Dog2vec: Self-Supervised Pre-Training for Canine Vocal Representation

Xingyuan Li[1], *Kenny Q. Zhu*[2*], *Mengyue Wu*[1*]

[1]X-LANCE Lab, School of Computer Science, Shanghai Jiao Tong University
MoE Key Lab of Artificial Intelligence, Jiangsu Key Lab of Language Computing, China
[2]University of Texas at Arlington, United States

xingyuan@sjtu.edu.cn, kenny.zhu@uta.edu, mengyuewu@sjtu.edu.cn

Abstract

Previous generalized biological voice models were trained on large amounts of data from multiple species. However, on average, there is very little training data on species-specific voices, while large differences between the vocalizations of species may even be a barrier to encoding vocal features. This leads to potentially large errors in using generic models for species-specific vocalization studies. We collected over 6000 hours of dog barking videos and presented the first animal-specific bioacoustic embedding model, Dog2vec[1]. The results indicate that Dog2vec outperforms species-independent pre-trained models and achieves state-of-the-art results on a series of dog-related tasks, including dog bark type recognition and dog sound event detection, and obtain a relative 8.2% performance increase.

Index Terms: bioacoustics, self-supervision, canine vocal representation

1. Introduction

Detecting acoustically active animals through their acoustic signals can provide a wealth of information that is important for conservation biology, ecology, evolutionary biology, animal behavior, and welfare [1]. In terms of passive acoustic monitoring (PAM), a method for studying and conserving animals and their habitats in a non-invasive manner [2], the bioacoustic signal is also one of the easiest indicators to pick up from vocalizing animals today. It reveals many mysteries about the animal itself, and indeed its surroundings. A substantial corpus [3, 4] of research has demonstrated that biological acoustic signals can encode a multitude of information, including individuality, age, gender, emotional state, and physiological condition. This provides a fundamental foundation for exploring biological and natural laws through acoustic signals.

Nonetheless, animal acoustic signal datasets are still scarce, with the difficulty of access and the high cost of labeling being one of the main reasons. Recently, rapid advances in machine learning have provided a path to solving this problem. Neural network models such as wav2vec2 [5], HuBERT [6], and BEATs [7] can be self-supervised trained on a large amount of unlabeled data and achieve good performance on a variety of tasks such as automatic speech recognition, sound classification, and sound event detection. The utilization of substantial unlabelled data constitutes a salient advantage of self-supervised training methodologies, with transformer-based approaches demonstrating a capacity to prioritize the inherent features of the data during the self-supervised training process. This approach can provide a good feature encoding for in-domain data, or even out-of-domain data, which can be cost-effectively migrated to multiple downstream tasks.

AVES [8] trained a transformer self-supervised model from a publicly available animal acoustic dataset with 360 hours of training data containing hundreds of different species. Robinson et al. [9] propose the use of contrastive language-audio pretraining for bioacoustics. Their model was trained on the AnimalSpeak [9] dataset, which contains over a million text-captioned audios spanning over 25,000 species. All of the above work was trained on sound data from multiple species but with less average training data per species. Average training data per species is much less than for humans (Librispeech [10]: an English automatic speech recognition corpus based on public domain audio books). This may result in suboptimal performance of the model in specific downstream tasks for particular species. In fact, the ablation experiments of AVES [8] have also shown the phenomenon that using more out-of-domain data for training is worse instead. A considerable body of research [11, 12, 13, 14, 15, 16] has demonstrated that a multitude of species possess highly intricate sound communication systems and modes of communication. These findings indicated a potential requirement for more extensive and sufficient data to facilitate a more comprehensive modeling of animal acoustic signals. In fact, many works [17, 2, 3] on animal sound signals are also based on specific animal sound datasets.

Canines are among the most prevalent animals and are considered to be excellent companions to humans. Many previous works [4, 3, 18, 19, 20, 21, 22] have identified that canine acoustic signals similarly possess intricate information, which may encompass gender, individuality, emotional state, and so on. It is hoped that the rules and finer-grained meanings present in dog barks can be further delineated. However, the majority of these studies utilize a restricted number of smaller datasets, which may impede their capacity to decode a greater volume of information contained within the barks of canines.

