Northwestern

interactive audio lab

I-SED: an Interactive Sound Event Detector

[B. Kim and B. Pardo, IUI 17]

Bongjun Kim

PhD Candidate, Interactive Audio Lab

EECS, Northwestern University

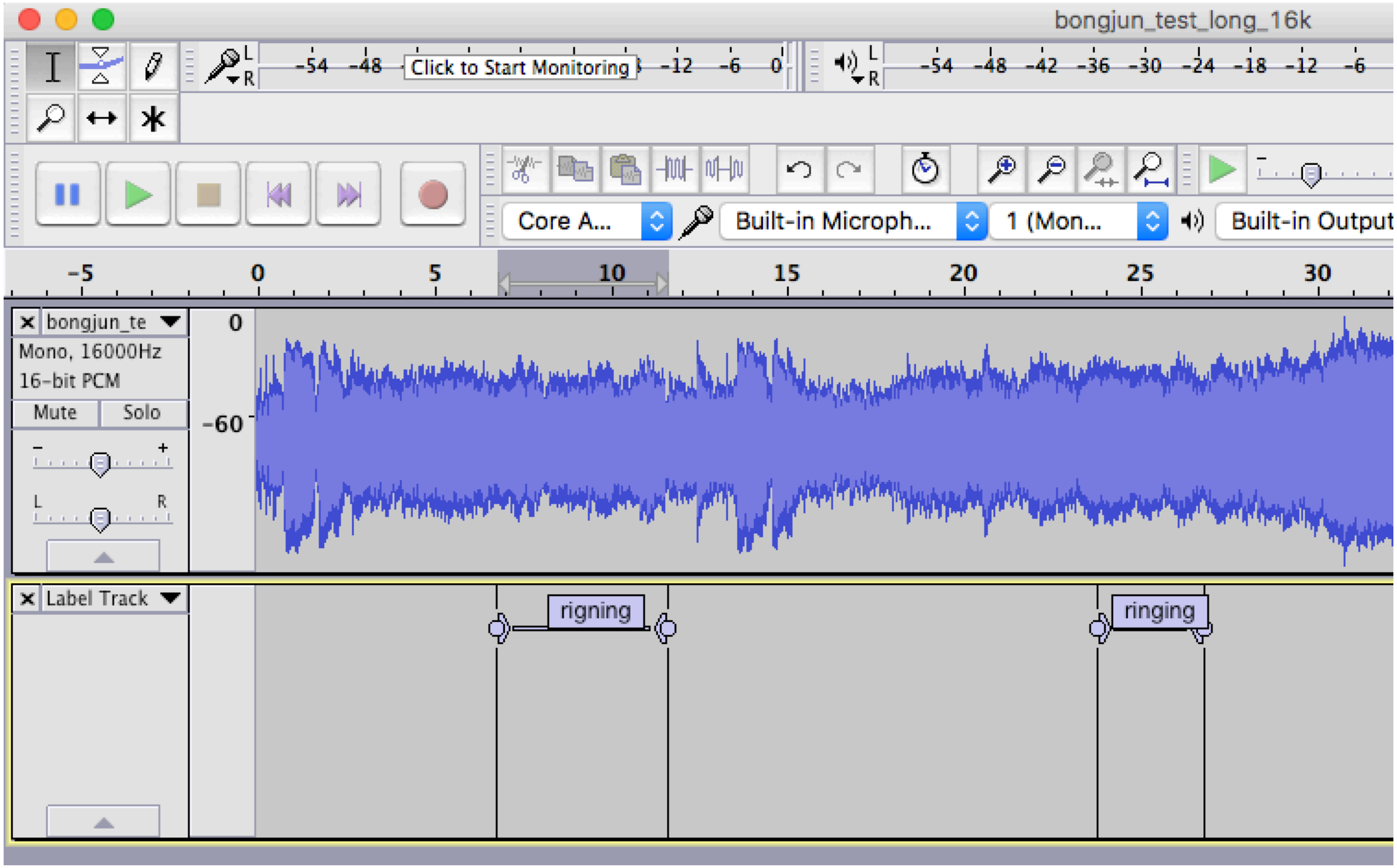
Simons Institute Workshop on Interactive Learning

Feb 16, 2017

Motivation Scenarios

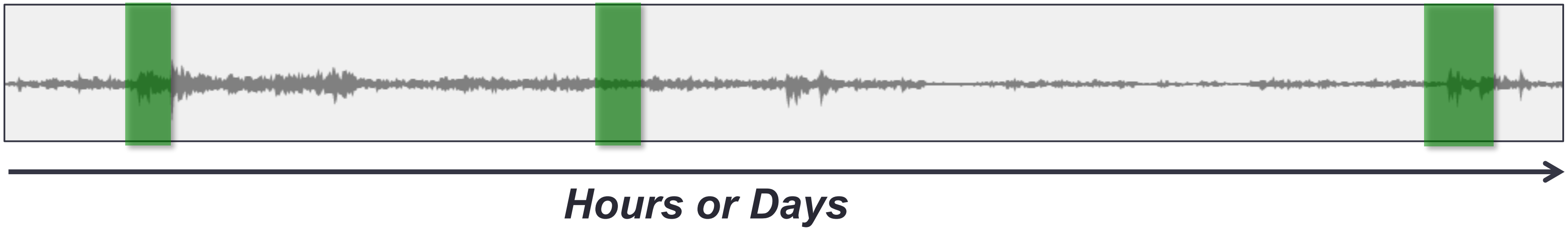
- Speech and language pathologist wants to analyze relationship between kid's language development and their listening environment

Manual Annotation



Audacity [Li, 2006]

