Vocal Characterizer Dataset

Why do we need the Vocal Characterizer Dataset?

Based on Mehrabian's communication theory, personal communication is 7% verbal, 38% paralinguistic, and 55% nonvocal. However, Most of the existing human vocal dataset is concentrated on human speech and a number of researches have been done for years. Furthermore, only a limited amount of human nonverbal vocal sound dataset is available in spite of the fact that nonverbal vocalization plays a significant role in communication. The Vocal Characterizer Dataset is the first human-made nonverbal sound dataset to support the attempt to comprehend human nonverbal sound.

What's inside the Vocal Characterizer Dataset?

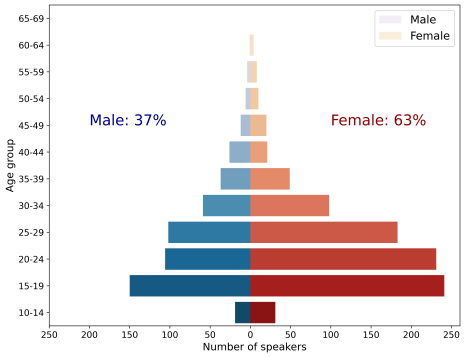
The Vocal Characterizer Dataset is a ***human nonverbal vocal sound dataset*** consisting of ***56.7 hours*** of short clips from ***1419 speakers***, crowdsourced by the general public in South Korea. Also, the dataset includes metadata such as age, sex, noise level, and quality of utterance. ***16 classes*** of Included human nonverbal sound contain ***'teeth-chattering'***, ***'teeth-grinding'***, ***'tongue-clicking'***, ***'nose-blowing'***, ***'coughing'***, ***'yawning'***, ***'throat clearing'***, ***'sighing'***, ***'lip-popping'***, ***'lip-smacking'***, ***'panting'***, ***'crying'***, ***'laughing'***, ***'sneezing'***, ***'moaning'***, and ***'screaming'***.

Every clip of the dataset is labeled with its level of noise and level of quality so that you can filter our dataset in accordance with the purpose of the research. The level of noise is defined as a two-leveled hierarchy in which are ***'noiseless'*** and ***'noisy'***. Moreover, the level of quality is also defined as a two-leveled manner in which are ***'high'*** and ***'low'***.

Every recorded clip in the dataset is a 44,100 Hz, 16-bit PCM, 1 channel wav file, and recorded with Android(operating system) devices. And each recording is resampled to 16,000Hz and saved into a single h5 file with its metadata. Refer to the following instruction to play around with the h5 file.

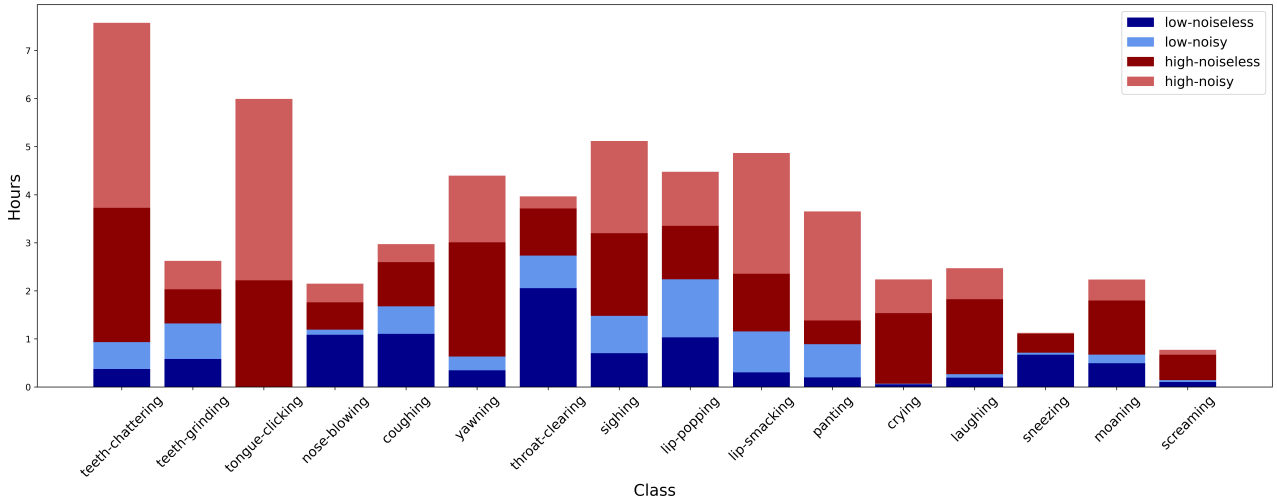
Information & Statistics

Figure 1 demonstrates the demographic information about the speakers of the dataset. The participating speakers comprise 37% male and 63% female. And due to the characteristics of the online-mannered data gathering, almost 70% of the speakers are within the range from 15 to 24.



<Figure 1. Age distribution by sex>

Figure 2 describes the total length(hours) of each class. The different colors in each bar graph refer to the combinations of the level of noise and quality(we'll refer to the combinations as 'filter' later in this paper). For instance, the whole recordings of the class 'teeth-chattering' add up to 7.57 hours consisting of 0.37 hours of 'low-noiseless' part, 0.56 hours of 'low-noisy' part, 2.79 hours of 'high-noiseless' part, and 3.85 hours of 'high-noisy' part.



