Transcribing Vocal Communications of Domestic Shiba Inu Dogs

Jieyi Huang[1], **Chunhao Zhang**[2], **Mengyue Wu**[3*], **Kenny Q. Zhu**[4*]
[1,2,3,4]Shanghai Jiao Tong University, Shanghai, China
[1]xsiling1@gmail.com, [2]forest_zch@sjtu.edu.cn
[3]wumengyue@cs.sjtu.edu.cn, [4]kzhu@cs.sjtu.edu.cn

Abstract

How animals communicate and whether they have languages is a persistent curiosity of human beings. However, the study of animal communications has been largely restricted to data from field recordings or in a controlled environment, which is expensive and limited in scale and variety. In this paper, we take domestic Shiba Inu dogs as an example and extract their vocal communications from a large amount of YouTube videos of Shiba Inu dogs. We classify these clips into different scenarios and locations, and further transcribe the audio into phonetically symbolic scripts through a systematic process. We discover consistent phonetic symbols among their expressions, which indicates that Shiba Inu dogs can have systematic verbal communication patterns. This reusable framework produces the first-of-its-kind Shiba Inu vocal communication dataset that will be valuable to future research in both zoology and linguistics.

1 Introduction

It has long been an interesting interdisciplinary scientific challenge to understand the languages of animals (Hockett, 1959; Radick, 2007; Von Glasersfeld, 1974). Dogs, who are arguably the best friends of humans, have drawn particular attention. Learning what dogs want to express has broad and profound significance, such as in understanding biological evolution (Pongrácz, 2017), for applying their languages to information technology, or sometimes just for satisfying the our curiosity.

Vocal expressions of dogs, being their chief means of communication, have been studied previously. Here we define vocal expressions as all the sounds that a dog can make vocally, including bark, whine, whimper, howl, huff, growl, yelp, and yip. It has been shown that dogs can recognize the scenes and express their understandings of the outer world as well as their inner states by their voices (Molnár et al., 2008; Hantke et al., 2018). The limitations of previous works are from two aspects. On the one hand, previous research treats this task as a simple classification problem, which means that an audio segment containing barks will be straightforwardly sent into one model to get a particular label such as emotion (happy or sad). Although the results of them have shown that dogs have consistent sound patterns for different purposes, they provided little insight into exploring whether dogs have structural languages. The potential linguistic patterns beyond the dog's vocal expressions are dramatically ignored. On the other hand, previous datasets are collected by recording the voices of dogs under certain controlled environments. Such methodology is costly in practice, and the data thus produced is limited in size and variety (as we will show later in Table 1). In this way, it's hard to infer the latent linguistic patterns. The patterns and semantic meanings of some environments not covered in these datasets will not be investigated as well.

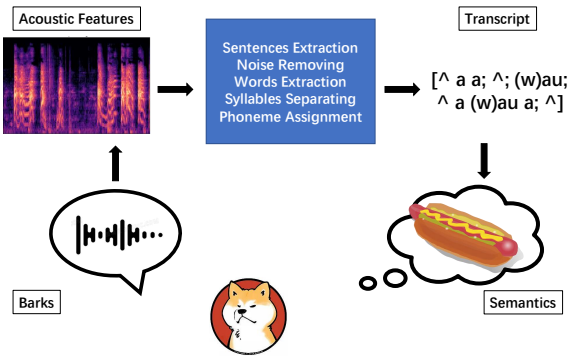


Figure 1: We aim at matching the barks of dogs with its semantic meanings. In our approach, the barks of dogs will be transcribed into symbols.

Even though it is still highly debatable whether animals, or dogs in this case, have languages at all, in this paper, we present an approach pipeline

* Corresponding authors.

to treat dog sounds as a kind of language, similar to human languages. During this process (Figure 1), the specific patterns found in their vocal expressions imply that their barking sounds can carry corresponding semantic meanings just as humans use fixed sound patterns to signify. In this paper, we present a dataset of phonetically symbolic transcripts of Shiba Inu dog barks [1] called *ShibaScript*, which ameliorates some of the aforementioned challenges. We pick Shiba Inu as the subject because it is a popular breed around the world and there are a large number of their videos on the web. Meanwhile, we provide preliminary phonetic analysis on this dataset. We believe that this work is a first step toward investigating whether dogs have sound-actuated language just as humans with speech.

ShibaScript contains barks coming from 16 different Shiba Inu dogs, corresponding transcripts with timestamps of their barks, among which consistent sound patterns are found. These 16 dogs are respectively from 16 families who post these dogs' videos on YouTube. The dataset has a total length of over 4 hours of pure dog sound production, 4469 sentences, and 7761 words. There are in total 9 distinct syllables in these transcripts. Note that due to the ever-evolving nature of social media, the dataset-construction methodology we propose in this paper can be applied to YouTube continuously and yields a dataset that is growing in size and variety. We believe this dataset will help with future research on canine communication as well as any general audience who are interested in learning what dogs want to express.

Our contributions lie in three aspects:

1. we introduce a reusable framework for transcribing animal voices from social media like YouTube, the framework is the first to assign phonetic symbols on dog barks as well as describe dogs' vocal communication in a formal way;

2. we release a novel Shiba Inu voice transcription dataset [2], which is the first of its kind in the CL community;

3. we present some preliminary statistical findings from this dataset. 9 consistent phonetic symbols are discovered, with phonemes/words/sentences being existent, The consistent sound patterns found over the these dogs reveal that dogs may have structural vocal communication patterns.

2 Approach

We now describe the method of constructing ShibaScript. To collect these clean Shiba Inu barks and endow them with corresponding transcripts, a six-step process is used. These steps, in sequence, are getting videos related to Shiba Inu dogs, extracting barks as "sentences" removing barks with noise, extracting barks as "words", separating syllables, and clustering to assign appropriate phonemes based on their acoustic features.

