Paper Instructions and Template for INTERSPEECH 2023

Anonymous submission to INTERSPEECH 2023

Abstract

This paper presents a preliminary investigation on how human language environment influences pet dog barking sound. We first present a new dataset of Shiba Inu dog barks from YouTube, which contains 9,200 audio pairs, each with two audio clips under matching scenarios from either English or Japanese host language environments. (Kenny: It's a bit strange to mention confounding factors here.) These scenarios are confounding factors including activity, location, scene that might give rise to difference in dog barking. With a classification task and prominent factor analysis, we discover significant acoustic differences on the dog barks from two language environments. We further identify some acoustic features that are potentially related to human language phonetic patterns.

Index Terms: dog barks, human language influence, bark dataset EJShibaVoice

1. Introduction

Understanding the language of animals is an interesting interdisciplinary scientific challenge, in particular pet dogs, who are closely integrated with the humans. Previous research endeavours to comprehend dog language for a number of reasons, such as for better understanding animal biological evolution[**?**], applying their language to information technology, or just curiosity about dogs' intention when they bark [**?**, **?**]. However, this task is challenging not only due to unknown language pattern of dogs, but also the lack of high-quality dataset.

In this paper, we investigate dogs' spoken language, which is their barking sounds, one of the main communication channels. One ubiquitous feature of any natural language is that it does evolve with the interaction between the environment and creatures around it[**?**]. Previous research has demonstrated that dog's language indeed reflects their intrinsic characteristics [**?**, **?**], emotional expression [**?**, **?**, **?**] and scene understanding [**?**, **?**]. (Kenny: What do you mean by reflecting scene understanding?) However, little research has looked into the influence arising from the interaction between human hosts and dogs. In our work, we hypothesize that the language of dogs can be shaped by such interaction, particularly the host's language. (Kenny: Because our final conclusion is that dog's language is highly "correlated" with human's linguistic features, we might not be able to say that dog language is "shaped" by the host language.) To verify that, we explore the barking difference of a particular dog species (ShibaInu) from two different host language environments (English vs. Japanese) (Figure 1).

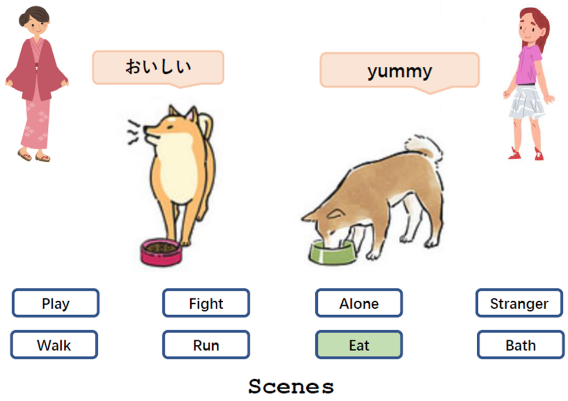


Figure 1: *(Kenny: Reconsider this caption: During interaction between dogs and their hosts from different language environments, dog barking sounds are possible to be shaped.)*

Since there is no existing dataset, we first construct a dataset called "EJShibaVoice." We choose to study Shiba Inu dogs because there are a large number of Shiba Inu dogs in Japanese and English-speaking communities. As YouTube features a large number of Shiba Inu dogs videos, we designed a framework to crawl Shiba Inu audio clips from both English and Japanese speaking host families with barking sound extraction, and tag them with extra information including their language environment, activity scene and location. (Kenny: Do we have locations?)

Then, we conducted classification experiments to check if there exists any interesting accoustic properties that distinguish dog barks from the one language environment to the other. The classification experiment is performed on the clips paired with similar context to exclude some of the confounding factors which are non-linguistic. To explore the prominent acoustic features, multiple commonly-used audio features are utilized, including filterbank [**?**], ComParE [**?**], GeMAPS, eGeMAPS [**?**], PLP [**?**] and MFCC [**?**].

Finally, to discover the most prominent factors that distinguish dog barks by their language environment, we perform an importance analysis of different factors using Shapley values. The fact that several most important factors have substantial correlations with the acoustic characteristics of hosts' language (i.e., English or Japanese) supports our hypothesis that the host language environment does have (Kenny: a profound influence) on dogs' language.

