

User Study for paper "Target Speaker Extraction through Comparing Noisy Positive and Negative Audio Enrollments"

May 22, 2025

User Study Instructions

Thank you for participating in this user study. The goal of this user study is to collect human labeling on when a specific speaker talks in an audio mixture. The human labeling will be used to construct audio input to our target speaker extraction model. Specifically, we aim to assess how potential inaccuracies in human labels may affect the effectiveness of our system.

Labeling requirements:

1. You should label **all** and **only** those timesteps during which the target speaker (identified by the instruction below) is speaking. You should primarily rely on listening to identify the target speaker's **characteristics** in the mixture, and use the waveform shown by the audio editing platforms (e.g., Audacity, GarageBand, audio-preview plugin in VSCode) to get more accurate timestep labeling on when the target speaker start and stop speaking. **The start and stop timesteps usually happen between peaks in the waveform.**
2. **The target speaker may speak in more than one segments, please separately label all these segments without including the gaps between them.** You may ignore those brief pauses shorter than 0.5 seconds without explicitly excluding them in the labels. **Please provide your labels in seconds, to one decimal place (e.g., [2.5, 5.8], [10.2, 12.7]).**
3. **If the instruction is hint words spoken by the target speaker, You should label the full segment where the people is talking, not just the time steps where the speaker spoke the given hint words.**
4. You may need to listen to the audio multiple times to identify the most likely speaker from a set of candidates with similar voice characteristics in the enroll_[n].wav files. If you still couldn't decide whether the target speaker spoke in some challenging timesteps, you may include these timesteps in the labeling.

Example:

demo.wav contains 2 speakers' voice. Suppose the task is "label the male speaker's voice", then the labeling should be [1.4, 6.4], because a male voice is clearly audible from 1.4s to 6.4s.

For each enroll_{n}.wav audio, please label the following target speakers respectively:

1. **enroll_1.wav:** label the same speaker heard in hint_1.wav.
2. **enroll_2.wav:** label the same speaker heard in hint_2.wav.
3. **enroll_3.wav:** label the female speaker.

4. `enroll_4.wav`: label the female speaker.
5. `enroll_5.wav`: label the same speaker heard in `hint_5.wav`.
6. `enroll_6.wav`: label the male speaker.
7. `enroll_7.wav`: label the same speaker heard in `hint_7.wav`.
8. `enroll_8.wav`: label the same speaker heard in `hint_8.wav`.
9. `enroll_9.wav`: label the same speaker heard in `hint_9.wav`.
10. `enroll_10.wav`: label the same speaker heard in `hint_10.wav`.
11. `enroll_11.wav`: label the male speaker.
12. `enroll_12.wav`: label the people who spoke: " ... for the first time, the top 10 ...".
13. `enroll_13.wav`: label the people who spoke: " ... the other day, assuming a president ...". You should label the full segment where the people is talking, not just the time steps where the speaker spoke the given hint words.
14. `enroll_14.wav`: label the people who spoke: " ... spending the better part of 2020 ...". You should label the full segment where the people is talking, not just the time steps where the speaker spoke the given hint words.
15. `enroll_15.wav`: label the male speaker.
16. `enroll_16.wav`: label the people who spoke: " ... is a hero ... by looking straight ...". You should label the full segment where the people is talking, not just the time steps where the speaker spoke the given hint words.
17. `enroll_17.wav`: label the female speaker who is trying to speak over the crowd.
18. `enroll_18.wav`: label the female speaker who tried to interrupt the other's speech but failed.
19. `enroll_19.wav`: label the people who spoke: "... four years ... exhausting, imaging ...". You should label the full segment where the people is talking, not just the time steps where the speaker spoke the given hint words.
20. `enroll_20.wav`: label the people who spoke: "... only word I can think of ... ". You should label the full segment where the people is talking, not just the time steps where the speaker spoke the given hint words.

If you have any questions during the process, please feel free to reach out to the study coordinators.