2.1 Collecting Data

In this work, we aim at investigating the language patterns of Shiba Inu dogs. Previous works (Ide et al., 2021; Ehsani et al., 2018; Molnár et al., 2008; Hantke et al., 2018) endeavor to understand dog language patterns conduct experiments on datasets (Table 1) which have limited sizes and scenes. Their frequent approach is to get several dogs and record their barks when dogs are put into the context of different events and in various kinds of places. The disadvantages of this method are three-fold. First, the number of dogs is limited by the budget and practical conditions of these experiments. Second, such an approach can only include several "typical scenarios", and is almost impossible to cover all of the situations that dogs might experience in their daily lives. Third, field study like this is costly in terms of humans, machines, and time. Therefore it is hard to transfer the research to other animal species.

To solve these problems, we make use of the abundant resources from online social media. Each year, millions of videos are uploaded to YouTube, which is the largest video-sharing site around the world. These videos include large amounts of Shiba Inu dogs videos of different scenes uploaded by those who keep them. There are even people who set up an account specifically for dogs and upload hundreds or thousands of their videos. Collecting data from such Shiba Inu enthusiasts can substantially enlarge the number of dogs, cover more scenes, and reduce the cost. And most importantly, researchers can adapt this methodology to other dog breeds or even animal species, which

[1] Here we refer to "barks" in its broadest sense, which includes any vocal expressions coming from a dog.

[2] The complete code and dataset are available at https://github.com/XSiling/ShibaScript.

Name	Type	# of Dogs	Scenes	Activities	Size
Full Dataset (Ide et al., 2021)	video, audio, sensor	-	simulated disaster sites	-	2825s
DECADE (Ehsani et al., 2018)	video, audio, sensor	1	indoors and outdoors	-	4864s
Unknown (Molnár et al., 2008)	audio	14	mostly indoors, street	-	6,646 barks
EmoDog (Hantke et al., 2018)	audio	12	7 fixed types	-	9,447s
ShibaScript	audio, link	16	37⋆	44⋆	14,702s⋆

⋆: The number of scenes and activities in ShibaScript is not fixed and can be expanded as the dataset is continuously collected.

Table 1: Dog-voice data sources used previously. Existing datasets are collected by manual recording. The first two contain videos of various lengths, while the latter two contain a certain amount of pure barks with pauses.

means this approach is highly reusable.

We select 16 users who have uploaded plentiful Shiba Inu dogs videos and have relatively good recording conditions. These videos are the raw data.

2.2 Extracting Sentences

What we care and label transcripts for are the moments when dogs make any vocal expressions. Similar to humans, it is possible to define the sentence in the sound system of dog expressions. The definition is as below: In a sentence, dogs bark continuously on the granularity of seconds. Barks here represent the sounds dogs generate through vocal cord vibration.

In the videos we obtain from different YouTube users, there are a lot of irrelevant and silent frames when the concentrations of videos are not dogs or the dog in the current frame is not barking.

In order to extract the video clips containing vocal expressions of dogs. We use PANNs (Kong et al., 2020), a pretrained large sound event detection model including as many as 527 sound classes that can output audio tagging results as well as events' on- and off- timestamps. Those frames which are tagged with "bark" in the top 10 results are considered to contain barks. We manually labeled 300 samples and compared them with PANNs output, a precision of 0.92 is observed.

2.3 Removing Noises

In constructing the dataset, there is an apparent advantage of recording the audio of dogs in reality: the background noise and the conditions of the recording device can be better controlled. In this work, since we pursue better coverage of the dataset and use the resources from public social media, the problem of noises in the audio samples is inevitable.

To generate the scripts and statistical results more accurately, we have tried our best to produce clean dog bark samples from two aspects: first in Section 2.1, we have selected the users who uploaded videos with less noise and better recording conditions; and second, we use the following approach to significantly remove the noise from our data.

From artificial sampling, we find that the majority of the noises come from either the background music which the user edited into the video, or the human talking while the dog was barking. In order to remove this kind of noise, we make use of the result of PANNs as well. Those frames which are tagged with "speech" or "music" in the top 10 results are considered noisy frames. Sentences that contain noisy frames are filtered out.

2.4 Extracting Words

In the vocal expressions of dogs, there are mainly long pauses and short pauses. A long audio sample can be divided into several sentences with long pauses in between; a sentence can be further divided into several words with short pauses in between like in Figure 2. We can define "words" in dogs statistically: In a word, dogs bark continuously on the granularity of microseconds.

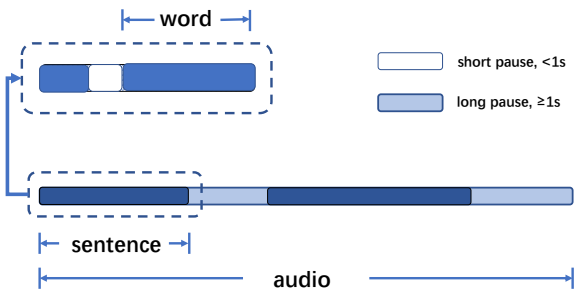


Figure 2: The result of sentence-level and word-level split of a complete audio sample.

As mentioned in Section 2.2, the pre-trained model PANNs (Kong et al., 2020) performed well on the task of sound event detection. Besides the small-grid pauses, there may also be some noise that failed to be filtered in the previous step. To eliminate such small-grid pauses and noise, here

we directly detect the "barking" event from the sentences, and do the word-level splitting based on it. In Hershey et al. (2021), The authors picked out a subset of audios from the original AudioSet (Gemmeke et al., 2017) and assigned "strong" labels to them(about 0.1 sec resolution). The strong-labeled subset of AudioSet results in improved model performance.

