1 Reviewer 3

Single-sourced data from YouTube might be limited, however, this is the biggest video platform and we have tried to balance our data. For example, we use single-breed focus to ensure data unity and we are currently working on extension to other breeds and animals. We will update the challenges of extracting meaningful sounds and context from videos in next version.

Data Construction We mainly concentrate on the youtubers who only own one dog so there is no mixed dog vocalisations. This is indeed an assumption and has been partially supported by (Huang et al. 2023). As a full dog sound is long, filled with pauses and several isolated vocalisations, we propose a more fine-grained structure of sentence, word and subword. We will add this in the next version.

To ensure the accuracy and consistency of our definition in dog vocalizations, we mainly use models and human evaluation.

Methodology We choose Shiba Inu because it is widely adopted at home and plenty of videos is available on YouTube as mentioned in Introduction. This framework mainly focuses on the dog itself. We think large amount of data can mitigate the influence of human interactions effectively. Each youtuber may present a specific preference, but large amount of users will even this bias.

Results This variation in phonetic symbols for the same word type may elucidate why one word type can convey multiple distinct meanings, akin to polyseme.

Given the dataset's web-based origin, determining breed, age, or individual personality presents challenges. We only focus on two available factors, location and activity. Our dataset presents an overall image of Shiba Inu instead of their respective characters.

Our work can help further dividing current word types into finer ones that each type conveys an exact semantic meaning. For example, an dog sound interpretation application may be developed and hosts can use it to understand their dogs'.

Our findings have consistently aligned with prior research and initial assumption. Our findings are listed in Table 2.

2 Reviewer 4

Most of our segmentation framework is based on existing methods, however this is because we disassemble a novel problem (deciphering animal language) into sub-steps that can be partially solved by SOTA methods, for the purpose of leveraging extensive research on humans. However, we don't think this weakens the novelty in our work as we are the first work to implement a web-data-driven approach to explore semantics of animal language. The pipeline we proposed is workable and reusable and we have promising results.

Question 1 A subword is akin to a syllable. A syllable is minimally composed of a local sonority maximum, typically represented by a vowel, and optionally, it may include less sonorous sounds in the onset and coda (Räsänen, Doyle, and Frank 2018). "Sonority fluctuation" means that the changes in the wave envelope of a sound with time caused by the variations in the sonority of a syllable.

Question 2 We select the topmost prediction. Our newest activity classification accuracy is 61.4% and this is not low, since the classification is on 14 different dog activities, with some certain exhibiting similarities between each other as shown in Appendix Figure 1 (b). A full activity list contains "Mount Or Hump (beg)", "Play With People", "Sit", "Lay Down", "Walk", "Sniff", "Eat", "Stand", "Take a Shower", "Run", "Be Touched", "Unknown", "Fight With Dogs" and "Show Teeth or Bit". The reason we utilized a fine-grained activity class taxonomy is because we would like to investigate precise correlations between vocalization and certain activity. If we further integrate similar activities of dogs and define only 10 categories, its accuracy will grow to 74.3%.

The dog's activity may undergo several changes in a short period of time, but we only label one activity as ground truth in the test set. The model may also notice other actions. Also, the prediction of dog activity is relatively difficult compared with human activity recognition. The shooting angle is often high, resulting in little variation in the movements of dogs. The distinction between dog movements is less obvious because dogs have shorter limbs. Due to the nature of YouTube videos, we also encounter low resolution of videos and frequent camera movement and transition.

3 Reviewer 6

Question 1 Please refer to Reviewer 4, Question 2.

Question 2 This pipeline can be easily reused to any other animal which makes it a valuable tool for other animal researchers. Our approach offers a new opportunity to explore animal language through large amount web-based data.

We explore "word sequence" to analyse the semantic interconversion of "words". We also try to decipher the meaning of dog words, and this work indicates the possibility of dividing words into finer semantic units. Our dataset is the biggest dog semantic dataset and will facilitate dog language understanding in the future. This work could help enhancing our communication with dogs.

Question 3 The task is single-label. We admit that these labels may not be totally mutually exclusive, but in each scenario, there will must be only one main location and one main activity. We decide the label of location by considering the occupying proportion of the image. In Appendix Figure 5, we give pictures to illustrate these scenarios.

Question 4 The findings is expected to generalize to other dog species. We are consider other species. In this work, we mainly concentrate on Shiba Inu.

References

Huang, J.; Zhang, C.; Wu, M.; and Zhu, K. 2023. Transcribing Vocal Communications of Domestic Shiba lnu Dogs. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics.

Räsänen, O.; Doyle, G.; and Frank, M. C. 2018. Pre-linguistic segmentation of speech into syllable-like units. *Cognition*, 171: 130–150.