Towards Dog Bark Decoding: Leveraging Human Speech Processing for Automated Bark Classification

Artem Abzaliev[1], **Humberto Pérez Espinosa**[2], **Rada Mihalcea**[1]
[1,3]University of Michigan, [2]National Institute of Astrophysics, Optics and Electronics (INAOE)
abzaliev@umich.edu, humbertop@inaoep.mx, mihalcea@umich.edu.com

Abstract

Similar to humans, animals make extensive use of verbal and non-verbal forms of communication, including a large range of audio signals. In this paper, we address dog vocalizations and explore the use of self-supervised speech representation models pre-trained on human speech to address dog bark classification tasks that find parallels in human-centered tasks in speech recognition. We specifically address four tasks: dog recognition, breed identification, gender classification, and context grounding. We show that using speech embedding representations significantly improves over simpler classification baselines. Further, we also find that models pre-trained on large human speech acoustics can provide additional performance boosts on several tasks.

Keywords: animal vocalizations, semi-supervised learning, audio processing

1. Introduction

Until recently, "what humans do" has been considered the most widely accepted definition of intelligence (Tomasello, 2019), but a large body of recent work has demonstrated that there are numerous other forms of non-human intelligence (Bridle, 2022; Call and Carpenter, 2001; Biro et al., 2003). While there are several new studies demonstrating plant intelligence (Wohlleben, 2016; dos Santos et al., 2024) most of the research to date has focused on the intelligence of animals (De Waal, 2016; Grandin and Johnson, 2009). Forms of animal intelligence range from memory (Matzel and Kolata, 2010) and problem-solving (Seed and Call, 2010), all the way to the use of tools (St Amant and Horton, 2008) and communication (Seyfarth and Cheney, 2003; López, 2020).

Like humans, animals use both verbal and non-verbal forms of communication, including audio signals such as calls, songs, or hisses; visual signals such as facial expressions, tail moves, or postural gestures; chemical cues; tactile cues; and bioluminescence. In general, the study of animal communication has been mainly addressed in fields such as biology, ecology, and anthropology, including, for instance, prairie dogs (Slobodchikoff et al., 2009), birds (Thorpe, 1961) or body movement in bees (Al Toufailia et al., 2013). Only recently we have started to see research that leverages advances in machine learning (Bergler et al., 2019; Jasim et al., 2022; Maegawa et al., 2021).

Focusing specifically on animal vocal communication, a recent study (Andreas et al., 2022) highlighted three main questions to be answered to increase our understanding of how animals communicate: (1) What are the phonetic and perceptual units used by animals? (phonemes); (2) What are the composition rules used to combine those units? (morphology, syntax); and (3) Do those units carry meaning and, if so, how do we map the sound units to their meaning? (semantics, pragmatics).

In this work, we explore the third question and specifically attempt to understand the semantics of dog vocalizations. We use a state-of-the-art human speech representation learning model and show that such models can predict the context of a bark.

This paper makes three main contributions. First, we introduce a dataset and a set of tasks for dog bark classification. We draw parallels between human speech classification tasks and dog bark classification tasks, including dog recognition, breed recognition, gender identification, and context grounding. Second, through several experiments, we show that we can leverage models developed for human speech processing to explore dog vocalizations and demonstrate that these can be used to significantly enhance performance on several dog bark classification tasks. Finally, through this work, we hope to open new opportunities for research in the area of animal communication, which can leverage the extensive expertise available in the NLP community.

2. Related Work

Animal Communication Datasets. Compared to human languages, there are significantly fewer datasets available for animal communication. The largest library of animal vocalizations is the Macaulay Library at the Cornell Lab of Ornithology,[1] which includes audio, photos, and videos of 2,674 species of amphibians, fish, mammals, and more, with the main focus of the library on birds. Another large library of animal vocalizations is the Animal Sound Archive[2], which covers 1,800 bird

[1] https://www.macaulaylibrary.org
[2] https://www.tierstimmenarchiv.de/webinterface/