We first trained a uniform model from PANNs for sound event detection on the strong-labeled subset of AudioSet. Then to extract words out of the sentence, we annotated strong labels on the event "barking", for 246 sentences with a total length of 715 seconds by the phonetic analysis tool Praat (Boersma and Van Heuven, 2001) and fine-tuned the pre-trained model. As shown in Figure 3, the finetuned model is used to detect the "barking" event and based on the onset and offset of the event, we can extract words from sentences and eliminate the short pauses.

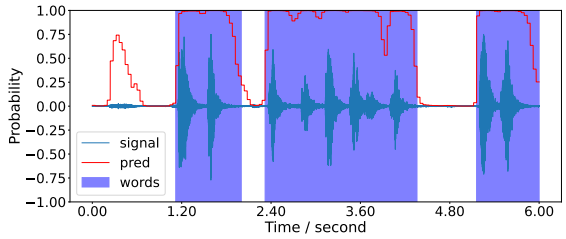


Figure 3: The SED model predicts whether the event "barking" existed in one frame. Words are extracted from the sentences by the onset and the offset.

2.5 Separating Syllables

In human speech, we have the minimum unit as a phoneme that can construct syllables and words, based on which we form sentences with grammatical rules. We retain this setting in exploring dog language and define their barking sounds from the minimal unit, phonetic symbols (Rohrmeier et al., 2015). However, as dogs have different articulatory anatomy from humans, the sounds can be vastly different. We try to label dog sound excerpts with International Phonetic Alphabet (IPA).

In Räsänen et al. (2018), the authors introduce that it is possible to do syllabification even when no priory linguistic knowledge exists. The way to segment speech into syllable-like units depends on sonority to show the edges of syllables (Figure 4). Considering the fact that current dog voices are without any known language patterns, we can adopt this method to separate syllables in one word.

2.6 Clustering and Phonemes Assignment

Given all these syllables and the assumption that dogs have a special system of syllables, we can do clustering and matching up to find a coexisting alphabet for Shiba Inu dogs. As these 16 dogs have different sex, ages, and physical conditions, we conduct Spectral Clustering (Von Luxburg, 2007) on syllables from one certain dog respectively. The feature we use is Filterbank (Strang and Nguyen, 1996). Generally, we set the number of clusters according to the number of videos of each dog, from 10 to 20 (the more videos, the more clusters). The clustering results after dimension reduction can be seen in Figure 5:

After clustering, we have found that compared to human languages, dogs have fewer phonetic categories, which is understandable because humans have a more complex vocal system. Aggregating all the clustering results together, we refer to IPA for illustration and find nine consistent syllables (Table 2). After setting up the syllables dictionary, we can reversely get the words transcripts with short pauses, sentences transcripts with long pauses and audios transcripts with pauses.

Symbol	Description
[a]	Steady pronunciation
[^]	Contains strong sounds
[i]	Stridulation
[u]	Lasts for short
[u:]	Lasts for long
[k]	Sounds like knocking
[en]	Sounds like [ng] in English
[au]	Ends with sounds like [o]
[(w)au]	Starts with [w]

Table 2: The nine types of syllables as well as the syllables description of Shiba Inu. Every description is a clickable hyperlink to an actual sound sample.

A typical symbolic transcript of one audio sentence can be in Figure 6.

3 Dataset

3.1 Data Scale

With the hierarchical structure as audios, sentences, words and syllables, we have given each of the barks of Shiba Inu dogs symbolic transcripts. The distributions of each tier are shown in Table 3. As the whole videos are got from open public media YouTube, they contain a large excess of non-labeling fragments, when the dog doesn't bark or some noise such as human speech and background music. What we concern more are those barking

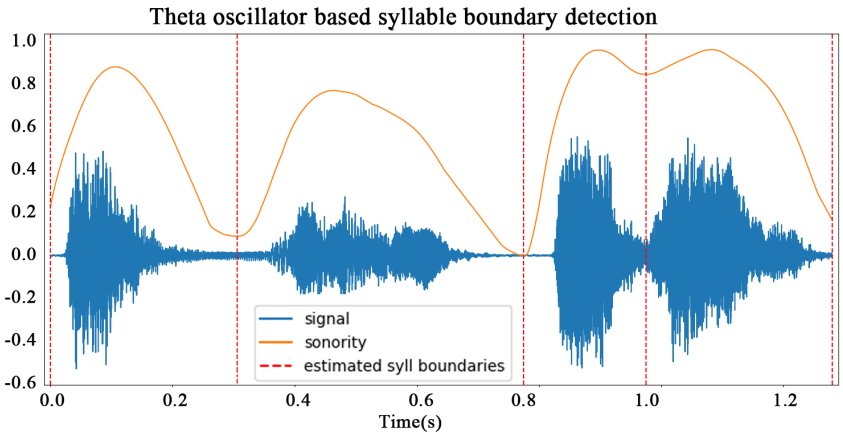


Figure 4: Here a word is separated into four syllables based on the sonority. The complete transcript of this dog is shown in Figure 6.

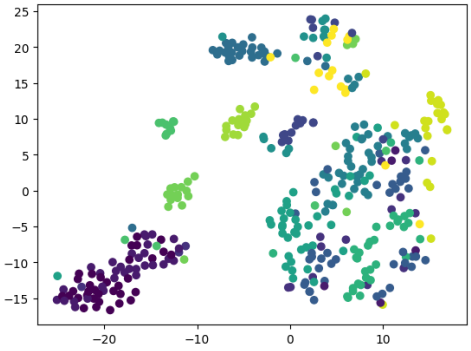


Figure 5: 2-D Visualization of spectral clustering of one dog's data using t-SNE. The complete clustering of all dogs can be checked in Section A.