To address the various challenges mentioned above, we proposed a HuBERT-based, self-supervised model Dog2vec which pre-trained on a large amount of dog barking signal data. HuBERT [6] was initially trained on the LibriSpeech [10] dataset, which contains 960 hours of audio. This training resulted in Hubert achieving the best performance on multiple tasks in SUPERB [23], particularly the automatic speech recognition task. We collected more than 6,000 hours of dog barking video and audio data from the large online social platform (YouTube) containing the six common dog breeds (Chihuahua, Husky, Shiba Inu, Pitbull, Labrador, and German Shepherd). Following a series of cleaning procedures, approximately 150 hours of dog barking was obtained for the training of the model. Given that the data originates from publicly accessible social media plat-

*Corresponding authors.

[1]The model, data, and code are available at https://github.com/fispresent/dog2vec.

forms, the data source comprises a diverse sample of individuals from various geographical locations worldwide. The utilization of a diverse array of data has been demonstrated to enhance the robustness of the model, thereby ensuring that experimental outcomes are not influenced by minority bias. The model has the capacity to encode dog vocalizations and can be applied to a variety of canine-related downstream tasks. Furthermore, it can be used to further parse the rules and meaning of dog vocalizations.

We validated Dog2vec on multiple downstream tasks and from multiple perspectives. Compared to past generalized models, we achieved better performance on a dog vocalization-related downstream task. Dog2vec can provide better feature encoding for multiple downstream tasks related to dog barking, while also demonstrating its potential for exploring finer-grained dog vocalization rules.

2. Method

Inspired by Huang [24] and Wang [25], we collected more dog-related data[2] (over 6,000 hours) from the large social platform (YouTube), covering six common dog breeds (Chihuahua, Husky, Shiba Inu, Pitbull, Labrador, and German Shepherd).

2.1. Data processing

Before training models, we need to clean the data in order to retain as pure dog barking data as possible for subsequent self-supervised model training. HuBERT was initially trained on top of the clean, high-concentration human speech dataset Librispeech [10]. In order to maximize the representational power of HuBERT, we also need to clean the data. Indeed, the experiments in [8] have shown that training HuBERT with more noisy data leads to performance degradation. After our examination, there are two main types of noise present in the dataset: one is noise mixed with dog barking, from which the dog barking needs to be separated; and the other is pure noise, which can be removed directly.

2.1.1. Separating dog barkings from mixed noise

The goal of this section is to remove the first noise mentioned above. Due to the great diversity in the sources of the data, there is no control over the inclusion of only clean sounds. In order to retain as much clean data as possible, we used AudioSep [26] to separate the dog barking from the noise. AudioSep is a foundation model for open-domain audio source separation with natural language queries [26]. AudioSep is pre-trained on large-scale multimodal datasets which are more than ten thousand hours, including the AudioSet [27], VGGSound [28], and AudioCaps [29] datasets, etc. We used "Dogs" as text input to extract clean dog barks from the noise audios.

2.1.2. Remove noise clips

The goal of this step is to remove the second noise mentioned above. This noise is pure noise, does not contain any barking fragments, and can be removed directly, as long as we know where the noise starts and where it ends. In this work, we used the fine-tuned DCASE2023 challenge task 4 baseline model to get dog sentences. We manually labeled about 2.5 hours of data for fine-tuning the model and achieved an F1 score of 0.8556 on the test set. Note that some of the uncleaned noise data retained by AudioSep will also be removed during this process.

[2]The data is available at `https://github.com/fispresent/dog2vec`.

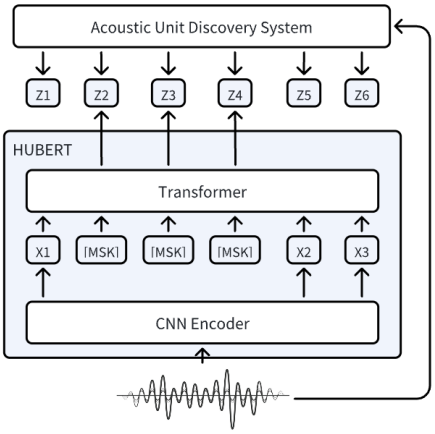

Figure 1: *The Dog2vec pretraining process.*

2.2. Self-supervised transformer model pretraining

For each step of data cleaning, we examined the data to ensure that the dataset was clean enough. After data cleaning, we obtained over 150 hours of clean dog barking data.