LREC-COLING 2024, pages 16480–16486

Context	# segments	Duration (sec)
Very aggressive barking at a stranger (L-S2)	2,843	2,778.66
Normal barking at a stranger (L-S1)	2,772	2,512.92
Barking due to assault on the owner (L-A)	829	956.58
Negative grunt (during the presence of a stranger) (GR-N)	637	746.60
Negative squeal (during the presence of a stranger) (CH-N)	298	546.72
Sadness/anxiety barking (L-TA)	288	200.27
Positive squeal (during gameplay) (CH-P)	91	150.49
Barking during play (L-P)	76	51.21
Barking due to stimulation when walking (L-PA)	62	84.06
Barking in fear at a stranger (L-S3)	54	45.08
Positive grunt (during gameplay) (GR-P)	51	79.56
Barking arrival of the owner at home (L-H)	24	26.20
Barking that is neither playful nor strange (L-O)	9	9.50
Non-dog sounds (voices, TV, cars, appliances, etc.) (S)	8,755	14,304.05
TOTAL	16,789	22,491

Table 1: 14 types of dog vocalizations together with the corresponding number of segments and duration.

species and 580 mammal species.

There are also several datasets related to marine mammals. Ness et al. (2013) presented a large dataset of over 20,000 recordings of Orca vocalizations. The Watkins Marine Mammal Sound Database[3] contains 15,000 annotated sound clips for more than 60 species of marine mammals.

Specifically for dog vocalizations, one of the most popular datasets was introduced by Pongrácz et al. (2005). It includes twelve Mudi dogs and consists of 244 recordings. Another dataset is the UT3 database Gutiérrez-Serafín et al. (2019), with 74 dogs and 6,000 individual audios. Neither of these datasets is publicly available.

Computational Approaches to Animal Communication Analysis. Several studies have applied machine learning to animal communication, most of which used Convolutional Neural Networks (CNNs) to classify bird calls Maegawa et al. (2021); Jasim et al. (2022), primate species Pellegrini (2021); Oikarinen et al. (2019), multi-species classification of birds and frogs LeBien et al. (2020), or orca sounds (Bergler et al., 2019). Ntalampiras (2018) used various methods to transfer the signal from music genre identification to bird species identification.

Specifically for dogs, there have been several studies studying dog vocalizations (Pérez-Espinosa and Torres-García, 2019; Pérez-Espinosa et al., 2015; Gutiérrez-Serafín et al., 2019). Our work is more closely related to (Yin and McCowan, 2004), where the contexts in which barking occurs are predicted, along with individual dog recognition. The results of the experiments of Hantke et al. (2018) confirmed that one can predict the context of the bark. Molnár et al. (2009) also finds that the barks include information about the individual dog, as well as information about the context. However, no pre-trained models for dog vocalizations are currently available. To our knowledge, we are the first to use neural acoustic representations for tasks on dog vocalizations, and we are also the first to explore the use of human speech pre-training.

[3] https://cis.whoi.edu/science/B/whalesounds/index.cfm

3. Dataset

We use a dataset consisting of recordings of 74 dogs, collected in Tepic (Mexico) and Puebla (Mexico), at the homes of the dogs' owners. A subset of this dataset was previously used by Pérez-Espinosa et al. (2018). The dog vocalizations were recorded while being exposed to different stimuli (e.g., stranger, play, see Table 1). The recordings were conducted using a video camera Sony CX405 Handycam; in this work, we only use the audio recordings, obtained using the built-in microphone on the camera. The audio codec is A52 stereo with a sampling rate of 48,000 Hz and a bit rate of 256 kbps. The protocol for obtaining the dog vocalizations used in this study was designed and validated by experts in animal behavior from the Tlaxcala Center for Behavioral Biology in Mexico.

The dataset includes recordings of 48 female and 26 male dogs, mostly of three breeds: 42 Chihuahua, 21 French Poodles, and 11 Schnauzer. For mixed breeds, we first selected the breed mentioned. We focused on these breeds since they are among the most common domestic breeds in Mexican households. Given time and resource constraints during the data collection process, these breeds allowed for a broader choice of participants. The dog's average age is 35 months, ranging between 5 to 84 months old.