Manual Annotation



Automated annotation

- Automated annotation tool
  - *TotalRecall* [Kubat, 2007], *Sonic Visualizer* [Cannam, 2006]
  - *ASAnnotation* [Boqaards, 2008], *LENA* [Xu, 2009]
  - **Issue 1: Predetermined sound classes (or acoustic features)**
  - **Issue 2: Too unreliable (for mission critical tasks)**
    - LENA agrees with human annotators only 76% of the time on a four-way forced choice labeling task

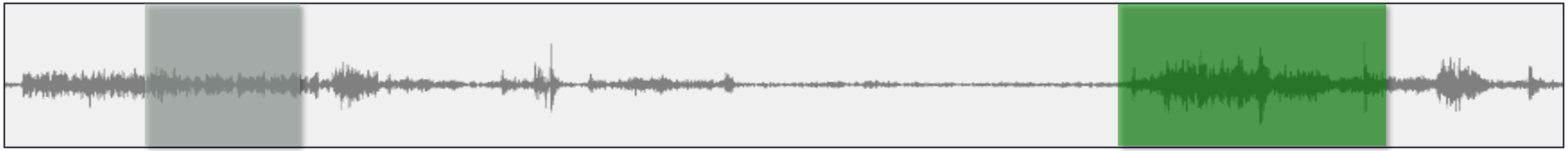
- Training a new model
  - **Issue 3: We do not have enough labeled training examples of the particular sound class (even hard to search).**

- Crowdsourcing
  - **Issue 4: The audio is credential (medical data), and we also need expert-level ground truth annotation.**

Going back to manual annotation....

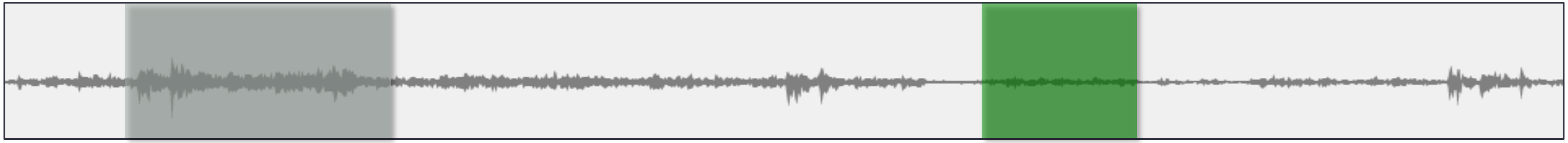
- We need a tool (an interface) that..
  - Speeds up my manual annotation of audio
  - Allows us to define a target sound class on-the-fly
  - Does not require any knowledge about machine learning and audio signal processing