Our contributions are summarized as follows:

- We define a new task to discover the human linguistic influence on the vocal expression (barks) of Shiba Inu dogs, which can inspire further research on animal languages. (Section 2)
- We construct a pairwise Shiba Inu barks dataset

EJShibaVoice with 9,200 audio clips, produced by dogs from two different language environments: English and Japanese. Each sample contains two dog barking audio clips from the same or different language environments, but recorded under the same scenarios. (Section 2)

- We discovered prominent acoustic differences of dogs from different language environments: Shiba Inus from Japan have lower frequency, while those from English environments (Kenny: Don't we ditch unvoice seg length already?) have longer unvoiced segment length, which correlates with corresponding host languages. (Section 4.1)

2. Problem and Dataset

In this paper, we raised two research questions: 1) do pet dogs from different human language environment bark differently? 2) if so, is their barking related to their host's language in anyway? To answer these questions, the prerequisite is a dataset that contains pure dog barking voices from at least two different countries with distinctly different languages. Considering Japanese and English are two common languages with much difference, and Shiba Inu dogs are widely kept as pets in these countries, they became the targets of this dataset.

To find out if dogs in different cultures bark differently, we built **EJShibaVoice** dataset [1]which is composed of pairwise bark samples from two language environments. To eliminate confounding factors such as locations and activities of the dogs, we select two audio clips of barks with similar context to be paired carefully. Our full pipeline in constructing the EJShibaVoice dataset is introduced as follows.

2.1. Sourcing Dog Barks Online

Since Youtoube contains a large amount of user-uploaded videos about their pet dogs, our first step is to source relevant clips from two countries via web crawling. To further ensure the comparison fairness, we defined several dog activity scenes and compare barking sound under the same scene. By setting eight keywords (play, fight, alone, stranger, walk, run, eat, bath) related to Shiba Inu and their actions, we respectively accessed the videos on YouTube. Accordingly, Shiba Inu barks are extracted from the audio tracks of these videos.

2.2. Dog Barking Extraction

Need to be rewritten.

2.3. Pairing the Barks

In addition to language environment, many confounding factors can influence barking sound such as dog activities and their locations. To constrain these factors, we paired two barks under similar contexts with *keywords*, *locations* (obtained via a scene classification model) and *video content* (generated from image caption and visual question answer models). Specifically, the *keywords* are what we used to search on YouTube. The *locations* are inferred from Inception-ResNet-V2 model[**?**] trained on AI Challenge 2017 Scene Classification dataset, which achieves 94.3% top-3 accuracy. While for video content, we apply image caption and visual question answer (VQA) models from OFA [**?**] to first generate a caption for the image extracted from one clip and then ask the model "what is the dog doing in the image?". The caption results are transformed into word embeddings with BERT pretrained model[**?**] and a cosine similarity is computed to determine whether two captions are semantically similar or different. Two clips from different language environments are considered matching only on the condition of same keywords and locations, and a caption cosine similarity higher than 0.95. Some samples from this dataset are shown in Figure 2.

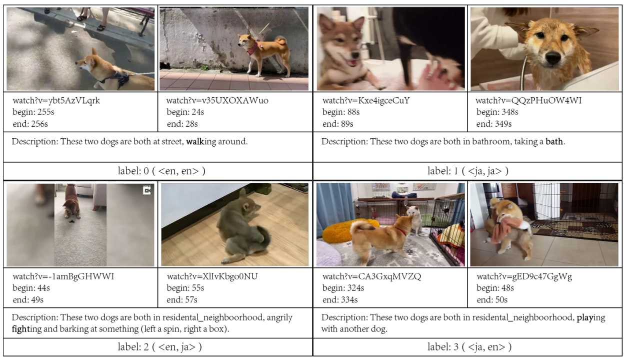


Figure 2: *Examples from the proposed EJShibaVoice dataset.*

3. Experimental Setup

To verify our assumptions that dogs from different human language environments bark differently and the barking difference is related to their host's language, we conduct classification-based experiments and use Shapley value to interpret prominent acoustic features.

3.1. Classwise Classification

Note that in EJShibaVoice, each sample includes a pair of bark clips with similar context and a label (En-En, Ja-Ja, En-Ja, Ja-En). We conduct a four-class classification experiment to validate whether the dog barking differences can be distinguished. If the classification has a high accuracy, we can conclude that dog barks of different host language environments are indeed different. Acoustic features including eGeMAPS, GeMAPS, CompareE, filterbank, PLP, ComParE are used. These features contain different levels of acoustic characteristics. Our classification models include xgboost, KNN, Logistic Regression and Random Forest. All experiments are conducted with 5-fold cross validation.

3.2. Prominent Factors Analysis

To ascertain the influence of host language on dog barks, we analyse prominent factors that distinguish Japanese and English dogs' sounds. Shapley value is commonly adopted to explain feature importance for a given machine learning model, which can help determine the prominent features influencing dog barking.

GeMAPs[**?**] is a widely-utilized acoustic feature-set containing, consisting of statistical computation on acoustic features, which are low-level descriptors that exhibit high interpretability. For its universality and explainability, it is selected as the input features to compute Shapley values and determine our prominent influencing acoustic features.

