EmotionalCanines: A Dataset for Analysis of Arousal and Valence in Dog Vocalization

Tuan M. Dang
The University of Texas at Arlington
Arlington, Texas, USA
txd0904@mavs.uta.edu

Theron S. Wang
The University of Texas at Arlington
Arlington, Texas, USA
sxw7663@mavs.uta.edu

Hridayesh Lekhak
The University of Texas at Arlington
Arlington, Texas, USA
hxl7195@mavs.uta.edu

Kenny Q. Zhu
The University of Texas at Arlington
Arlington, Texas, USA
kenny.zhu@uta.edu

Abstract

This study centers on the creation of a novel dog bark emotion dataset, EmotionalCanines, capturing the emotional spectrum of canine vocalizations. In the current literature on animal communication and its intersection with machine learning, there is a limited amount of open-sourced data available to facilitate research, mainly due to constraints in animal subjects and recording conditions. To address this gap, we propose a framework that enables the collection of reliable arousal and valence labels in animal emotional state at scale. Through its application, we built a dataset of 1,400 dog bark sequences with corresponding arousal and valence labels, the largest of its kind, for the Husky and Shiba Inu dog breeds. By constructing this dataset, we provide a foundation for decoding dog bark patterns and advancing animal communication research.

CCS Concepts

• **Computing methodologies → Machine learning**; **Supervised learning**; **Natural language processing**; **Speech recognition**.

Keywords

bioacoustics; animal communication; animal emotions; dogs

ACM Reference Format:
Tuan M. Dang, Theron S. Wang, Hridayesh Lekhak, and Kenny Q. Zhu. 2025. EmotionalCanines: A Dataset for Analysis of Arousal and Valence in Dog Vocalization. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3746027.3758286

1 Introduction

Animal vocalization processing and understanding is a rapidly growing field of research that is moving beyond traditional acoustic methods and into modern machine learning approaches. In dogs, an animal that has been evolving alongside humans for thousands of years, we believe there is much unknown about the communication patterns of their vocalization and aim to have a better understanding of this enigma. One of the main usages of communication in humans is to convey emotions. Some past studies have found a crossover of this function in animals [11]. Some studies have found correlations between acoustic features (such as pitch, tempo, tonality, etc.) in vocalization and emotional state in animals such as pigs [33, 36, 56], horses [12, 35], goats [13], and more specifically in dogs [22, 27, 49]. However, there has not been a study that properly attempted to breakdown dog vocalization and investigated the bark patterns that appear when they convey specific emotional states. Furthermore, there is a shortage of datasets available in dogs (and other animals) that can facilitate this research problem.

In this paper, we describe the process of building the EmotionalCanines dataset, the largest of its kind, to facilitate research focused on showing evidence of emotional expression in dog vocal communication. Our framework can be applied to videos of the same or other dog breeds and animal species, resulting in a dataset that can continuously grow in size without limitations in animal subjects or recording conditions, moving beyond traditional datasets derived from coordinated scenarios and solving the data shortage problem with the target animal. Furthermore, since this dataset is multi-breed, it can also facilitate the study of dog emotion-vocalization differences among different breeds. Some breeds could be more expressive at showing emotions, while others could be more "shy".

In the rest of this paper, we use the following terminology:

- **Emotional state** is a short term affective state or inner state as a reaction to an event and is quantified by two measurements in this study: arousal and valence. Traditionally, these two attributes are continuous values on a Likert scale (i.e. from 1-5 or 1-7, etc.), but we are choosing to discretize this range into three classes for this dataset. More details on the literature and this decision are discussed in **Section** 3.1.
- **Arousal** refers to the level of movement or energy exhibited by the dog, classified into three labels: `Low`, `Medium`, and `High`.
- **Valence** refers to the level of "pleasantness" or "unpleasantness" associated with a stimulus and its effect on the dog, classified into three labels: `Negative`, `Neutral`, and `Positive`.

From video clips containing dog vocalizations, we had annotators label the arousal and valence of the dog based on the aforementioned classes. Human annotators were given clear guidelines when labelling these clips (More details in **Section** 3.1).

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3758286

Our contributions are as follows:

- We created the largest emotion-vocalization dataset in dogs in the current literature, consisting of two breeds (Husky and Shiba Inu), totalling 1,400 pure bark sequences and 35 minutes in duration. [1]
- We are introducing a framework to collect dog video data and generate dog emotional state annotations at scale, not limited by animal subjects or recording conditions, built upon the current literature on animal affective science.
- We provide analysis on baseline models and features for an emotion classification task to highlight the characteristics and differences of emotion-vocalization from Huskies and Shiba Inus, further demonstrating the challenge of understanding dog emotional state from vocalization.

2 Related Work

In the current literature, there are two popular databases created to enable analysis of dog vocalizations: the Mudi dog database [49] and the Mescalina database [46]. Multiple research works have been conducted with the use of datasets or subsets of datasets created from these databases, as described below. The latest versions of these datasets include Mudi (6,614 barks, 12 Mudi dogs), Mescalina 2015 (6,077 barks, 37 dogs, nine Mexican breeds), and Mescalina 2017 (6,948 barks, 65 dogs) [26]. These datasets support research on the information about individual, breed, age, sex, and context inferred from dog barks.

