A configurable speech recognition pipeline for social robots

Creating a fusion framework for analyzing speech and meta data

---

Robert Feldhans

11.7.2019

Master Thesis

Supervisiors: Sven Wachsmuth
Florian Lier
Birte Richter

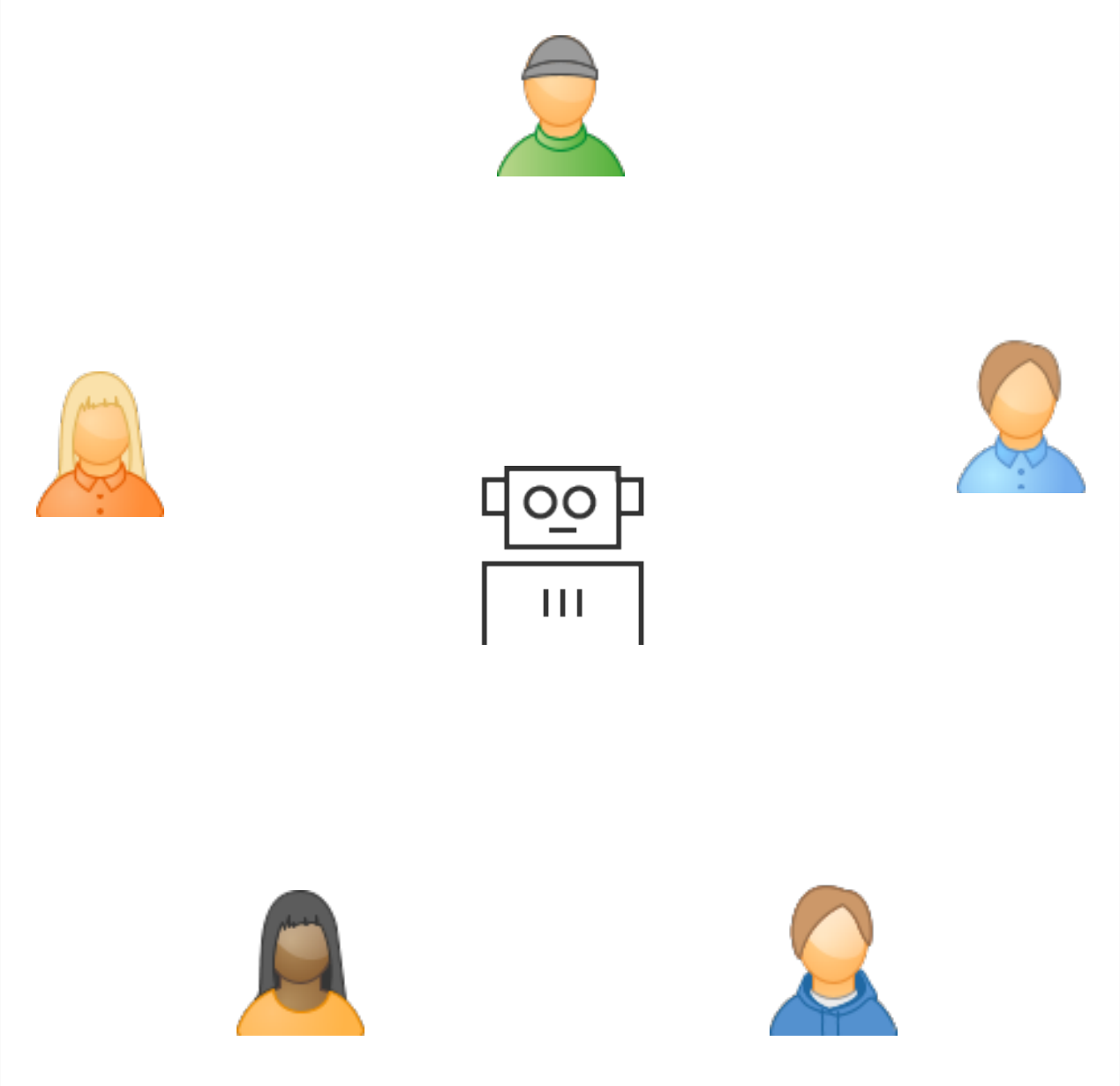
Reviewers: Sven Wachsmuth
Florian Lier

Content

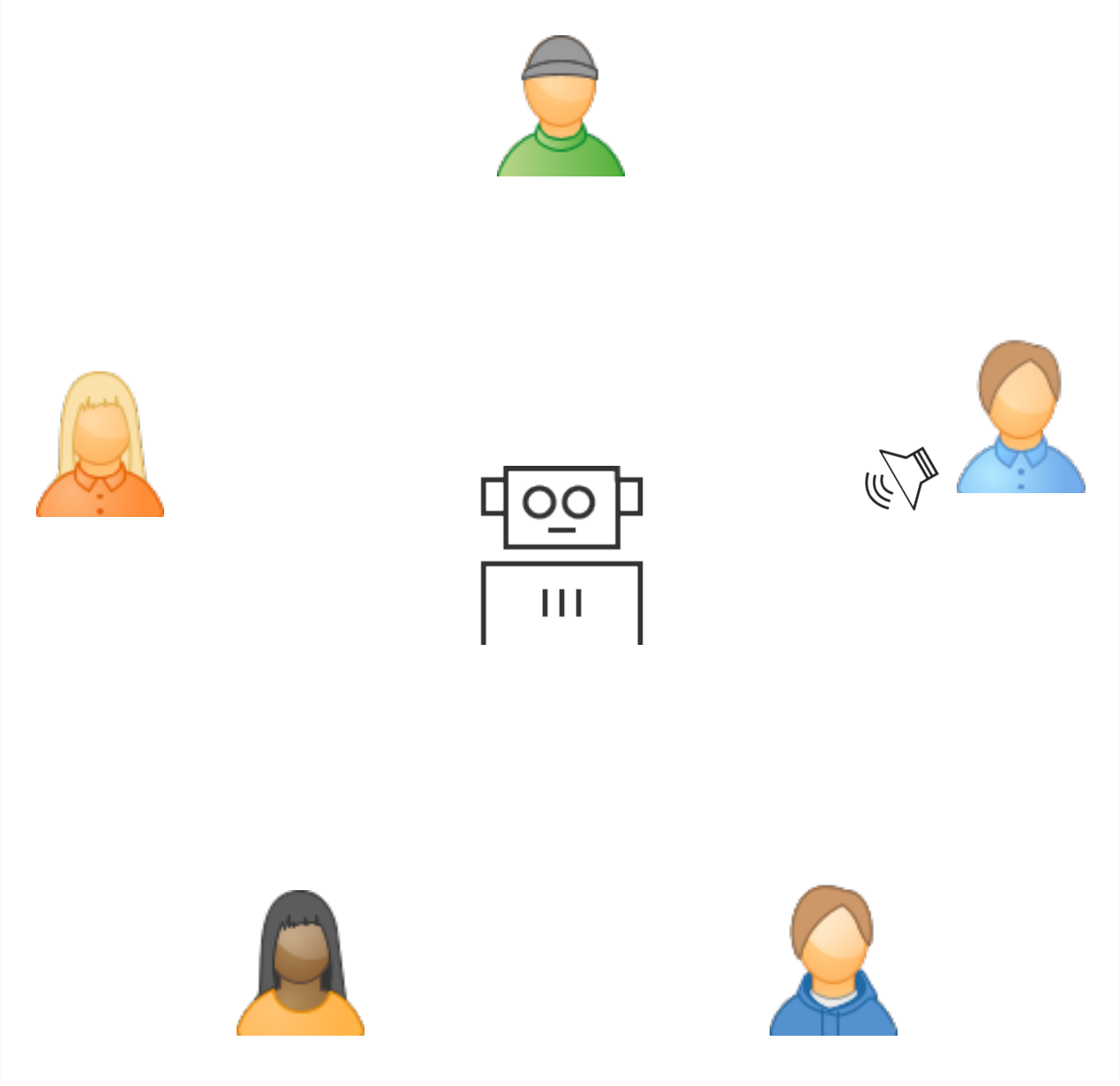
1. Motivation
2. Pipeline
3. Planned Evaluation
4. Current State

