DogSpeak: A Canine Vocalization Classification Dataset

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

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

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

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

Abstract

Progress in understanding real-world canine vocal communication is constrained by datasets lacking scale and 'in-the-wild' diversity. We introduce DogSpeak, a large-scale public dataset of 77,202 Barkseqs (33.162 hours) from 156 dogs (5 breeds), uniquely sourced from online social media with accurate dog ID, sex, and breed labels. DogSpeak, one of the largest of its kind, addresses prior limitations. Benchmark tasks (sex, breed, individual dog recognition) demonstrate its utility and highlight how its inherent real-world challenges necessitate and foster research into more robust bioacoustic models, preprocessing, and feature representation.

CCS Concepts

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

Keywords

Canine Vocalization; Dog Bark Analysis; Animal Communication; Vocalization Dataset; Machine Learning; Computational Bioacoustics; Bioacoustics Classification

ACM Reference Format:
Hridayesh Lekhak, Theron S. Wang, Tuan M. Dang, and Kenny Q. Zhu. 2025. DogSpeak: A Canine Vocalization Classification Dataset. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3746027.3758298

1 Introduction

Domesticated dogs have evolved alongside humans for millennia, developing sophisticated vocal communication patterns to interact among themselves and with humans [26, 30]. Understanding the nuances of these vocalizations is crucial not only for deciphering canine interactions but also for gaining broader insights into animal communication, cognition, and the evolution of complex signaling systems. Like humans who exhibit early vocal acoustic development [18], and other species such as sperm whales showing 'babbling-like' phases [10] or wolves demonstrating maturation in call features [9], dogs' vocal repertoires may hold significant clues about their development, internal states, and adaptive strategies.

Recent research has demonstrated that dog vocalizations possess complex structural patterns amenable to digital representation and clustering into potential phonetic units [33]. Acoustic features derived from these vocalizations have enabled machine learning models to achieve some success in classifying canine sex [11, 20], breed [11], individual identity [11, 20, 25, 36], and the context of vocalizations [11, 20, 25, 31, 36]. Further work has focused on the discovery of a canine phonetic alphabet and lexical structures from vocalization data [34]. Similar successes in leveraging vocal features for classification are evident in studies of other animals, including cats [19, 28, 32], mice [17], and birds [2, 22]. These advancements suggest that decoding vocal signals is a promising avenue for a deeper understanding of animal communication.

Despite these promising results, progress is often hampered by the limitations of available datasets. Existing resources, such as the Mudi dog dataset [20] (800 barks, 8 dogs) and the Mescalina 2015/2017 datasets (with Pérez-Espinosa et al. [29] detailing a specific version of 6,103 barks from 36 dogs, alongside broader project estimates of approx. 6000-7000 barks from 37-65 dogs), while valuable, are typically recorded in controlled environments. This can restrict the diversity of contexts and vocal expressions captured. Furthermore, their limited scale may hinder the development and generalization of more complex machine learning and deep learning models, which often require vast amounts of data to learn robust representations.

The unique challenge of analyzing data from real-world, uncontrolled settings remains largely unaddressed. To address these gaps, we introduce DogSpeak, a novel, large-scale canine vocalization dataset. The DogSpeak dataset is uniquely sourced from tens of thousands of dog videos on online social media platforms like YouTube and TikTok. This approach allows us to capture vocalizations from a wide array of natural, organic interactions and environments, far exceeding the situational diversity of lab-collected data. While this online data provides unprecedented scale and contextual richness, it also introduces a significant challenge: the presence of noise and variability inherent in 'in-the-wild' recordings. We have developed a machine learning pipeline to automate data cleaning, annotation, and audio extraction from these varied sources.

DogSpeak, with its extensive collection of vocalizations, enables the study of automatic classification of dog vocalizations at a more comprehensive scale than previously possible. The inherent noisiness and diversity of the data, while challenging, also present a

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3758298

unique opportunity to develop more robust and generalizable models capable of handling real-world acoustic scenes. To facilitate this, we provide not only the dataset but also benchmark results for three key classification tasks: identifying canine sex, breed, and individual identity using only vocal cues.