Previous research works showed some evidence that humans can categorize barks by context [49] but struggle to identify individual dogs [41]. Building on this work, Molnár et al. [42] attempted the context and individual classification task on the Mudi dataset. A different approach revisited context classification [45], extracting 6,552 features with openSMILE, reduced to 500 via Relief in Weka. They trained SVMs and tested under four validation schemes (OMPD, 10FCV, Resample, LODOV), concluding that MFCC was the most effective feature.

Another study extended research on dog vocal classification to include sex and age classification [31]. They extracted acoustic measures and tested with four supervised learning methods (naive Bayes, classification trees, k-NN, logistic regression) and three feature selection strategies, evaluated via 10-fold cross-validation.

Using the Mescalina dataset, researchers tackled individual, breed, age, sex, and context classification [26, 46]. Machine learning models (CNNs, J48, SVM, Random Forest, Bagging, Naive Bayes) used MFCC, Mel spectrograms, and Low-Level Descriptors, with CNNs outperforming the other classifiers.

Another study attempted to advance this area by utilizing transfer learning with the self-supervised Wav2Vec2 model, pre-trained on human speech, to classify 8,034 bark segments across 14 contexts [1]. Fine-tuned for dog recognition, breed, gender, and context, it outperformed models trained solely on dog barks.

On the other hand, a very limited amount of research has been done in the area of emotion in dog vocalization. Some studies investigated whether human listeners can categorize dog barks based on the situations in which they were recorded and associate them with emotional content [22, 49]. The study aimed to explore acoustic differences in barks and the influence of listeners' experience with dogs on their ability to interpret these vocalizations.

A total of 72 barks recorded in six contexts from 19 Mudis were used. Each listener was asked to rate the emotionality (aggressiveness, fearfulness, despair, playfulness, happiness) and categorize the situation from the barks. All groups categorized barks significantly above chance (16.67%), with accuracies of approximately 39-41%. Listeners consistently associated specific emotions with certain situations. Stranger and schutzhund barks were rated high in aggressiveness. Alone barks were rated high in despair and fearfulness, with low happiness and playfulness scores. Play barks were rated high in playfulness and happiness. Walk and ball barks were harder to distinguish emotionally.

Emotion classification was explored using the Emotional Dog Corpus (EmoDog), built on the Mudi dataset [27]. Six trainers rated 226 bark sequences from 12 dogs for emotions (Aggression, Despair, Fear, Fun, Happiness). Three feature sets (eGeMAPS, COMPARE, BoAW) were tested with SVMs for classification and SVR for intensity prediction. COMPARE excelled in context classification, eGeMAPS in emotion classification.

Although these research works have contributed greatly in the area of dog and animal vocalization, there are still some gaps and flaws that could be improved upon.

Except for EmoDog [27], no dog bark datasets have been built solely for dog emotions, but mainly for contexts. The study focuses on specific contexts (e.g., Alone, Play, Stranger) but lacks exploration of a broader range of stimuli and emotional annotations, potentially limiting generalizability to real-world scenarios and missing other significant stimuli (e.g., interaction with other dogs or animals) that could influence bark characteristics. Neither the Mudi nor Mescalina datasets record dogs interacting with each other or with other animals in their contexts, although those situations have high potential for the dogs to express a purpose to their vocalization. These datasets also have limitations with small amounts of training data and number of dogs. Furthermore, the 1-5 Likert scale for emotions is subjective and may be influenced by personal biases. The study lacks validation of emotional categories against physiological or behavioral dog data, assuming human-perceived emotions reflect the dog's state, introducing anthropomorphic bias, and risking misalignment with actual dog states.

In humans, over the past few decades, there has been extensive research on human speech emotion in the intersection between Psychology, Affective Computing, and Linguistics, among other areas. Multiple datasets have been created to facilitate this research such as IEMOCAP [15], FAU Aibo [55], GEMEP-CS [6], MSP-PODCAST [34], etc. The field of Linguistics provide a different prospective on the relationship between human speech and human emotion, focusing more on how the words spoken in human speech can give us information on their emotional state, mental health diagnosis, educational level, social class, etc. This includes the Linguistic Inquiry and Word Count (LIWC) method and the numerous research works that have applied them to empirical studies [44, 57].

In affective science, there has been numerous theories on how human emotions should be defined, categorized, and measured. Many of these views can complement or contradict the others. One popular school of thought is that human emotions can be categorized into discrete classes such as "Happy", "Sad", "Afraid", "Angry",

[1] Our dataset is available at: https://github.com/tmdang1101/EmotionalCanines

etc [8, 20, 43, 47]. Another popular school of thought is that human emotions can be categorized using dimensions of “Arousal” and “Valence” [7, 52]. These two view points do not necessarily contradict each other; in fact, discrete emotion classes can be placed on a two-dimensional graph of arousal and valence scales (i.e. “Happy” can have Medium Arousal and Positive Valence, while “Angry” can have High Arousal and Negative Valence).

Many of these ideas have carried over into animal emotion research as well [11, 38, 39]. The EmoDog dataset went with the discrete emotion view and used classes such as (Aggression, Despair, Fear, Fun, Happiness). We decided to adopt the two-dimensional arousal-valence space view as described by Mendl et al. [38] to construct our dog emotion dataset. We aim to avoid the anthropomorphic bias by not annotating dog emotional states using discrete classes, but by using the dog's behavior and physiological signals to derive arousal and valence (more details to be discussed in **Section** 3.1). We believe by adopting the two-dimensional arousal-valence framework, we can represent a larger range of nuanced situations and contexts, which are often absent if only a few selected, controlled environments are conducted. This will allow us to scale our dataset with the amount of data that we scraped from the Internet, where the generalizability of this framework will help with the limitless amount of contexts in dog videos.

