WHAT IS MY DOG TRYING TO TELL ME? THE AUTOMATIC RECOGNITION OF THE CONTEXT AND PERCEIVED EMOTION OF DOG BARKS

Simone Hantke[1,2]*, Nicholas Cummins*[1]*, Björn Schuller*[1,3]

[1]Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[2]Machine Intelligence & Signal Processing Group, Technische Universität München, Germany
[3]Group on Language, Audio, and Music, Imperial College London, UK

simone.hantke@informatik.uni-ausgburg.de

ABSTRACT

A wide range of research disciplines are deeply interested in the measurement of animal emotions, including evolutionary zoology, affective neuroscience and comparative psychology. However, only a few studies have investigated the effect of phenomena such as emotion on the acoustic parameters of (non-human) mammalian species. In this contribution, we explore if commonly used affective computing-based acoustic feature sets can be used to classify either the context, the emotion, or predict the emotional intensity of dog bark sequences. This comparison study includes an in-depth analysis of obtainable classification performances. Results presented indicate that the tested feature representations are suitable for the proposed recognition tasks. Of particular note are results that demonstrate machine learning-based acoustic analysis can achieve above human level performance when classifying the context of a dog bark.

Index Terms— affective biology, canine emotion, affective computing, acoustic analysis, bag-of-audio-words

1. INTRODUCTION

Affective biology, the measurement of animal emotions, is starting to gain considerable interest in a wide range of research disciplines including evolutionary zoology, affective neuroscience and comparative psychology [1]. To date, very little research attention has focused on using affective computing techniques to recognise emotions in mammals other than humans. However, such approaches – if successful – would be of great benefit to veterinarian and animal welfare science.

Speech-based emotion detection is a well-established and mature area of research within affective computing. Many of the concepts that underlie this research field are transferable to other mammals. Firstly, human vocalisations share many similar aspects with mammal vocalisations in terms of acoustic, physiology and neural control [2]. All mammals generate their primary acoustic signal at a source, typically rapid pulses of air, generated in the lungs, being forced through the vocal folds. The combined action of the vocal tract and the articulators filters this source signal to produce vocalisations.

Further, all mammals have similar neurophysiological responses to emotional stimuli, e. g., changes in the brain activity or in the heart rate [1]. Accordingly, changes in the emotional state of an animal should effect the muscular systems used to control the vocal apparatus altering the acoustic properties of the vocalisation [1]. Therefore, the work presented in this paper explores if acoustic feature representations, designed and developed for human-based affective computing purposes, can be used to recognise context and emotions in dog barks.

1.1. Related Work

Compared to other vocalisations such as growling or howling, dog barks are highly variable and are used in various situations such as care or contact solicitation [3]. Research in the literature reveals that in certain *contexts*, barks have distinct acoustic properties [4, 5, 6]. For example, results present in [4] indicate that barks elicited from a *disturbance* situation had proportionally more energy at lower frequencies, while barks elicited from *play* have a harmonically rich structure.

Perception tests reveal that human listeners have the ability to categorise dog barks and growls accurately in regards to the original recording situation and associate them with an appropriate emotionality [6, 7, 8, 9]. In these tests, acoustic parameters including tonality, pitch and inter-bark time intervals had a strong effect on how human listeners described the emotionality of these dog vocalisations [6]. In another set of perception tests [10], low pitched barks were described as aggressive, while tonal and high pitched barks were scored as either fearful or desperate.

Initial machine learning analysis has shown that acoustic descriptors including spectral roll off, spectral flatness and formant features have shown to be suitable for classifying bark context achieving a classification efficiency of 43% [5]. These findings all suggest that dog barks have bark-specific acoustic features; similar findings have also been reported in several other social mammal species [11, 12].

1.2. Contributions of this Work

This paper explores, if acoustic feature representations, commonly used in affective computing, can be used to recognise the context and perceived emotion of dog barks. While previous work has focused on the recognition of single bark sounds [5], this work focuses on the recognition of sequences of dog barks and performs three experiments: (i) classifying the context, (ii) classifying the perceived emotion, (iii) predicting the perceived emotional intensity.