Our contributions are threefold:

- We present DogSpeak [1], the largest and most comprehensive dog vocalization dataset to date, comprising over 77,202 bark sequences (a.k.a. *Barkseqs*) from 156 dogs of 5 breeds, totaling 33.162 hours of pure dog barks (see Sec. 2 and Sec. 3).
- We propose three benchmark classification tasks on this dataset: identifying the *sex*, *breed*, and *individual dog* from a given barkseq (see Sec. 4).
- Our preliminary experiments using both traditional and advanced acoustic methods indicate that current pure acoustic representations are insufficient to master these tasks satisfactorily. This underscores the complexity of the challenge presented by our dataset and opens avenues for research into more sophisticated structural, prosodic, and potentially linguistic features of dog vocalizations (see Sec. 4).

The DogSpeak dataset is poised to significantly advance the scientific understanding of animal communication. By providing a large and challenging benchmark, it aims to inspire research into robust preprocessing techniques, novel data representation learning, and advanced modeling architectures tailored for complex bioacoustic data. Insights derived from exploring DogSpeak could have implications for understanding canine cognition and behavior [14, 27, 35], potentially revealing how computational methods can uncover the communicative intricacies in dogs and, by extension, other animal species. Ultimately, this work serves as a catalyst for deeper exploration into the rich world of canine vocal communication.

2 Dataset Creation

Although previous efforts were made to collect dog vocalization data from a controlled environment, the scale of the data is usually limited and the content may be biased due to the limited scenarios researchers can simulate. Videos from online social media, on the other hand, offer a larger and more diverse pool of data, capturing dog behaviors and activities in many more contexts. We believe these videos represent more authentic and diverse communication patterns. Next, we describe the steps to create DogSpeak dataset.

2.1 Seed Videos Collection

By using five different breed names, i.e., *Husky*, *Chihuahua*, *German Shepherd*, *Pitbull*, and *Shiba Inu*, as search queries, we were able to get a long list of raw video clips about these breeds. Not all of these videos are valid. For example, the query "Shiba Inu" returns many promotional videos about the Shiba coin cryptocurrency (see Figure 1a). To filter out invalid videos, we trained a binary classifier using both BERT and ViT. The ViT model takes the thumbnails of the videos as input, while the BERT model uses a combination of metadata of the videos such as the title, description, and comments as input. We label 1100 positive videos and 1100 negative videos, with the labels encompassing multiple breeds, including mostly the 5 breeds that are present in the dataset, and train the BERT model and ViT model separately. Our BERT model achieves an accuracy of 95.4%, while the ViT model achieves 92.6%. When the two models agree that a video clip is valid then the video is valid. By using the above method, we are able to get a clean list of URLs to seed videos without downloading all the raw videos. An example of a relevant, correctly classified video depicting a Shiba Inu dog is shown in Figure 1b.

[1]Our dataset is available at: https://huggingface.co/datasets/ArlingtonCL2/DogSpeak_Dataset

2.2 Dog ID, Sex, and Breed Annotation

From URLs of the seed videos, we are able to identify a list of YouTube or TikTok channels that feature the five breeds of dogs that we are interested in. We remove channels that feature multiple dogs or carry fewer than 50 videos. At this point, all the videos from a channel will be about one dog, so the Dog IDs are guaranteed. Next we determine the sex and breed by manually going through the channel description, video title and comments, and even external social media pages linked from the channel. If the owner of a YouTube or TikTok channel explicitly states the breed and gender of the dog in any of these sources, we use this as a reliable confirmation. A channel is removed if we cannot confidently verify the sex or breed of the subject. At this stage, we have narrowed down our selection to 156 channels. We then employ a web crawler to collect about 59,700 videos, totaling approximately 1,270 hours of footage from these channels. To this end, every video will contain an accurate dog ID, sex, and breed labels.

2.3 Dog Barks Preprocessing

The term dog *bark sequence* or *barkseq* refers to a continuous sequence of dog barks separated by long pauses (> 0.5 seconds) [16, 33]. Each barkseq, along with its corresponding ID, sex, and breed labels, constitutes a data sample in our dataset.

