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

Artem Abzaliev, Humberto Pérez Espinosa, Rada Mihalcea

Research question

Can we leverage speech representation models pre-trained on human speech to improve dog vocalization recognition?

Motivation

- Data scarcity in animal vocalization studies.
- Even unlabelled data is scarce!
- Maybe we can leverage other data sources/modalities?

Contributions

- We introduce a dataset with 4 tasks for dog vocalizations
- We show that human speech pre-training can improve model performance on several of those tasks

Dataset

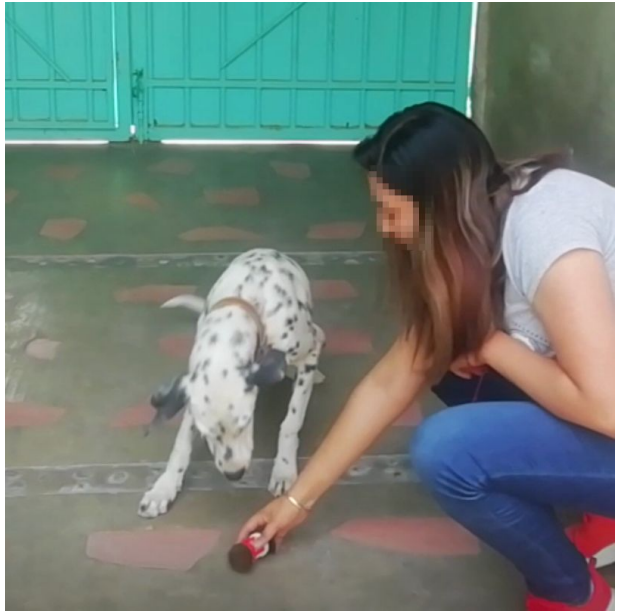
- Mescalina Bark ID Database, 2017 version. The dog vocalizations were “induced” by the data collectors in different contexts, 14 contexts
- 48 female and 26 male dogs, Chihuahua, French Poodles, and Schnauzer

1. Espinosa, Humberto & Pérez-Martínez, José & Durán-Reynoso, Joséjosé´ & Reyes-Meza, Veronica. (2015). Automatic Classification of Context in Induced Barking.
2. Espinosa, Humberto & Reyes-Meza, Veronica & Aguilar-Benitez, Emanuel & Sanzón-Rosas, Yuvila. (2018). Automatic individual dog recognition based on the acoustic properties of its barks.

Stimuli

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

“Playing with toy” stimuli



Dog Bark Classification Tasks

- Breed classification (3 classes) ~ human accent recognition
- Gender classification (2 classes)
- Context classification (4 classes) ~ human language grounding
- Individual dog recognition (74 dogs) ~ identifying speakers in audio processing

Baseline approach - wav2vec2

- Two options:
  - Randomly initialized
  - Pretrained - ininitialize the weights from the model pre-trained on human speech (Librispeech corpus)
- Validation: we selected 7-8 dogs as a test, and the rest as train (repeat 10 fold cross validation)
- Individual dog: stratified 10-fold CV

Results | Breed Prediction

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.

Results | Individual Dog Recognition

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

Table 2: Accuracy for the dog recognition task.

Results | Gender Prediction

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.

Results | Context Prediction

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.

Conclusion

- Pre-trained self-supervised speech representation models can improve the performance of animal vocalization models
- We propose 4 different animal vocalization tasks
- The dataset is publicly available by request from `humbertop@ccc.inaoep.mx`

Limitations & Discussion

- Is it just better initialization?
- Wav2vec2 includes vector quantization component, maybe there is a smart way to utilize this for unsupervised discovery?
- Why gender prediction didn’t work?
- More modalities are necessary