

Does My Dog “Speak” Like Me? The Acoustic Correlation between Pet Dogs and Their Human Owners

Anonymous submission

Abstract

How hosts language influence their pets’ vocalization is an interesting yet underexplored problem. This paper presents a preliminary investigation into the possible correlation between domestic dog vocal expressions and their human host’s language environment. We first present a new dataset of Shiba Inu dog vocals from YouTube, which provides 7500 clean sound clips, including their contextual information of these vocals and their owner’s speech clips with a carefully-designed data processing pipeline. The contextual information includes the scene category in which the vocal was recorded, the dog’s location and activity. With a classification task and prominent factor analysis, we discover significant acoustic differences in the dog vocals from the two language environments. We further identify some acoustic features from dog vocalizations that are potentially correlated to their host language patterns¹.

1 Introduction

Understanding animals’ verbal expressions is an interesting interdisciplinary scientific challenge. This is particular true with pet dogs, who closely interact with humans. Previous research endeavors to comprehend dog vocal sounds for a number of reasons, such as for a better understanding of animal biological evolution (Pongrácz 2017), applying their language to information technology, or just curiosity about dogs’ intention when they make a sound (Pongrácz et al. 2011; Jéghe-Czinege, Faragó, and Pongrácz 2020). However, this task is challenging not only due to the unknown acoustic pattern of dogs but also the lack of a suitable and high-quality dataset.

Previous researchers have demonstrated that a dog’s vocalization indeed reflects their individual characteristics (Pongrácz, Molnár, and Miklósi 2010; Larranaga et al. 2015), emotional expression (Thorndike and Bruce 2017; Hantke, Cummins, and Schuller 2018; Paladini 2020) and perception of outside world (Larranaga et al. 2015; Molnár et al. 2008). However, despite the fact that dogs are human’s most familiar animals, little research has looked into the influence on dogs’ communication arising from their interaction with human hosts. As a matter of fact, dogs exhibit many modes of communication ranging from behavioural patterns to vocalizations, in this paper we pay attention to their vocal sounds,

which serve as one of the most important communication channels (Siniscalchi et al. 2018). In our work, we hypothesize that a dog’s acoustic characteristics may be correlated with such interaction, particularly the host’s spoken language. To verify that, we explore the vocal difference of a particular dog breed (Shiba Inu) from two different host language environments (English vs. Japanese) (Figure 1). Shiba Inu dogs are chosen as the subject of this study because Shiba Inus are very popular among dog owners and there is an abundance of their audio/video resources available online. Moreover, we choose to work on dogs in English and Japanese language environments because these two widely-spoken languages have very different phonetic systems, and at the same time, Shiba Inu dogs are very popular in both Japanese and English-speaking households (i.e., in Japan and in the US).

One may argue that the host language is not the only factor that might influence how dogs sound in the online videos. For example, social norms and customs play also play apart in these countries/cultures. We believe these factors are inter-related with the languages that are in question in this paper. Having correlation between dog vocals and human speeches does not nullify the correlation between dog vocals and social norms and cultural behaviors, and vice versa. In fact, these two kinds of correlations can strengthen each other.



Figure 1: A dog’s vocalization may be acoustically correlated with its host language, so that a “Japanese” dog should vocalize differently than an “American” dog under the same context such as “eating on the lawn”.

We verify the above hypothesis via a two-stage pipeline. First, we conduct classification experiments to investigate the possible existence of interesting acoustic properties that distinguish dog vocalizations from one language environment

¹The data, code and a live demo are available at <https://anonymous.4open.science/r/EJSHIBAVOICE-0DC5>.

to the other. The classification experiment is performed on pairs of Japanese and English dog vocal clips under the same *context* which is composed of the *scene category*, *location*, and *activity* of the dog during the recording, so as to exclude these confounding non-linguistic factors, which may affect how dogs voice out.

In stage two, to discover the most prominent factors that distinguish dog vocals by their language environments, we perform an importance analysis of different factors using Shapley values. A similar analysis is also performed on host languages, to study the acoustic difference between English and Japanese. Moreover, we compute the Pearson correlation between dog vocals and their host speech. The fact that several most important acoustic characteristics differentiating dogs have substantial correlations with those of their host language (i.e., English or Japanese) supports our hypothesis that the domestic dog vocal expressions do share a few acoustic similarities with its host language. It is possible that the host language environment has an influence over the dog’s vocalization.