To extract these Barkseqs, we follow the approach of Wang et al. [33]. Raw audios are first denoised using AudioSep [23] with the prompt "Dogs". Subsequently, a sound event detection (SED) framework employing BEATs [4] as the audio encoder is utilized. This BEATs-SED model was fine-tuned on 9,000 seconds of manually labeled dog bark data for 5.5 hours on two Nvidia RTX 4090 GPUs with a batch size of 4, achieving an F1-score of 0.86 on a test set, and is then used to detect Barkseqs from the denoised audio. These detected Barkseqs form the input for the subsequent segmentation stage.

In the next section, we present the statistical characteristics of our dataset.

3 Dataset Statistics

Our DogSpeak dataset contains 77,202 Barkseqs from 156 distinct dogs. The exact distribution of the data on sex and breed is shown in Table 1. From these statistics we can see that Huskies and Shiba Inus form the two biggest breeds in this dataset, but even the smallest breed (Pitbull) accounts for 7.9% of the data. The dataset is also fairly balanced between the two sexes.

4 Benchmark Results

In this section, we show the results of applying acoustic methods to three benchmark tasks using the DogSpeak dataset, namely, *sex*

(a) Illustration of an irrelevant video example, as might be returned by YouTube's search algorithm for "Shiba Inu" (e.g., cryptocurrency content).

(b) Illustration of a relevant video example, depicting a Shiba Inu dog, which is the target content for our dataset.

Figure 1: Examples of video thumbnails illustrating (a) irrelevant content and (b) relevant content for dataset creation.

Table 1: Overall Statistics of DogSpeak dataset.

Breed	Dogs		Barkseqs		Dur. (h)	
	M	F	M	F	M	F
Chihuahua	14	11	4,789	3,366	1.725	0.933
Shiba Inu	22	18	6,806	12,612	2.684	4.771
Pitbull	13	15	1,681	4,429	0.721	1.873
Husky	11	20	25,392	10,576	11.819	4.965
German Shepherd	19	13	3,331	4,220	1.386	2.285
Total	**79**	**77**	**41,999**	**35,203**	**18.335**	**14.827**

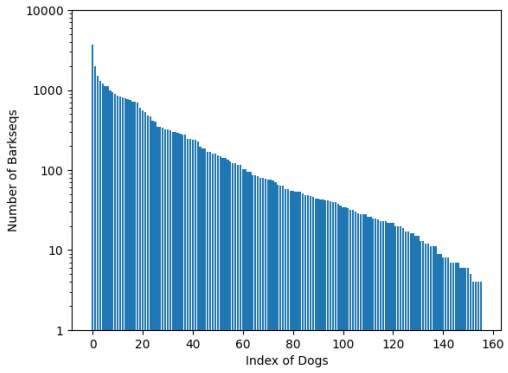

Figure 2: Distribution of Number of Barkseqs Over 156 Dogs (y-axis is in log scale).

classification, *breed classification*, and *individual dog recognition*. Before that, we first introduce the various acoustic features used in this evaluation.

4.1 Benchmark Preparation

For the tasks of sex and breed classification, the dataset Table 1 was balanced to ensure roughly equal male-to-female ratios and total Barkseqs. Dogs with fewer than 100 barks were placed in the test set. For these benchmark tasks, the goal is to take the acoustic features of a Barkseq as input and predict the corresponding dog's individual ID (n-way classification task for individual dog recognition, where n is the number of individual dogs taken from the dataset), sex (a 2-way classification task for sex classification), or breed (a 5-way classification task for breed classification).

Table 3 shows the final train-test split used for canine sex and breed classification.

Importantly, we ensured that the same dog was only included in either the train or test split, not both, to prevent data leakage. This step was taken to eliminate potential data overlap between the training and test sets, ensuring the classification task genuinely focused on breed and sex classification.

To provide context for our dataset within the broader landscape of canine vocalization research, Table 2 offers a comparison of its statistics with other key related datasets.

