Chart, box and whisker chart

Description automatically generated

Participant id: 6CcWU1oa6YeaL1hletunsAHLiOx2

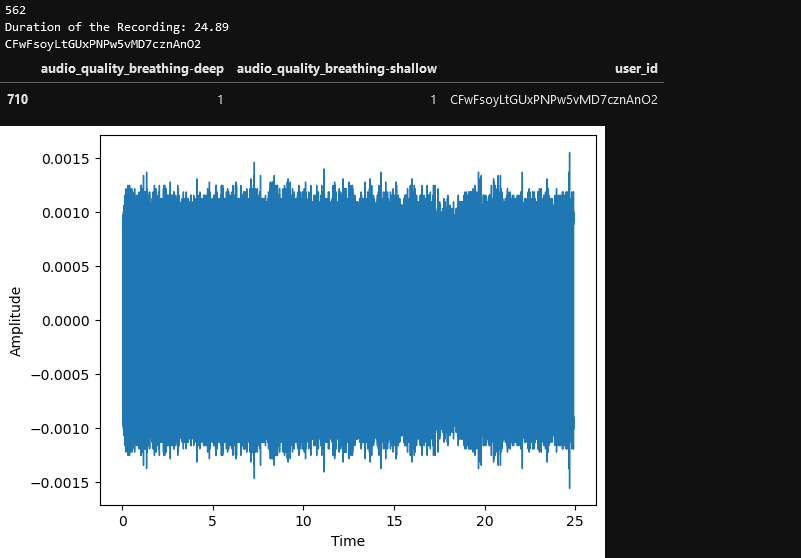
Participant index (2023\_03\_13\_logmel\_combined\_breaths\_6s\_FFT2048\_fmax5500\_112logmel.pickle dataset): 239

Combined breaths (concatenated)

Audio quality according to Coswara people (2 is best quality):

- 2/2 for heavy breath  
- 2/2 for shallow breathing

WHAT THE FUCK



Pure noise only!!! No breath

Graphical user interface

Description automatically generated

# Solution:

* Remove manually found bad audio by collecting ids in an excel file that is read at the start and all the bad ids are filtered out (there are separate sheets for breathing/cough/vowel because the bad recordings may vary across the types of recording). The same goes for creating new feature sets. Those IDs wont be created for those recordings. This saves memory as well.
* Additionally, for combined recordings (heavy cough + shallow cough for example) it suffices if one of them has a quality rating of 0 (which is the worst while 2 is the best – quality rating supplied by the Coswara people). This goes for both when loading a feature set and when creating one. This way old feature sets may be used without the bad audio as well.
* There is an option in the Dataset Class where you can additionally remove audio that has a rating below 2 (if it’s a combined recording, again one of them having a rating of 1 suffices).
* The combined recordings had the problem, if there was silence in one recording and signal in the other, the silence was not trimmed because the trimming happened for each audio file separately before. It trimmed audio below a dB threshold relative to its maximum signal. To solve this, it was trimmed separately but also again after combining them.
* Normalization was applied (separately and after combination). Shallow and heavy breathing often had a strong difference in signal power. Normalization in general might be reasonable and I don’t know why it wasn’t there (I am sure I used to normalize the data before at some point…. Somehow it got removed from the code)

The dataset now is a lot smaller, especially when only quality 2 is used.  
This new kind of feature processing is applied starting may 2023