Speech recordings

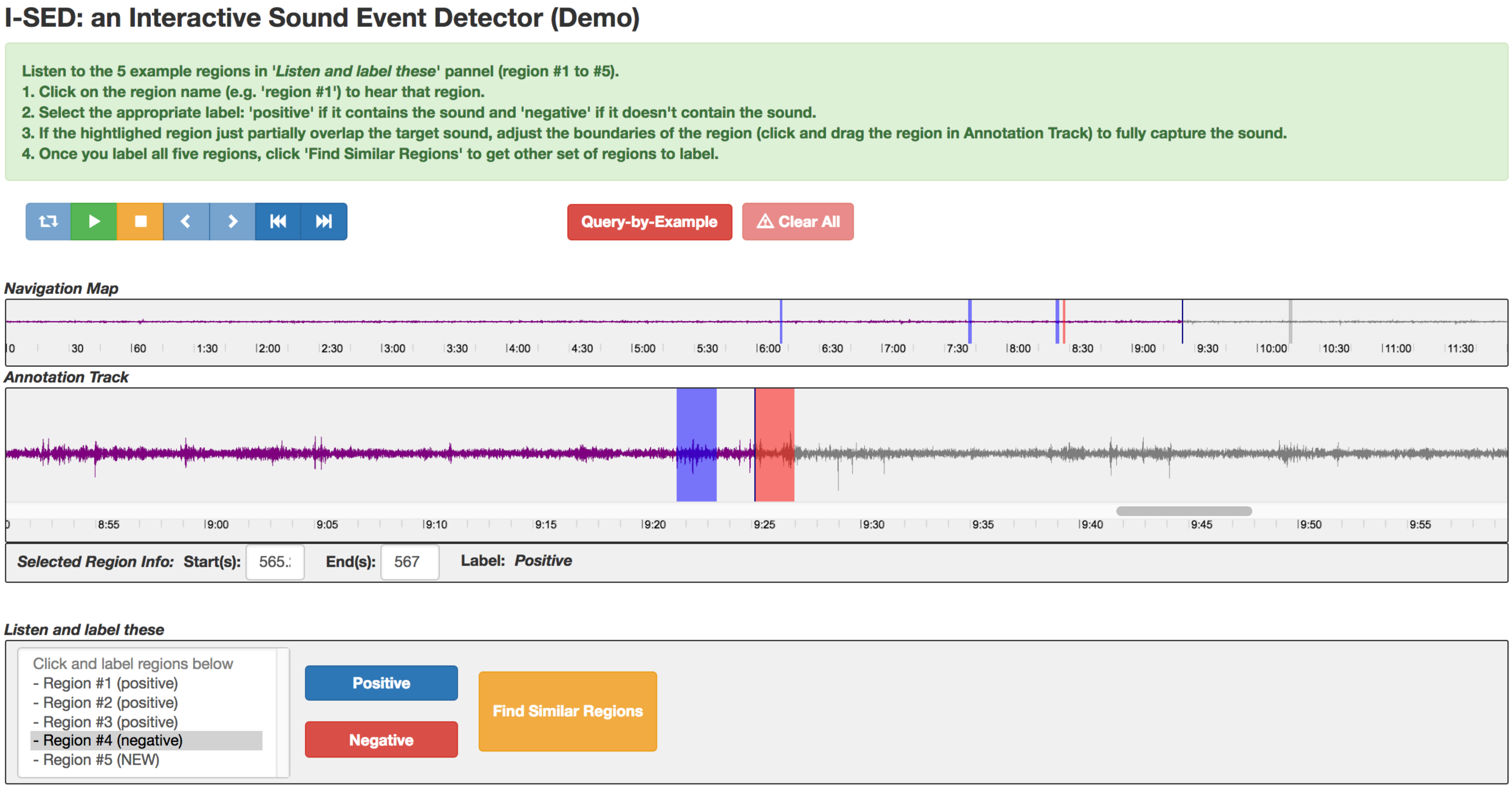


Bob **Sue**

Nature recordings



I-SED: an Interactive Sound Event Detector



I-SED: an Interactive Sound Event Detector (Demo)

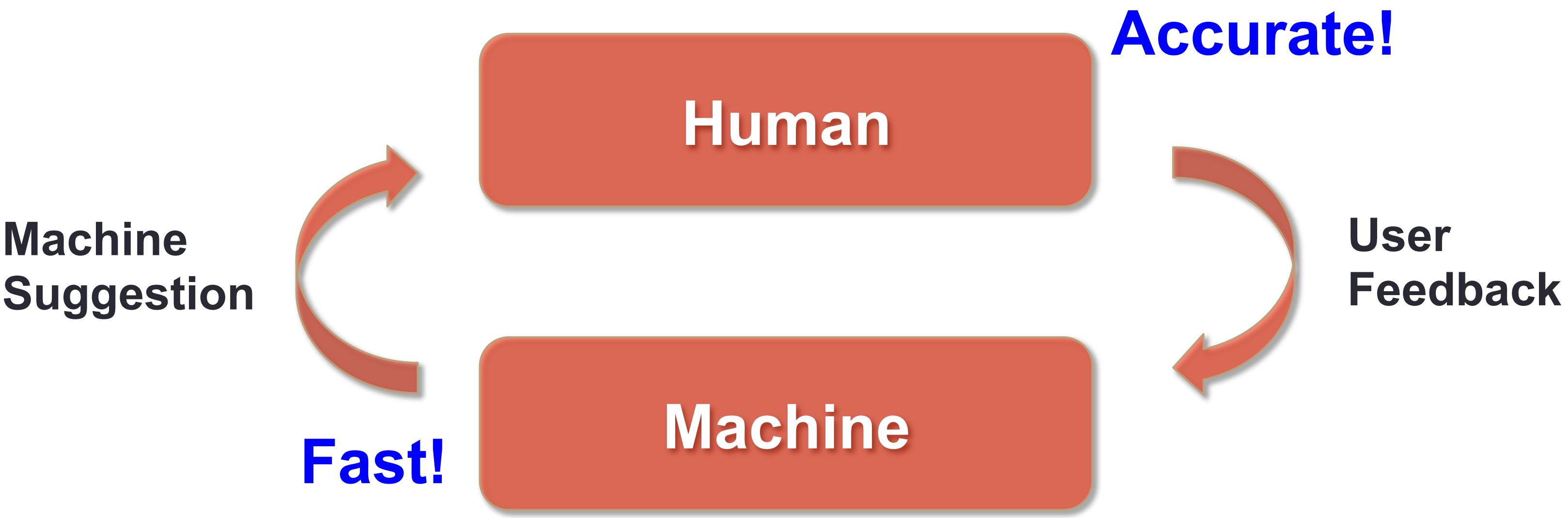
Listen to the 5 example regions in '*Listen and label these*' pannel (region #1 to #5).
1. Click on the region name (e.g. 'region #1') to hear that region.
2. Select the appropriate label: 'positive' if it contains the sound and 'negative' if it doesn't contain the sound.
3. If the hightlighed region just partially overlap the target sound, adjust the boundaries of the region (click and drag the region in Annotation Track) to fully capture the sound.
4. Once you label all five regions, click 'Find Similar Regions' to get other set of regions to label.

Selected Region Info: **Start(s):** 565.: **End(s):** 567 **Label:** ***Positive***

Listen and label these

Click and label regions below
- Region #1 (positive)
- Region #2 (positive)
- Region #3 (positive)
- Region #4 (negative)
- Region #5 (NEW)

Interactive annotation



Accurate!