4.2 Features Used

We utilize a set of acoustic features, including Mel-Frequency Cepstral Coefficients (MFCCs), filterbanks (256-d), GeMAPS, eGeMAPS, and HuBERT embeddings. These features capture various aspects of the vocalizations, from spectral characteristics to high-level learned representations.

4.2.1 Mel-Frequency Cepstral Coefficients (MFCCs). MFCCs [6] capture spectral and cepstral characteristics, reflecting the timbre and frequency distribution of vocalizations. We extracted 40-dimensional MFCCs (base coefficients, deltas, and delta-deltas) per frame using Librosa [24]. This was done with a Fast Fourier Transform (FFT) window of 512 samples (equivalent to 32ms given a 16kHz sampling rate, assuming this from the 160 sample hop length for 10ms) and a hop length of 160 samples (10ms). For each barkseq, we then calculate the mean and the standard deviation for each of these 40 coefficients across all frames in the audio. This results in a 80-dimensional feature vector for each barkseq.

4.2.2 Filterbanks. Filterbanks [6] (256-d) capture the energy present in different frequency bands, providing a representation of the spectral shape of the audio signal.

4.2.3 GeMAPS and eGeMAPS Features. We utilized the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) and its extended version, the 88-dimensional eGeMAPS v02 functional feature set [7]. These were extracted for each Barkseq using the openSMILE toolkit [8]. The eGeMAPS set, which includes the original GeMAPS features, provides a standardized collection of prosodic, spectral, and voice quality features commonly used in paralinguistic analysis. These

Table 2: Comparison of DogSpeak with Other Key Dog Vocalization Related Datasets.

Feature	DogSpeak (Ours)	Mudi Dataset [20]	Mescalina [29]	Wang et al. [33]
Number of Dogs	**156**	8	36	>1,300 users (distinct dogs not specified)
Number of Vocalizations	**77,202** (Barkseqs)	~800 barks	6,103 barks	37,919 'sentences'
Number of Breeds	5 (Husky, Chihuahua, German Shepherd, Pitbull, Shiba Inu)	1 (Mudi)	Multiple (Chihuahua, French Poodle, Schnauzer, mixed)	Diverse (YouTube sourced), Primarily Shiba Inu and Husky
Total Duration	**33.162 hours**	Not explicitly stated	Not explicitly stated	>23 hours
Recording Conditions	Online (YouTube, TikTok), diverse	Controlled, specific communicative situations	Owners' homes, induced stimuli, consistent equipment	Online (YouTube), diverse
Key Annotations / Discoveries	Dog ID, sex, breed	Sex, age, context, Dog ID	Dog ID, (inferred) context, breed, sex	Discovered 'phones' (140 types), 'words' (phone seq.), activity correlations

Table 3: Balanced Train and Test Datasets for Sex and Breed Classification.

Breed	Dogs		Barkseq		Dur. (h)	
	Train	Test	Train	Test	Train	Test
Chihuahua	13	12	4427	385	1.45	0.10
Shiba Inu	22	18	8110	482	3.38	0.15
Pitbull	9	19	2075	761	0.91	0.35
Husky	19	12	8769	276	4.12	0.15
German Shepherd	11	21	3514	548	1.71	0.23
Total (M)	37	42	13328	1180	5.75	0.48
Total (F)	37	40	13567	1272	5.82	0.51

features have been previously employed in dog bark feature extraction [12].

4.2.4 HuBERT Features. Features were extracted using a HuBERT [15] base model pre-trained on dog vocalisations [33]. Embeddings from the 11th transformer layer yielded a sequence of 768-dimensional embeddings per Barking Unit (BU). Each of these embeddings corresponds to a 20ms audio segment.

To ensure interpretability, we use Logistic Regression [13] (LR), Random Forest [3] (RF), and XGBoost [5] (XGB) as classifiers to analyze the acoustic features and identify prominent acoustic properties for each task. Beyond classic machine learning models, we also fine-tune HuBERT using a linear neural network to enhance feature representation.

4.3 Sex Classification

Table 5 presents the performance of various features in classifying dog sex. The fact that all the F1 scores and accuracies are greater than 0.5 shows the acoustic features are useful in determining the sexes, but only marginally.