To compare the relationships between dog barking and host language, we also included features extracted from hu-

[1] The data is available at `https://anonymous.4open.science/r/Ani-3040/`.

man speech (English and Japanese corpus). The speech sources include CommonVoice [**?**] and human speech extracted from the audios from which the dog barks. Similar procedures are conducted on human language and the prominent factors are later compared with those inferred from dog barking (Section 4.2). CommonVoice is a publicly available multilingual speech dataset contributed by volunteers around the world. From its latest Japanese and English corpus, we respectively extract 4000 clips. Regarding human speech from the same videos as the dog barks, we applied PANNs to extract these clean speech clips.

4. Results and Analysis

In this section we present the classification results from different machine learning models with different features extracted from our dataset EJShibaVoice. Then we compare Shapley values on GeMAPs feature to find the prominent factors.

4.1. Classification Results

	xgboost	KNN	LR	RF
filterbank	0.9827	0.9057	0.6374	0.9603
ComParE	0.9868	0.5717		0.9520
GeMAPs	0.9836	0.7317	0.6230	0.9474
eGeMAPs	0.9840	0.7432	0.6901	0.9567
PLP	0.9733	0.7701	0.4375	0.9123
MFCC	0.9828	0.9161	0.5441	0.9587

Table 1: *Classification results on EJShibaVoice dataset. Note that the barking excerpts are under pairing scenarios, excluding other confounding factors.*

The classification results are presented in Table 1, where six different audio features are compared under four classification models. Results indicate that xgboost exhibited the highest classification accuracy, with all six features revealing an accuracy higher than 0.90. High distinctiveness of dog barking from different host language environments is observed.

4.2. Results of Prominent Factors

The results of SHAP values can be seen in Figure 3 and details of the selected dimensions are listed in Table 2.

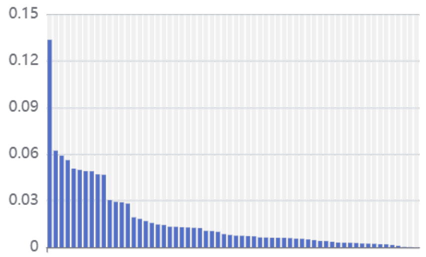


Figure 3: *The average SHAP values (absolute value) of GeMAPs sorting from high to low. Features with SHAP values higher than 0.05 (above the red line, the highest 40% SHAP values) are considered prominent.*

Dimension	Area	SHAP	Pearson
loudness_sma3_amean	spec	0.1336	2.3e-4
F0semitoneFrom27.5Hz_sma3nz_percentile50.0†	temp	0.0622	0.54
loudness_sma3_meanRisingSlope†	spec	0.0590	0.36
logRelF0-H1-A3_sma3nz_stddevNorm†	ener	0.0561	0.62
loudnessPeaksPerSec*	spec	0.0507	0.79
F0semitoneFrom27.5Hz_sma3nz_percentile80.0*	freq	0.0498	0.98
hammarbergIndexV_sma3nz_stddevNorm†	freq	0.0491	0.55
slopeV0-500_sma3nz_amean	freq	0.0490	2.0e-11
loudness_sma3_percentile80.0	spec	0.0470	5.4e-3
slopeV500-1500_sma3nz_stddevNorm*	temp	0.0467	0.93

Table 2: *Details of Prominent Dimensions. SHAP columns refer to the average SHAP values. In the independent t-test of barks from different host language environments, features with* * *have p-values <0.05, those with* † *<0.1, both in bold style, while others are in the range from 0.1 to 0.7. Pearson column refers to the correlation between barks and corresponding speech, which will be discussed below.*

The ten prominent factors above are grouped into four categories: spectral, temporal, energy and frequency according to original GeMAPS definition. Most of the prominent factors fall into Spectral parameters, including dimensions about HammerbergIndex and slope. Respectfully HammerbergIndex represents for ratio of the strongest peak in the 0-2kHz region to the strongest peak in the 2-5kHz region; Slope represents the linear regression slope of the spectral power spectrum within the given band;

F0semitone-related factors describe the pitch, which is also highly related to frequency. The factors about segment length are in the category of temporal parameters. The energy related parameter loudness, which represents for the estimation of sound intensity, is usually largely influenced by the recording device and environment. Here we take it as the audio sampling difference during recording.

Considering these factors, the results show that dog barks from two host language environments have distinctive differences on their energy distribution over frequency. The segment length plays an important role as well. In quantitative analysis, barks from Japanese language environment reveal lower than those from English in frequency and contain longer unvoiced segment length.