As we are classifying bark sequences, we test the suitability of feature representations which extract supra-segmental information from low-level frame wise features. Namely, the small but tailor made for emotion recognition *extended Geneva Minimalistic Acoustic Parameter Set* (eGeMAPS) [13]; the large brute-force *Interspeech Computational Paralinguistics Challenge features set* (COMPARE) [14]; and a range of *Bag-of-Audio-Words* (BoAW) representations [15]. These feature representations can be considered state-of-the-art for human emotion detection [13, 15, 16].

Given that results presented in the literature demonstrate that humans apply similar rules when processing the emotion of both dogs' or humans' vocalisations [17], and the similarities in the anatomical nature and structure of human and dog vocal apparatus, we speculate our chosen feature representations should be suitable for recognising the context or perceived emotion in dog barks.

2. EMOTIONAL DOG CORPUS

The *Emotional Dog Corpus* (EmoDog) contains vocalisations – bark sequences – of dogs recorded in different standardised situations. For consistency, only the barks of the Mudi, a herding dog breed from Hungary, were collected for this corpus. The working style of this breed is characterised by its extensive use of barking. A total of 226 bark sequences were recorded from 12 different Mudi dogs. The average length of the bark sequences is 41.8 seconds with a standard deviation of 42.8 seconds.

The barks were obtained from one of seven different contextual situations as follows: (i) *Alone*: The owner tied the dog to a tree with a leash and walked out of sight of the dog; (ii) *Ball*: The owner held a ball at a height of approximately 1.5 m in front of the dog; (iii) *Fight*: A professional dog trainer encouraged the dog to bark aggressively and to bite a padded glove on the trainer's arm whilst the owner kept the dog on a leash; (iv) *Food*: The owner provided the dog with food; (v) *Play*: The owner played a typical game such as chasing or wrestling; (vi) *Stranger*: An experimenter appeared outside the home of the dog in the absence of the owner; (vii) *Walk*: The owner behaved as if they were preparing to take the dog for a walk. The distribution of the context situations can be seen in Table 1.

Table 1. Distribution of the different contextual situations used to generate all barks in the EmoDog Corpus.

Alone	Ball	Fight	Food	Play	Stranger	Walk
18	50	22	41	22	44	29

2.1. Annotations

In earlier work, six professional dog trainers assigned emotions to the bark sequences [6, 10, 5]. The annotators were asked to rate each bark in terms of either *Aggression*, *Fear*, *Despair*, *Fun*, or *Happiness* on a scale from 1 to 5. Taking the mean of these score, we are able to assign an intensity score per emotion to each of the bark sequences (cf. Table 2). Each bark sequence was then assigned a single emotion label, using that sequences' maximum mean annotation score.

To gain perspective into how well our systems performed *contextual classification*, we conducted human classification tests through our gamified crowdsourcing platform iHEARu-PLAY[1] [18]. Each listener was asked to classify each sequence in terms of the reason for barking i. e., *Alone*, *Ball*, *Fight*, *Food*, *Play*, *Stranger*, or *Walk*. The listeners had the possibility to repeat listening to each sequence as often as required, before submitting their final answer. Overall, five listeners labelled the 227 bark sequences for the reason for barking.

3. EXPERIMENTAL SETTINGS

3.1. Acoustic Feature Sets

We investigate two different acoustic feature sets – the 88 dimensional *extended Geneva Minimalistic Acoustic Parameter Set* (eGeMAPS) [13] and the 6373 *Interspeech Computational Paralinguistics Challenge* (COMPARE) features set [19] – for the efficacy in our classification and prediction tests both of which have consistently shown to capture emotion information in human speech [19, 20, 21, 22, 23, 24].

Given results in the literature showing that prosodic and spectral cues may both be important when analysing dog barks [4, 6, 5], we also test two subsets of the COMPARE features separately: (i) COMPARE prosodic features only (COMPARE Pros.), and (ii) COMPARE spectral and cepstral features only (COMPARE Spec.).