The performance metrics on our dataset also highlight its challenging nature, particularly when considered alongside benchmarks from datasets recorded in more controlled environments and with potentially different train-test methodologies. For instance, research on the Mescalina dataset [29] focused primarily on individual dog recognition, reporting a high F1-score of 90.50% using SVMs. While that study did not center on sex classification as a primary reported outcome with detailed benchmarks comparable to ours, other research leveraging Mescalina-derived data, such as Abzaliev et al. [1], reported 68.90% accuracy for gender identification with Wav2Vec2 on a related dataset.

Our results for sex classification (e.g., HuBERT F1-score of 0.562 as shown in Table 5) might appear lower when compared to some benchmarks from datasets recorded in more controlled settings. This difference can be attributed to several factors. Firstly, our dataset, sourced from diverse online social media, could be inherently noisier and more variable than data from controlled environments like the Mescalina dataset (recorded in owners' homes with induced stimuli and consistent equipment [29]). Secondly, our strict train-test split, ensuring no dog appears in both sets for sex and breed classification, provides a rigorous evaluation of generalization to unseen individuals. Methodologies where the same dog's vocalizations (even different samples) might be present in both train and test splits could lead to models learning individual-specific patterns rather than purely sex-indicative features, potentially impacting performance metrics. The complexities of our naturalistic dataset and stringent evaluation protocol highlight the need for advanced, robust models.

If we delve into the specific acoustic features of GeMAPS, we find that the most prominent features for XGB and RF, presented in Table 4, include both frequency-related features, such as `F0` and `F3`, and slope-based features, indicating that these features are particularly relevant for distinguishing between the sexes of dogs, as they may reflect anatomical differences, such as the size and shape of the vocal cords.

Table 4: Top 4 Gemaps features used by XGB and RF models respectively for dog sex classification.

Gemaps Features	Modal
`slopeV500-1500_sma3nz_amean`	XGB
`MeanVoicedSegmentLengthSec`	XGB
`F3amplitudeLogRelF0_sma3nz_amean`	XGB
`F0semitoneFrom27.5Hz_sma3nz_percentile80.0`	XGB
`slopeV500-1500_sma3nz_amean`	RF
`slopeUV500-1500_sma3nz_amean`	RF
`slopeV500-1500_sma3nz_stddevNorm`	RF
`F0semitoneFrom27.5Hz_sma3nz_percentile80.0`	RF

The 768-dimensional HuBERT feature suggests that it encapsulates rich, high-dimensional representations of audio data. However, its relatively lower accuracy in XGB and RF compared to LR may be attributed to feature correlation within the HuBERT representation. Since HuBERT incorporates contextual information during training, its features exhibit stronger inter-frame dependencies. In contrast, traditional acoustic features may be more orthogonal to each other, facilitating easier training for models like XGB and RF. LR, being a simpler model, may benefit from the dense information contained within HuBERT embeddings.

4.4 Breed Classification

To address the class imbalance in the dog breed dataset, a weighted loss function was applied, allowing the model to better handle

Table 5: Macro F1 and Accuracy for Dog Sex Classification.

Feature	LR		XGB		RF	
	F1	Acc	F1	Acc	F1	Acc
HuBERT	**0.562**	**0.563**	0.544	0.546	**0.549**	0.553
Gemaps	0.543	0.548	**0.555**	**0.561**	0.541	0.550
eGemaps	0.540	0.543	0.549	0.554	0.546	**0.555**
MFCC	0.530	0.533	0.550	0.552	0.542	0.546
Filterbanks	0.468	0.510	0.516	0.519	0.534	0.538

under-represented breeds during training. Table 6 presents the performance of various features in classifying dog breeds.