Human

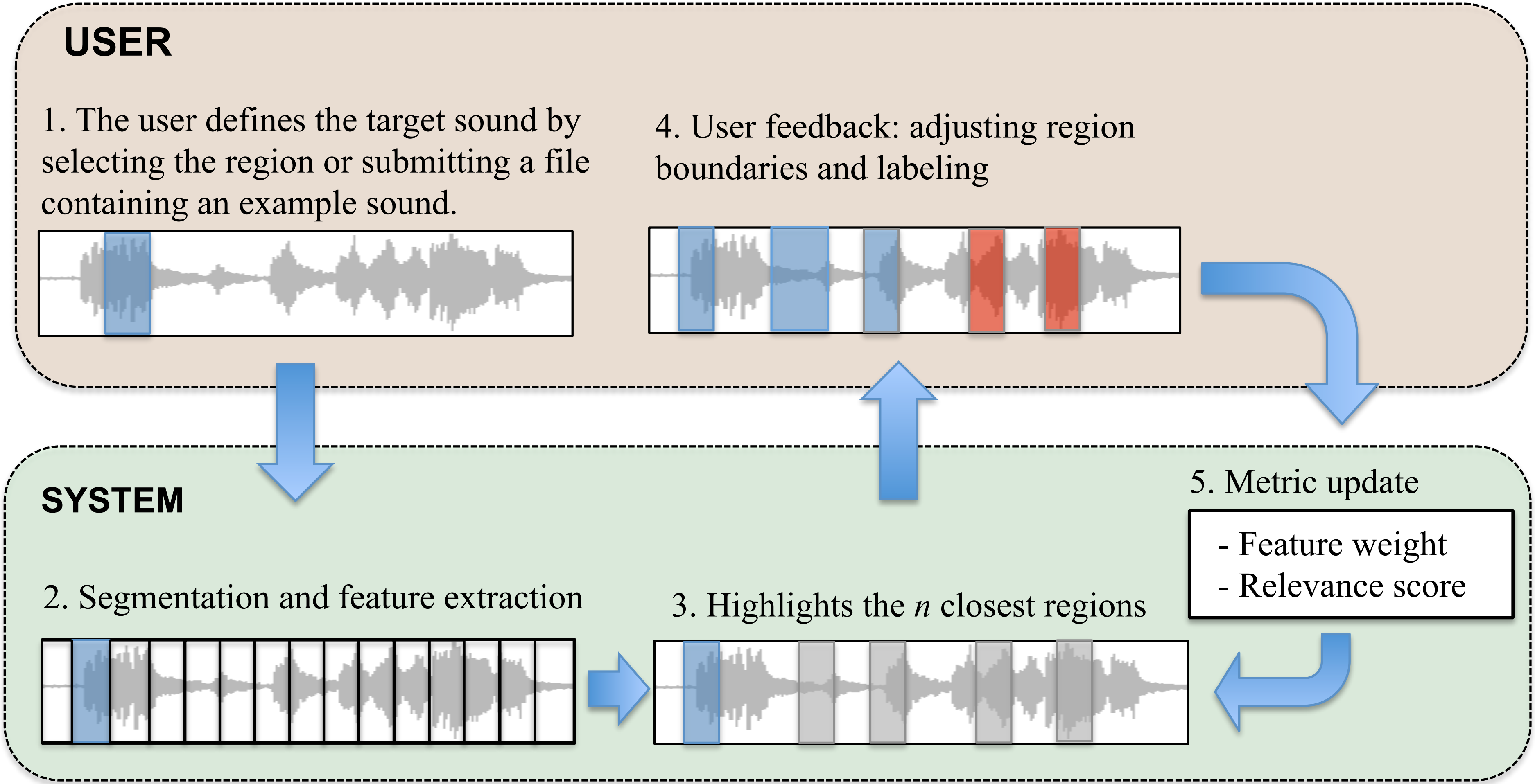
Machine Suggestion

User Feedback

Machine

Fast!

System Overview



USER

1. The user defines the target sound by selecting the region or submitting a file containing an example sound.

4. User feedback: adjusting region boundaries and labeling

SYSTEM

2. Segmentation and feature extraction

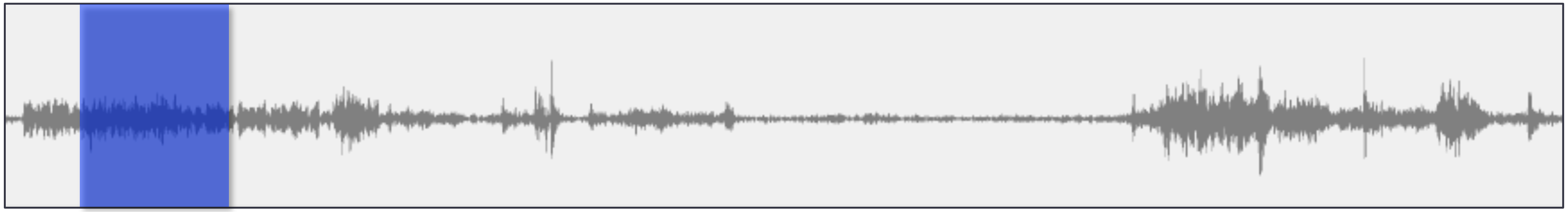
3. Highlights the *n* closest regions

5. Metric update

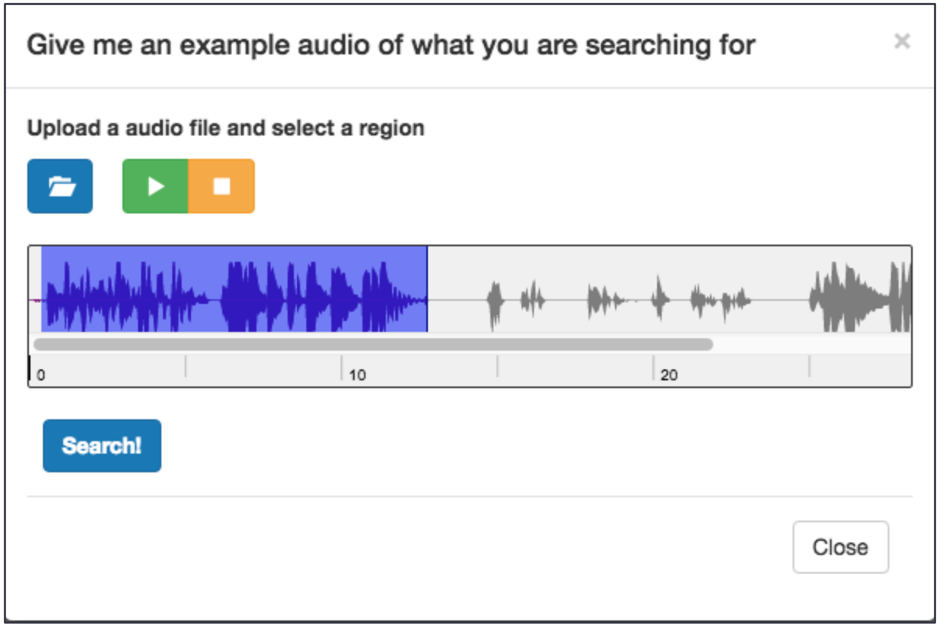
- Feature weight
- Relevance score

Defining the target sound

Method 1. Selecting the sound by dragging a mouse over the region



Method 2. Submitting a file containing an example sound (and select a region)