The training process for Dog2vec (Figure 1) is similar to HuBERT. The average duration of dog clips is less than 2 seconds, which is much lower than that of humans, so we modified the model parameters to make it more suitable for the dog barking dataset. We used the K-means cluster algorithm as the acoustic unit discovery system to generate frame-level pseudo-labels. The audio is passed through the CNN encoder to obtain a sequence of frame-level CNN features. Then use a similar method to wav2vec2 to mask part of the continuous segment and input it into the transformer to get the features E^T, where T is the number of frames. E^T is that the extracted features can be migrated to multiple downstream tasks. In the pretraining process, we only calculate the cross-entropy loss for masked partial predictions and pseudo-labeling Z^T:

$$L_m = \sum_{t \in M} p(z_t | E^T, t), \tag{1}$$

where M is the location of the masked portion. p contains two operations: computing the cosine similarity between $f(E)$ and codeword, and then going through softmax, where f is a projection layer.

3. Experiments

3.1. Training details

For the pretraining Dog2vec model, we continued training on top of the HuBERT-base, which had been trained on 960 hours of human speech. We trained two stages with the number of clusters being 100 and 200 respectively. Set mask length to 6. The minimum and maximum sample lengths are set to 5600 and 80000, respectively. These hyperparameters are designed to be set exactly to the characteristics of the dog barking data. Compared to human speech, dog barks are generally shorter.

	Training set	Validation set	Test set	Total (hours)
Breed recognition	23,921	2,986	2,986	15.84
Individual identification (10)	5,708	709	709	3.45
Individual identification (20)	6,655	825	825	4.18
BEANS (dogs)	414	139	139	2.45
Bark type recognition	7,943	990	990	2.19
Detection	1,096	361	361	2.99

Table 1: *The datasets used to evaluate the models.*

3.2. Evaluation setup

In order to provide a more comprehensive evaluation of the performance of the models, The models were evaluated through the implementation of a suite of classification and detection tasks, which are two of the most common tasks considered in the bioacoustic literature [8].

In the classification tasks, the models were evaluated from three perspectives: dog breed recognition, dog individual identification, and dog bark type recognition. In the detection task, a sound event detection task was employed to assess the capacity of the model to capture barking edge features, in addition to its performance under the influence of non-barking noise present in the surrounding environment.

The control models involved in the experiment were all in operation and had been trained on a large dataset. The datasets utilized in the evaluation experiments partly consist of public datasets or public benchmarks, and we have also produced some higher-quality datasets to participate in the evaluation.

A proportion of the data is retained for the purpose of preventing potential data leakage. This data is not utilized in the training of Dog2vec but rather serves to generate the dataset employed for the evaluation process.

3.2.1. Finetuning details

In order to fairly evaluate the ability of different models to model dog barking, we froze all pre-trained models. It means they only extract features of audio without updating parameters. The classification header contains only two fully connected layers, the first of which has a hidden dimension of 1024. It also contains relu as an activation layer and dropout with a probability of 0.3. The model is trained for up to 500 epochs, using an Adam learning rate with parameter 2e-4. The batch size for the classification and detection tasks were 4096 and 512, respectively.

All datasets were divided into training, validation, and test sets. We selected the weights that performed best on the validation set during the training of the model to infer the test set and show the metrics on the test set.

3.2.2. Models of Comparison

The comparison models primarily utilize two types of pre-trained models that are currently demonstrating superior performance, are readily available to the public, and are frequently employed for a variety of downstream tasks. They are trained on a large human speech dataset and a large biological speech dataset respectively:

- **wav2vec2** [5] is a framework for self-supervised learning of representations from raw audio data. Wav2vec2 is a transformer-based self-supervised one that has been pre-trained on 960 hours of human speech data (Librispeech [10]). The model has been widely migrated to a variety of human speech downstream tasks.
- **HuBERT** [6] is a speech representation learning approach that relies on predicting K-means cluster assignments of masked segments of continuous input. It is also trained in Librispeech. We used it to compare the ability of Dog2vec to get better features of dog barking by continuing to train on top of it.
- **AVES** [8] is a self-supervised, transformer-based audio representation model for encoding animal vocalizations for downstream bioacoustic tasks. AVES is currently one of the most commonly used migration models in bioacoustics. It contains four models, all pre-trained on a large amount of bioacoustic data. One of the best-performing models was pre-trained on 360 hours of bioacoustic data. It also includes dog barking data in its training data. We evaluated all four of its models.
- **BioLingual** [9] use of contrastive language-audio pretraining for bioacoustics. It was trained on a large dataset of more than one million text-captioned audios covering 25,000 species. BioLingual sets a new state-of-the-art on nine tasks in the Benchmark of Animal Sounds [30].