Besides HuBERT features dominating the LR and XGB columns, MFCCs also outperform other features in these two methods and even surpass HuBERT features in the RF method. This suggests that for dog breed classification, certain breeds may exhibit distinguishable vocal characteristics, such as specific frequency patterns, which MFCCs capture more effectively than other features. We identify MFCC 5, 1, 6, 4, 3, and 8 as the most prominent dimensions in XGB and RF. MFCC 3, 4, 5 (300-1200 Hz) are in the lower-mid frequency range, which indicates different breeds may have unique resonant frequencies due to differences in vocal tract anatomy. And Lower-order MFCCs (1-10) tend to contain more useful phonetic and spectral energy information than higher-order MFCCs.

These results highlight the overall effectiveness of HuBERT in capturing vocal characteristics that differentiate dog breeds, even though the improvement over other features is not as substantial as in sex classification.

Table 6: Macro F1 and Accuracy for Dog Breed Classification.

Feature	LR		XGB		RF	
	F1	Acc	F1	Acc	F1	Acc
HuBERT	**0.460**	**0.462**	**0.411**	**0.413**	0.304	0.325
Gemaps	0.301	0.310	0.304	0.306	0.273	0.278
eGemaps	0.305	0.312	0.316	0.318	0.271	0.279
MFCC	0.405	0.412	0.379	0.381	**0.312**	**0.327**
Filterbanks	0.268	0.318	0.324	0.332	0.294	0.308

4.5 Individual Dog Recognition

For this task, we selected dogs that have at least 150 Barkseqs, resulting in a total of 51 dogs with random 150 Barkseqs each. We split each dog's Barkseqs into 10 equal parts and do 10-fold cross-validation on the splits. That is, we train on $135 \times 51 = 6885$ Barkseqs and test on $15 \times 51 = 765$ Barkseqs for each fold. We can do this because we have a perfectly balanced dataset for this task. Table 7 shows the performance of various features for this task.

Table 7: Macro F1 scores and Accuracy for Individual Dog Recognition by 10-fold Cross Validation.

Feature	LR		XGB		RF	
	F1	Acc	F1	Acc	F1	Acc
HuBERT	**0.549**	**0.550**	**0.508**	**0.515**	**0.460**	**0.482**
Gemaps	0.235	0.280	0.371	0.372	0.360	0.381
eGemaps	0.237	0.279	0.376	0.384	0.353	0.377
MFCC	0.433	0.458	0.470	0.478	0.441	0.467
Filterbanks	0.196	0.203	0.443	0.449	0.410	0.433

Because individual voice recognition is a task that only relies on small amount of training data, we first test to see if the above result changes when we vary the amount of training data. We modify the above 10-fold cross-validation by training 9 different models using HuBERT with 10%, 20%, up to 90% of the data for each fold, and test the 9 models on the 10% test data. This gives us a curve of macro F1-score vs amount of training data. We average the curves obtained from 10 folds to plot Figure 3. The plot is interesting because it shows that the gain of increasing training data is almost linear, up until 40%, when the curve saturates to almost a flat line of 66% F1 score. It suggests that even with the most advanced acoustic feature like HuBERT, there is diminishing returns in terms of how much it can learn from the dog vocalization. One might want to look beyond acoustic techniques to better model the dog communication patterns.

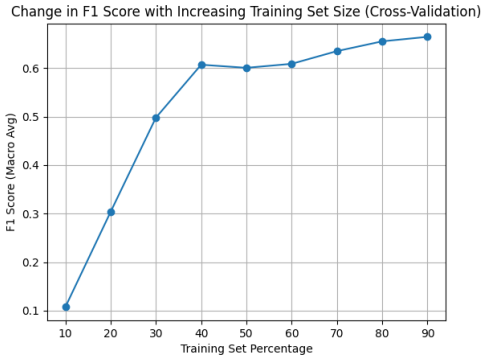

Figure 3: Change of HuBERT F1 Score with Increasing Training Data on Individual Recognition Task.

In the final results, HuBERT features dominate across all three methods, suggesting that rich, high-dimensional representations facilitate the capture of individual differences. HuBERT embeddings encode deep phonetic and prosodic features, which may be crucial for distinguishing individual dogs based on their unique vocal characteristics.