Previous dog vocalization datasets are mostly collected under controlled environment, i.e. the researchers raised several dogs and recorded their physical as well as vocal behaviours (Ide et al. 2021; Ehsani et al. 2018; Molnár et al. 2008; Hantke, Cummins, and Schuller 2018). However, such data acquisition methods are not only costly but also lacking generalization capability as the number of dogs is highly restricted. Since webly data provides a large amount of dog vocalizations with plenty of metadata featuring the dog breed, language environment and contextual information, we derive a pipeline to obtain and filter the vocals from social media, leading to a large-scale of dog vocalization dataset “EJShibaVoice”. The number of families for Japanese and English respectively stand at 219 and 275, much larger than any other similar studies previously reported. Moreover, we use the context information tagged with the vocal clips to eliminate the confounding factors in our experiments. Specifically, we develop a framework that crawls Shiba Inu audio clips from both English and Japanese-speaking host families, extracts the vocal clips, segments the clips into contiguous singular sounds, and tags them with contextual information.

Our main contributions are summarized as follows:

- A newly-defined task to uncover the human linguistic influence on the vocal expressions of domestic Shiba Inu dogs via a unique data-driven and computational approach, which can inspire further research on animal languages. (Section 2)
- We construct a large-scale Shiba Inu vocalization dataset **EJShibaVoice** containing clean audio clips produced by dogs from two different language environments: English and Japanese, including dogs’ host speech clips. The dog vocal and human speech data undergo a systematic pipeline, which extracts clean dog voices and the corresponding host speech from the social media videos. (Section 2)
- We discover prominent acoustic differences between dogs from different language environments: Shiba Inus vocalizations from English-speaking households have a *lower*

frequency, while those from Japanese environments have *faster speed*, which correlates with these two human languages, respectively. (Section 4.1)

2 Problem and Dataset

In this paper, we ask two research questions:

1. Do pet dogs from different human language environments sound differently?
2. If so, is their vocalization related to their host’s language in any way?

To answer these questions, we construct a dataset “EJShibaVoice”, which is composed of Shiba Inu vocal samples from Japanese and English language environments and their corresponding hosts speech. Additionally, each audio sample is also tagged with metadata such as the begin and end timestamps, and the context under which the vocal sound is made. Additionally, host’s speech from the same video is extracted using a similar approach as we process dog vocalizations. This procedure will be described in detail in Section 2.4.

To ensure the high quality of the dataset, we develop a rigorous processing pipeline to extract pure dog sounds in a variety of environments, with little or no extraneous noises such as human speech or background music. We segment dog audio clips into minimal units, which is a full singular bark, for the purpose of fine-grained acoustic comparison between different language environments. We further provide detailed extraneous context for each vocal clip, as there are other variables other than language environments that could potentially influence dog’s sound, for example whether a dog is eating or playing, outdoor or indoor (Larranaga et al. 2015; Molnár et al. 2008). We make this fine distinction on the context of the dog sounds so as to constrain confounding factors other than the linguistic environment of the dogs. Hereby, with the sound segmentation and context tagging, each clean singular vocalization has a corresponding context.

Next, we present the full data processing pipeline (Figure 2) for constructing the EJShibaVoice dataset.

2.1 Sourcing Dog Vocalization Sounds Online

Since YouTube contains a large amount of user-uploaded videos about their pet dogs, our first step is to source relevant clips from two language environments via searching for the keyword “Shiba Inu” and a scene category term on YouTube. In this study, we use a total of eight scene categories, which are defined in Table 1. These eight categories were proposed in previous work about dog activity classification (Larranaga et al. 2015). For example, to download videos about Shiba Inu eating, we search for the query “Shiba Inu Eat -Coin” for English and the corresponding Japanese of this query for vocals under Japanese host language environment. Here “-Coin” is used as there are a large number of Shiba Inu Coin videos, we add this term to filter out those irrelevant videos. All eight scenes are searched under the same keyword patterns.

Because YouTube videos are not typically tagged with original language or locale, we infer the language environment of a video by the languages of its caption or title. For example, if the title contains Japanese characters, we consider

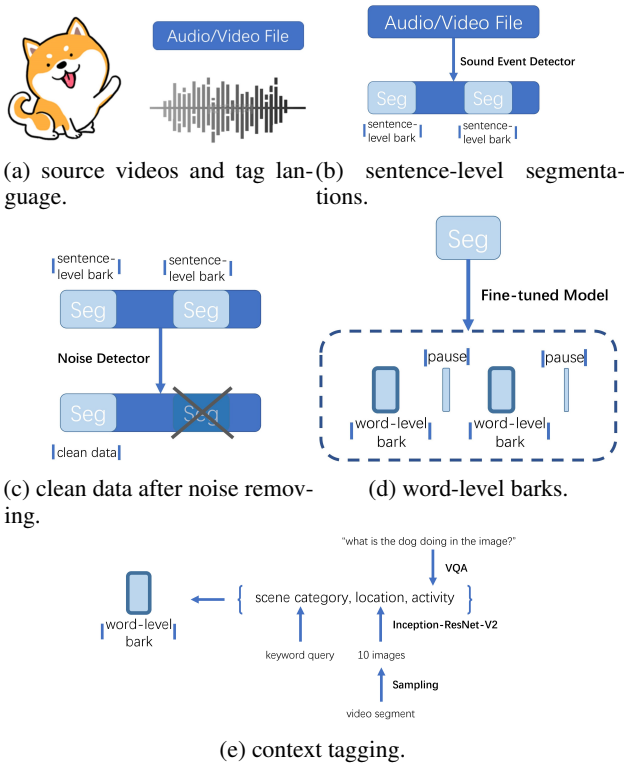


Figure 2: The procedure of dataset construction.