3.2. Bag-of-Audio-Words

We also test the *Bag-of-Audio-Words* (BoAW) paradigm [15], that has been shown to be suitable for a range of speech-based emotion recognition tasks [20, 25].

The BoAW representations were formed using our open-source openXBOW toolkit [15]. An extensive iterative search was performed to identify the *codebook size* ($Cs \in \{10, 20, 50, 100, 200, 500, 1\,k\}$) and *number of assignments*

[1] https://www.ihearu-play.eu

Table 2. Distribution of the emotion intensities in the EmoDog Corpus in terms of the minimum (Min.), mean, maximum (Max.) and standard Deviation (Std. Dev.); also shown are the number of sequences (# Seq.) to each of the emotion categories. The emotion scores were assigned by six expert annotators.

	Agg.	**Dsp.**	**Fear**	**Fun**	**Hap.**
Min.	0.50	0.33	0.33	0.67	0.50
Mean	2.76	3.08	2.60	2.50	2.51
Max.	5.00	5.00	5.00	4.83	4.83
Std. Dev.	1.01	0.82	0.92	0.97	0.94
# Seq.	71	16	14	52	29

($Na \in \{10, 20, 50, 100, 200, 500\}$), with random assignments being used to generate all codebooks.

3.3. Classification and Prediction Setup

Both the eGeMAPs and ComParE feature sets, as well as all LLDs used in our BoAW test, were extracted using the *openSMILE* feature extraction toolkit [19]. Six different LLDs feature sets were used in conjunction with BoAW: (i) *Mel Frequency Cepstral Coefficients* (BoAW MFCC) 1 – 12 and the logarithmic signal energy, extracted using 25 ms long frames, with a frame rate of 10 ms, and a preemphasis filter (k = 0.97); (ii) the same set of MFCCs appended with the corresponding first and second order derivative features (BoAW MFCC + deltas); (iii) eGeMAPs LLDS (BoAW eGeMAPS); (iv) ComParE LLDs (BoAW ComParE); (v) ComParE Prosodic LLDs (BoAW ComParE Pros.); and, (vi) ComParE Spectral and Cepstral LLDs (BoAW ComParE Spec.).

All *classification* tests were performed with a linear *Support Vector Machine* (SVM) implemented using the open-source *LIBLINEAR* toolkit [26], with the cost parameter being tuned separately for each experiment using a search space of $C \in \{1, 2, 5\} \cdot 10^{-6}$ to $\{1, 2, 5\} \cdot 10^{2}$. All *prediction* tests were performed using an *epsilon–Support Vector Regression* (SVR) implemented via the open-source *LIBSVM* toolkit [27]. A grid search was undertaken to find the optimal C (same range as for the classification tests) and an epsilon ($\epsilon \in \{1\} \cdot 10^{-6}$ to $\{1\} \cdot 10^{2}$) parameters.

3.4. Evaluation

As the *classification* evaluation measurement of the performance of the different feature sets, we employed a *leave-one-dog-out* cross fold validation scheme and all results are reported in terms of *Unweighed Average Recall* (UAR). The motivation to consider UAR rather than other measures is that it can better reflect the overall accuracy in the presence of imbalanced classes as well as for more than two classes, and is widely used for emotion, and even other computational paralinguistics recognition tasks [14, 28]. All *prediction* tests are evaluated with respect to the *Root Mean Square Error* (RMSE).

Table 3. Comparison of different acoustic feature representations, known to capture human emotions, for classifying either the context (7 classes) or the perceived emotion (5 classes) of dog bark sequences. All results are given in terms of Unweighed Average Recall (UAR).