3.2.3. Tasks

The evaluation tasks were divided into two main categories, containing five classification tasks and one detection task. The 5 classification tasks evaluated the model from 3 perspectives: Dog breed recognition, Dog individual identification, and Dog bark type recognition. Table1 presents information on the evaluation datasets used for each task.

- **Dog breed recognition**: This is a classification task. We used data not involved in the training of Dog2vec to make a six-categorized dataset (Chihuahua, Husky, Shiba Inu, Pitbull, Labrador, and German Shepherd).
- **Dog individual identification (10)**: This is a classification task. We produced a 10-category dataset using data that was not involved in training Dog2vec. Data for each of the 10 categories came from different individual dogs.
- **Dog individual identification (20)**: This is a classification task. All configurations are the same as **Dog individual identification (10)**, except that the number of categories changes to 20, representing data from 20 different individual dogs.
- **Dogs in BEANS [30]**: This is a classification task. BEANS [30] (the BEnchmark of ANimal Sounds) is a collection of bioacoustics tasks and public datasets, specifically designed to measure the performance of machine learning algorithms in the field of bioacoustics. There is a dog barking classification task in beans (a 10-classification task for individual dog identification).
- **Dog bark type recognition**: This is a classification task. In AudioSet [27], dog barking is categorized into six categories: Bark, Yip, Howl, Bow-wow, Growling, and Whimper (dog). We cut these six categories of data from the AudioSet-Strong-Unbalanced [31] dataset for this task.
- **Dog sound event detection**: This is a sound event detection task, similar to [32]. The purpose is to tag out dog vocalization as well as detect the on- and off-sets of the event. We manually labeled 1,818 pieces of data as the dataset (more than 2.99 hours) for this task.

3.2.4. Metrics

For the classification tasks, we used the F1 score. For the detection task, instead of using mAP like BEANS [30], we used the more common metric Segment-based F1 score in sound event

	Breed		Individual(10)		Individual(20)		BEANS (dogs)		Bark type	
	micro	macro	micro	macro	micro	macro	micro	macro	micro	macro
wav2vec2	0.6527	0.4825	0.5557	0.5395	0.4230	0.4009	0.7842	0.7457	0.7525	0.5027
HuBERT	0.6333	0.4317	0.5458	0.5239	0.3903	0.3730	0.7482	0.6745	0.7242	0.4410
AVES-core	0.6875	0.5169	0.6587	0.6409	0.5188	0.5110	0.8129	0.7692	0.8101	0.6318
AVES-bio	0.7194	0.5776	0.6812	0.6650	0.5588	0.5454	0.8489	0.8101	0.8051	0.6116
AVES-nonbio	0.7063	0.5429	0.6728	0.6546	0.5503	0.5388	0.8417	0.8165	0.8040	0.6264
AVES-all	0.7056	0.5457	0.6855	0.6637	0.5539	0.5384	0.8201	0.7838	0.7949	0.6088
BioLingual	0.6962	0.5347	0.6178	0.5955	0.4897	0.4712	0.8921	0.8790	0.8091	0.5674
Dog2vec	**0.7793**	**0.6499**	**0.7362**	**0.7191**	**0.6218**	**0.6094**	**0.9137**	**0.8948**	**0.8434**	**0.7138**

Table 2: *The results of the classification tasks. The best metrics are highlighted.*

	wav2vec2	HuBERT	AVES-core	AVES-bio	AVES-nonbio	AVES-all	BioLingual	Dog2vec
Dog detection	0.7788	0.7732	0.7914	0.8028	0.8012	0.7866	0.8097	**0.8287**

Table 3: *The results of the detection task. The best metrics are highlighted.*

detection tasks. Because dog barks are usually very short (less than 1 second), it is more appropriate to use the Segment-based F1 score. The threshold for detection was set at 0.5.

