A Canine Language Lexical Analysis System

Sinong Wang[1], Xingyuan Li[2], Hridayesh Lekhak[3], Mengyue Wu[4], Kenny Q. Zhu[5]

[1,3,5]University of Texas at Arlington, Arlington, Texas, USA
[2,4]Shanghai Jiao Tong University, Shanghai, China
[1]sxw7663@mavs.uta.edu
{[2]xingyuan, [4]mengyuewu}@sjtu.edu.cn
[3]hxl7195@mavs.uta.edu
[5]kenny.zhu@uta.edu

Abstract

Whether animal vocal communications constitute a "language," which is characterized by fixed linguistic patterns is a long-standing scientific curiosity. In this paper, we present a web-based interactive canine language lexical analysis system which automatically processes dog vocalizations using HuBERT, transcribes them into sequences of distinct phonemes, and further segments them into a sequence of words from a vocabulary discovered by statistic analysis. These vocalizations are either extracted from YouTube videos or uploaded by the users. Further the system allows to visualize dog "sentences," in terms of the audio spectralgrams and the transcripts in both phonemes and words. This system is a first step toward canine language inference and understanding, and can be used as a platform for verification of canine language processing algorithms as well as an annotation tool for understanding the semantics of canine lexicons.

1 Introduction

Whether animals other than human beings possess languages has long been a contentious scientific topic Snowdon (1990), with debates revolving around whether animal communications contains regular lexicon, grammar and semantics Rendall (2021). Better understanding of animal languages can significantly contribute to animal welfare, agriculture, disaster control, resource extraction and ultimately understanding of our planet. To date, the research on animal communications has been hindered by lack of quality data. Audio recordings of animal communications, especially by amateurs, are often laced with background noise and interferences, which poses significant obstacles to the advancement of this research endeavor. Currently, there is no available interactive system, but such a system is crucial for researchers to directly and visually observe and compare the connections and differences between animal vocalization segments.

We chose dogs as our research subjects due to their widespread integration into human society, with many people keeping dogs as pets. However, even so, we had to make trade-offs in terms of the generality and accuracy of the data. Accurate, low-noise dog videos recorded with sophisticated equipment are rare, let alone videos of other animals that humans encounter less frequently. However, large datasets often come with a compromise in terms of the presence of noise or audio quality.

Furthermore, evaluating the transcription results of canine language poses a challenging task. Without a ground truth for canine language, we can only infer the performance of the transcription results through observations, listening, comparing the consistency of canine sounds, and their occurrence contexts. To facilitate researchers in conveniently and thoroughly studying canine language transcription, we have introduced for the first time an interactive and scalable canine lexical discovery system [1].

The system is an interactive and scalable canine sound transcription and vocabulary discovery system. In addition to transcribing canine sounds into symbolic sequences as shown as Figure 1, the system also provides clickable functionality for listening to canine sound phonemes, facilitating users or researchers in identifying the consistency of the same phonemes in canine sounds. Furthermore, the system can continue to expand its transcription methods beyond our proposed baseline method, allowing for the adoption of more accurate and efficient methods in the future.

Additionally, the platform can present a two-dimensional mapping of cluster centers for canine phoneme audio along with audio examples, aiding users or researchers in exploring the relationships between phonemes with similar sound characteristics and the accuracy of phoneme classification. To further analyze the meaning of canine language, the

[1] The demo is available at http://202.120.38.146:8079/

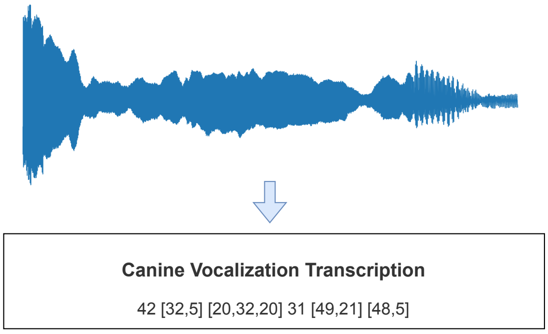


Figure 1: Canine vocalization Transcription Example

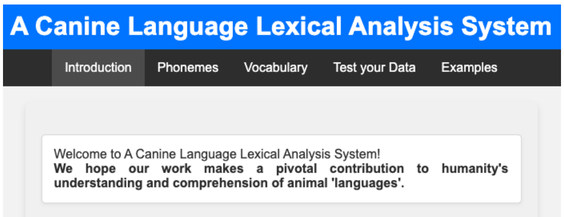
platform also provides a canine vocabulary derived from NLP techniques along with corresponding video and audio examples, facilitating researchers in observing, distinguishing, and validating the relationships between canine words and environmental or situational contexts. Overall, the system is an intuitive canine sound transcription platform that is ready-to-use and scalable, designed for researchers, enabling them to verify the accuracy of each step in the transcription process through a visual canine word discovery system.