% UAR	**Context**	**Emotion**
eGeMAPS	24.4	**28.5**
ComParE	31.3	25.7
ComParE Pros.	30.0	27.5
ComParE Spec.	**32.9**	25.8
BoAW MFCC	19.2	24.4
BoAW MFCC + deltas	21.1	23.8
BoAW eGeMAPS	19.1	25.1
BoAW ComParE	16.7	21.9
BoAW ComParE Pros.	15.9	22.9
BoAW ComParE Spec.	16.6	21.8
Human Performance	23.7	–
Chance	14.3	20.0

4. RESULTS AND DISCUSSION

4.1. Context Classification

For classifying *Bark Context*, we calculated the human performance as a fusion of the five (crowdsourced) annotators which achieved a UAR of 23.7% (cf. Table 3). The eGeMAPS and ComParE feature sets were able to match or outperform the annotators at this task. The ComParE spectral and cepstral features gave the strongest performance with a UAR of 32.9%. This matches with results in [4, 5], which showed that barks elicited from different contextual situations have different frequency distributions.

For the BoAW representation, the best contextual classification was achieved with a UAR of 21.1% (cf. Table 3), using BoAW MFCC + deltas features ($Cs = 1000$, $Na = 1$). The weaker performance of this representation compared to the other feature sets was surprising. We speculate this might be due to the distribution of acoustic events – as captured by ComParE functionals – containing more relevant contextual information compared to the 'frequency' of audio events as modelled by BoAW.

4.2. Perceived Emotion Classification

When classifying the *emotions* of the dog barks, the eGeMAPS feature set gave the best performance with 28.5% UAR (cf. Table 3). This result is not surprising; the eGeMAPS feature set was explicitly designed to have a high level of robustness for human emotion recognition [13]. The ComParE feature sets performs slightly below eGeMAPS

Table 4. Comparison of different acoustic feature representations, known to capture human emotions, for predicting the intensity of five different emotions, namely Agg(ression), Des(pair), Fear, Fun and Hap(piness), in dog bark sequences. All results are given in terms of Root Mean Square Error (RMSE) where the emotion intensity scores ranged 1-5.

RMSE	Agg.	Des.	Fear	Fun	Hap.
eGeMAPS	.888	.832	.891	**.876**	**.882**
ComParE	.885	.834	.890	.928	.920
ComParE (Pros.)	.855	.837	.902	.954	.932
ComParE (Spec.)	.891	.833	.896	.937	.922
BoAW MFCC	**.847**	.814	.907	.925	.922
BoAW MFCC + deltas	.869	.819	.905	.931	.924
BoAW eGeMAPS LLDs	.910	.853	.937	.957	.929
BoAW ComParE LLDs	.987	.818	.927	.982	.944
BoAW ComParE Pros. LLDs	.918	**.775**	**.877**	.977	.922
BoAW ComParE Spec. LLDs	.988	.815	.927	.982	.944

obtaining a UAR of 25.7%. It can also be observed that the splitting of the feature set into prosodic and spectral features does not bring any advantages for the emotion task compared to the content task; similar observations have been made for humans where the combination of different acoustic feature types is used to capture the effects of emotion in speech [19, 13].

The BoAW representations are more competitive for the emotion classification task (cf. Table 3); the best result achieved was an UAR 25.1% found with eGeMAPS LLDs ($Cs = 1000$, $Na = 1$). Similar results have also been observed in human emotion classification [20], where ComParE features outperformed a BoAW representation in a 5-class emotion classification paradigm.

4.3. Perceived Emotional Intensity

When predicting the *intensity* of aggression, despair, fear, fun, and happiness in the bark sequences (cf. Table 4), the results indicate that using our chosen feature representation, the (perceived) intensity of the negative emotion aggression, despair and fear, is easier to estimate than in the positive emotions fun and happiness. As in the emotion classification tests, the eGeMAPS feature set performs very well at this task achieving the lowest RMSE in the comparison of all feature sets for fun (.876, $C = 10$) and for happiness (.882, $C = 5$). As already stated, eGeMAPS was specifically designed for (human) emotion recognition, therefore, the strong results on all bark emotion tasks are not surprising.

We also applied the BoAW paradigms to this emotional intensity prediction task (cf. Table 4). Observing these results we can see that BoAW MFCC ($C = 50$, $Cs = 100$, $Na = 5$) performed with a RMSE of .847, which is best for the aggression task and achieved the best result for this emotion class. This BoAW MFCC setup has shown state-of-the-art performance for arousal tracking in human speech [25]. The strong aggression result presented matches with this finding; high aggression is associated with high arousal.