it as a video recorded in an Japanese-speaking environment, whereas a video with purely English title is considered to be recorded in an English-speaking environment.

After all the videos are downloaded and tagged, 16,161 videos in total are collected. Among them we randomly sample 1,000 clips to verify the host language tagging accuracy by watching the video clips. The accuracy is 93.2%, which indicates that the majority of the Shiba Inu videos are tagged with the correct host language environments, and fit our research purpose.

2.2 Dog Vocalizations Extraction

The original audios downloaded from YouTube contain long uninformative segments where dogs are silent or background noises muffling the vocalizations. To ensure the quality of our dataset, we have adopted a systematic and rigorous pipeline of three steps to extract pure and singular vocal segments from the audios.

Step 1: sentence-level segmentations extraction As a preliminary step, we extract long segments in which contains a continuous series of dog vocalizations. Such long segments, which we call a dog "sentence", can be detected because they are preceded or followed by significant periods of silence. We apply PANNs (Kong et al. 2020), a pre-trained large sound event detection model including as many as 527 sound classes for recognizing the sound events. The continuous segments detected by PANNs with the event class "barking" are considered as the sentence-level bark segmentations.

Step 2: noise elimination From our practical experience, we have found that sometimes vocalizations will be unclear due to the background music and human talking. Thus in the second step, among those sentence-level segmentations, those with co-existing event "speech" and "music" detected by PANNs are removed from the dataset.

Step 3: word-level vocalizations extraction Another challenge is that the coarse-grained sound clips may have some short pauses in the middle. To remove these pauses, we finetune a sound event detection model to determine the start and end time of singular vocalizations, which we call "words" in this paper.

PANNs is used as the pretrained model, which is trained on AudioSet (Gemmeke et al. 2017) for all the sound classes first. We then manually made framewise labels for the event "barking" on 246 audio clips with a total length of 715 seconds and fine-tuned this pretrained model. This fine-tuned model can precisely detect dog vocals in the audio with the start time and end time. Based on the precise start time and end time, we use the fine-tuned model to remove the pause segments between barkings.

At the end of these three steps, only a singular dog vocalization is left in each clip. In total, 7,500 clips from 1,551 raw videos remained. While we acknowledge the fact that the data noise induced by recording devices and audio ambience is difficult to eliminate via web data, such factors/biases should be well mitigated with the large and diverse data source coming from either side of the Pacific that we curated on YouTube.

2.3 Context Metadata Tagging

We list the range of possible values for *scene*, *location* and *activity*, three components of the context in Table 1. The set of values for locations values is a subset of the location set from Inception-ResNet-V2 model.

Context	Possible values
Scene	Alone, Bath, Eat, Fight, Play, Run, Stranger, Walk
Location	76 possible locations, such as lawn, street, office, recreation_room and so on
Activity	a 768-dimensional real valued vector

Table 1: Description of the context metadata.

For locations and activities, we sample 10 images (equi-spaced) from the video segment that coincides with the duration of the audio bark sample, and classify the images into one of the location labels listed in Table 1. The locations are inferred from Inception-ResNet-V2 model (Szegedy et al. 2017) trained on AI Challenge 2017 Scene Classification dataset. We apply image caption and visual question answering (VQA) models from OFA (Wang et al. 2022) to first generate a caption for an image and then ask the model "what is the dog doing in the image?". The caption results are transformed into word embeddings with pre-trained BERT model (Devlin et al. 2019), and the 10 word embeddings of the 10 images are averaged to obtain the overall activity embedding of that bark sample. Besides, the timestamp of each

bark clip in the original video is provided. Some samples from this dataset are shown in Figure 3.

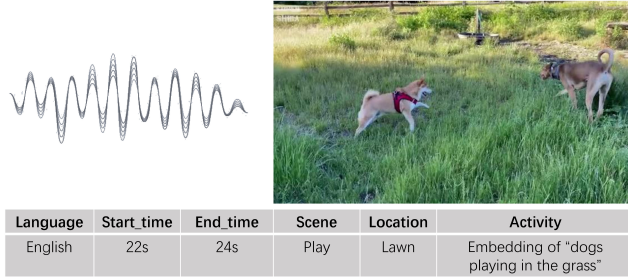


Figure 3: One audio sample in EJShibaVoice with its meta-data of language environment, timestamps, and context.