Stimuli. Dog vocalizations were induced by exposing them to several stimuli, with the participation of the owner and/or an experimenter. To illustrate, the following represent examples of situations used during the data collection: the experimenter repeatedly rings the home doorbell and knocks the door hard; the experimenter simulates an attack on the owner; the owner speaks affectionately to the dog; the owner stimulates the dog using the objects or toys with which the dog normally plays; the owner performs the normal routine that precedes a walk; the owner ties the dog on a leash to a tree and

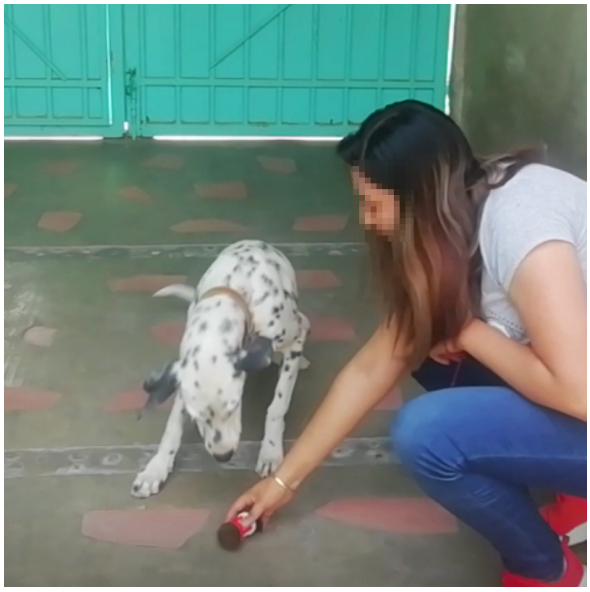

Figure 1: Data collection for the stimulus "playing with toy"; the owner stimulates the dog using the toys with which the dog normally plays.

walks out of sight (see Figure 1 as an example); and others. The dogs are recorded while reacting to these stimuli, resulting in recordings lasting between 10 sec to 60 min.

Data Processing and Annotation. The recordings are automatically segmented into shorter segments ranging between 0.3 to 5 sec in length. The segmentation is performed using a threshold to separate between sound and silence or background noise; the threshold was identified using the short-time energy and spectral centroid aspects of the acoustic signal. Only the sound segments are used for the experiments. Each of the resulting segments was manually annotated using the information associated with the stimulus. One of the fourteen contexts was assigned to each segment; if the audio did not have any dog-related sounds, it was assigned a Non-dog sound label.

Table 1 shows the fourteen labels used in the annotation, along with the corresponding statistics for the number of segments and total duration.

4. Dog Bark Classification Tasks

Using acoustic representations of dog barks, we explore several fundamental tasks, including the recognition of individual dogs; the identification of the breed of a dog; the identification of a dog's gender; and the grounding of a dog bark to its context. These tasks have counterparts in human speech analysis, such as speaker identification or grounded language analysis.

Leveraging human speech for acoustic dog bark representations To create acoustic representations of the dog vocalizations in the dataset, we fine-tune a pre-trained state-of-the-art self-supervised speech representation model. We use Wav2Vec2 (Baevski et al., 2020), which uses a self-supervised training objective to predict masked latent representations, pre-trained on the Librispeech corpus (Panayotov et al., 2015). Wav2vec2 uses 960 hours of unlabeled human-speech data to learn how to represent audio signals as a sequence of discrete tokens. Some of those discrete tokens are masked, similar to the process used to train the BERT contextual embedding model (Devlin et al., 2018). In Wav2Vec2, the learning of discrete units and unmasking are happening simultaneously.

We use an open-source implementation of Wav2Vec2 from HuggingFace (Wolf et al., 2019). We experiment with two model versions: (1) a model trained from scratch, using the dog vocalizations dataset in Section 3; (2) a model pre-trained on 960 hours of unlabeled human speech data, and fine-tuned on dog vocalizations.