For the emotion classes of despair and fun, the best results were obtained using the BoAW ComParE Prosodic LLDs (cf. Table 4), resulting in .775 RMSE for despair ($C = 200$, $Cs = 100$, $Na = 50$) and .877 for fear ($C = 20$, $Cs = 200$, $Na = 50$). This result matches with previous results presented in [6, 10], showing prosodic features including tonality, pitch and inter-bark time intervals had a strong effect on how human listeners described the emotionality of these dog vocalisations.

5. CONCLUSION AND OUTLOOK

We investigated if commonly used affective computing based acoustic feature sets can be used to classify either the context, the perceived emotion, or the intensity of a dog bark. Our results show that acoustic feature sets purposely designed to capture human emotions can be used to classify these three tasks accordingly. The strong performance of the eGeMAPS features set in the emotion classification tasks was to be expected. This feature set was hand crafted especially to capture emotion information in human vocalisations [13]. All achieved results indicate the suitability of human based features for characterising dog barks. This result supports theories that due to similarities in emotion responses and vocalisation apparatus mammalian emotional changes produce similar acoustics effects across mammals [1, 2].

Future work will include testing alternative feature representation and more advanced machine learning approaches, in particular long-short-term-memory based Recurrent Neural Networks. We will also consider transfer learning analyses including human emotional data to gain further insights into this research field of dog vocalisations.

6. ACKNOWLEDGEMENT

The research leading to these results has received funding from the European Community's Seventh Framework Programme under grant agreement No. 338164 (ERC Starting Grant iHEARu). We thank all iHEARu-PLAY users for donating their annotations.

7. REFERENCES

[1] E. F. Briefer, "Vocal Expression of Emotions in Mammals: Mechanisms of Production and Evidence," *Journal of Zoology*, vol. 288, pp. 1–20, 2012.

[2] W. Tecumseh Fitch, "The Evolution of Speech: A Comparative Review," *Trends in Cognitive Sciences*, vol. 4, pp. 258 – 267, 2000.

[3] J. Cohen and M. Fox, "Vocalizations in Wild Canids and Possible Effects of Domestication," *Behavioural Processes*, vol. 1, pp. 77–92, 1976.

[4] S. Yin and B. McCowan, "Barking in Domestic Dogs: Context Specificity and Individual Identification," *Animal Behaviour*, vol. 68, pp. 343–355, 2004.

[5] C. Molnár, F. Kaplan, P. Roy, F. Pachet, P. Pongrácz, A. Dóka, and A. Miklósi, "Classification of Dog Barks: A Machine Learning Approach," *Animal Cognition*, vol. 11, pp. 389–400, 2008.

[6] P. Pongrácz, C. Molnár, A. Miklósi, and V. Csányi, "Human Listeners are able to Classify Dog (Canis Familiaris) Barks Recorded in Different Situations," *Journal of Comparative Psychology*, vol. 119, pp. 136–144, 2005.

[7] P. Molnár, C.and Pongrácz, A. Dóka, and A. Miklósi, "Can Humans Discriminate Between Dogs on the Base of the Acoustic Parameters of Barks?," *Behavioural Processes*, vol. 73, pp. 76–83, 2006.

[8] P. Pongrácz, A. Miklósi, and V. Csányi, "Owner's Beliefs on the Ability of Their Pet Dogs to Understand Human Verbal Communication: A Case of Social Understanding," *Current Psychology of Cognition*, vol. 20, pp. 87–108, 2001.

[9] T. Faragó, N. Takács, Á. Miklósi, and P. Pongrácz, "Dog Growls Express Various Contextual and Affective Content for Human Listeners," *Royal Society Open Science*, vol. 4, pp. 170134, 2017.

[10] P. Pongrácz, C. Molnár, and A. Miklósi, "Acoustic Parameters of Dog Barks Carry Emotional Information for Humans," *Applied Animal Behaviour Science*, vol. 100, pp. 228–240, 2006.