3 Dataset

3.1 Dataset Generation

Our starting dataset was created from the process described in previous works [30, 32, 59, 60], which resulted in 306,233 dog bark sequences of 6 breeds (Chihuahua, German Shepherd, Husky, Labrador, Pitbull, and Shiba Inu), ranging from 0.5 to 5 seconds in length, totalling 152 hours of pure dog vocalizations. From this dataset, we randomly selected a smaller subset of bark sequences to be annotated for arousal and valence labels. Each sequence was annotated by three people and decided by majority agreement to add validity to the labels.

The annotators were 12 people which included 1 university professor, 5 graduate students, and 6 undergraduate students. Each annotator is instructed to watch and listen to the video clip that corresponds with the bark sequence (padded by 3 seconds at the beginning and end to include more context), making sure to match the barks to the right dog in the clip, and assign one label for each of the arousal and valence attributes based on the behavior that the dog display while vocalizing the barks. For example, if the video shows a dog playing with its owner, moving energetically with a lot of fast, rapid barking, then an annotator would label this sample “High” arousal and “Positive” valence.

Annotators have an option to mark a video as “Invalid” if they see specific signs such as “there is no dog in the video”, “the dog that barked is not in the video”, “multiple dogs are barking and overlapping”, etc. A list of scenarios for invalid videos is provided to the annotators. If two out of three people mark a video as “Invalid”, we would exclude the matching bark sequence from the dataset.

We understand there will be concerns about the validity of human annotation of dog emotional states by watching a video clip of the dog, which is a topic that we take seriously and will address here. We will first discuss some limitations.

It is virtually impossible to directly measure subjective emotional state in animals purely by facial or body observations (which is also the case in humans). There has been some research to test whether humans can discriminate between an animal's emotional states based on visual and vocal cues, with inconclusive or unsatisfactory accuracies from human testers [4, 22, 24, 48, 49, 54, 58]. However, by using behavioral, physiological, neural, and cognitive changes, it is possible to measure objective emotional state [9, 10, 14, 23, 39].

Our method to prevent subjectivity in annotating emotional state is twofold: (1) we specifically instruct the annotators to pay attention to objective signs in the video including the context, the stimuli, and the dog's behavior in reaction to the stimuli, and (2) we only add an annotated sample into our dataset only if there were majority agreement among the three annotators for that sample. Annotators were given a list of objective behavioral signals to look for (details in **Table** 1). These behavioral signals came from the literature on dog behavior research, most notably the works of Miklósi [40]. Annotators were also given an arousal-valence space graph (inspired by the works of Russell [52] and Mendl et al. [38]) to give them complementary guidance (see **Figure** 1).

To evaluate inter-rater reliability before the main data annotation process, we conducted a preliminary study in which 100 samples were annotated. We calculated Fleiss' Kappa scores [25] to assess agreement between the three annotators: Arousal (0.30) and Valence (0.42). Fleiss' Kappa ranges from -1 to 1, where values of 0.30 and 0.42 indicate fair agreement. As a point of reference, in the creation of the MSP-PODCAST human emotion speech corpus [34], they reported agreement scores Arousal (0.426) and Valence (0.459). This shows evidence that our annotation for arousal and valence in dogs are comparable to that in humans. Majority agreement was achieved for 91% of samples for Arousal and 94% for Valence individually, and 87% for both in combination. Most disagreements arose when the annotators disagree whether the audio clip is valid. In the main data annotation process, we only include samples that achieve majority agreement for both arousal and valence. By including instructions for clearly-defined, objective behavioral signals and strict majority agreement, we ensure the validity of our dog emotional state annotation.

Table 1: Characteristics of Arousal and Valence States.

Attribute	**Characteristics**
Arousal States	
High Arousal	Engaged, loud/long/rapid vocalization, fast and energetic movements, high intensity, tense face and body, tail wagging
Medium Arousal	Moderate amount of movement and energy
Low Arousal	Nonchalant, low/short/slow vocalization, few/slow/low-energy movements, low intensity, sleepy
Valence States	
Positive Valence	Playful, friendly, social
Neutral Valence	No bias toward positive nor negative
Negative Valence	Aggression, bared teeth, growling, snapping, biting, tense face and body

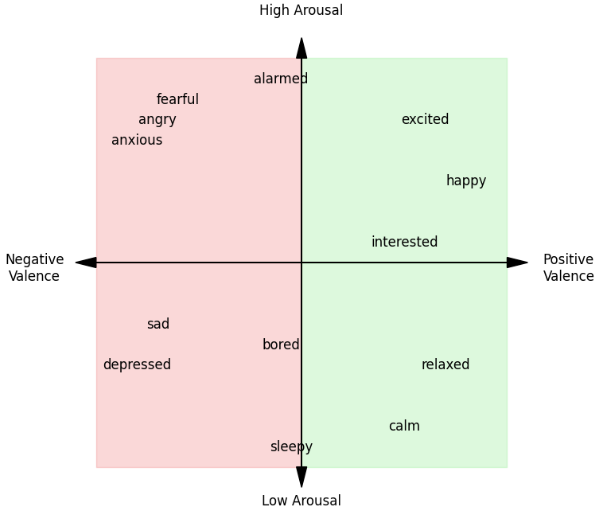


Figure 1: Two-dimensional arousal-valence space.