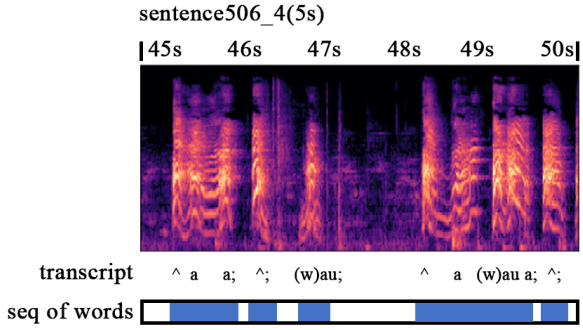


Figure 6: The script of the sentence in the introduction, containing the id of this sentence, the source audio id, the time of this sentence in the audio, the 5 words and their information in this sentence. Each word in the "transcript" is splitted by ";".

fragments, that is the "sentences." We can take the length of sentences as the length of our dataset. At the same time, because we obtain our data from YouTube, the dataset size can grow over time with more users uploading videos.

3.2 Data Variety

Shiba Inu is a very common and lively breed of dog, many people like them and keep them as pets. Those hosts live with their dogs and record their daily lives through videos. As the dataset ShibaScipt is transcribed from the audios extracted from life recording videos on YouTube, the dogs may appear in a variety of common and even uncommon scenes rather than a limited set of scenes, and they may be doing many activities. Therefore, ShibaScript covers a very diverse set of scenes and activities, including 37 different scenes and 44 different activities for dogs. What's more, unlike other datasets which record audios in fixed scenes or manually, the scenes and activities covered by ShibaScript can be expanded as the dataset is continuously collected.

Figure 7 shows the scenes and activities covered by ShibaScript. We can find that there is a subset of the activities that appear in the vast majority of users' videos. For pet dogs, their daily activities such as walking, running, and sleeping are essential and common, and their owners may also record these activities, so these daily activities are covered by most of the users. This holds for the statistical results of scenes as well, that common scenes in daily life like "quilt", "road", "bedroom", "dog bowl" appear in the vast majority of users' videos. Benefiting from the large number of videos used to transcribe the dataset, ShibaScript covers the vast majority of everyday scenes and activities.

Besides, there are some activities and scenes that appear rarely in the statistics. These activities and scenes are shown as "others" in Figure 7. There are

DogID	Sentence		Word	
	Num	Duration (s)	Num	Duration (s)
0	40	135	65	27
1	52	157	77	47
2	55	171	123	65
3	56	224	107	94
4	87	299	134	101
5	118	350	324	144
6	115	374	217	98
7	158	514	241	129
8	130	562	257	147
9	135	566	316	143
10	188	570	320	203
11	255	795	408	157
12	346	1107	577	363
13	553	1643	847	469
14	993	2930	1719	749
15	1188	4306	2029	1372
avg.	**279**	**919**	485	269

Table 3: The basic statistical information of ShibaScript.

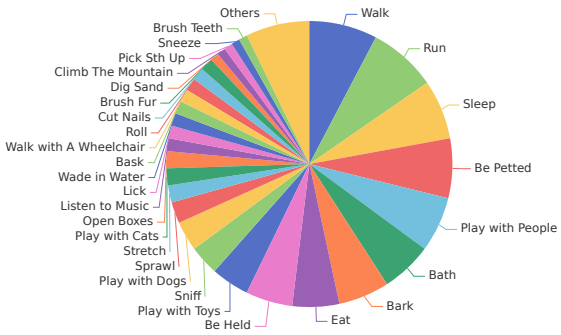
two possible reasons why an activity or scene appears infrequently. First of all, it is highly possible that this activity or scene is related to the personal characteristics of the user. For example, a dog has to wear a cone collar to prevent the dog from licking the wound, so the activity "wear a cone collar" appears only when the dog has had surgery, and this event is not a common one. The second reason is that users have different shooting habits, and a user may only record videos in certain scenes or activities. For example, some users only take indoor videos, so some outdoor activities and scenes like "dig sand" and "beach" are not possible to be covered in their videos, even if the dog actually participated in those activities or scenes. These activities and scenes with personal characteristics greatly expand the diversity of ShibaScript, so that it can cover some non-daily activities and scenes. Benefiting from the wide range of dogs, we can investigate a universal sound pattern of dogs, as they are extracted via them doing various activities under different scenes.

4 Analysis

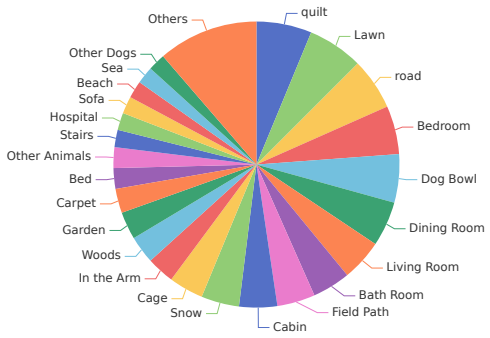
We present preliminary statistical findings from ShibaScript, including lexical analysis and transcribing accuracy evaluation.

4.1 Lexical Analysis

During the transcribing, there are in total 11 types of tokens, in which 9 types are phonetic symbols (Table 2), the other two are short pauses between words and long pauses between sentences.



(a) Activities covered by ShibaScript. The activities which occurrences is less than 5 were merged into "others", see Section C for details.



(b) Scenes covered by ShibaScript. The scenes which occurrences is less than 1 were merged into "others", see Section C for details.

Figure 7: The activities and scenes that covered by ShibaScript. The area of the patches represents the number of dogs producing this symbol.

Similar to humans, the length of these tokens contain ample information. The exact lengths of tokens are kept in ShibaScript for concrete analysis. Because the long pauses are largely determined by the scene at that time, the numerical analysis of it will not be included here.

The mean and variance of each token length can be seen in Table 4. In which we find that almost every phonetic symbol has a similar length of 0.35s or so. Except for the phonetic symbol `[u:]`, which is a prolonged sound owning an average length of 0.45s. While phonetic symbol `[k]` is a relatively short-lived symbol, only having 0.24s average length.