This paper presents a canine language lexical analysis system, our contributions include:

- We designed and implemented the first interactive animal language segmentation and analysis system, that allows transcribing dog videos or audio files, demonstrating preliminary dog language phonemes and vocabulary.
- This interactive and scalable system provides users with the functionality to upload their own recorded canine video or audio files and validate their relationship with the environment, actions, and consistency of canine vocalizations.
- The phonemes obtained through our proposed pipeline exhibit significant consistency in acoustic features. The words discovered in the vocabulary demonstrate basic temporal characteristics typical of regular words. Additionally, the quality of some transcriptions has been validated as excellent.

2 System Demonstration

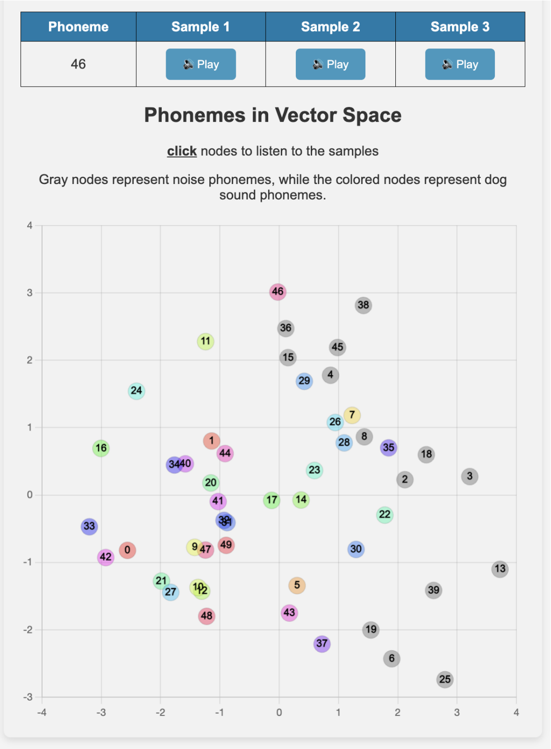
The system as shown as Figure 2 comprises four primary modules: phoneme module, vocabulary module, test your data module, and example module. Phoneme module displays a 2D distribution graph of the 50 canine phoneme cluster nodes derived from self-supervised learning and clustering, alongside corresponding audio examples; Vocabulary module showcases the vocabulary list filtered through plausible score, presenting words with associated video and audio examples; Test your data module enables users to upload videos or audio files for transcription results, allowing them to listen to phoneme sounds by clicking on the waveform; Example module displays the transcription results of 10 randomly selected dog vocalization clips.



A Canine Language Lexical Analysis System

Welcome to A Canine Language Lexical Analysis System!
We hope our work makes a pivotal contribution to humanity's understanding and comprehension of animal 'languages'.

Figure 2: Instruction Page in Demo



Phoneme	Sample 1	Sample 2	Sample 3
46	🔈 Play	🔈 Play	🔈 Play

Phonemes in Vector Space

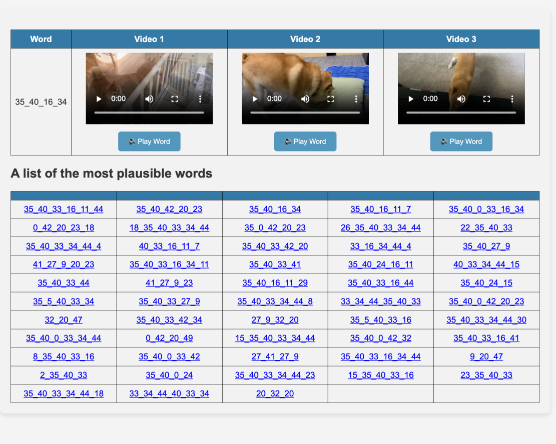
click nodes to listen to the samples

Gray nodes represent noise phonemes, while the colored nodes represent dog sound phonemes.