2.4 Host Speech Extraction

To address the second research question, which is to find the correlation between the barks and their host languages, we need the audio speech data from dogs’ hosts as well. The human speech in the original audio track is also extracted by a strict pipeline. The first few steps are similar to dog barking extraction: we start by getting those continuous segments tagged as “speech” by PANNs and remove those tagged as “music” from them to reduce the noises. As it is normal for several people talking simultaneously, we further apply a speaker diarization model Pyannote(Bredin et al. 2020; Bredin and Laurent 2021) to separate them from each other and retain speech of a single person in more fine-grained segments.

2.5 The Final Product

The dataset consists of two parts: the barks and the corresponding host speech. After the whole processing, 7,500 clean bark sounds and 15,197 corresponding host speech clips finally remain in our EJShibaVoice dataset (Table 2).

	# of Clips	Avg Len (s)	Var of Len (s^2)	English Per(%)
Bark	7500	0.61	0.288	46.72
Speech	15197	1.56	2.134	44.01

Table 2: EJShibaVoice Statistics.

As the data statistics indicate, we include roughly equal number of audio clips from English and Japanese environments, leading to a balanced dataset that can well support our investigation on the influence of language environment.

The number of clips of each scene differs (Figure 4). As we use the same method to process each scene, the possible reason for imparity is that Shiba Inu barks diversely under different conditions. For example, dogs tend to bark more to show their strength when fighting, therefore the number of clips under “fight” is the largest.

3 Methodology

To verify our assumptions that dogs from different human language environments have different vocals and the difference

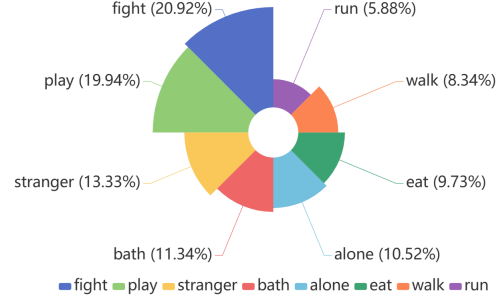


Figure 4: The percentage of bark clips for different scenes.

is related to their host’s language, we conduct classification-based experiments and analyze the Shapley value to find which features are important in distinguishing their vocals.

3.1 To Answer the First Question: Pairwise Classification

To better control context and language environment, we adopt a 4-way classification for pairwise comparison. Specifically, we pair the dog vocalizations from the same context but under different language environments in 4 ways: En-En, Ja-Ja, En-Ja, and Ja-En. Two clips are considered to be from the same context if they have the same scene category, the same location, and their activity vectors with a cosine similarity of 0.95 or above. The problem is then defined as the classification of these pairs into any of the above four classes. Our classification models include xgboost, KNN, Logistic Regression, and Random Forest. Multiple acoustic features including spectral features (filterbank (Strang and Nguyen 1996), PLP (Hermansky 1990), MFCC (Davis and Mermelstein 1980)) and handcrafted feature-sets (eGeMAPs, GeMAPs (Eyben et al. 2015), ComParE (Schuller et al. 2013)) are utilized to feed into the model.

Dogs may also differ by their age and sex, but these attributes are hard to obtain from the YouTube data even manually. Given the large size of our dataset, we believe a significantly higher than random classification accuracy will show that we can distinguish English dogs with Japanese dogs just by acoustic features.

An alternative method is to group all fine-grained clips by the contexts, and then do a two-way (English or Japanese environment) classification in each group using only acoustic features. However given the complex combination of scene, location and activity, few samples from both languages share exactly the same context hence we conducted the more difficult four-class experiment.

3.2 To Answer the Second Question: Correlation on Prominent Factors

To ascertain the influence of the host language on dog vocalizations, we analyze 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 vocalizations.

To compare the relationships between a dog vocalization and the host language, we also include features extracted from human speech (English and Japanese corpus). The speech sources include 8,000 clips from CommonVoice (Ardila et al. 2019), which is an open source multilingual speech dataset contributed by volunteers around the world, and host speech from EJShibaVoice. We adopt two different sets of speech data because they provide distinctive features. Speech from EJShibaVoice has direct relation with vocalizations, so that we can conduct Pearson value analysis between them, while CommonVoice is purer and more common to help us find universal feature of human speech. Similar procedures are conducted on human language and the prominent factors are later compared with those inferred from dog vocalizations (Section 4.2). Furthermore, to ascertain the correlation between vocalizations and their host speech in a statistical way, we analyze the Pearson correlation between them. In the meantime, the Pearson correlation between vocalizations and random speech is shown to compare (Section 4.2).