3.2 Dataset Description

The Siberian Husky is a medium-sized working breed developed by the Chukchi people of northeastern Siberia, Russia, over thousands of years [3]. They bred these dogs for pulling sleds, transporting goods, and hunting assistance in extreme Arctic conditions. Huskies are not excessive barkers. They are more vocal through howling, whining, or “talking” (unique vocalizations that sound like attempts to mimic human speech). They may bark to alert or when excited, but howling is their primary form of communication, often triggered by environmental stimuli or to express themselves.

The Shiba Inu is a small-to-medium Japanese breed, bred thousands of years ago for hunting small game in mountainous regions [2]. Shiba Inus are now popular worldwide, especially in urban and suburban settings. Shiba Inus are not frequent barkers. They are generally quiet but may bark to alert owners of strangers or unusual activity. They are known for their “screams”, a dramatic, high-pitched vocalization used when excited, stressed, or unhappy.

We chose these two breeds due to their drastic differences in barking habits and vocal characteristics, which could facilitate our research by showcasing the effects of these differences. Both breeds also have a long history of being domesticated by humans. There has been some research to provide evidence that the process of domestication has improved the capability for cooperation and communication in dog-human relationships [51, 53].

Our dataset has 700 bark sequences for each breed, totalling 1400 bark sequences. We split the data in a train and a test set, with 600 and 100 bark sequences respectively, making sure the two sets do not contain the same dogs. We perform experiments separately between breeds, since it is possible to compromise experimental validity to train on data from two breeds considering their different characteristics. The train set is imbalanced, while the test set is balanced. See **Tables** 3 and 4 for details.

The distribution of the three labels for each attribute is fairly balanced (**Table** 3). The imbalance in the train set can be seen when we look at the distribution of the combination between the two attributes (**Table** 4). These distributions directly reflect each breed’s characteristics. For example, when huskies are in a positive emotional state, they tend to exhibit a medium to high level of arousal. They also tend to be a positive dog breed in general, which might explain the lack of negative samples and the prevalent of positive samples. The discrepancy between low and medium to high arousal samples could be due to the husky breed having a high level energy and activity. As for Shiba Inus, they tend to exhibit high levels of arousal when they are either in a positive or negative emotional state, and low arousal when they are in a neutral state, due to the fact that they are generally a calm and quiet breed.

Lastly, we also want to mention that there is inherently a level of imbalance with video data scraped from the Internet. For example, YouTubers are more likely to post a dog video if the dog shows some interesting and fun behaviors, leading to higher engagement for the YouTube channel. This could potentially lead to videos with either high positive or high negative emotional states being more prevalent than others. This is a limitation to keep in mind in regards to our data generation method.

Table 2: Dataset Statistics (Duration in mm:ss).

Breed	Split	Dogs	Num. Bark Seqs	Duration
Husky	Train	44	600	15:30
	Test	18	100	2:44
Shiba Inu	Train	101	600	14:23
	Test	36	100	2:25
Overall		199	1400	35:02

Table 3: Data Distribution for Attributes Separately.

Attributes	Labels	Husky	Shiba Inu
Arousal	Low	187 (26.71%)	258 (36.86%)
	Medium	296 (42.29%)	204 (29.14%)
	High	217 (31.00%)	238 (34.00%)
Total		700	700
Valence	Negative	171 (24.43%)	192 (27.43%)
	Neutral	222 (31.71%)	252 (36.00%)
	Positive	307 (43.86%)	256 (36.57%)
Total		700	700

3.3 Potential Applications

This dataset, comprising dog bark audio clips with corresponding arousal and valence labels, offers significant potential for advancing research in animal communication. It can be used to investigate the structure of canine vocalizations, enabling researchers to decode how arousal and valence—key dimensions of emotional expression—manifest in specific bark patterns. Additionally, this dataset will allow researchers to explore parallels between dog barks and vocalizations among different dog breeds, to identify universal or breed-specific emotional cues.

Table 4: Data Distribution for Attributes in Combination.

Arousal	Valence	Husky	Shiba Inu
Low	Negative	55 (7.86%)	62 (8.86%)
Low	Neutral	76 (10.86%)	139 (19.86%)
Low	Positive	56 (8.00%)	57 (8.14%)
Medium	Negative	68 (9.71%)	60 (8.57%)
Medium	Neutral	98 (14.00%)	69 (9.86%)
Medium	Positive	130 (18.57%)	75 (10.71%)
High	Negative	48 (6.86%)	70 (10.00%)
High	Neutral	48 (6.86%)	44 (6.29%)
High	Positive	121 (17.29%)	124 (17.71%)

The framework used to create this dataset holds promising applications in advancing the study and interpretation of canine emotions. It could be extended to other species, facilitating comparative studies of animal communication and emotional expression across taxa, thus enriching ethological research.

Lastly, understanding dog and animal emotions through vocalizations is crucial for advancing animal welfare, as it provides a window into their internal states and needs. Vocalizations such as whines and growls often convey emotions like fear or distress, enabling caregivers, veterinarians, and shelter staff to identify signs of stress, pain, or discomfort that might otherwise go unnoticed.

