment, slight movement, watching surroundings.
200 95	0.06	0.02	Indoor residential space, occasional drinking and feeding.
207 95	0.07	0.04	Outdoor play space, engaged in exploring and moving.
103 95	0.10	0.06	Passive indoor setting, slight interaction with items.
13 95	0.07	0.03	Observing indoors near furniture, occasional social interaction.
8 95	0.07	0.02	Outdoor setting, slight movement, observing humans.
239 95	0.06	0.00	Outdoor space near plants, occasional interaction.
90 95	0.07	0.02	Indoor furniture setting, passive with slight movement.
174 95	0.08	0.04	Outdoor play area, engaged in social interaction.
153 95	0.07	0.02	Indoor residential space, engaged in observation and drinking.
50 95	0.08	0.03	Indoor space with minor interactions, some object interaction.
0 95 0	0.14	0.09	Indoor near furniture, passive with some movement.
69 62	0.15	0.07	Residential indoor space, occasional feeding and interaction.

Continued on next page

Canine Word	Mean Duration	Std Duration	Description
95 21	0.09	0.05	Passive indoors, minor movement and object interaction.
95 180	0.09	0.03	Outdoor greenery, slight water interaction.
229 95	0.06	0.01	Indoor furniture-related, minor pet space activity.
12 95	0.10	0.05	Passive outdoor observation, light drinking.
151 95	0.07	0.02	Indoors near pet spaces, engaging with objects.
66 95	0.07	0.01	Active near windows, slight exploration.
148 188 203 98 230 19	0.23	0.03	Passive in domestic settings, engaged with walls and windows.
95 3	0.10	0.09	Indoors near transportation areas, slight movement.
60 95	0.08	0.03	Indoors near lighting sources, some feeding activity.
95 32	0.13	0.05	Near plants indoors, exploring and interacting.
19 95 19	0.19	0.10	Active outdoors, moderate social interactions.
46 95 150 0 52 108 68 155 229 95	0.28	0.03	Outdoor observation, slightly engaged in activities.
93 95	0.14	0.13	Passive around plants, light exploration.
122 95	0.09	0.07	Indoor near pet spaces, slightly attentive.
121 95	0.10	0.05	Observing outdoors, interacting with animals.
0 95 0 95	0.15	0.06	Passive indoors with slight object interaction.
238 95	0.06	0.01	Water-related area with slight movement and watching.