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. We further compare Shapley values on the GeMAPs feature to find the prominent factors. We also try several different classification models on both spectral and handcrafted acoustic features.

4.1 Results of Pairwise Classification

	xgboost	KNN	LR	RF
filterbank	0.9827	0.9057	0.6374	0.9603
PLP	0.9733	0.7701	0.4375	0.9123
MFCC	0.9828	0.9161	0.5441	0.9587
ComParE	0.9868	0.5717	0.6004	0.9520
GeMAPs	0.9836	0.7317	0.6230	0.9474
eGeMAPs	0.9840	0.7432	0.6901	0.9567

Table 3: 4-class classification accuracy on the vocal pairs.

In total, we form 9,200 pairs of vocal clips. Among these, we formed 4 different pairs, EN-EN, JA-JA, EN-JA, and JA-EN, with 2,300 clips in each set respectively. And we perform 5-fold cross-validation on this paired dataset. The overall classification results are presented in Table 3, where six commonly-adopted audio features are compared using four different classification models. Among these features, filterbank, PLP, and MFCC are extracted from the spectral transformation and have 24, 13, and 13 dimensions respectively. By contrast, ComParE, GeMAPs, and eGeMAPs are human-crafted features and have 6373, 62, and 88 dimensions respectively. These features are easier to explain from perception perspectives, however, they might be less informative than direct spectral features, resulting in relatively lower accuracy in some models.

Our first observation is that no matter which feature set or which model is used, dog vocals from different host language environments can be clearly distinguished, because the accuracies are all clearly higher than 0.25 (the random guess). In other words, this indicates that there is a certain acoustic difference between the dog vocals in these two language environments. Specifically, the accuracy of ComParE significantly drops while using KNN, which is due to the curse of dimensionality of this 6,373 dimension feature. XG-Boost shows the highest classification accuracy, with all six features demonstrating an accuracy higher than 0.90. This result affirms that **dog vocals under English environments are distinctly different from those under Japanese environments**.

Next, we will find out what are the crucial features that distinguish the English dogs from the Japanese dogs.

4.2 Results of Correlation on Prominent Factors

Prominent Analysis on GeMAPs The results of SHAP values can be seen in Figure 5 and details of the selected dimensions are listed in Table 4.

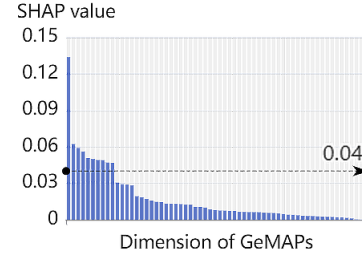


Figure 5: The average SHAP values (absolute value) of GeMAPs sorting from high to low. Features with SHAP values higher than 0.04 are considered prominent because of the sudden drop in SHAP values to the right those features.

Dimension	Dim Type	SHAP
loudness_sma3_amean	Energy	0.1336
F0semitoneFrom27.5Hz_sma3nz_percentile50.0	Frequency	0.0622
loudness_sma3_meanRisingSlope	Energy	0.0590
logRelF0-H1-A3_sma3nz_stddevNorm	Frequency	0.0561
loudnessPeaksPerSec	Temporal	0.0507
F0semitoneFrom27.5Hz_sma3nz_percentile80.0	Frequency	0.0498
hammerbergIndexV_sma3nz_stddevNorm	Spectral	0.0491
slopeV0-500_sma3nz_amean	Temporal	0.0490
loudness_sma3_percentile80.0	Energy	0.0470
slopeV500-1500_sma3nz_stddevNorm	Temporal	0.0467

Table 4: Details of Prominent Dimensions in GeMAPs.

The ten prominent factors in Table 4 fall into four categories: *spectral*, *temporal*, *energy*, and *frequency*, according to the original GeMAPs definition. Most of the prominent factors are spectral parameters, including dimensions of HammerbergIndex and slope. HammerbergIndex represents the ratio of the strongest peak in the 0-2kHz region to the

strongest peak in the 2-5kHz region, while Slope represents the linear regression slope of the spectral power spectrum within the given band. F0semitone-related factors describe the pitch, which is highly related to the fundamental frequency. The factors about segment length are temporal parameters. The energy-related parameter loudness, which represents the sound intensity, is largely affected by the recording device and environment.

Considering these factors, the results show that dog vocals from two host language environments have distinctive differences in their energy distribution over frequency. In quantitative analysis, vocals from the Japanese language environment have a higher frequency than those from English.

In order to better compare the vocals with human language, we conduct a SHAP analysis on human speech datasets: open public language corpus CommonVoice and the host speech extracted from the same videos as the dog vocals. To find out the difference between these two languages, we use xgboost as well to classify human speech into English or Japanese and then compute Shapley values, with results presented in Figure 6.