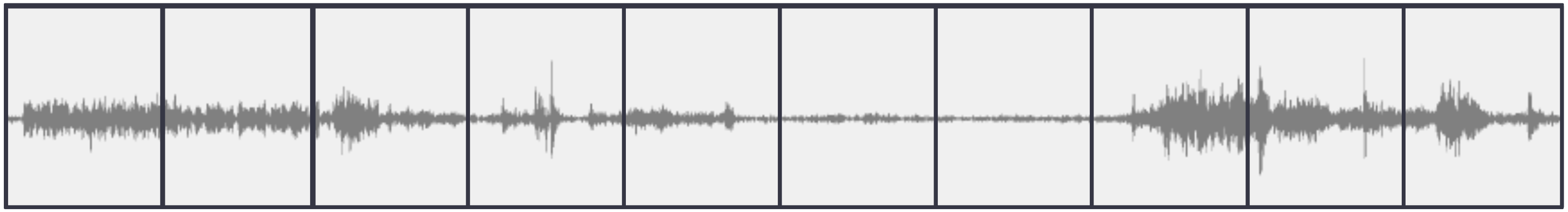


Give me an example audio of what you are searching for ×

Upload a audio file and select a region

Segmentation and feature extraction

Segmentation



Feature extraction: Mel Frequency Cepstral Coefficients (MFCCs)

➔Each segment is represented as 52-dimensoinal feature vector

➔ Distance between segments can be computed in the feature space

Relevance score and feature re-weight

- Relevance score
  - Measuring how relevant it is to the target sound
  - A nearest neighbor approach used in [Giorgio, 2007]

d: weighted Euclidean distance

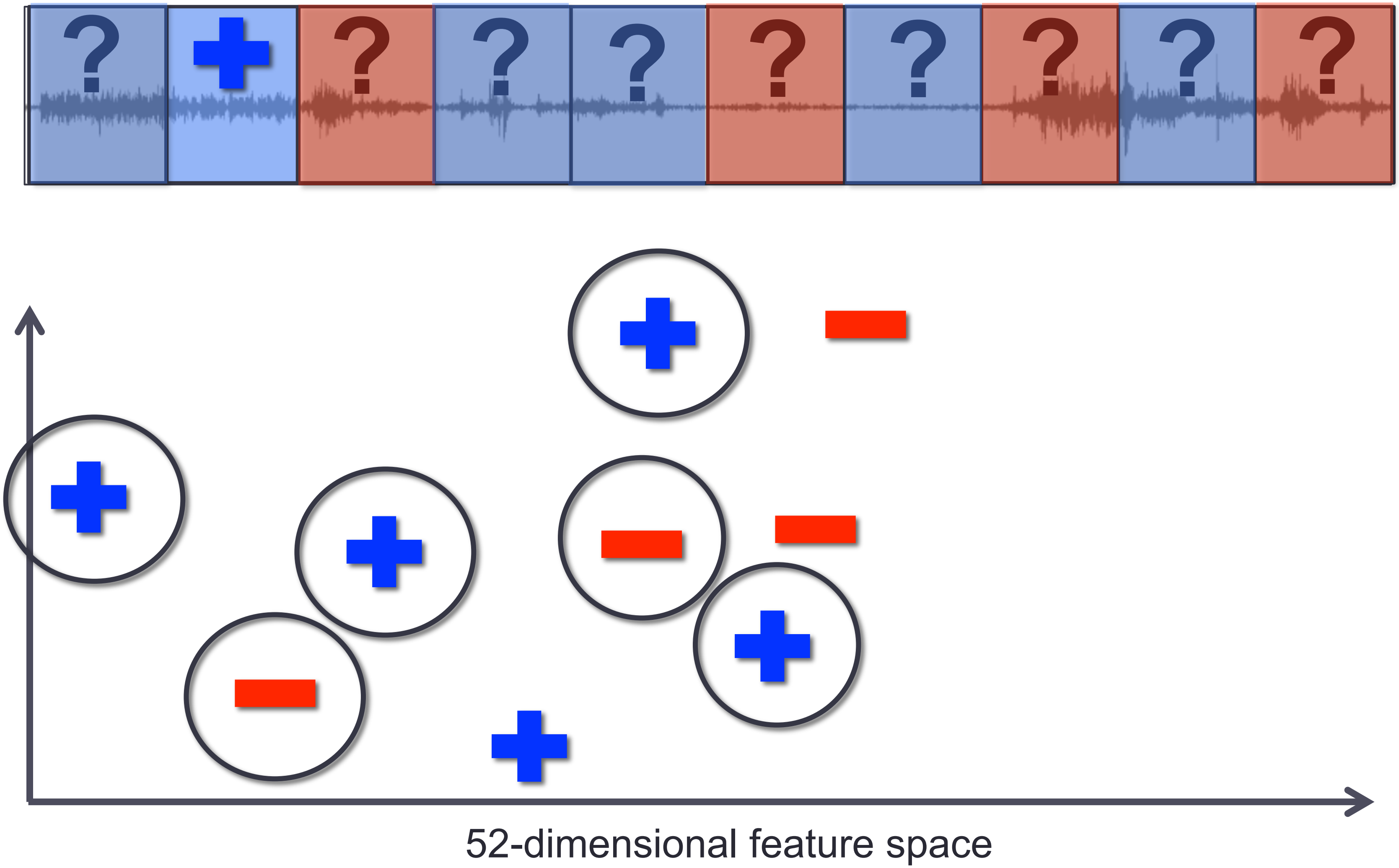
$$Rel(s) = \frac{d(s, s_n)}{d(s, s_n) + d(s, s_p)}$$