Considering the monogram (Figure 8) of ShibaScript, we can find that the most frequent symbol is `[en]`, which reaches to 3478 times in ShibaScript, the following two are `[au]` and `[a]`, which reaches 1981 and 2011 times respectively.

Symbol	Mean len (s)	Variance (s)
[au]	0.35	0.022
[a]	0.35	0.017
[^]	0.34	0.017
[u:]	0.45	0.054
[u]	0.35	0.030
[i]	0.33	0.020
[k]	0.24	0.009
[(w)au]	0.34	0.018
[en]	0.36	0.032
short pause	0.57	0.335

Table 4: The mean and variance of the duration of 9 phonetic symbols and short pauses between words.

One of the reasons why [en] exceeds much, which is counterintuitive, is that symbols such as [a], [au], [(w)au] are divided up. The least frequent symbol is [k], which only reaches 15 times. This is because dogs seldom produce air-sounds like [k].

At the same time, we can find that these phonetic symbols exist in multiple dogs' sounds, showing that these 9 symbols are universal.

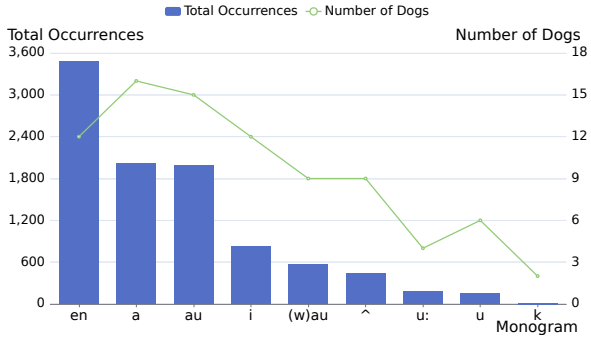


Figure 8: The occurrences of each monogram. The blue bars show the occurrences across the whole dataset of each monogram in ShibaScript, the green lines show the numbers of dogs producing the symbols, from 1 to 16.

After analyzing the monogram, we come to find the relationship between symbols, as well as the bigram (Figure 9) of ShibaScript. Among these bigrams, several appear extremely frequently. It shows a possibility that they are associated with some common semantic meanings. We will dive into that in the future works. Due to space constraints, the detailed information of bigram is shown in Section B.

4.2 Accuracy of Transcription

In this paper, we discover the certain phonetic pattern of Shiba Inu dogs and assign a vocal dictionary of 9 symbols, which is a first-step trial in this

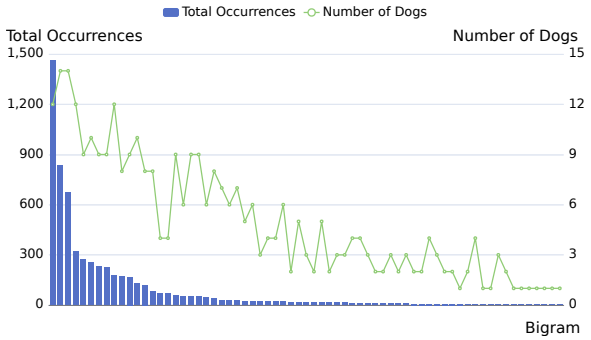


Figure 9: The occurrences of each bigram. The blue bars show the occurrences across the whole dataset of each monogram in ShibaScript, the green lines show the numbers of dogs producing the symbols, from 1 to 16.

area. To better evaluate the phonetic symbols set as well as the integral accuracy of our transcribing, we have done an evaluation test on these two aspects. The evaluation metric is 5-level Mean Opinion Score (Viswanathan and Viswanathan, 2005). Three raters will give scores to either one syllable or one word according to Table 5.

Score	Description
5	The label exactly matches up.
4	Some difference exists between the label and the sound. Humans are sometimes hard to distinguish.
3	Difference exists between the label and the sound. Humans can tell the difference immediately.
2	Although the label is obviously wrong, there is similarity between the label and the sound.
1	The label is totally wrong.

Table 5: The evaluation metric of rating, which is similar to MOS in speech synthesis evaluation metric.

4.2.1 Phonetic Symbol Accuracy Evaluation

For each syllable category, we select 50 syllables randomly. The rating result is in Figure 10. The Fleiss Kappa (Kılıç, 2015) between three annotators is 0.609.

4.2.2 Word Accuracy Evaluation

For the word accuracy evaluation, we select 30 words for each dog randomly and find the same person who rates for phonetic symbols to score for them. The rating result is in Figure 11. The Fleiss Kappa between three annotators is 0.516.

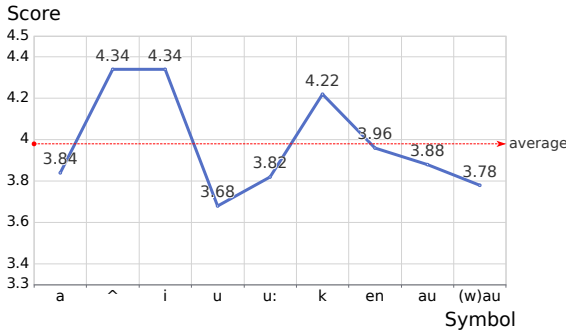


Figure 10: The evaluation result of 9 phonetic symbols.

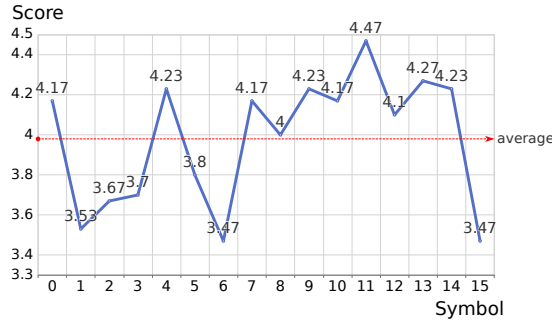


Figure 11: The evaluation result of words for 16 different dogs.