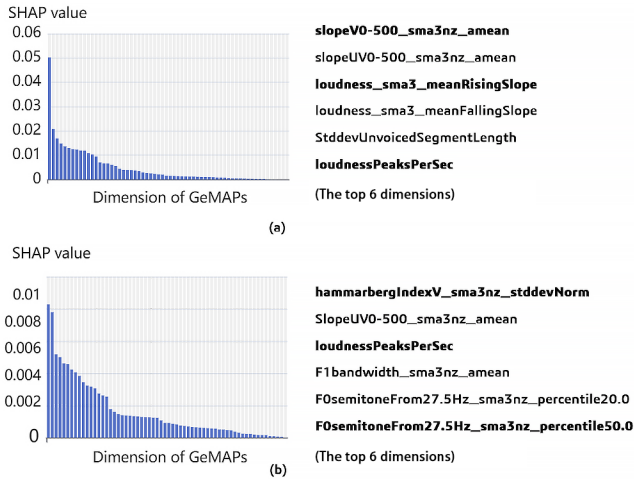


Figure 6: (a) and (b) are the results of SHAP analysis ranked by descending SHAP values on open corpus and extracted human speech audios, respectively. The top six dimensions are shown on the right, and those that overlap with prominent factors in vocals of dogs are bolded.

In the human speech analysis, the difference between English and Japanese concentrates on the slope, loudness, and F0semitone. From the above, we can conclude that the difference between vocals coming from different host language environments is mainly related to frequency. In the meantime, from our analysis of prominent acoustic factors, the vocals of dogs and voices of humans share several common prominent factors, suggesting that the host human language have a correlation with the vocals of dogs.

Furthermore, to ascertain the correlation between vocals and speech more directly besides the overlap of their prominent factors, we calculate the Pearson’s correlation between

the two from the same videos in the ten prominent dimensions selected (Table 5).

Language Speed Comparison Dog vocals under English and Japanese host language environments differ not only in these acoustic dimensions above, but a more intuitive dimension: *speed*.

A formal measure of the speed of speech is the number of syllables in a given time duration. For speech from the same video as the dog voices, we apply an automatic speech recognition (ASR) model in Whisper (Radford et al. 2022), one of the state-of-the-art ASR models which smoothly transcribes the extracted speech into either English or Japanese texts. At the same time, CommonVoice already provides corresponding texts for their audio. The segmentation of text into sequences of syllables is done by some subfunction of *eS-peak* (Duddington and Dunn 2012), which is a compact open source software speech synthesizer for English and other languages. However, as there is no concept of “syllable” nor alphabet for Shiba Inu barks, we adopt another similar way for syllablizing dog clips. The dog clips are divided into sequences of syllable-like units according to an oscillator-based algorithm (Räsänen, Doyle, and Frank 2018).

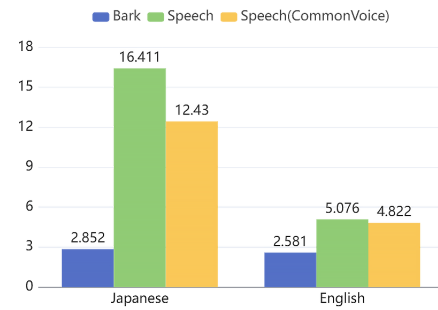


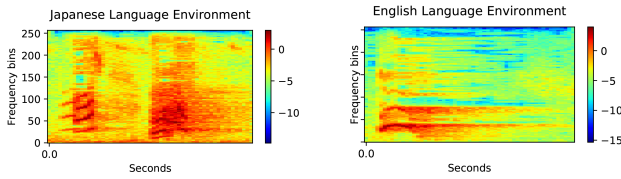
Figure 7: The speed comparison for barks, host speech from same video and speech from CommonVoice.

The number of syllables or syllable-like units per second is shown in Figure 7. We make the following discoveries: First, dog vocals from Japanese environment and Japanese human speech clips are generally faster than dog vocals from English environment and English speech clips. This suggests that when it comes to speed, there is a correlation between dog vocals and human speech. Second, it appears that human speech in EJShibaVoice dataset is faster than that of CommonVoice. The difference in speed can be attributed to the recording style of CommonVoice, which tends to be slower than typical speech.

Here we illustrate the spectrograms of two typical dog vocal samples from the two language environments in Figure 8. It is clear that in the same time length, the vocal under Japanese language environment has more syllable-like units than that under English language environment while the frequency of those vocals in the English environment is closer to the low-frequency region.