***S**$_n$*: Nearest negatively labeled segment

***S**$_p$*: Nearest positively labeled segment

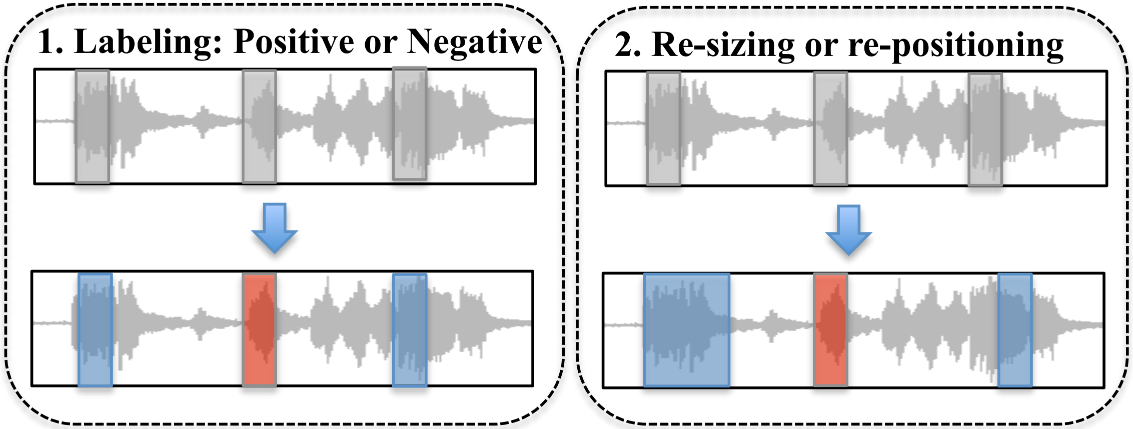
- Feature re-weight
  - Features are re-weighted based on labeled segments
  - Fisher’s criterion: More weights on the feature that contributes to the relatively better discrimination between positive and negative examples

Relevance score and feature re-weight

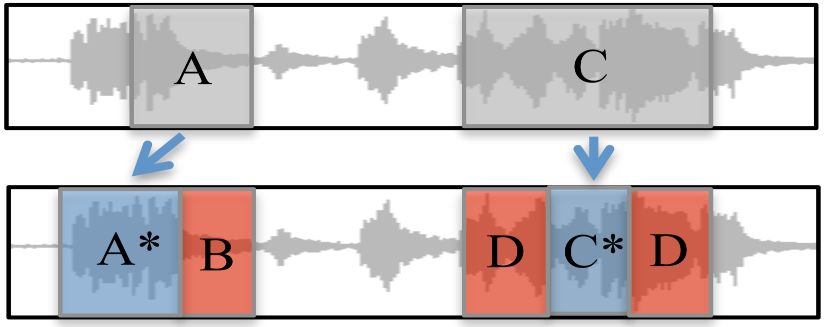


User feedback

- Users listen to the machine's suggestion (*n* segments) and provide two kinds of feedback to the system


1. Labeling: Positive or Negative
2. Re-sizing or re-positioning

- System automatically collects additional negative examples from users' boundary adjustments



DEMO

Watch the demo video and try the system out here:
http://www.bongjunkim.com/ised/

Evaluation

The two interfaces compared

- The interactive annotator
  - The initial target sound file is given to participants
  - The system presents 5 most relevant regions to user at each round.
- The manual annotator
  - The identical interface to the interactive annotator except for the removal of the recommendations from the system.
  - Listening to the track, every time they detect the target sound, they drag a mouse over the region containing the target sound.

Q1) Which interface enable participants to label given audio faster?

Q2) Do participants prefer the proposed interactive annotator to manual annotator?