4 Baselines

In this section, we will explore the details of input features and machine learning models used for the classification task. The input consists of an audio clip featuring a sequence of dog barks, and the output is one of the arousal and valence labels, assigned separately.

4.1 Features

In our classification task, we utilize a combination of acoustic features to capture diverse aspects of the bark sequence input for predicting arousal and valence labels. Feature vectors include Mel-Frequency Cepstral Coefficients (MFCCs) [17], the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [21], and advanced representation learning feature vectors from HuBERT [29] and Whisper [50]. Complementing these, raw audio waveforms and spectrograms provide time-domain and time-frequency representations, respectively, to capture the temporal and spectral dynamics of dog barks. An example of a husky bark spectrogram is shown in **Figure** 2. The example includes complementary details such as the scales, axis, and titles, which are removed during training, leaving only the spectrogram itself.

4.2 Models

We employ a diverse set of machine learning models to leverage the feature vectors (MFCCs, eGeMAPS, HuBERT, and Whisper), audio waveforms, and spectrogram images. For the feature vectors, we utilize Logistic Regression (LR), Support Vector Machines (SVM), Random Forests (RF), and XGBoost (XGB) as classifiers.

To process the raw audio and spectrogram inputs, we fine-tune pre-trained deep learning models. The Wav2Vec 2.0 model, originally pre-trained on human speech, is fine-tuned on our dog bark audio clips [5]. For spectrograms, we fine-tune ResNet18, a convolutional neural network (CNN) [28], and a Vision Transformer (ViT) [19], both are pre-trained on ImageNet [18].

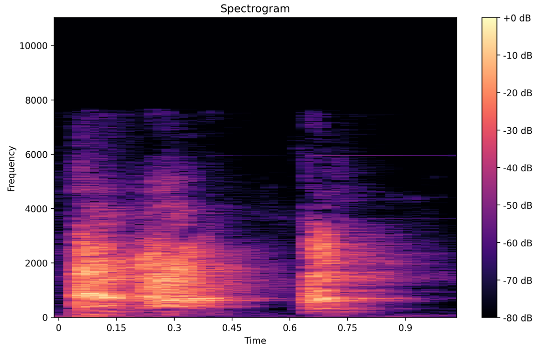


Figure 2: Example spectrogram of a husky bark sequence.

To enhance the robustness and generalization of our models, we implement a comprehensive training strategy. For audio inputs, we apply data augmentation techniques using the Librosa library [37], including adding noise, time stretching, and pitch shifting, thereby increasing the diversity of the training set. For spectrogram inputs, we employ augmentation methods such as flipping (horizontal or vertical), rotation, color jitter, and random resizing and cropping. To address class imbalance across all input types (feature vectors, audio, and spectrograms), we use the Synthetic Minority Oversampling Technique (SMOTE) [16] to generate synthetic samples for underrepresented classes.

We use 10-fold cross-validation to assess model performance, reporting mean and standard deviation of validation accuracy scores. Early stopping is applied to halt training when validation performance plateaus, preventing overfitting. ResNet18 and ViT models are trained using cross-entropy loss. All training was conducted on two NVIDIA GeForce RTX 4090 GPUs.

4.3 Results

We use two standard metrics: accuracy to measure overall classification performance, and confusion matrices to analyze per-class performance and identify misclassification patterns.

For the Husky breed, the eGeMAPS feature set demonstrated superior results for both attributes, outperforming other approaches. Specifically, eGeMAPS achieved the highest accuracies for arousal classification with LR and RF models, both reaching 53%, while for valence, eGeMAPS with RF and XGB yielded 50% and 51%, respectively. The stronger performance of eGeMAPS likely stems from its design to capture emotionally relevant acoustic features in humans, which seemed to transferred well to the expressive nature of dog barks. Arousal classification generally outperformed valence, possibly due to clearer acoustic cues like intensity being more distinguishable than the subtler, context-dependent patterns of valence.

For the Shiba Inu breed, both eGeMAPS and Whisper feature sets excelled, with MFCC showing comparable results, while spectrograms lagged slightly behind. The best arousal classification

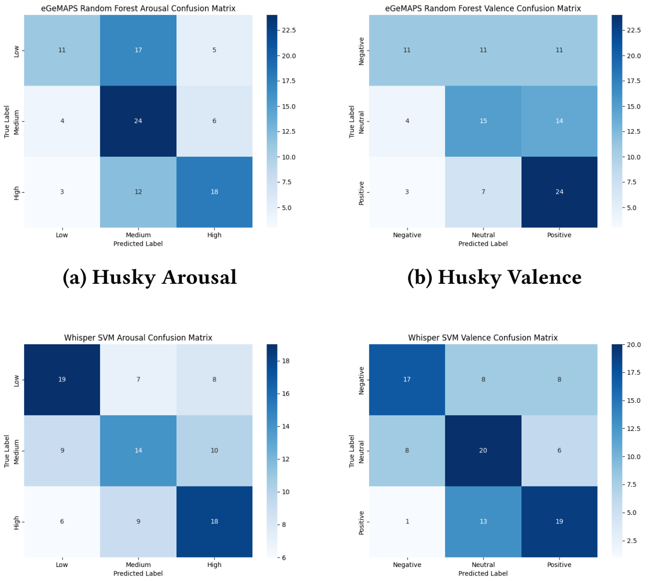
Table 5: Classification Results for Husky / Shiba.