Experimental Setup. All the experiments use a ten-fold cross-validation setup. Specifically, for the tasks of breed identification, gender identification, and grounding, we use *grouped* ten-fold validation, with individual dogs being the group variable. That is, we leave 7-8 dogs as a test dataset and train on the remaining dogs' vocalizations, to control for any confounding information. For the dog recognition task, given the goal to recognize individual dogs, the model has to see each class (i.e., each dog) during the training, and thus all 74 individual dogs have to be present in both the training and test datasets. We note that this particular way of cross-validating might enable easier learning for the model and does not prevent shortcut learning, which is a common drawback for all author identification tasks. Therefore even Wav2Vec2 pretrained from scratch performs relatively well, and the performance boost is more pronounced than for other tasks.

4.1. Dog Recognition

We formulate this task as classifying a single audio segment as belonging to one of the 74 dogs in the dataset. According to (Molnár et al., 2006) humans struggle to discriminate between individual dog barks, but machine learning methods, even unsupervised, can perform rather well (Yin and McCowan, 2004). This task is similar to identifying speakers, where many datasets (Nagrani et al., 2017; Chung et al., 2018) and methods (Huang et al., 2023; Ding et al., 2020) already exist.

Table 2 shows the results, where we apply the Wav2Vec2 model to dog identification. Our results are in line with the results from (Pérez-Espinosa et al., 2018; Molnár et al., 2009), and demonstrate

that effectiveness of acoustic representations to discriminate between individual dogs. Further, we find that a model pre-trained on human speech significantly outperforms the model trained from scratch.

Given the differences between human speech and animal vocalizations, we still need more work to understand how pre-training on human speech improves the performance on dog vocalization tasks. We believe that the pre-training on human speech enables the model to learn abstract vocalization structures, which in turn are beneficial for understanding animal vocalizations. This hypothesis is supported by previous studies showing that pre-training on seemingly unrelated tasks can be beneficial, for instance, pretraining on symbolic music data and applying it to natural language data provides significant performance improvements due to the ability of the neural networks used by the model to represent abstract syntactic structure (Papadimitriou and Jurafsky, 2020). Similarly, in computer vision, pre-training on ImageNet (object recognition) data is found to improve radiography processing (Jabbour et al., 2020).

Method	Accuracy
Majority	5.03%
Wav2Vec2 (from scratch)	23.74%
Wav2Vec2 (pre-trained)	**49.95%**

Table 2: Accuracy for the dog recognition task.

4.2. Breed Identification

In this task, we aim to predict the breed of a dog. Our dataset contains mostly three breeds: Chihuahua, French Poodle, and Schnauzer. We hypothesize that different breeds have different pitches so the acoustic model should be able to identify those differences, independent of the context. This experiment is related to previous work (LeBien et al., 2020; Oikarinen et al., 2019). The task is similar to human accent recognition, where given audio files in a single language (i.e., English) the goal is to classify the accent of a speaker (e.g., USA vs. UK vs. India), with several approaches proposed in previous work (Ayrancı et al., 2020; Honnavalli and Shylaja, 2021; Sun, 2002).

The results are shown in Table 3. Wav2Vec2 trained from scratch outperforms most baselines. As before, we obtain an additional significant boost in performance when pre-training on human speech data. The variation in individual breeds can be explained by the unbalanced number of observations per breed, with Chihuahua being the most common breed in our dataset (57%), followed by French Poodle (28%) and Schnauzer (15%).

4.3. Gender Identification

The goal of this task is to probe whether it is possible to predict the gender of a dog from vocalizations. This is a task analogous to the prediction of demographics (e.g., age, gender) from language or speech, with many previous studies conducted on this topic (Qawaqneh et al., 2017; Saraf et al., 2023; Gupta et al., 2022; Welch et al., 2020).

Table 5 shows the results. The Wav2Vec2 model trained from scratch performs better than the baseline model, with no further improvements obtained with Wav2Vec2 pre-trained on human speech. Interestingly, we do see an improvement brought by human speech pre-training on the female class, for which we have significantly more data in our dataset (67.95% female vs 32.04% male by total duration). We found that gender identification is the most difficult task among all the tasks we propose. We hypothesize that the model trained from scratch focuses on learning acoustic features, while the pre-trained wav2vec attempts to learn shortcuts and overfits quickly. We noticed that it often predicts just the majority class (female) so that F1 increases for female and decreases for male, while the overall accuracy is almost the same as for the majority baseline.