Dimension	With Host Speech		With Random Speech	
	Pearson Coefficient	<i>p</i> -value	Pearson Coefficient	<i>p</i> -value
loudness_sma3_amean	0.537	2.73e-83	0.025	0.405
F0semitoneFrom27.5Hz_sma3nz_percentile50.0	0.105	4.76e-4	-0.012	0.691
loudness_sma3_meanRisingSlope	0.159	1.08e-7	0.043	0.155
logRelF0-H1-A3_sma3nz_stddevNorm	0.002	0.941	4.25e-4	0.989
loudnessPeaksPerSec	0.124	3.75e-5	1.20e-3	0.968
F0semitoneFrom27.5Hz_sma3nz_percentile80.0	0.162	6.40e-8	-0.011	0.714
hammarbergIndexV_sma3nz_stddevNorm	9.13e-3	0.762	6.47e-3	0.830
slopeV0-500_sma3nz_amean	0.580	7.37e-100	0.029	0.334
loudness_sma3_percentile80.0	0.436	3.03e-52	0.030	0.322
slopeV500-1500_sma3nz_stddevNorm	0.017	0.573	-4.40e-3	0.884

Table 5: The correlation analysis is on two groups of data. In the first group, the analysis is on the correlation between dog vocals and their host speech, in the second group the correlation is between dog vocals and random speech. The dimensions of high correlation with *p*-value lower than 0.05 are bold.



(a) A Bark under Japanese Env. (b) A Bark under English Env.

Figure 8: The spectrogram of two audio samples which are from different language environments and similar situations.

4.3 Summary of Findings

Through all the above analysis, we can find that **the key acoustic features that distinguish Shiba Inus from two language environments are frequency and speed, which correlate well with human speech.**

Our findings corroborate with previous literature (Coupé et al. 2019; Graham 2014) that the syllable speed of Japanese surpasses that of English and many other languages. At the same time, the frequency of Japanese is relatively higher than that of English. All these findings show that the vocals under English and Japanese language environments are different and in two dimensions of frequency and speed. To explain in detail, English dog vocals at a lower frequency than Japanese dogs, but Japanese dog vocals faster than English dogs. The same phenomenon can be observed with human speakers of these two languages.

4.4 Human Evaluation on Acoustic Features

Alongside acoustic analysis, we assess cross-linguistic dog vocals via surveys. We pick 20 vocal pairs with distinct language settings from Section 3. Each of the 20 questionnaires, given to 30 participants, makes 600 in total. Pair sequence is randomized. When spotting differences, participants gauge from four dimensions below: (1)urgency, which represents the speed (2)pitch (3)duty ratio, which is the ratio of the duration of the voiced segments to that of the unvoiced segments (4)loudness.

Among four above dimensions, 74.01%, 79.09%, 67.23% and 63.56% of the questionnaires which says that there are

differences between the two clips report dimension difference respectively. This shows that from a human perspective the differences are most revealed by pitch, which is highly frequency related. This is in line with our acoustic analysis.

4.5 Restrictions

Compared to previous studies, we adopt a first-of-its-kind method of collecting data from the Internet. Some may argue that former studies have a more controllable environment to eliminate confounding factors. However, it is very unlikely to cultivate exactly the same growing environment as subjectivity always exists. On the contrary, with the large amount of data from various families, we can remedy the confounding factors in a statistical way. Our pipeline is hence beneficial as we can scale up the data in the future and further validate our argument.

5 Conclusion

In this paper, we define a new problem of discovering the linguistic influence of the host on the sounds of their pet dogs. Experiments have shown that there is a significant difference in audio frequency and speed between the voice of dogs in the Japanese language environment and the English environment. Specifically, English dogs bark at a lower pitch than Japanese dogs, while Japanese dogs bark faster than English ones. The phenomena can be observed in humans as well.

We sourced our data from YouTube, which, though noisy, ensures the quantity and variety and has rarely been covered in previous studies. The fact that we removed quite a lot of data from the raw videos is due to the rigorous pipeline we developed for quality purposes. Eventually, our EJShibaVoice dataset, which contains a large number of Shiba Inu sound clips and their corresponding host speech under various scenes will facilitate future research in this field.

Future direction can be a larger dataset for a more general investigation, i.e. more breeds and more sound clips, as well as more various language environments in addition to Japanese and English in this study. Note that the pipeline we proposed is language-agnostic and hence can be applied to other studies with similar purposes.

