Speech Commands Dataset Enhanced for Direction-of-Arrival Estimation

This dataset can serve as a training and evaluation corpus for the task of training keyword detection with speaker direction estimation (keyword direction of arrival - KWDOA).

It was created by processing the existing Speech Commands dataset [1] with the PyroomAcoustics library so that the resulting speech recordings simulate the usage of a circular microphone array with 4 microphones having a distance of 57 mm between adjacent microphones. Such design of a simulated microphone array was chosen in order to match the existing physical microphone array from the Seeeduino series.

The original dataset was cleaned before creating the simulated recordings.

The following parameters were being altered during the simulations:

- room dimensions [X,Y,Z] (corresponding to [width, length, height]), ranging from [3.0, 3.0, 2.2] to [8.0, 8.0, 5.0] metres
- RT60 reverberation time, ranging from 0.15 to 0.5 seconds
- signal-to-noise ratio (SNR) of randomly added noises
    - 15 simulations are performed for each original recording. 4 of them are without added noise, 4 have SNR around 22dB, 4 around 16dB and the last 3 have SNR around 10dB
- location of the microphone array in the room

In the original dataset, the recording length ranges from 0.4 seconds to 1.0 second. After simulation, each recording, including the simulated noises, has a length of 1.0 second.

The records are divided into folders according to their use: *train*, *validate*, *test* and *background_noise*. The names of the subfolders correspond to the commands that are pronounced in the individual records.

The *background_noise* folder contains individual noises for debugging and other purposes. These noises were also used during simulations of single commands.

There are 20 000 records for each of the following background noises: brushing_teeth, can_opening, cat, clapping, coughing, crying_baby, dog, doing_the_dishes, door_knock, door_wood_cracks, drinking_sipping, dude_miaowing, exercise_bike, footsteps, keyboard_typing, laughing, mouse_click, pouring_water, running_tap, sneezing, snoring, toilet_flush, vacuum_cleaner, washing_machine, wind.

The dataset contains a total of 1183567 keywords for training, 142620 keywords for validation and 157587 keywords for testing.

Example illustrating the file naming convention:

The substrings contained in the file name:

11-7.94 7.75 3.31-5.64 5.7 1.91-3.2 4.4 2.6-2.85 208.0 76.0-0.25-9376.0-bab36420 nohash 0.flac have the following meaning:

- 11 - number of simulation run
- 7.94 7.75 3.31 - dimensions of the room: X=7.94[m];Y=7.75[m]; Z=3.31[m]
- 5.64 5.7 1.91 - location of the source: X=5.64[m];Y=5.70[m]; Z=1.91[m]
- 3.2 4.4 2.6 - location of the microphones: X=3.20[m];Y=4.40[m]; Z=2.60[m]
- 2.85 208.0 76.0 - polar coordinates: r=2.85[m];θ=208.0[°]; ϕ=76.0[°]
- 9376.0 - estimated length of the command: τ=9376.0[1]
- 0.25 - reverb time: RT60=0.25[s]
- bab36420 nohash 0 - name of the original audio file

[1] Warden, Pete. “Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition.” ArXiv.org, 2018, arxiv.org/abs/1804.03209