5 Related Work

Early works on understanding animal communications have never reached a point of maturity, which have direct connections between their vocal or literal expressions and their meanings. In these works, researchers attempted to interpret animals in a certain aspect through classifications. Among animals, dogs are popular as research subjects. Considering their vocal expressions, these researches can be divided into mainly three kinds: activity understanding (Ide et al., 2021; Ehsani et al., 2018; Molnár et al., 2008), emotion understanding (Hantke et al., 2018; Paladini, 2020) and individual understanding (Larranaga et al., 2015).

The situation above comes from two reasons. The first is that we are short of ample dataset related to the expressions of dogs, and the second reason is that we have never mastered, or seldom investigated the language patterns of dogs.

In some datasets (Parkhi et al., 2012; Iwashita et al., 2014; Abu-El-Haija et al., 2016) related to visual information of dogs, abundant data was collected from the Internet, which saved the cost and made the data extensible. Compared to that, previous vocal-related datasets depended on manual recordings, which limits the context and costs a lot. Given this, a thought is that we can utilize data on the Internet when collecting vocal-related data if we design a systematic process to extract useful fragments.

In the meantime, previous research adopted a straight-forward classification method, thus lacked enough investigation into the potential sound patterns of dogs. While lexical analysis (Yule, 2022) is the fundamental step for language processing, another thought is that we can set up an own "alphabet" for dogs and transcribe barks of dogs into readable tokens for further research.

6 Conclusion

In this work, we introduce an unprecedented approach for transcribing vocal communications of Shiba Inu dogs and release a corresponding dataset ShibaScript. Compared to the former approaches, it can save a lot of cost and make the dataset extensible. The approach can be transferred to other animals easily. And most importantly, the method is the first-step into investigating the vocal patterns of dogs, bringing inspiration to the field of animal understanding.

We also make some preliminary statistical evaluation and analysis on ShibaScript. The evaluation shows that our symbol assignments in those transcripts are consistent. In the analysis part, we have shown some interesting findings related to the lexical distribution. For future work, we can further research the semantic meanings of dog vocal expressions because we have obtained the corresponding videos of dog vocal expressions.

Limitations

Dataset Noise As the audios are obtained from the videos on YouTube, the quality of the videos will have an impact on the quality of the final transcript. For example, inferior recording equipment may affect the quality of the sound, although we have done noise removal to keep the quality, the presence of background noise will cause some losses in the transcribing process.

Relationship Between Transcripts and Scenes In this work we get the transcript of Shiba Inu dogs, and we also find that the dataset covers a variety of activities and scenes. There may be an interesting relationship between the dog vocal units and the environment including the scene and activity. However, we did not quantitatively analyze the relationship. Considerably more work will need to

be done to discover semantic information in dog barks.

Phoneme Labeling Accuracy In Section 2.6 we cluster the syllables and assign phonetic symbols to them. Then in Section 4.2.1 we evaluate the result by MOS. It can be seen in Figure 10 that the accuracy score is not very high, which can be improved in our future work.

Ethics Statement

This paper makes use of only open-source video data from YouTube. During the transcribing we only focus on the dog barkings, make no use of the personal information of the users, so the released dataset ShibaScript does not contain any personal information, hence doesn't breach the privacy of any persons.

Acknowledgments

Kenny Q. Zhu was supported by the CMB Credit Card Center & SJTU joint research grant, and Meituan-SJTU joint research grant.

References

Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*.

Paul Boersma and Vincent Van Heuven. 2001. Speak and unspeak with praat. *Glot International*, 5(9/10):341–347.

Kiana Ehsani, Hessam Bagherinezhad, Joseph Redmon, Roozbeh Mottaghi, and Ali Farhadi. 2018. Who let the dogs out? modeling dog behavior from visual data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4051–4060.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE.

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.

Shawn Hershey, Daniel PW Ellis, Eduardo Fonseca, Aren Jansen, Caroline Liu, R Channing Moore, and Manoj Plakal. 2021. The benefit of temporally-strong labels in audio event classification. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 366–370. IEEE.

Charles F Hockett. 1959. Animal" languages" and human language. *Human Biology*, 31(1):32–39.

Yuta Ide, Tsuyohito Araki, Ryunosuke Hamada, Kazunori Ohno, and Keiji Yanai. 2021. Rescue dog action recognition by integrating ego-centric video, sound and sensor information. In *International Conference on Pattern Recognition*, pages 321–333. Springer.

Yumi Iwashita, Asamichi Takamine, Ryo Kurazume, and Michael S Ryoo. 2014. First-person animal activity recognition from egocentric videos. In *2014 22nd International Conference on Pattern Recognition*, pages 4310–4315. IEEE.

Selim Kılıç. 2015. Kappa testi. *Journal of Mood Disorders*, 5(3).

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. 2020. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894.

Ana Larranaga, Concha Bielza, Péter Pongrácz, Tamás Faragó, Anna Bálint, and Pedro Larranaga. 2015. Comparing supervised learning methods for classifying sex, age, context and individual mudi dogs from barking. *Animal cognition*, 18(2):405–421.

Csaba Molnár, Frédéric Kaplan, Pierre Roy, François Pachet, Péter Pongrácz, Antal Dóka, and Ádám Miklósi. 2008. Classification of dog barks: a machine learning approach. *Animal Cognition*, 11(3):389–400.

Aleida Paladini. 2020. The bark and its meanings in inter and intra-specific language. *Dog behavior*, 6(1):21–30.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. 2012. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE.

Péter Pongrácz. 2017. Modeling evolutionary changes in information transfer: Effects of domestication on the vocal communication of dogs (canis familiaris). *European Psychologist*, 22(4):219.