Task procedure

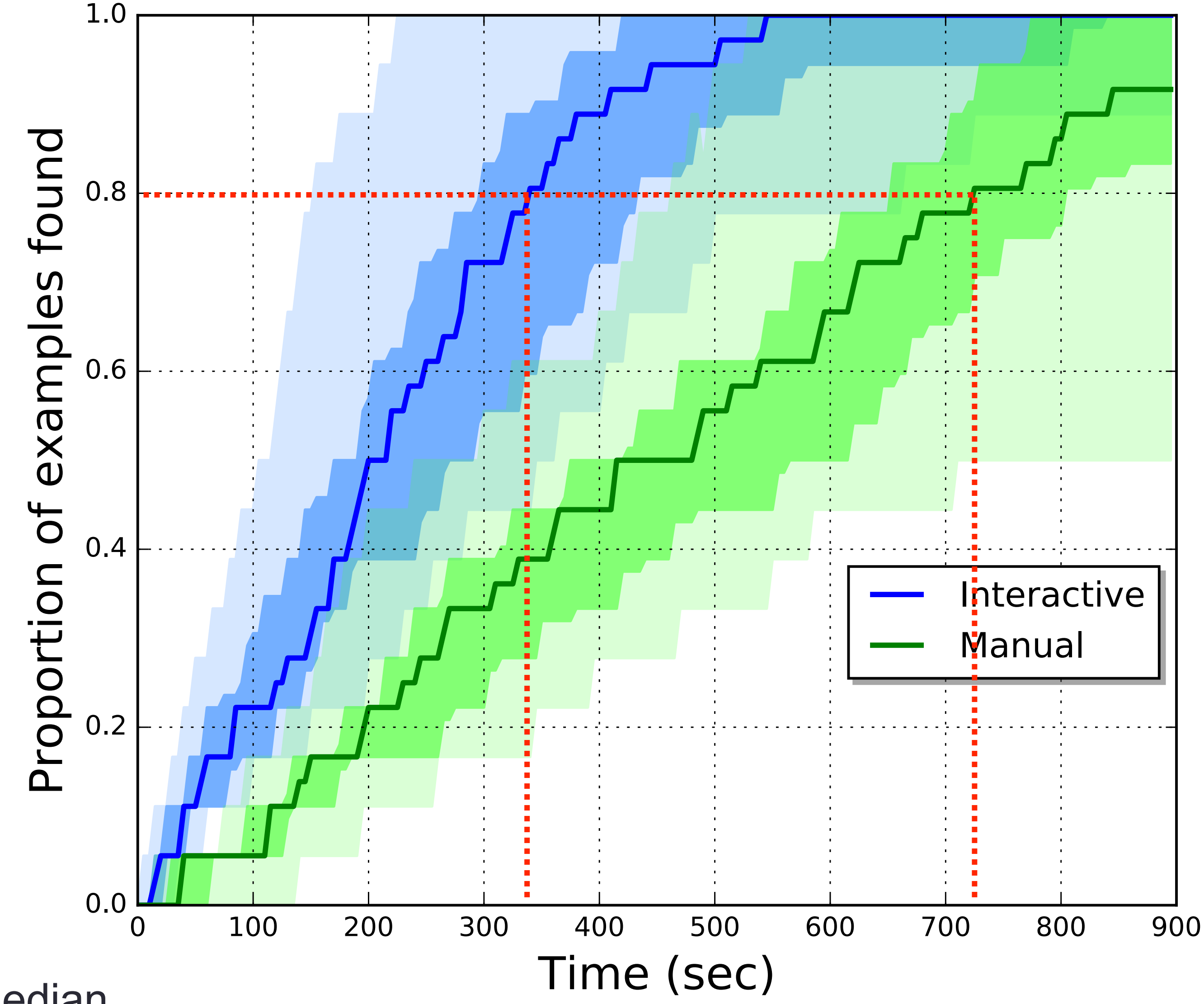
- Each subject has participated in one session.
  - In one session, a participant tried the two interfaces.
  - The manual annotator | the interactive annotator

- Counter-balanced design
  - Two interfaces: Interactive | manual
  - Two tasks: labeling door knock | human speech
    - Sound events in the first task are randomly reordered in the second task.
  - 20 participants were divided into 4 groups:
    - User group 1: Manual, Task 1 ➔ Interactive, Task 2
    - User group 2: Manual, Task 2 ➔ Interactive, Task 1
    - User group 3: Interactive, Task 1 ➔ Manual, Task 2
    - User group 4: Interactive, Task 2 ➔ Manual, Task 1

Task procedure

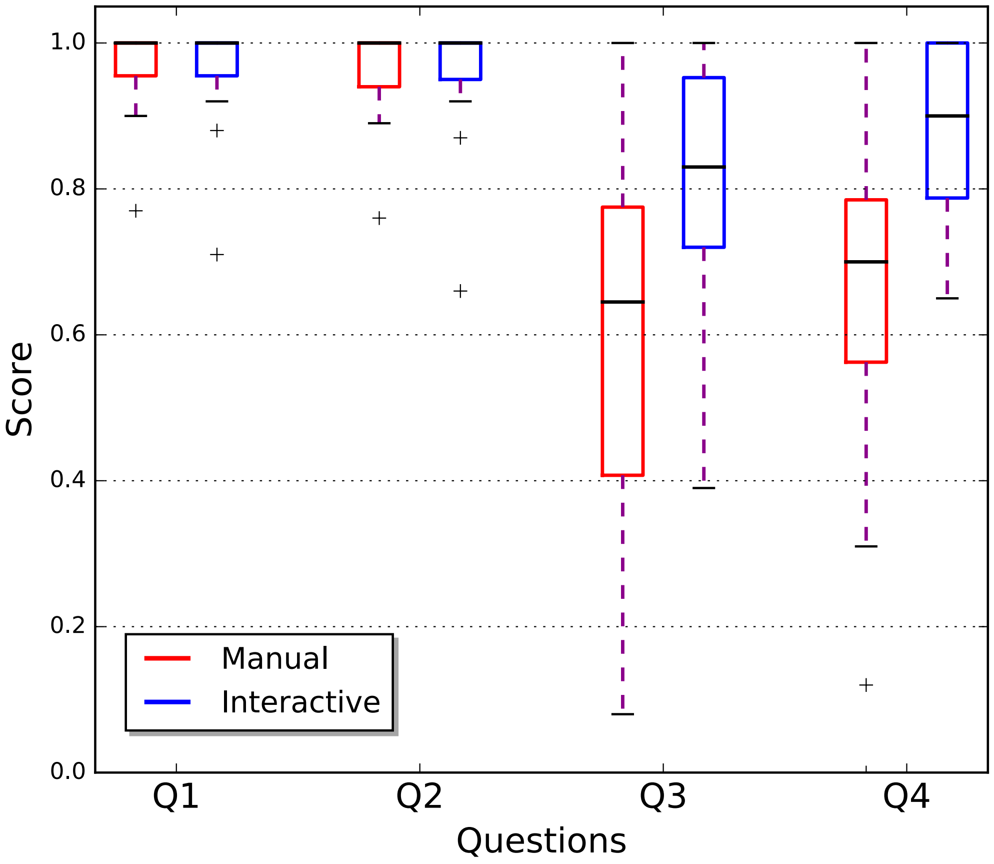
- Training session
  - Before the actual task, participants learn and practices how to use each interface for the task for at least 4 minutes.
- Two actual tasks
  - Task: find as many regions containing the target sound as they could in 15 minutes.
  - There are 18 events for each target sound in 12 minute long recording (participants did not know how many events to find).
- Questionnaire
  - After each task, participants were asked to report their experience with each interface.