[11] M. Manser, R. Seyfarth, and D. Cheney, "Suricate Alarm Calls Signal Predator Class and Urgency," *Trends in Cognitive Sciences*, vol. 6, pp. 55–57, 2002.

[12] R. Seyfarth, D. Cheney, and P. Marler, "Monkey Responses to Three Different Alarm Calls: Evidence of Predator Classification and Semantic Communication," *Science*, vol. 210, pp. 801–803, 1980.

[13] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. André, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Transaction on Affective Computing*, vol. 7, pp. 190–202, 2016.

[14] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity & Native Language," in *Proc. of INTERSPEECH*, San Francisco, USA, September 2016, ISCA, pp. 2001–2005.

[15] M. Schmitt and B. Schuller, "openXBOW – Introducing the Passau Open-Source Crossmodal Bag-of-Words Toolkit," *Journal of Machine Learning Research*, vol. 18, 2017.

[16] B. Schuller and A. Batliner, *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*, Wiley, November 2013.

[17] T. Faragó, A.Andics, V. Devecseri, A. Kis, M. Gcsi, and A. Miklsi, "Humans Rely on the Same Rules to Assess Emotional Valence and Intensity in Conspecific and Dog Vocalizations," *Biology Letters*, vol. 10, pp. 5, 2014.

[18] S. Hantke, F. Eyben, T. Appel, and B. Schuller, "iHEARu-PLAY: Introducing a Game for Crowdsourced Data Collection for Affective Computing," in *Proc. of WASA 2015, co-located with ACII*, Xi'an, P. R. China, September 2015, IEEE, pp. 891–897.

[19] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor," in *Proc. of ACM MM*, Barcelona, Spain, October 2013, ACM, pp. 835–838.

[20] N. Cummins and S. Amiriparian and G. Hagerer and A. Batliner and S. Steidl and B. Schuller , "An Image-based Deep Spectrum Feature Representation for the Recognition of Emotional Speech," in *Proc. of ACM MM*, Mountain View, CA, October 2017, ACM, pp. 478–484.

[21] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic, "AV+EC 2015 - The First Affect Recognition Challenge Bridging Across Audio, Video, and Physiological Data," in *Proc. of AVEC'15, co-located with ACM MM*, Brisbane, AU, October 2015, ACM, pp. 3–8.

[22] Z. Huang, B. Stasak, T. Dang, K. G. Wataraka, P. Le, V. Sethu, and J. Epps, "Staircase Regression in OA RVM, Data Selection and Gender Dependency in AVEC 2016," in *Proc. of AVEC '16*, Amsterdam, NL, 2016, ACM, pp. 19–26.

[23] D. Le, Z. Aldeneh, and E. Mower Provost, "Discretized Continuous Speech Emotion Recognition with Multi-Task Deep Recurrent Neural Network," in *Proc. INTERSPEECH 2017*, Stockholm, Sweden, August 2017, ISCA, pp. 1108–1112.

[24] P. Laukka, H. A. Elfenbein, N. S. Thingujam, T. Rockstuhl, F. K. Iraki, Wa. Chui, and J. Althoff, "The Expression and Recognition of Emotions in the Voice Across Five Nations: A Lens Model Analysis Based on Acoustic Features," *Personality and Social Psychology*, vol. 111, pp. 686–705, 2016.

[25] M. Schmitt, F. Ringeval, and B. Schuller, "At the Border of Acoustics and Linguistics: Bag-of-Audio-Words for the Recognition of Emotions in Speech," in *Proc. of INTERSPEECH*, San Francisco, CA, USA, September 2016, ISCA, pp. 495–499.

[26] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin, "LIBLINEAR: A Library for Large Linear Classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[27] C. Chang and C. Lin, "LIBSVM: A Library for Support Vector Machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 1–27, 2011.

[28] B. Schuller, S. Steidl, and A. Batliner, "The Interspeech 2009 Emotion Challenge," in *Proc. of INTERSPEECH*, Brighton, UK, September 2009, ISCA, pp. 312–315.