Gregory Radick. 2007. *The simian tongue: the long debate about animal language*. University of Chicago Press.

Okko Räsänen, Gabriel Doyle, and Michael C Frank. 2018. Pre-linguistic segmentation of speech into syllable-like units. *Cognition*, 171:130–150.

Martin Rohrmeier, Willem Zuidema, Geraint A Wiggins, and Constance Scharff. 2015. Principles of structure building in music, language and animal song. *Philosophical transactions of the Royal Society B: Biological sciences*, 370(1664):20140097.

Gilbert Strang and Truong Nguyen. 1996. *Wavelets and filter banks*. SIAM.

Mahesh Viswanathan and Madhubalan Viswanathan. 2005. Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (mos) scale. *Computer speech & language*, 19(1):55–83.

Ernst Von Glasersfeld. 1974. The yerkish language for non-human primates. *American Journal of Computational Linguistics*.

Ulrike Von Luxburg. 2007. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416.

George Yule. 2022. *The study of language*. Cambridge university press.

A Clustering Visualization

The full results of clustering can be seen in Figure 12.

B Bigram Statistical Result

Because of the space restrictions, we don't show the detailed results in the main paper. The complete result is in Table 6.

Bigram	Freq.	Co.	Bigram	Freq.	Co.
en en	1464	12	a a	835	14
au au	674	14	i i	319	12
au en	274	9	en au	256	10
(w)au (w)au	231	9	^ ^	224	9
a en	175	12	au a	173	8
a au	168	9	en a	130	10
i en	120	8	en i	80	8
u: u:	70	4	(w)au en	68	4
a i	55	9	en (w)au	52	6
i au	52	9	au i	50	9
u u	43	6	i a	41	8
^ a	29	7	a ^	28	6
a (w)au	25	7	en u	22	5
(w)au a	21	6	au u:	21	3
u en	20	4	(w)au au	20	4
i ^	20	6	u: en	17	2
^ en	16	5	au u	15	3
au (w)au	15	2	^ i	14	5
^ au	14	2	i u:	14	3
u: au	13	3	en ^	11	4
(w)au i	10	4	u: i	9	3
en u:	8	2	i u	8	2
^ (w)au	8	3	^ u:	8	2
u au	7	3	k k	6	2
au ^	6	2	(w)au ^	6	4
i (w)au	6	3	u i	6	2
(w)au u	6	2	u: a	6	1
u a	5	2	u (w)au	5	4
a u:	5	1	u ^	4	1
^ u	4	3	a u	4	2
k a	2	1	k en	1	1
au k	1	1	a k	1	1
en k	1	1	k au	1	1
u: u	1	1			

Table 6: The frequency and coverage number of 16 dogs' bigrams. Here Freq. represents for the frequency of one certain bigram, Co. represents for the numbers of dogs who have made this bigram.

C Activities and Scenes Covered by ShibaScript

44 activities and 37 scenes are covered by ShibaScript. The full statistics of them are in Table 7.

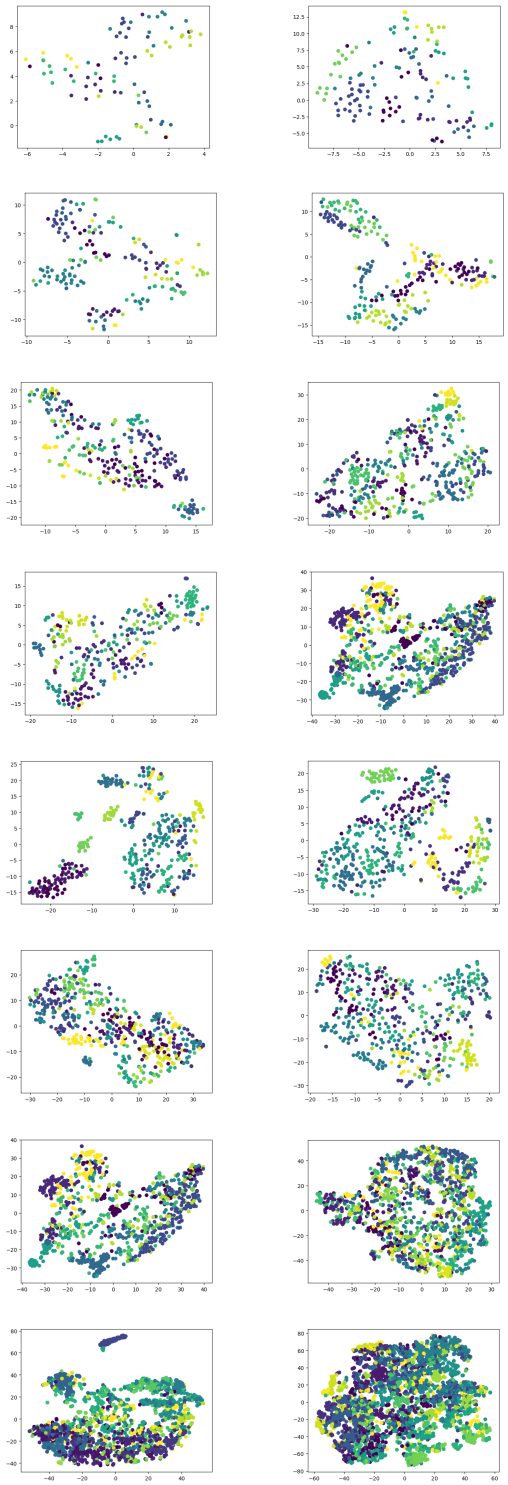


Figure 12: Visualization of Spectral Clustering after TSNE of 16 dogs. The dog's IDs are increasing from left to right, up to down. Phonetic symbols are assigned to different clusters.