4.4. Grounding

In this task, we predict the context of the bark; i.e., we determine the association between a dog vocalization and its surrounding. Because of the highly skewed label distribution (see Table 1), we focus on the contexts for which more examples are available: very aggressive barking at a stranger (L-S2); normal barking at a stranger (L-S1); negative squeal (in the presence of a stranger) (CH-N); negative grunt (in the presence of a stranger) (GR-N). We do not include barking due to assault on the owner (L-A) because in early experiments we found that the model cannot distinguish it from the very aggressive barking at a stranger (L-S2).

Human language grounding is the mapping of language symbols such as words to their corresponding objects in the real world. There have been several works showing that animals ground their vocalizations as well. For instance, the vocalizations of prairie dogs are grounded and used to transmit the characteristics of the predators (e.g., color or size) (Slobodchikoff et al., 2009). Other work has also demonstrated that it is possible to predict call types for marmoset monkeys (Oikarinen et al., 2019) also shows. We hypothesize that dog vocalizations are related to their context, and therefore can be grounded.

Table 4 shows the results. Similar to the previous experiments, both Wav2Vec2 models outperform the majority baseline, with the Wav2Vec2 pre-trained on human speech leading to the most accurate results.

5. Conclusion

In this paper, we explored the use of pre-trained self-supervised speech representation models to

Method	Acc.	F-1 measure		
		Chihuahua	French Poodle	Schnauzer
Majority	58.76%	61.49%	6.59%	6.78%
Wav2Vec2 (from scratch)	60.18%	74.42%	14.96%	5.79%
Wav2Vec2 (pre-trained)	**62.28%**	**74.47%**	**36.11%**	**14.88%**

Table 3: Accuracy and F-1 measure for dog breed identification.

Method	Acc.	F-1 measure			
		L-S2	CH-N	GR-N	L-S1
Majority	56.37%	41.31%	0.00%	0.00%	30.39%
Wav2Vec2 (from scratch)	58.45%	49.26%	21.26%	78.20%	48.64%
Wav2Vec2 (pre-trained)	**62.18%**	**49.66%**	**45.26%**	**90.70%**	**51.13%**

Table 4: Accuracy and F-1 measure for context grounding.

Method	Acc.	F-1 measure	
		Female	Male
Majority	68.70%	74.73%	7.76%
Wav2Vec2 (from scratch)	**70.07%**	76.54%	**19.29%**
Wav2Vec2 (pre-trained)	68.90%	**79.31%**	7.10%

Table 5: Accuracy and F-1 measure for dog gender identification.

address dog barking classification tasks. We specifically addressed four tasks that find parallels in human-centered speech recognition tasks: dog recognition, breed recognition, gender identification, and context grounding. We showed that acoustic representation models using Wav2Vec2 can significantly improve over simpler classification baselines. Additionally, we found that a model pre-trained on human speech can further improve the results. We hope our work will encourage others in the NLP community to start addressing the many research opportunities that exist in the area of animal communication. The dataset used in this work, along with the baselines that we introduced, are publicly available by request from `humber-top@ccc.inaoep.mx`.

6. Limitations

In this work, we focused on only one species, domestic dogs, and only three breeds. More species are required to understand how modern computational methods can be used for studying animal vocalization. In the future, we are planning to extend our work to birds and marine mammals, since those species have a large amount of data available.

We also focused on only one neural network architecture, Wav2Vec2. While it is a popular architecture for human speech processing, other architectures might be more suitable for studying animal vocalizations. Also, we used supervised learning in this work, since the dataset was manually annotated. The majority of the datasets are not annotated and thus would require semi-supervised or unsupervised learning, which is more challenging.

Hasan Al Toufailia, Margaret J Couvillon, Francis LW Ratnieks, and Christoph Grüter. 2013. Honey bee waggle dance communication: signal meaning and signal noise affect dance follower behaviour. *Behavioral Ecology and Sociobiology*, 67:549–556.