For the task of individual dog recognition, deep learning-based embeddings, like HuBERT, generally outperform traditional handcrafted acoustic features. This is because these learned embeddings are adept at capturing the complex, high-dimensional, and subtle phonetic or prosodic variations that constitute an individual's unique vocal signature. Handcrafted features, while useful for characterizing broader acoustic properties, often lack the fine-grained specificity required to reliably distinguish among a large number of individuals.

Compared to sex classification (binary) and breed classification (5-way), individual recognition is inherently a more fine-grained task. General handcrafted features (MFCCs, eGeMAPS, GeMAPS) may not be sufficient to differentiate among a large number of individuals, whereas deep embeddings provide a richer feature space that enhances individual recognition.

5 Related Work

The computational analysis of animal vocalizations, particularly canine barks, has evolved from foundational studies to the application

of sophisticated machine learning techniques. Early research established that dog barks convey meaningful information, with studies demonstrating context specificity, individual identification [36], and human listeners' ability to classify barks and infer emotional content [30, 31]. Machine learning subsequently enabled automated classification of bark contexts and individual Mudi dogs [25], as well as sex, age, and context [20]. More recent work has focused on recognizing context and perceived emotion [12].

Progress in this field is critically dependent on well-annotated datasets. Initial efforts often utilized smaller datasets from controlled settings, such as the Mudi dataset [20, 25] or versions of the Mescalina dataset [29], which featured more dogs and breeds but in semi-controlled environments. While valuable, these datasets often have limitations in scale, diversity, and the range of natural vocal expressions captured. Gómez-Armenta et al. [11] employed deep learning on a larger, though likely still controlled, dataset of 19,643 barks from 113 dogs. The dataset introduced in this paper significantly expands on previous work by providing 77,202 Barkseqs from 156 dogs across 5 breeds, uniquely sourced from 'in-the-wild' online social media. This scale and naturalistic data source address prior limitations but also introduce challenges related to noise and variability, as discussed by general bioacoustic best practices.

The advent of deep learning and advanced acoustic features, particularly self-supervised learning (SSL) models pre-trained on human speech (e.g., HuBERT [15], BEATs [4]), has become a key trend. This approach, leveraging transferable representations [21], is crucial in bioacoustics where annotated animal data can be scarce. Recent work, including by Wang et al. [33] on phonetic and lexical discovery and Huang et al. [16] on transcribing Shiba Inu communications, reflects this trend. A more recent paper has introduced an iterative algorithm for the automatic discovery of a canine phonetic alphabet and lexical structures [34].

Our methodology, employing BEATs for sound event detection and HuBERT for feature extraction, aligns with this state-of-the-art. The 'in-the-wild' nature of our dataset, processed with tools like AudioSep [23] for denoising, aims to spur the development of models robust to real-world complexities, an area where current methods still face significant challenges as indicated by our benchmark results.

6 Conclusion

In this paper, we introduce **DogSpeak**, a large-scale dataset of canine vocalizations, comprising of 77,202 Barkseqs from 156 individual dogs across 5 different breeds. This dataset is one of the largest and most comprehensive dataset in the field of animal vocalization, providing detailed annotations for each clip, including the dog's breed, individual ID, and sex. Along with the dataset, we present three benchmark classification tasks: *canine sex classification*, *canine breed classification*, and *individual dog recognition*. The size and accuracy of the DogSpeak dataset enable future research in the field of canine communication, providing an invaluable resource for improving classification methodologies and deepening our understanding of dog vocalizations. Despite the apparent simplicity of tasks such as sex, breed, and individual identification, our preliminary benchmark experiments demonstrate the complexity of these challenges. The current acoustic methods employed in our study do not handle these tasks as effectively as expected. This highlights the need for further exploration into more sophisticated approaches.