Results



- Line: median,
- Dark and light bands: 75th, 25th percentile

Results – Self-reported performance



Q1: I had a clear understanding of the task.
Q2: I understood how to use this interface to achieve the given goal.
Q3: I was satisfied with using this interface.
Q4: I was able to label target sound events easily.

- Responses ranging from 0 (strongly disagree) to 1 (strongly agree)

- No difference between two interfaces for Q1 and Q2

- Significant difference between two interfaces for Q3, Q4. ($p<0.05$)

- Participants were more satisfied with the interactive annotator than the manual one.

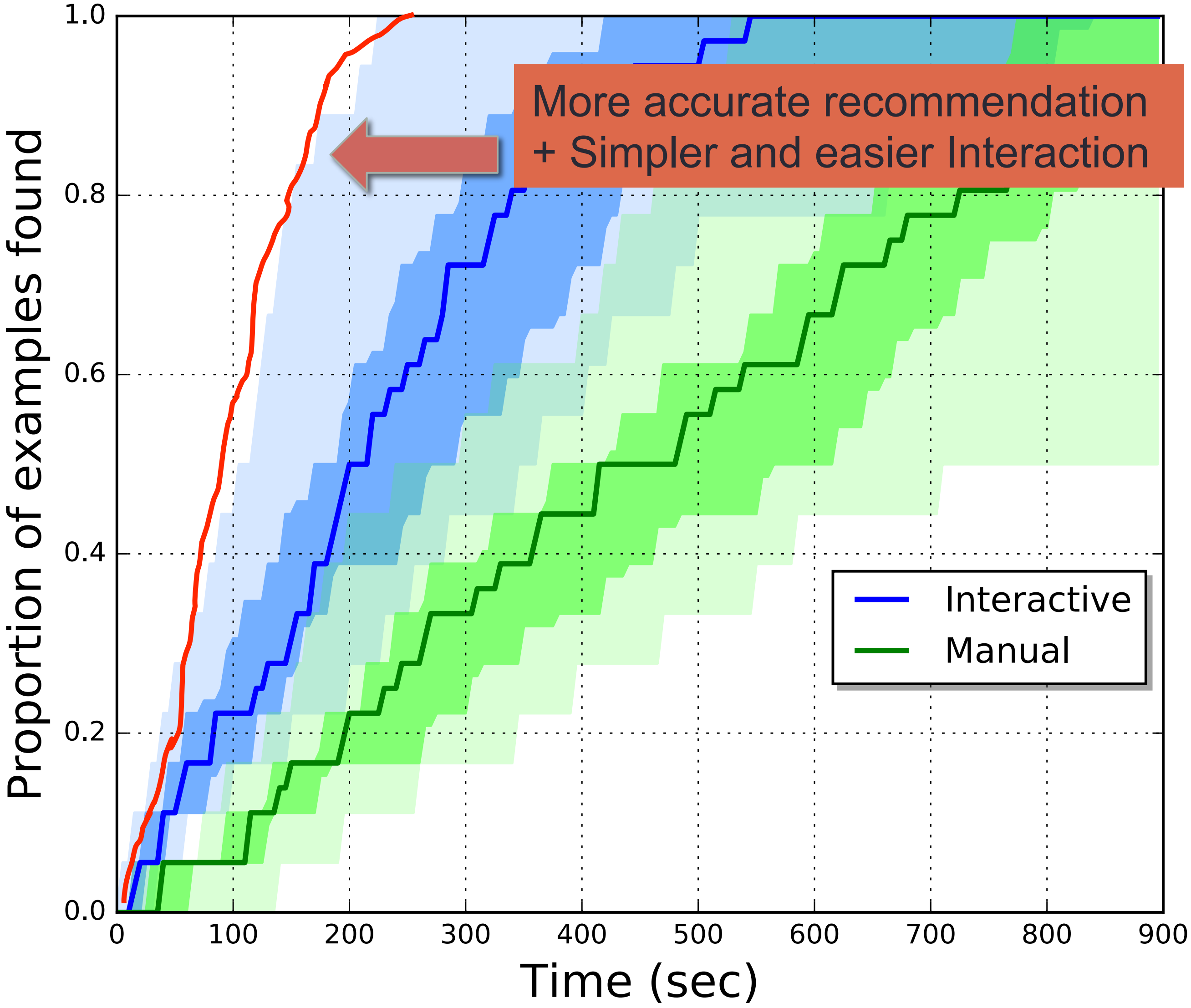
Conclusions

- A new approach for environmental sound event annotation using interactive learning by user's relevance feedback.

- The log data from the experiment showed that the proposed interface lets users find sparsely-distributed target sounds roughly twice as fast as manually labeling the target sounds.

- From the survey response data, it seems that most participants were more satisfied with the interactive annotator against the manual annotator.

Future works: improving this speed-up even more?



More accurate recommendation
+ Simpler and easier Interaction

Future works

- Improving this speed-up by exploring alternate feature representations and classifiers
- Developing a systematic stopping criterion
- Designing workflow for multiclass labeling problems

Thanks

http://www.bongjunkim.com/ised