Figure 3: Phoneme Module Demonstration

2.1 Phoneme Module

As Figure 3 shown, the Phoneme module aims to illustrate the spatial distribution and relative positioning of the 50 distinct phoneme nodes. Users can click on each cluster node to listen to examples.

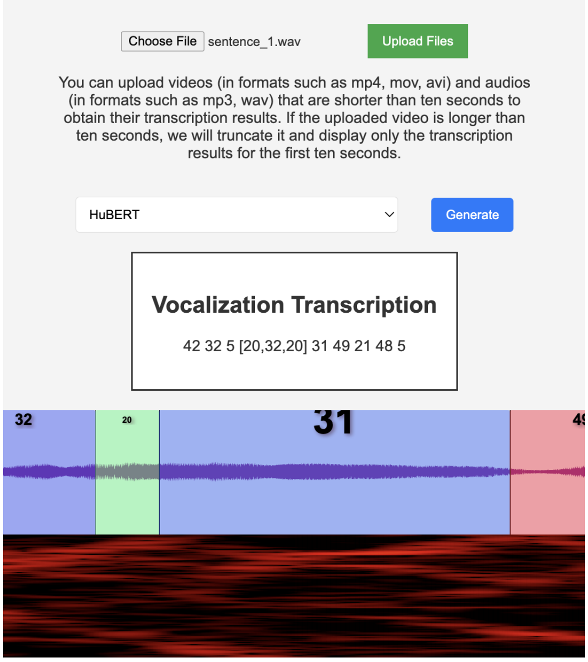


Word	Video 1	Video 2	Video 3
35_40_16_34	0:00 Play Word	0:00 Play Word	0:00 Play Word

A list of the most plausible words

35_40_33_16_11_44	35_40_42_20_23	35_40_16_34	35_40_16_11_7	35_40_0_33_16_34
0_42_20_23_18	18_35_40_33_34_44	35_0_42_20_23	26_35_40_33_34_44	22_35_40_33
35_40_33_34_44_4	40_33_16_11_7	35_40_33_42_20	33_16_34_44_4	35_40_27_9
41_27_9_20_23	35_40_33_16_34_11	35_40_33_41	35_40_24_16_11	40_33_34_44_15
35_40_33_44	41_27_9_23	35_40_16_11_29	35_40_33_16_44	35_40_24_15
35_5_40_33_34	35_40_33_27_9	35_40_33_34_44_8	33_34_44_35_40_33	35_40_0_42_20_23
32_20_47	35_40_33_42_34	27_9_32_20	35_5_40_33_16	35_40_33_34_44_30
35_40_0_33_34_44	0_42_20_49	15_35_40_33_34_44	35_40_0_42_32	35_40_33_16_41
8_35_40_33_16	35_40_0_33_42	27_41_27_9	35_40_33_16_34_44	9_20_47
2_35_40_33	35_40_0_24	35_40_33_34_44_23	15_35_40_33_16	23_35_40_33
35_40_33_34_44_18	33_34_44_40_33_34	20_32_20		

Figure 4: Vocabulary Module Demonstration



You can upload videos (in formats such as mp4, mov, avi) and audios (in formats such as mp3, wav) that are shorter than ten seconds to obtain their transcription results. If the uploaded video is longer than ten seconds, we will truncate it and display only the transcription results for the first ten seconds.

HuBERT

Vocalization Transcription

42 32 5 [20,32,20] 31 49 21 48 5

Figure 5: Test your data Module Demonstration

Color-coded centroids represent canine phonemes, while gray nodes represent noise phonemes. Despite noise reduction techniques, complete elimination of noise remains unattainable. Users can identify similar phonemes by listening to audio examples of adjacent cluster nodes.

2.2 Vocabulary Module

As Figure 4 shown, the Vocabulary module functions similarly to the Phoneme module, providing the most probable canine lexical vocabulary sorted by plausible score and represented by cluster centroid numbers. To better explore the context of each word, we offer not only example audio but also accompanying videos extended by 0.5 seconds before and after for user observation.

2.3 Test Your Data Module

As Figure 5 shown, in this module, users can upload videos or audio files and choose the transcription method to obtain the transcription results of the canine audio within. However, it's important to note that to ensure smooth server operation, videos or audio longer than 10 seconds will be truncated, retaining only the first 10 seconds for transcription. The page will display the translation results alongside the corresponding audio waveform, allowing users to click on segments corresponding to phonemes to listen to the audio.