Acknowledgments

This work was partially supported by NSF Award No. 2349713. Our gratitude goes to Mengyue Wu for her 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. *arXiv preprint arXiv:2404.18739* (2024).
[2] Markus Boeckle, Georgine Szipl, and Thomas Bugnyar. 2018. Raven food calls indicate sender's age and sex. *Frontiers in zoology* 15 (2018), 1–9.
[3] Leo Breiman. 2001. Random forests. *Machine learning* 45 (2001), 5–32.
[4] 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* (2022).
[5] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785–794.
[6] 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.
[7] 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.
[8] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*. 1459–1462.
[9] Tamás Faragó, Simon Townsend, and Friederike Range. 2013. The information content of wolf (and dog) social communication. In *Biocommunication of animals*. Springer, 41–62.
[10] Shane Gero, Hal Whitehead, and Luke Rendell. 2016. Individual, unit and vocal clan level identity cues in sperm whale codas. *Royal Society Open Science* 3, 1 (2016), 150372.
[11] 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.
[12] 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.
[13] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer.
[14] Jessica M Hoffman, Dan G O'Neill, Kate E Creevy, and Steven N Austad. 2018. Do female dogs age differently than male dogs? *The Journals of Gerontology: Series A* 73, 2 (2018), 150–156.
[15] 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.
[16] 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.
[17] Alexander Ivanenko, Paul Watkins, Marcel AJ van Gerven, K Hammerschmidt, and Bernhard Englitz. 2020. Classifying sex and strain from mouse ultrasonic vocalizations using deep learning. *PLoS computational biology* 16, 6 (2020), e1007918.
[18] Raymond D Kent and Ann D Murray. 1982. Acoustic features of infant vocalic utterances at 3, 6, and 9 months. *The Journal of the Acoustical Society of America* 72, 2 (1982), 353–365.
[19] Maksim Kukushkin and Stavros Ntalampiras. 2021. Automatic acoustic classification of feline sex. In *Proceedings of the 16th International Audio Mostly Conference*. 156–160.
[20] Ana Larranaga, Concha Bielza, Péter Pongrácz, Tamás Faragó, Anna Bálint, and Pedro Larranaga. 2015. Comparing supervised learning methods for classifying sex, age, context and individual Mudi dogs from barking. *Animal cognition* 18, 2 (2015), 405–421.
[21] Xingyuan Li, Kenny Zhu, and Mengyue Wu. 2025. Dog2vec: Self-Supervised Pre-Training for Canine Vocal Representation. In *Proceedings of the 26th Interspeech*

Conference.

[22] Zeying Li, Tiemin Zhang, Kaixuan Cuan, Cheng Fang, Hongzhi Zhao, Chenxi Guan, Qilian Yang, and Hao Qu. 2022. Sex detection of chicks based on audio technology and deep learning methods. *Animals* 12, 22 (2022), 3106.

[23] 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* (2023).

[24] 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.

[25] 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 (2008), 389–400.

[26] 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.

[27] Yunbi Nam, Michelle White, Elinor K Karlsson, Kate E Creevy, Daniel EL Promislow, Robyn L McClelland, and Dog Aging Project Consortium. 2024. Dog size and patterns of disease history across the canine age spectrum: Results from the Dog Aging Project. *PLoS One* 19, 1 (2024), e0295840.

[28] Stavros Ntalampiras, Luca Andrea Ludovico, Giorgio Presti, Emanuela Prato Previde, Monica Battini, Simona Cannas, Clara Palestrini, and Silvana Mattiello. 2019. Automatic classification of cat vocalizations emitted in different contexts. *Animals* 9, 8 (2019), 543.

[29] 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.

[30] 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.

[31] 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.

[32] Emanuela Prato-Previde, Simona Cannas, Clara Palestrini, Sara Ingraffia, Monica Battini, Luca Andrea Ludovico, Stavros Ntalampiras, Giorgio Presti, and Silvana Mattiello. 2020. What's in a meow? A study on human classification and interpretation of domestic cat vocalizations. *Animals* 10, 12 (2020), 2390.

[33] 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.

[34] 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.

[35] Marina M Watowich, Evan L MacLean, Brian Hare, Josep Call, Juliane Kaminski, Ádám Miklósi, and Noah Snyder-Mackler. 2020. Age influences domestic dog cognitive performance independent of average breed lifespan. *Animal cognition* 23 (2020), 795–805.

[36] Sophia Yin and Brenda McCowan. 2004. Barking in domestic dogs: context specificity and individual identification. *Animal behaviour* 68, 2 (2004), 343–355.