3.3. Results

Table2 and Table3 show the results of the models on the classification tasks and the detection task, respectively. Overall, Dog2vec is the best performer in all downstream tasks, outperforming AVES [8] and BioLingual [9] trained on a large bioacoustic dataset, as well as HuBERT [6] and wav2vec2 [5] trained on a large human speech dataset. The experimental results demonstrate that the model is capable of extracting features from canine sounds and migrating effectively to various downstream tasks.

In addition, we also note that the models pre-trained on the human speech dataset all perform worse than the models pre-trained on the biological sound signals. This illustrates the gap between human speech and dog vocalizations, demonstrating the vast differences in vocalizations between species. Thus, it also further illustrates the need for a model trained for the vocalizations of a particular species if the vocalizations of that species are to be studied.

3.4. Ablation on the feature of different layers of Dog2vec

Figure 2 shows how the features we extracted from the different layers of Dog2vec perform on the downstream task. The experimental results show that Dog2vec does not take the best performance in the last layer. In fact, this result is not unexpected, the same is true for HuBERT [6] trained on human speech, which may be related to the fact that the last layer features of the model are fitted to the pseudo-labels during the self-supervised training process. Instead, the features of the previous layers have higher quality, and at the same time, these features have phonetic information (on human speech). This also shows the potential of Dog2vec to further explore finer-grained information in dog vocalizations.

3.5. Limitation

Despite the large amount of data we obtain from social platforms, however, there is still a large gap between the durations of data (more than 150 hours) retained after cleaning and human speech data (960 hours). This gap may result in the inability of the model to model dog barks as well as it does human

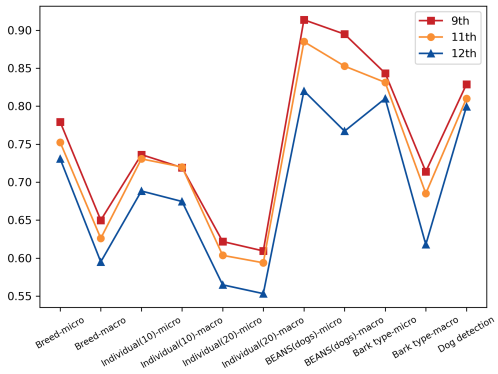

Figure 2: *The results on the features of 9th, 11th and 12th layers of Dog2vec.*

voices. Nevertheless, compared to the generic model, Dog2vec is trained on more dog barking datasets and has better representational capabilities.

4. Conclusion

In this work, we proposed a HuBERT-based, self-supervised model pre-trained on a large amount of dog barking signal data. Dog2vec can provide good features for dog barking. We evaluated Dog2vec in several downstream tasks from multiple perspectives, and Dog2vec can perform better in various downstream tasks related to canines compared to generalized biological sound models. Since Dog2vec is trained on a single creature (dog) voice that is not affected by the vocalizations of other animals, coupled with the strengths of HuBERT in encoding fine-grained speech, Dog2vec is not only simple to migrate to a variety of other canine downstream tasks but also provides a basis for further parsing of canine vocalizations at a finer level of granularity.

5. Acknowledgement

This work was supported by the Key Research and Development Program of Jiangsu Province, China (No.BE2022059), Guangxi major science and technology project (No. AA23062062), and NSF Award No. 2349713.

6. References

[1] A. Kershenbaum, Ç. Akçay, L. Babu-Saheer, A. Barnhill, P. Best, J. Cauzinille, D. Clink, A. Dassow, E. Dufourq, J. Growcott *et al.*, "Automatic detection for bioacoustic research: a practical guide from and for biologists and computer scientists," *Biological Reviews*, 2024.

[2] B. Ghani, T. Denton, S. Kahl, and H. Klinck, "Global birdsong embeddings enable superior transfer learning for bioacoustic classification," *Scientific Reports*, vol. 13, no. 1, p. 22876, 2023.

[3] A. Abzaliev, H. P. Espinosa, and R. Mihalcea, "Towards dog bark decoding: Leveraging human speech processing for automated bark classification," *arXiv preprint arXiv:2404.18739*, 2024.

[4] A. Larranaga, C. Bielza, P. Pongrácz, T. Faragó, A. Bálint, and P. Larranaga, "Comparing supervised learning methods for classifying sex, age, context and individual mudi dogs from barking," *Animal cognition*, vol. 18, no. 2, pp. 405–421, 2015.

[5] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[6] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.

[7] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, and F. Wei, "Beats: Audio pre-training with acoustic tokenizers," *arXiv preprint arXiv:2212.09058*, 2022.