References

- Ardila, R.; Branson, M.; Davis, K.; Henretty, M.; Kohler, M.; Meyer, J.; Morais, R.; Saunders, L.; Tyers, F. M.; and Weber, G. 2019. Common Voice: A Massively-Multilingual Speech Corpus. *CoRR*, abs/1912.06670.
- Bredin, H.; and Laurent, A. 2021. End-to-end speaker segmentation for overlap-aware resegmentation. In *Proc. Interspeech 2021*.
- Bredin, H.; Yin, R.; Coria, J. M.; Gelly, G.; Korshunov, P.; Lavechin, M.; Fustes, D.; Titeux, H.; Bouaziz, W.; and Gill, M.-P. 2020. pyannote.audio: neural building blocks for speaker diarization. In *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*.
- Coupé, C.; Oh, Y. M.; Dediu, D.; and Pellegrino, F. 2019. Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Science Advances*, 5(9): eaaw2594.
- Davis, S.; and Mermelstein, P. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4): 357–366.
- Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 4171–4186. Association for Computational Linguistics.
- Duddington, J.; and Dunn, R. 2012. eSpeak text to speech. Web publication: <http://espeak.sourceforge.net>.
- Ehsani, K.; Bagherinezhad, H.; Redmon, J.; Mottaghi, R.; and Farhadi, A. 2018. Who let the dogs out? modeling dog behavior from visual data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4051–4060.
- Eyben, F.; Scherer, K. R.; Schuller, B. W.; Sundberg, J.; André, E.; Busso, C.; Devillers, L. Y.; Epps, J.; Laukka, P.; Narayanan, S. S.; et al. 2015. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2): 190–202.
- Gemmeke, J. F.; Ellis, D. P.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R. C.; Plakal, M.; and Ritter, M. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 776–780. IEEE.
- Graham, C. 2014. Fundamental frequency range in Japanese and English: The case of simultaneous bilinguals. *Phonetica*, 71(4): 271–295.
- Hantke, S.; Cummins, N.; and Schuller, B. 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)*, 5134–5138. IEEE.
- Hermansky, H. 1990. Perceptual linear predictive (PLP) analysis of speech. *the Journal of the Acoustical Society of America*, 87(4): 1738–1752.
- Ide, Y.; Araki, T.; Hamada, R.; Ohno, K.; and Yanai, K. 2021. Rescue dog action recognition by integrating ego-centric video, sound and sensor information. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part III*, 321–333. Springer.
- Jégh-Czinege, N.; Faragó, T.; and Pongrácz, P. 2020. A bark of its own kind – the acoustics of ‘annoying’ dog barks suggests a specific attention-evoking effect for humans. *Bioacoustics*, 29(2): 210–225.
- Kong, Q.; Cao, Y.; Iqbal, T.; Wang, Y.; Wang, W.; and Plumbley, M. D. 2020. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28: 2880–2894.
- Larranaga, A.; Bielza, C.; Pongrácz, P.; Faragó, T.; Bálint, A.; and Larranaga, P. 2015. Comparing supervised learning methods for classifying sex, age, context and individual Mudi dogs from barking. *Animal cognition*, 18(2): 405–421.
- Molnár, C.; Kaplan, F.; Roy, P.; Pachet, F.; Pongrácz, P.; Dóka, A.; and Miklósi, Á. 2008. Classification of dog barks: a machine learning approach. *Animal Cognition*, 11(3): 389–400.
- Paladini, A. 2020. The bark and its meanings in inter and intra-specific language. *Dog behavior*, 6(1): 21–30.
- Pongrácz, P. 2017. Modeling evolutionary changes in information transfer: Effects of domestication on the vocal communication of dogs (*Canis familiaris*). *European Psychologist*, 22(4): 219.
- Pongrácz, P.; Molnár, C.; Dóka, A.; and Miklósi, Á. 2011. Do children understand man’s best friend? Classification of dog barks by pre-adolescents and adults. *Applied animal behaviour science*, 135(1-2): 95–102.
- Pongrácz, P.; Molnár, C.; and Miklósi, Á. 2010. Barking in family dogs: an ethological approach. *The Veterinary Journal*, 183(2): 141–147.
- Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.
- Räsänen, O.; Doyle, G.; and Frank, M. C. 2018. Pre-linguistic segmentation of speech into syllable-like units. *Cognition*, 171: 130–150.
- Schuller, B.; Steidl, S.; Batliner, A.; Vinciarelli, A.; Scherer, K.; Ringeval, F.; Chetouani, M.; Wenginger, F.; Eyben, F.; Marchi, E.; et al. 2013. The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*.
- Siniscalchi, M.; d’Ingeo, S.; Minunno, M.; and Quaranta, A. 2018. Communication in dogs. *Animals*, 8(8): 131.

Strang, G.; and Nguyen, T. 1996. *Wavelets and filter banks*. SIAM.

Szegedy, C.; Ioffe, S.; Vanhoucke, V.; and Alemi, A. A. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*.

Thorndike, L.; and Bruce, D. 2017. *Animal intelligence: Experimental studies*. Routledge.

Wang, P.; Yang, A.; Men, R.; Lin, J.; Bai, S.; Li, Z.; Ma, J.; Zhou, C.; Zhou, J.; and Yang, H. 2022. OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework. 162: 23318–23340.