Jacob Andreas, Gašper Beguš, Michael M. Bronstein, Roee Diamant, Denley Delaney, Shane Gero, Shafi Goldwasser, David F. Gruber, Sarah de Haas, Peter Malkin, Nikolay Pavlov, Roger Payne, Giovanni Petri, Daniela Rus, Pratyusha Sharma, Dan Tchernov, Pernille Tønnesen, Antonio Torralba, Daniel Vogt, and Robert J. Wood. 2022. Toward understanding the communication in sperm whales. *iScience*, 25(6):104393.

Ahmet Aytuğ Ayrancı, Sergen Atay, and Tülay Yıldırım. 2020. Speaker accent recognition using machine learning algorithms. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 1–6. IEEE.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations.

Christian Bergler, Manuel Schmitt, Rachael Xi Cheng, Andreas K Maier, Volker Barth, and Elmar Nöth. 2019. Deep learning for orca call type Identification-A fully unsupervised approach. In *INTERSPEECH*, pages 3357–3361. isca-speech.org.

Dora Biro, Noriko Inoue-Nakamura, Rikako Tonooka, Gen Yamakoshi, Cláudia Sousa, and Tetsuro Matsuzawa. 2003. Cultural innovation and transmission of tool use in wild chimpanzees: evidence from field experiments. *Animal Cognition*, 6:213–223.

James Bridle. 2022. *Ways of being: Beyond human intelligence*. Penguin UK.

Josep Call and Malinda Carpenter. 2001. Do apes and children know what they have seen? *Animal Cognition*, 3:207–220.

Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. Voxceleb2: Deep speaker recognition. In *Interspeech*.

Frans De Waal. 2016. *Are we smart enough to know how smart animals are?* WW Norton & Company.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Shaojin Ding, Tianlong Chen, Xinyu Gong, Weiwei Zha, and Zhangyang Wang. 2020. Autospeech: Neural architecture search for speaker recognition.

Luana Silva dos Santos, Victor Hugo Silva dos Santos, and Fabio Rubio Scarano. 2024. Plant intelligence: history and current trends. *Theoretical and Experimental Plant Physiology*.

Temple Grandin and Catherine Johnson. 2009. *Animals in translation: Using the mysteries of autism to decode animal behavior*. SUNY Press.

Tarun Gupta, Duc-Tuan Truong, Tran The Anh, and Chng Eng Siong. 2022. Estimation of speaker age and height from speech signal using bi-encoder transformer mixture model.

Benjamín Gutiérrez-Serafín, Humberto Pérez Espinosa, Juan Martínez-Miranda, and Ismael Edrein Espinosa-Curiel. 2019. Classification of barking context of domestic dog using high-level descriptors. *Res. Comput. Sci.*, 148(3):23–35.

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

Dweepa Honnavalli and SS Shylaja. 2021. Supervised machine learning model for accent recognition in english speech using sequential mfcc features. In *Advances in Artificial Intelligence and Data Engineering: Select Proceedings of AIDE 2019*, pages 55–66. Springer.

Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. 2023. Masked autoencoders that listen.

Sarah Jabbour, David Fouhey, Ella Kazerooni, Michael W. Sjoding, and Jenna Wiens. 2020. Deep learning applied to chest x-rays: Exploiting and preventing shortcuts.

Hasan Abdullah Jasim, Saadaldeen R Ahmed, Abdullahi Abdu Ibrahim, and Adil Deniz Duru. 2022. Classify bird species audio by augment convolutional neural network. In *2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, pages 1–6. ieeexplore.ieee.org.

Jack LeBien, Ming Zhong, Marconi Campos-Cerqueira, Julian P Velev, Rahul Dodhia, Juan Lavista Ferres, and T Mitchell Aide. 2020. A pipeline for identification of bird and frog species in tropical soundscape recordings using a convolutional neural network. *Ecol. Inform.*, 59:101113.