Input	Model	Arousal	Valence
MFCC	LR	34% / 51%	38% / 38%
MFCC	SVM	49% / 46%	42% / 45%
MFCC	RF	45% / 52%	47% / 45%
MFCC	XGB	42% / 57%	42% / 43%
eGeMAPS	LR	53% / 53%	42% / 41%
eGeMAPS	SVM	51% / 51%	42% / 47%
eGeMAPS	RF	53% / 50%	50% / 49%
eGeMAPS	XGB	52% / 51%	51% / 42%
HuBERT	LR	35% / 39%	37% / 36%
HuBERT	SVM	38% / 49%	41% / 44%
HuBERT	RF	35% / 42%	40% / 50%
HuBERT	XGB	45% / 43%	42% / 45%
Whisper	LR	43% / 41%	41% / 41%
Whisper	SVM	42% / 56%	36% / 56%
Whisper	RF	39% / 47%	41% / 47%
Whisper	XGB	41% / 49%	38% / 49%
Audio	Wav2Vec 2.0	49% / 55%	43% / 51%
Spectrogram	CNN	42% / 46%	34% / 45%
Spectrogram	ViT	41% / 49%	44% / 36%

accuracies were obtained with MFCC using XGB at 57% and Whisper with SVM at 56%, whereas Whisper with SVM topped valence classification at 56%. The weakest performances were observed with HuBERT and LR for arousal (39%) and both HuBERT with LR and ViT on spectrograms for valence (36%). The success of eGeMAPS and Whisper may be attributed to their robustness in extracting nuanced acoustic features from Shiba Inu barks, which are often sharper and more varied in pitch. Arousal again showed stronger results than valence.

Additionally, we conducted some experiments with ten-fold cross-validation to provide some insights in the training process. Results for using eGeMAPS and Whisper feature sets are included below, showing notably higher accuracies compared to the test set results, likely due to the absence of individual dog overlap between training and test sets. For eGeMAPS, the mean accuracies and standard deviations across models were: LR (0.5416 ± 0.0332), SVM (0.6552 ± 0.0338), RF (0.7905 ± 0.0223), and XGB (0.7854 ± 0.0312). For Whisper, the results were: LR (0.6298 ± 0.0235), SVM (0.6391 ± 0.0315), RF (0.7231 ± 0.0231), and XGB (0.7307 ± 0.0248). These elevated cross-validation accuracies suggest that models benefit from training and testing on similar distributions of data, but the drop in test accuracy highlights the importance of separating individual dogs between the two sets to maintain experimental validity.

The best models overall were RF trained on eGeMAPS for Huskies, and SVM trained on Whisper vectors for Shiba Inus. The confusion matrices of these models are shown in **Figure** 3 as an example. We can see that the majority of misclassifications occur in the middle, intermediate states: "Medium" label in the arousal dimension, and "Neutral" label in the valence dimension, likely due to the nuanced nature of the audio data. The "Medium" arousal category, representing moderate intensity, may lack distinct acoustic signatures making it harder for the models to distinguish overlapping features between low to medium, and medium to high, since these neighboring classes would inherently have similar features. Similarly, the "Neutral" valence label, which could be lacking the clear emotional cues of "Positive" or "Negative", likely exhibits subtler spectral and temporal characteristics. This ambiguity can confuse models, as the acoustic differences may be too fine-grained, especially in diverse recording conditions or across individual dogs. These challenges highlight the need for more refined feature extraction methods to better capture these intermediate emotional states.



(a) Husky Arousal

(b) Husky Valence

(c) Shiba Inu Arousal

(d) Shiba Inu Valence

Figure 3: Analysis of Confusion Matrices.

5 Conclusion

We believe the development and analysis of this dog bark emotion dataset will prove valuable for research into their vocal communication patterns. By providing a framework for obtaining arousal and valence labels, this work will enable researchers to decode emotional cues in dog vocalization at scale and create a foundation for comparative studies with other species.

Through our baseline results, we believe these features and methods alone are not enough to interpret emotional state in dog vocalization. We hypothesize that a tokenization method is necessary to create a dictionary of vocal units or tokens that will be a more reliable indicator of emotional state or other elements of semanticity in dog vocalization. Expanding the dataset to include a wider variety of dog breeds and a larger amount of dog barks per breed would provide more resource for analysis of bark patterns. Future works could explore multimodal signals (e.g., body language) to provide more behavioral contexts.

Acknowledgments

This work was partially supported by NSF Award No. 2349713. Our gratitude goes to Mengyue Wu, Wonjun Park, and Essam Abdelghany for their kind help and useful suggestions.

References

[1] Artem Abzaliev, Humberto Pérez Espinosa, and Rada Mihalcea. 2024. Towards Dog Bark Decoding: Leveraging Human Speech Processing for Automated Bark Classification. doi:10.48550/ARXIV.2404.18739

[2] American Kennel Club. 2025. Shiba Inu. https://www.akc.org/dog-breeds/shiba-inu/ Accessed: 2025-05-24.

[3] American Kennel Club. 2025. Siberian Husky. https://www.akc.org/dog-breeds/siberian-husky/ Accessed: 2025-05-24.