[8] M. Hagiwara, "Aves: Animal vocalization encoder based on self-supervision," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[9] D. Robinson, A. Robinson, and L. Akrapongpisak, "Transferable models for bioacoustics with human language supervision," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 1316–1320.

[10] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[11] R. S. Payne and S. McVay, "Songs of humpback whales: Humpbacks emit sounds in long, predictable patterns ranging over frequencies audible to humans." *Science*, vol. 173, no. 3997, pp. 585–597, 1971.

[12] P. C. Bermant, M. M. Bronstein, R. J. Wood, S. Gero, and D. F. Gruber, "Deep machine learning techniques for the detection and classification of sperm whale bioacoustics," *Scientific reports*, vol. 9, no. 1, p. 12588, 2019.

[13] T. Bortolato, A. D. Friederici, C. Girard-Buttoz, R. M. Wittig, and C. Crockford, "Chimpanzees show the capacity to communicate about concomitant daily life events," *Iscience*, vol. 26, no. 11, 2023.

[14] M. Leroux, A. M. Schel, C. Wilke, B. Chandia, K. Zuberbühler, K. E. Slocombe, and S. W. Townsend, "Call combinations and compositional processing in wild chimpanzees," *Nature Communications*, vol. 14, no. 1, p. 2225, 2023.

[15] S. Engesser and S. W. Townsend, "Combinatoriality in the vocal systems of nonhuman animals," *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 10, no. 4, p. e1493, 2019.

[16] M. Berthet, C. Coye, G. Dezecache, and J. Kuhn, "Animal linguistics: a primer," *Biological reviews*, vol. 98, no. 1, pp. 81–98, 2023.

[17] J. Xie, Y. Zhong, J. Zhang, S. Liu, C. Ding, and A. Triantafyllopoulos, "A review of automatic recognition technology for bird vocalizations in the deep learning era," *Ecological Informatics*, vol. 73, p. 101927, 2023.

[18] S. Hantke, N. Cummins, and B. Schuller, "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, 2018, pp. 5134–5138.

[19] C. Molnár, F. Kaplan, P. Roy, F. Pachet, P. Pongrácz, A. Dóka, and Á. Miklósi, "Classification of dog barks: a machine learning approach," *Animal Cognition*, vol. 11, pp. 389–400, 2008.

[20] A. Paladini, "The bark and its meanings in inter and intra-specific language," *Dog behavior*, vol. 6, no. 1, pp. 21–30, 2020.

[21] R. L. Robbins, "Vocal communication in free-ranging african wild dogs (lycaon pictus)," *Behaviour*, pp. 1271–1298, 2000.

[22] P. Pongrácz, C. Molnár, and A. Miklosi, "Acoustic parameters of dog barks carry emotional information for humans," *Applied Animal Behaviour Science*, vol. 100, no. 3-4, pp. 228–240, 2006.

[23] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin *et al.*, "Superb: Speech processing universal performance benchmark," *arXiv preprint arXiv:2105.01051*, 2021.

[24] J. Huang, C. Zhang, M. Wu, and K. Zhu, "Transcribing vocal communications of domestic shiba lnu dogs," in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 13 819–13 832.

[25] T. S. Wang, X. Li, C. Zhang, M. Wu, and K. Q. Zhu, "Phonetic and lexical discovery of canine vocalization," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 13 972–13 983.

[26] X. Liu, Q. Kong, Y. Zhao, H. Liu, Y. Yuan, Y. Liu, R. Xia, Y. Wang, M. D. Plumbley, and W. Wang, "Separate anything you describe," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[27] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[28] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "Vggsound: A large-scale audio-visual dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 721–725.

[29] C. D. Kim, B. Kim, H. Lee, and G. Kim, "Audiocaps: Generating captions for audios in the wild," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 119–132.

[30] M. Hagiwara, B. Hoffman, J.-Y. Liu, M. Cusimano, F. Effenberger, and K. Zacarian, "Beans: The benchmark of animal sounds," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[31] S. Hershey, D. P. Ellis, E. Fonseca, A. Jansen, C. Liu, R. C. Moore, and M. Plakal, "The benefit of temporally-strong labels in audio event classification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 366–370.

[32] N. Shao, X. Li, and X. Li, "Fine-tune the pretrained atst model for sound event detection," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 911–915.