In order to better compare the barks with human language, we conduct SHAP analysis on human language: open public language corpus CommonVoice and the human speech extracted from the same videos as we extracted dog barking. To find the difference between these two languages, we use xgboost as well to classify human speech and then compute Shapley values, with results presented in Figure 4.

In the human speech analysis, the difference between English and Japanese concentrates on slope, loudness, segment length and F0semitone. From above, we can conclude that the difference between barks of different host language environments are mainly related to frequency. In the mean time, from our analysis on prominent acoustic factors, barks of dogs and voice of humans share several same prominent factors, which reveals the host human language influence on barks of dogs.

Furthermore, to ascertain the correlation between barks and speech more directly besides the overlap of their prominent factors, we calculate the Pearson correlation between barks and speech extracted from the same videos in the ten prominent dimensions selected . the Pearson correlation is shown in Pearson column in Table 2.

Here we list two typical samples from two language environments in Figure 5. It is clear that barks under English environment contain more unvoiced segments(the blue part) while the frequency of those barks under Japanese environment is

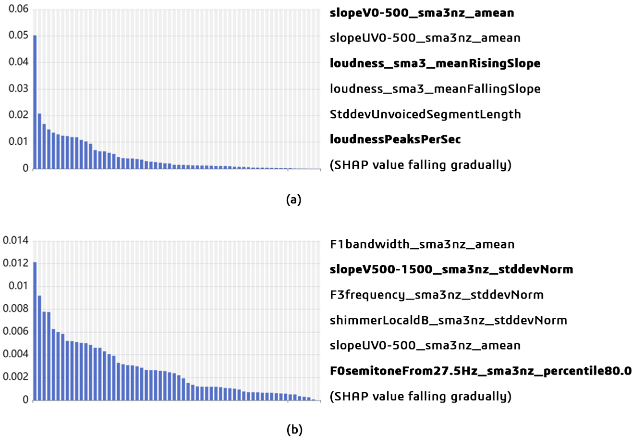
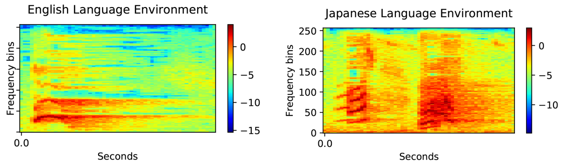


Figure 4: *(a) and (b) are respectfully results of SHAP analysis on open corpus and extracted human speech audios. The first six dimensions are labelled at right, and those who are overlapped with prominent factors in barks of dogs are bold.*

closer to the low frequency region.



(a) Barks under English Env.

(b) Barks under Japanese Env.

Figure 5: *The spectrogram of two audio samples which are from different language environments. There are in the similar situations.*

In addition, we conducted a subjective evaluation for human to bolster our results. The evaluation contains 20 pairs of barks from different host language environments, given out to 30 participants, making up of 600 items in total. Results indicated that 59.00% participants can distinguish differences between a pair, 79.09% agree that the difference lies in pitch, which is auditory perceptive expression of frequency. The results are on pair with our statistical analysis.

5. Conclusion

In this paper, we define a new problem about finding out the cultural influence of the host on the language of Shiba Inu dogs. Experiments have shown that there are significant difference between voice of dogs from Japanese language environment and English environment. By analyzing the acoustic prominent factors, we find that there is correlation between voice of dogs and voice of humans. Our EJShibaVoice dataset will facilitate future research in this field.

6. Acknowledgements

The authors would like to thank ISCA and the organising committees of past INTERSPEECH conferences for their help and for kindly providing the previous version of this template.

As a final reminder, the 5th page is reserved exclusively for references. No other content must appear on the 5th page. Appendices, if any, must be within the first 4 pages. The references may start on an earlier page, if there is space.

7. References

[1] R. K. Moore and L. Skidmore, “On the use/misuse of the term ‘Phoneme’,” in *Proc. INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association*, Graz, Austria, Sep. 2019, pp. 2340–2344.

[2] S. B. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[3] L. R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[4] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning – Data Mining, Inference, and Prediction*. New York: Springer, 2009.

[5] J. Smith, F. Lastname2, and F. Lastname3, “A really good paper about Dynamic Time Warping,” in *Proc. INTERSPEECH 2022 – 23rd Annual Conference of the International Speech Communication Association*, Incheon, Korea, Sep. 2022, pp. 100–104.

[6] R. Jones, F. Lastname2, and F. Lastname3, “An excellent paper introducing the ABC toolkit,” in *Proc. INTERSPEECH 2022 – 23rd Annual Conference of the International Speech Communication Association*, Incheon, Korea, Sep. 2022, pp. 105–109.