[4] Federica Amici, James Waterman, Christina Maria Kellermann, Karimullah Karimullah, and Juliane Bräuer. 2019. The ability to recognize dog emotions depends on the cultural milieu in which we grow up. *Scientific reports* 9, 1 (2019), 16414.

[5] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems* 33 (2020), 12449–12460.

[6] Tanja Bänziger, Marcello Mortillaro, and Klaus R Scherer. 2012. Introducing the Geneva Multimodal expression corpus for experimental research on emotion perception. *Emotion* 12, 5 (2012), 1161.

[7] Lisa Feldman Barrett. 2006. Solving the emotion paradox: Categorization and the experience of emotion. *Personality and social psychology review* 10, 1 (2006), 20–46.

[8] Paul H Barrett. 2016. *The Works of Charles Darwin: Vol 23: The Expression of the Emotions in Man and Animals.* Routledge.

[9] Tali Boneh-Shitrit, Shir Amir, Annika Bremhorst, Daniel S Mills, Stefanie Riemer, Dror Fried, and Anna Zamansky. 2022. Deep learning models for automated classification of dog emotional states from facial expressions. *arXiv preprint arXiv:2206.05619* (2022).

[10] Tali Boneh-Shitrit, Marcelo Feighelstein, Annika Bremhorst, Shir Amir, Tomer Distelfeld, Yaniv Dassa, Sharon Yaroshetsky, Stefanie Riemer, Ilan Shimshoni, Daniel S Mills, et al. 2022. Explainable automated recognition of emotional states from canine facial expressions: the case of positive anticipation and frustration. *Scientific reports* 12, 1 (2022), 22611.

[11] Elodie F Briefer. 2012. Vocal expression of emotions in mammals: mechanisms of production and evidence. *Journal of Zoology* 288, 1 (2012), 1–20.

[12] Elodie F Briefer, Anne-Laure Maigrot, Roi Mandel, Sabrina Briefer Freymond, Iris Bachmann, and Edna Hillmann. 2015. Segregation of information about emotional arousal and valence in horse whinnies. *Scientific reports* 5, 1 (2015), 9989.

[13] Elodie F Briefer, Federico Tettamanti, and Alan G McElligott. 2015. Emotions in goats: mapping physiological, behavioural and vocal profiles. *Animal Behaviour* 99 (2015), 131–143.

[14] Sofia Broomé, Marcelo Feighelstein, Anna Zamansky, Gabriel Carreira Lencioni, Pia Haubro Andersen, Francisca Pessanha, Marwa Mahmoud, Hedvig Kjellström, and Albert Ali Salah. 2023. Going deeper than tracking: A survey of computer-vision based recognition of animal pain and emotions. *International Journal of Computer Vision* 131, 2 (2023), 572–590.

[15] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42 (2008), 335–359.

[16] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[17] Steven Davis and Paul Mermelstein. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing* 28, 4 (1980), 357–366.

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition.* Ieee, 248–255.

[19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[20] Paul Ekman. 1994. The nature of emotion: Fundamental questions.

[21] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. 2015. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing* 7, 2 (2015), 190–202.

[22] Tamás Faragó, N Takács, Ádam Miklósi, and Peter Pongrácz. 2017. Dog growls express various contextual and affective content for human listeners. *Royal Society open science* 4, 5 (2017), 170134.

[23] Marcelo Feighelstein, Ilan Shimshoni, Lauren R Finka, Stelio PL Luna, Daniel S Mills, and Anna Zamansky. 2022. Automated recognition of pain in cats. *Scientific Reports* 12, 1 (2022), 9575.

[24] Piera Filippi, Jenna V Congdon, John Hoang, Daniel L Bowling, Stephan A Reber, Andrius Pašukonis, Marisa Hoeschele, Sebastian Ocklenburg, Bart De Boer, Christopher B Sturdy, et al. 2017. Humans recognize emotional arousal in vocalizations across all classes of terrestrial vertebrates: Evidence for acoustic universals. *Proceedings of the Royal Society B: Biological Sciences* 284, 1859 (2017), 20170990.

[25] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378.

[26] José Ramón Gómez-Armenta, Humberto Pérez-Espinosa, José Alberto Fernández-Zepeda, and Verónica Reyes-Meza. 2024. Automatic classification of dog barking using deep learning. *Behavioural Processes* 218 (2024), 105028.

[27] 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).* IEEE, 5134–5138.

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 770–778.

[29] 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 (2021), 3451–3460.

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

[31] Ana Larrañaga, Concha Bielza, Péter Pongrácz, Tamás Faragó, Anna Bálint, and Pedro Larrañaga. 2014. Comparing supervised learning methods for classifying sex, age, context and individual Mudi dogs from barking. *Animal Cognition* 18, 2 (Oct. 2014), 405–421. doi:10.1007/s10071-014-0811-7

[32] Xingyuan Li, Kenny Zhu, and Mengyue Wu. 2025. Dog2vec: Self-Supervised Pre-Training for Canine Vocal Representation. In *Proceedings of the 26th Interspeech Conference.*

[33] Pavel Linhart, Victoria F Ratcliffe, David Reby, and Marek Špinka. 2015. Expression of emotional arousal in two different piglet call types. *PloS one* 10, 8 (2015), e0135414.

[34] Reza Lotfian and Carlos Busso. 2017. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing* 10, 4 (2017), 471–483.