2.4 Example Module

In this module, we have prepared 10 randomly selected dog vocalization clips for users who do not currently have dog audio. Users can select a sound clip from the dropdown menu, listen to it, and observe the transcription results generated by two different methods.

3 Method

We present some technical details under the hood of this interactive canine lexical analysis system system. The method presented here are only preliminary, "various components" of this pipeline can be replaced and updated. To mitigate issues related to poor audio quality and severe noise contamination resulting from large datasets, we utilized deep learning-based audio denoising and audio event detection techniques to obtain the final canine sound sentences. Clustering was employed to label each audio frame, resulting in the identification of 36 canine sound phonemes and 14 noise phonemes, after manual filtering due to noise presence among the 50 phonemes. We designed and calculated the plausible score for each N-gram using NLP algorithms, obtaining the canine sound vocabulary, and subsequently utilized parsing algorithms to generate the final transcription of canine language. The whole pipeline is shown as Figure 6.

3.1 Phoneme Clustering

To obtain the phonemes in a sentence in the dog language, we use methods including audio clean-up by AudioSep, sentence extracting and phoneme recognition.

Audio Clean-up by AudioSep. To separate dog sounds from audios that contain both dog sounds

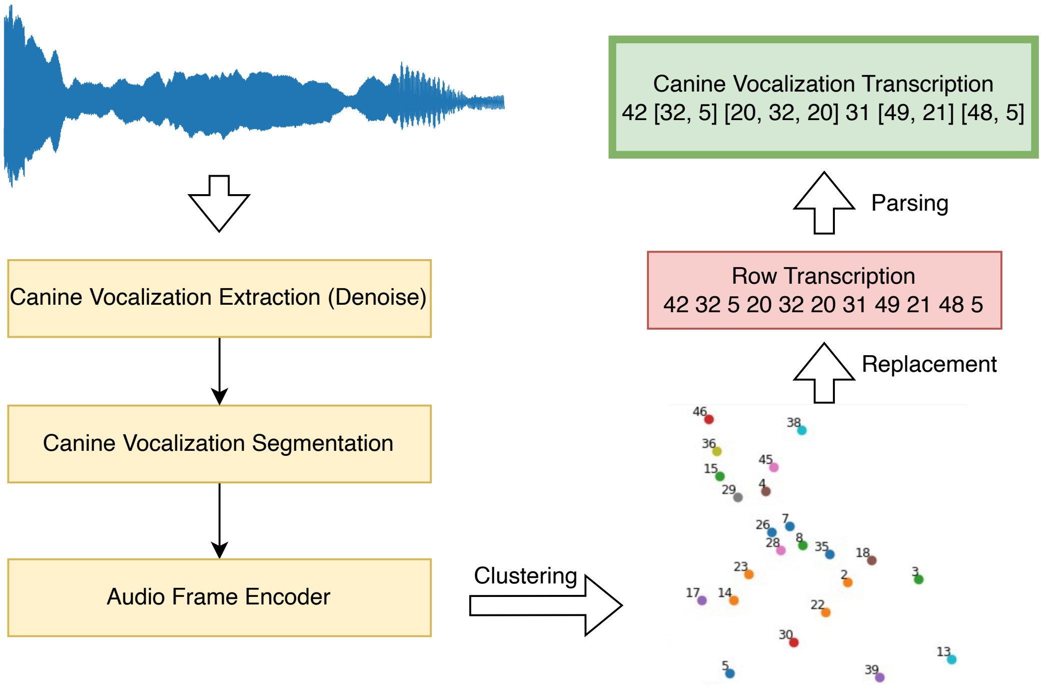


Figure 6: Transcription Example

and noises, we apply AudioSep (Liu et al., 2023), a audio source separation model that has been pre-trained on large-scale multimodal datasets, including AudioSet (Gemmeke et al., 2017) dataset, VGGSound (Chen et al., 2020) dataset, AudioCaps (Kim et al., 2019) dataset, etc. We apply PANNs (Kong et al., 2020) with 0.05 threshold to obtain shorter audio clips before applying AudioSep for further processing.

Sentence Extracting. In order to obtain cleaner data, we apply PANNs (Kong et al., 2020), a sound event detection model pre-trained on AudioSet (Gemmeke et al., 2017) dataset, to obtain dog sounds audio clips. Then use the same method and conditions to obtain and filter the dog “sentences”.