DogID	Scene Amount	Activity Amount	Scenes	Activities
0	18	20	bedroom, bathroom, dog bowl, cage, dining room, living room, stairs, hospital, quilt, in the arm, laun, road, other animals, shore, woods, field path, cabin	open boxes, bath, eat, walk, bark, sleep, pick sth up, roll, lick, stretch, play toys, play with dogs, sneeze, walk with a wheelchair, be held, listen to music, play with people
1	19	16	cage, quilt, in the arm, by the fire, dining room, in the arm, hospital, living room, dog bowl, bedroom, road, lawn, snow, stream, field path, other animals, beach, woods, cabin	play with people, eat, be medicated, die, walk with a wheelchair, be held, be petted, sleep, bark, walk, play with dogs, wears a muzzle, run, sniff, wade in water
2	16	12	bedroom, living room, other animals, sofa, quilt, dog bowl, hospital, dining room, stairs, snow, road, lawn, woods, field path, other animals, cabin	walk, run, eat, bark, be held, be petted, play with people, bath, sleep, play with dogs, lick, sniff
3	16	10	living room, carpet, quilt, cat tree, dining room, dog bowl, by the window, lawn, other dogs, road, field path, garden, woods, shore, beach, snow	be medicated, walk with a wheelchair, walk, run, play with cats, sleep, wade in water, bow, bark, stretch
4	18	15	quilt, bedroom, living room, cage, heating pad, dining room, bathroom, door, cage, dog bowl, in the arm, lawn, field path, woods, terrace, stream, snow, road	eat, sleep, sprawl, play with toys, play with people, bath, walk, bask, roll, watch fireworks, be petted, run, sniff
5	21	10	bedroom, living room, dog bowl, bed, quilt, cage, by the window, under the bed, dining room, bathroom, other animals, lawn, beach, shore, woods, lawn, heating pad, field path, road, hill, shrine	sprawl, play with cats, eat, run, bark, be held, open boxes, bath, dig sand, climb the mountain
6	14	13	bedroom, carpet, dining room, bathroom, bed, quilt, door, road, sightseeing bus, other dogs, lawn, snow, cabin, garden	be petted, be vacuumed, sprawl, bark, play with people, listen to music, bath, sleep, walk, run, play with dogs, dig sand, play with toys
7	15	12	carpet, cage, dining room, dog bowl, bedroom, sofa, stairs, door, quilt, in the arm, snow, road, lawn, field path, other animals	walk, run, play with people, play with toys, sleep, dig the snow, sniff, eat, has its teeth be brushed, be petted, be held, hum in the sleep
8	18	17	bedroom, quilt, dog bowl, living room, bed, bathroom, carpet, in the arm, hospital, cage, field path, lawn, snow, road, cabin, other dogs, woods	stretch, sleep, play with people, eat, run, be petted, bath, play with dogs, squat, bark, be held, cut nails, has its teeth be brushed, play with toys, walk, wear a cone collar, pick sth up
9	14	11	door, quilt, living room, carpet, stairs, bed, in the arm, bathroom, cabin, road, garden, lawn, other dogs, hill	play with people, bark, walk, run, sleep, be petted, play with cats, bath, play with toys, lick, climb the mountain
10	19	4	bedroom, dining room, dog bowl, in the arm, quilt, bathroom, cage, other animals, hospital, woods, lawn, field path, sea, beach, garden, cabin, road, mirror	be petted, eat, sniff, sleep, walk, run, cut nails, wade in water, play with people, bath, open boxes, listen to music, surf, stretch
11	13	13	carpet, living room, sofa, quilt, bathroom, by the fire, bedroom, cage, dog bowl, cabin, lawn, garden, snow	play with people, be petted, sleep, bath, has its fur be brushed, walk, run, play with toys, be held, sprawl, bark, wag the tail
12	17	14	living room, sofa, by the window, dining room, quilt, by the fire, bedroom, cage, dog bowl, bed, on the ice, road, lawn, cabin, garden, snow	sprawl, play with people, walk, run, cut nails, blow, eat, play with toys, be petted, stretch, be held, bask, open boxes, sneeze
13	13	15	sofa, bedroom, living room, dining room, bathroom, vacuum, quilt, dog bowl, stairs, road, lawn, cabin, garden	be massaged, play with dogs, bath, be held, play with toys, walk, run, sleep, be petted, pee, has its fur be brushed, be held, bark, roll, sniff
14	15	11	bedroom, carpet, dog bowl, dining room, bed, quilt, bathroom, road, snow, lawn, other dogs, cabin, garden, field path	sleep, walk, run, play with people, be petted, play with dogs, bath, has its fur be brushed, bark, be held, eat
15	15	11	bedroom, living room, in the arm, quilt, dog bowl, cage, dining room, bathroom, by the window, lawn, snow, sea, beach, field path, road	eat, walk, run, sleep, bark, be petted, be held, play with cats, bath, bask, play with people
total	39	44	bedroom, living room, dog bowl, bed, quilt, cage, by the window, under the bed, dining room, bathroom, other animals, stairs, hospital, in the arm, by the fire, cat tree, heating pad, sofa, carpet, door, lawn, beach, sea, woods, field path, road, hill, shrine, shore, cabin, stream, garden, snow, terrace, sightseeing bus, mirror, on the ice, vacuum, other dogs	open boxes, bath, eat, walk, run, bark, sleep, pick sth up, roll, lick, stretch, play with toys, play with dogs, sneeze, sniff, walk with a wheelchair, be held, be petted, listen to music, play with people, die, wears a muzzle, wade in water, be medicated, bow, bask, watch fireworks, play with cats, dig sand, climb the mountain, be vacuumed, sprawl, dig the snow, has its teeth be brushed, hum in the sleep, squat, cut nails, wear a cone collar, surf, wag the tail, blow, pee, be massaged, has its fur be brushed

Table 7: The full statistics for the scenes and activities appearing in each user. The order of the items in column "Scene" and "Activities" is not statistically significant