Bruno Díaz López. 2020. When personality matters: personality and social structure in wild bottlenose dolphins, tursiops truncatus. *Animal Behaviour*, 163:73–84.

Yuko Maegawa, Yuji Ushigome, Masato Suzuki, Karen Taguchi, Keigo Kobayashi, Chihiro Haga, and Takanori Matsui. 2021. A new survey method using convolutional neural networks for automatic classification of bird calls. *Ecol. Inform.*, 61:101164.

Louis D Matzel and Stefan Kolata. 2010. Selective attention, working memory, and animal intelligence. *Neuroscience & Biobehavioral Reviews*, 34(1):23–30.

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):76–83.

Csaba Molnár, Péter Pongrácz, Tamás Faragó, Antal Dóka, and Ádám Miklósi. 2009. Dogs discriminate between barks: The effect of context and identity of the caller. *Behavioural Processes*, 82(2):198–201.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. Voxceleb: A large-scale speaker identification dataset. In *Interspeech*.

Steven Ness, Helena Symonds, Paul Spong, and George Tzanetakis. 2013. The orchive : Data mining a massive bioacoustic archive.

Stavros Ntalampiras. 2018. Bird species identification via transfer learning from music genres. *Ecol. Inform.*, 44:76–81.

Tuomas Oikarinen, Karthik Srinivasan, Olivia Meisner, Julia B Hyman, Shivangi Parmar, Adrian Fanucci-Kiss, Robert Desimone, Rogier Landman, and Guoping Feng. 2019. Deep convolutional network for animal sound classification and source attribution using dual audio recordings. *J. Acoust. Soc. Am.*, 145(2):654.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Isabel Papadimitriou and Dan Jurafsky. 2020. Learning music helps you read: Using transfer to study linguistic structure in language models.

Thomas Pellegrini. 2021. Deep-learning-based central african primate species classification with MixUp and SpecAugment. In *Interspeech 2021*. ut3-toulouseinp.hal.science.

H. Pérez-Espinosa, V. Reyes-Meza, E. Aguilar-Benitez, and Y. M. &; Sanzón-Rosas. 2018. Automatic individual dog recognition based on the acoustic properties of its barks. *Journal of Intelligent & Fuzzy Systems*, 34(5):3273–3280.

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*, 100:63–74.

Humberto Pérez-Espinosa and Alejandro Antonio Torres-García. 2019. Evaluation of quantitative and qualitative features for the acoustic analysis of domestic dogs' vocalizations. *Journal of Intelligent & Fuzzy Systems*, 36(5):5051–5061.

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):136.

Zakariya Qawaqneh, Arafat Abu Mallouh, and Buket D Barkana. 2017. Deep neural network framework and transformed mfccs for speaker's age and gender classification. *Knowledge-Based Systems*, 115:5–14.

Amruta Saraf, Ganesh Sivaraman, and Elie Khoury. 2023. A zero-shot approach to identifying children's speech in automatic gender classification. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 853–859. IEEE.

AM Seed and Josep Call. 2010. Problem-solving in tool-using and non-tool-using animals. *Encyclopedia of animal behavior*, 2:778–785.

Robert M Seyfarth and Dorothy L Cheney. 2003. Signalers and receivers in animal communication. *Annual review of psychology*, 54(1):145–173.

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

Robert St Amant and Thomas E Horton. 2008. Revisiting the definition of animal tool use. *Animal Behaviour*, 75(4):1199–1208.

Xuejing Sun. 2002. Pitch accent prediction using ensemble machine learning. In *Seventh international conference on spoken language processing*. Citeseer.

William Homan Thorpe. 1961. Bird-song: the biology of vocal communication and expression in birds.

Michael Tomasello. 2019. *Becoming human: A theory of ontogeny*. Harvard University Press.

Charles Welch, Jonathan K. Kummerfeld, Verónica Pérez-Rosas, and Rada Mihalcea. 2020. Compositional demographic word embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4076–4089, Online. Association for Computational Linguistics.

Peter Wohlleben. 2016. *The hidden life of trees: What they feel, how they communicate—Discoveries from a secret world*, volume 1. Greystone Books.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

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