Phoneme Recognition. We apply a self-supervised approach, HuBERT (Hsu et al., 2021), to acquire dog sound units in this section.

3.2 Lexical Discovery

We believe that a word must satisfy the condition that it appears frequently enough in the transcription information to indicate its repeatability, and at the same time, we must ensure that the word is uttered by more than one dog to determine its universality. To this end, we designed a plausible score, the formula is as follows

$$Ps(gram_i^n) = f(gram_i^n) * \delta(gram_i^n),$$

where $Ps(\cdot)$ is the popularity score function, $f(\cdot)$ is a function to calculate the frequency of this gram, $\delta(\cdot)$ is a function to calculate the diversity of this word, $gram_i^n$ is a unique n-gram, n is the number of frames, i indicates the i_{th} unique ngram. The formula of f and δ is as follows:

$$\begin{aligned} f(gram_i^n) &= \frac{|gram_i^n|}{\sum_i |gram_i^n|)}, \\ \delta(gram_i^n) &= \left|\{x \in D : x \; contains \; gram_i^n\}\right|, \end{aligned}$$

where D is a set of different dogs in training dataset, $|\cdot|$ means to get the number of a set. The higher the popularity score of an ngram, the stronger the universality of the ngram, and its sound is uttered more times by more canines. Finally we first iterate through all dog vocalization clips and calculate the plausible score for each n-gram and determine 0.11 as the threshold of plausible score, where selected words can cover the most of the sentences and have a high diversity.

3.3 Sentence Parsing

After creating a dog language dictionary, we can parse sentences using the words in the vocabulary. Sentences in the language of dogs can consist of both words and noises, and may include multiple instances of each. In order to maximize the number of words in the sentence and use the longest possible ones, we proposed a dynamic programming algorithm (Algorithm 1).

Algorithm 1: Sentence Parsing

```
Input  :A sentence $S = (p_1, ..., p_n)$, A vocabulary
        $V = (w_1, ..., w_m)$.
Output :The result of parsing $R = (c_1, ..., c_k)$
$counts \leftarrow [0, n+1, ..., n+1]$;
$Rs \leftarrow [[], ..., []]$;
for $i \leftarrow 0$ to $n-1$ do
    foreach $w \in (V + set(S))$ do
        $l \leftarrow len(w)$;
        $f_1 \leftarrow S[i, ..., i+l] = w$;
        $f_2 \leftarrow counts[i+l] > counts[i] + 1$;
        if $f_1$ and $f_2$ then
            $counts[i+l] \leftarrow counts[i] + 1$;
            $Rs[i+l] \leftarrow Rs[i] + w$
        end
    end
end
$R \leftarrow Rs[n]$;
```

3.4 Implementation Details

For training HuBERT, in the first stage, we used 54 clusters, 100k training steps, and a learning rate of 0.0001. In the second stage, we used 100 clusters and 109k training steps. Finally, we used the features from the 12th transformer layer to train a K-Means model with 50 clusters.

4 Evaluation

Many factors influence the performance of our transcription system: noise present in the training data, performance degradation due to the lack of ground truth, among others. To validate whether our proposed baseline adequately transcribes canine sounds based on acoustic features, in this section, we demonstrate the performance of the baseline pipeline using our proposed canine lexical discovery system from three perspectives: phoneme, vocabulary, and transcription.

4.1 Phoneme Evaluation

Tester	dog voice label	total label
Tester 1	72.0%	66.67%
Tester 2	70.5%	64.89%
Agreement	80.5%	74.22%

Table 1: Accuracy and agreement result on testing the reliability of phonemes discovery

From an acoustic perspective, the performance of our method in phoneme discovery is reflected in the similarity of audio segments assigned to the same phoneme. We randomly selected 120 sets of audio clips from different dogs, each containing the same Phoneme, as well as 120 sets of audio clips from different dogs, each containing different Phonemes and assigned these test data to two testers after shuffling the data. Both two testers are graduate male students majoring in computer science and were tasked with annotating whether each set of audio clips belonged to the same Phoneme. The results are presented in the Table 1.

The "dog voice label" indicates that the test data only includes canine phonemes, while the "total label" encompasses all 50 phonemes. "Agreement" refers to the proportion of pairs where the testers reached a consensus, relative to the total test data. This result suggests that with a high level of agreement between the two testers, the phoneme results are relatively accurate. Additionally, the Phonemes page of the canine lexical discovery system provides a schematic diagram of phoneme nodes and randomly selected audio examples for researchers to assess the quality of phonemes.