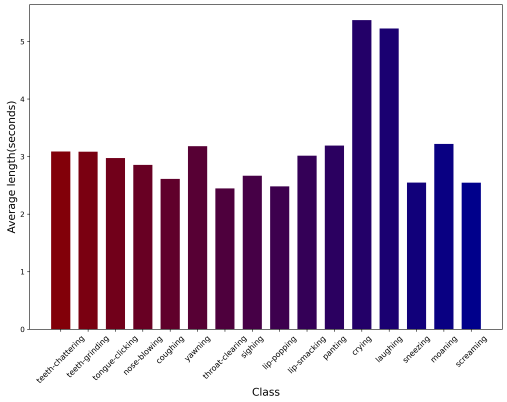
<Figure 2. Total length(hours) by class and filter>

As illustrated in Figure 3, the average length(in seconds) of the wav files of 14 classes(excluding 'crying' and 'laughing') is placed around 3 seconds, and the rest lies around 5 seconds. In Figure 4, we plotted length(in seconds) distribution by class. The plot suggests that the length distributions of the above-mentioned 14 classes are analogous to each other and those of the rest as well.

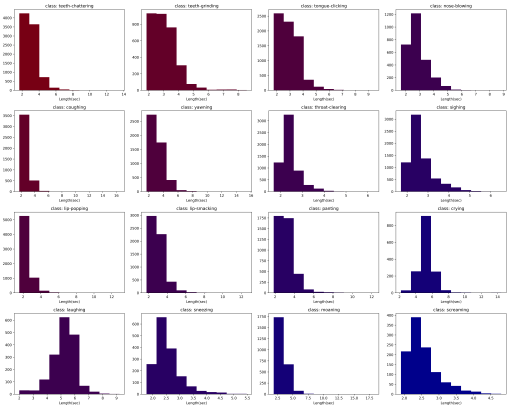
Figure 5 below illustrates how many distinctive speakers each class contains. Since, as shown above, the average length of recordings and the number of recordings submitted by the speaker doesn't differ drastically among each class, the distribution in Figure 5 is highly analogous to that of Figure 2.

Also, in Figure 6 is a histogram of which bin represents the number of classes covered by a single speaker. To clarify, for example, 432 speakers recorded only a single class according to the far-left bar in Figure 6 and 16 speakers recorded every class(16 classes) according to the far-right bar.

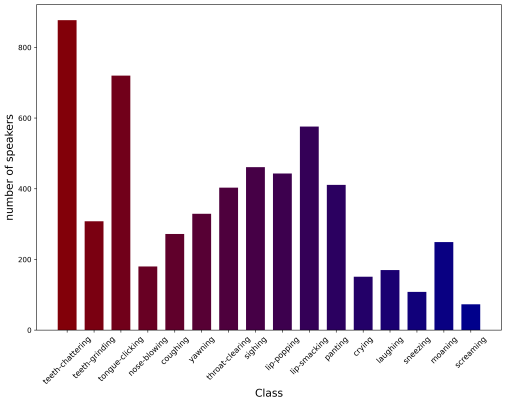
Lastly, 98.76% of the wav files were recorded indoors and 1.24% was recorded outdoors.



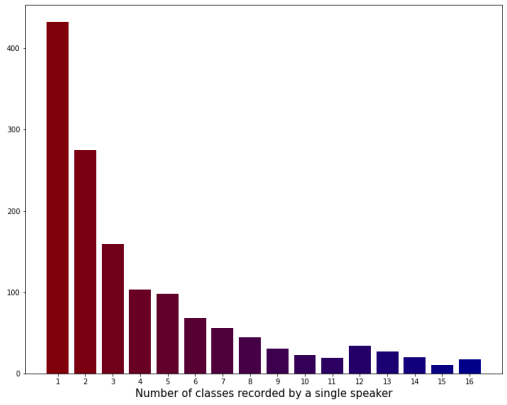
<Figure 3. Average length of recordings by class>



<Figure 4. Length of recordings distribution by class>



<Figure 5. Average number of speakers by class>



<Figure 6. Number of classes recorded by speaker>

Filename convention

Recorded wav files are named under following format:
{speaker ID}_{class}_{index}_{sex}_{age}_{location}_{quality}_{noise}.wav
Example: 3CHZ_9_8_0_29_0_0_1.wav

Speaker ID is a unique 4-digit alphanumeric code representing speaker, ***class*** is a digitized code indicating type of nonverbal sound, ***index*** is a digit used to identify different wav files of the same speaker in a certain class, ***location*** is a digitized code indicating where the recording took place, ***quality*** indicates the level of authenticity of the recording, ***noise*** indicates the level of noise at the time of recording, and sex and age are self-explanatory.

How to decode?

Class

0: ‘teeth-chattering’,
1: ‘teeth-grinding’,
2: ‘tongue-clicking’,
3: ‘nose-blowing’,
4: ‘coughing’,
5: ‘yawning’,
6: ‘throat-clearing’,
7: ‘sighing’,
8: ‘lip-popping’,
9: ‘lip-smacking’,
10: ‘panting’,
11: ‘crying’,
12: ‘laughing’,
13: ‘sneezing’,
14: ‘moaning’,
15: screaming’

Sex
{0: ‘Female’, 1: ‘Male’}

Location
{0: ‘indoor’, 1: ‘outdoor’}

Quality
{0: ‘High’, 1: ‘Low’}

Noise
{0: ‘Noiseless’, 1: ‘Noisy’}

License

Contact & Purchase

contact@deeplyinc.com

DEEPLY