[35] Anne-Laure Maigrot, Edna Hillmann, Callista Anne, and Elodie F Briefer. 2017. Vocal expression of emotional valence in Przewalski's horses (Equus przewalskii). *Scientific reports* 7, 1 (2017), 8779.

[36] Anne-Laure Maigrot, Edna Hillmann, and Elodie F Briefer. 2018. Encoding of emotional valence in wild boar (Sus scrofa) calls. *Animals* 8, 6 (2018), 85.

[37] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. *SciPy* 2015 (2015), 18–24.

[38] Michael Mendl, Oliver HP Burman, and Elizabeth S Paul. 2010. An integrative and functional framework for the study of animal emotion and mood. *Proceedings of the Royal Society B: Biological Sciences* 277, 1696 (2010), 2895–2904.

[39] Michael Mendl, Vikki Neville, and Elizabeth S Paul. 2022. Bridging the gap: Human emotions and animal emotions. *Affective Science* 3, 4 (2022), 703–712.

[40] Ádám Miklósi. 2014. *Dog behaviour, evolution, and cognition.* oUp Oxford.

[41] Csaba Molnár, Péter Pongrácz, Antal Dóka, and Ádám Miklósi. 2006. Can humans discriminate between dogs on the base of the acoustic parameters of barks? *Behavioural processes* 73, 1 (2006), 76–83.

[42] Csaba Molnár, Frédéric Kaplan, Pierre Roy, François Pachet, Péter Pongrácz, Antal Dóka, and Ádám Miklósi. 2008. Classification of dog barks: a machine learning approach. *Animal Cognition* 11, 3 (Jan. 2008), 389–400. doi:10.1007/s10071-007-0129-9

[43] Jaak Panksepp and Douglas Watt. 2011. What is basic about basic emotions? Lasting lessons from affective neuroscience. *Emotion review* 3, 4 (2011), 387–396.

[44] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of LIWC2015. (2015).

[45] Humberto Pérez-Espinosa, José Martın Pérez-Martınez, José Ángel Durán-Reynoso, and Verónica Reyes-Meza. 2015. Automatic classification of context in induced barking. *Research in Computing Science* 100 (2015), 63–74.

[46] Humberto Pérez-Espinosa, Verónica Reyes-Meza, Emanuel Aguilar-Benitez, and Yuvila M Sanzón-Rosas. 2018. Automatic individual dog recognition based on the acoustic properties of its barks. *Journal of Intelligent & Fuzzy Systems* 34, 5 (2018), 3273–3280.

[47] Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In *Theories of emotion*. Elsevier, 3–33.

[48] Péter Pongrácz, Csaba Molnár, Antal Dóka, and Ádám Miklósi. 2011. Do children understand man's best friend? Classification of dog barks by pre-adolescents and adults. *Applied animal behaviour science* 135, 1-2 (2011), 95–102.

[49] Péter Pongrácz, Csaba Molnár, Adám Miklósi, and Vilmos Csányi. 2005. Human listeners are able to classify dog (Canis familiaris) barks recorded in different situations. *Journal of comparative psychology* 119, 2 (2005), 136.

[50] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*. PMLR, 28492–28518.

[51] Friederike Range and Zsófia Virányi. 2015. Tracking the evolutionary origins of dog-human cooperation: the "Canine Cooperation Hypothesis". *Frontiers in psychology* 5 (2015), 1582.

[52] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39, 6 (1980), 1161.

[53] Hannah Salomons, Kyle CM Smith, Megan Callahan-Beckel, Margaret Callahan, Kerinne Levy, Brenda S Kennedy, Emily E Bray, Gitanjali E Gnanadesikan, Daniel J Horschler, Margaret Gruen, et al. 2021. Cooperative communication with humans evolved to emerge early in domestic dogs. *Current Biology* 31, 14 (2021), 3137–3144.

[54] Marina Scheumann, Anna S Hasting, Sonja A Kotz, and Elke Zimmermann. 2014. The voice of emotion across species: how do human listeners recognize animals' affective states? *PLoS One* 9, 3 (2014), e91192.

[55] Stefan Steidl. 2009. *Automatic classification of emotion-related user states in spontaneous children's speech*. Logos-Verlag Berlin, Germany.

[56] Céline Tallet, Pavel Linhart, Richard Policht, Kurt Hammerschmidt, Petr Šimeček, Petra Kratinova, and Marek Špinka. 2013. Encoding of situations in the vocal repertoire of piglets (Sus scrofa): a comparison of discrete and graded classifications. *PloS one* 8, 8 (2013), e71841.

[57] Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology* 29, 1 (2010), 24–54.

[58] Anna M Taylor, David Reby, and Karen McComb. 2009. Context-related variation in the vocal growling behaviour of the domestic dog (Canis familiaris). *Ethology* 115, 10 (2009), 905–915.

[59] Theron Wang, Xingyuan Li, Chunhao Zhang, Mengyue Wu, and Kenny Zhu. 2024. Phonetic and Lexical Discovery of Canine Vocalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 13972–13983.

[60] Theron S. Wang, Xingyuan Li, Hridayesh Lekhak, Tuan Minh Dang, Mengyue Wu, and Kenny Zhu. 2025. Toward Automatic Discovery of a Canine Phonetic Alphabet. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 9207–9219.