

Audio file (e.g. mp3, wav)



split into $\sim 25\text{ms}$ frames

Frame 1
[20 vals]

Frame 2
[20 vals]

Frame 3
[20 vals]

...

Frame N
[20 vals]



aggregate across frames

Mean across all frames: [20 values]

Std across all frames: [20 values]



concatenate

Audio embedding: [40 values]