4.2 Vocabulary Evaluation

	Sentences	Words	Phonemes
Pauses duration	873 ms	85 ms	39 ms
Length duration	2602 ms	206 ms	90 ms

Table 2: Average pauses duration time between sentences, words and phonemes in canine vocalization and average duration of sentences, words and phonemes.

To evaluate the discovered vocabulary without ground truth for dog language words, we assessed whether a word in canine language is usually indivisible into smaller units capable of independent use from an acoustic perspective. For this purpose, we calculated the average duration of non-word phonemes on both sides of each word in sentences parsed from the training data, representing the pauses Zellner (1994) duration for each word. Additionally, we computed the pauses duration for sentences and phonemes using the same method and the length duration of sentences, words and phonemes, as shown in the Table 2.

The pause duration on both sides of words can indirectly indicate the indivisibility of the combination of these phonemes. Furthermore, comparing sentences, we observe that for sentences with more pronounced pauses, the ratio of pause duration to length duration for words is similar to that of sentences. This confirms the reliability of the words in the vocabulary.

4.3 Transcription Evaluation

Tester	Transcription Quality
Tester 1	38.57%
Tester 2	48.70%
Agreement	68.50%

Table 3: Transcription quality result on testing the reliability of transcription

In the scenario where the vocabulary is determined, the performance of the parsing algorithm directly determines the quality of the transcription. We evaluated the performance by manually judging whether each word parsed in the sentence is acoustically complete, meaning it contains a complete energy peak. Two male graduate students majoring in computer science from university were randomly assigned 100 sentences from the training dataset as testers. They were asked to determine whether each word in each sentence was acoustically complete. Finally, we calculated the proportion of complete words to the total number of words in these 100 sentences, the results are shown in the Table 3.

After go through 100 canine sentences, testers finally examine a total of 854 words, with 68.5% agreement. The results indicate that the transcription quality is not high, which is consistent with the fact that there is still noise in the sentences even after the denoising process. Some of these noices were incorrectly recognized as phonemes and creep into the transcript as words. Nevertheless, the testers provided feedback indicating that nearly half of the words were deemed acoustically complete from the perspective of the energy plot, validating that the parsing method along with the vocabulary could reflect a portion of the true transcription results of canine language.

5 Conclusion

This paper introduces a canine language lexical analysis system where users or researchers can upload their own recorded canine videos or audios, interactively observe, listen to, and compare their transcription results. Additionally, they can use the Phoneme or Vocabulary modules to compare example video or audio scenes with the uploaded ones, thereby making significant contributions to verifying the structure of canine language and further understanding canine communication.

We also manually assessed the consistency of phoneme labels and the accuracy of parsing, corroborating the reliability of words through pauses duration. For future research, researchers can utilize this interactive system to explore the correlation between video context and canine vocalizations, discovering commonalities in canine vocal vocabulary meanings from a larger pool of videos, thereby laying a solid foundation for a deeper understanding of canine vocalizations.

Limitations

Despite the audio denoising process, our final canine sound sentences still contain some noise. This portion of noise was labeled as 14 noise phonemes during the phoneme labeling process. Moreover, due to the lack of ground truth for canine sound phonemes or vocabulary, we could only evaluate the transcription results from acoustic perspectives.

References

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. 2020. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE.

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.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132.

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.

Xubo Liu, Qiuqiang Kong, Yan Zhao, Haohe Liu, Yi Yuan, Yuzhuo Liu, Rui Xia, Yuxuan Wang, Mark D Plumbley, and Wenwu Wang. 2023. Separate anything you describe. *arXiv preprint arXiv:2308.05037*.

Drew Rendall. 2021. Aping language: Historical perspectives on the quest for semantics, syntax, and other rarefied properties of human language in the communication of primates and other animals. *Frontiers in psychology*, 12:675172.

Charles T Snowdon. 1990. Language capacities of nonhuman animals. *American Journal of Physical Anthropology*, 33(S11):215–243.

Brigitte Zellner. 1994. Pauses and the temporal structure of speech. In *Zellner, B.(1994). Pauses and the temporal structure of speech, in E. Keller (Ed.) Fundamentals of speech synthesis and speech recognition.(pp. 41-62). Chichester: John Wiley.*, pages 41–62. John Wiley.