Motivation

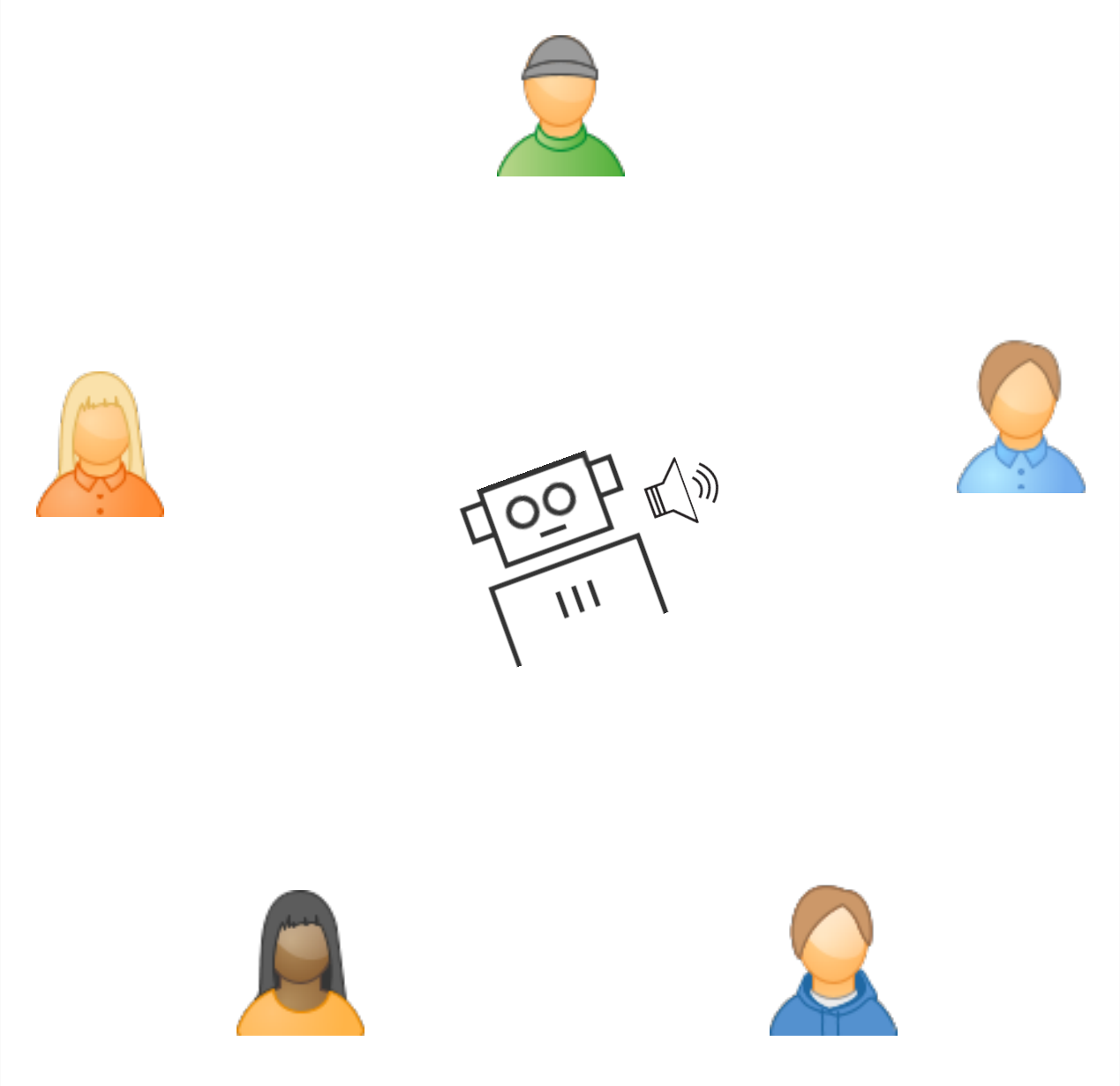
Foray RoboCup Speechrec Task



Foray RoboCup Speechrec Task



Foray RoboCup Speechrec Task



Foray RoboCup II

RoboCup

What are requirements for a robot regarding speech recognition? What is nice to have? What would be the optimum?

RoboCup

What are requirements for a robot regarding speech recognition? What is nice to have? What would be the optimum?

- Robustness (especially regarding noise)
- Beamforming / Signal enhancing
- Modularity
- Synchronization

Additional Thoughts

These requirements set, what can we get "for free"?

Additional Thoughts

These requirements set, what can we get "for free"?

- Gender Recognition
- Emotion Recognition
- Voice Recognition
- Signal Enhancing

Additional Thoughts

These requirements set, what can we get "for free"?

- Gender Recognition
- Emotion Recognition
- Voice Recognition
- Signal Enhancing

Synchronization can be used to fuse all this information!

Foray: Latency vs Signal Integrity

Latency

- Important for humans, stuttering makes recognition hard!
- Not that important for machines, as they can wait till they get the full signal

Foray: Latency vs Signal Integrity

Latency

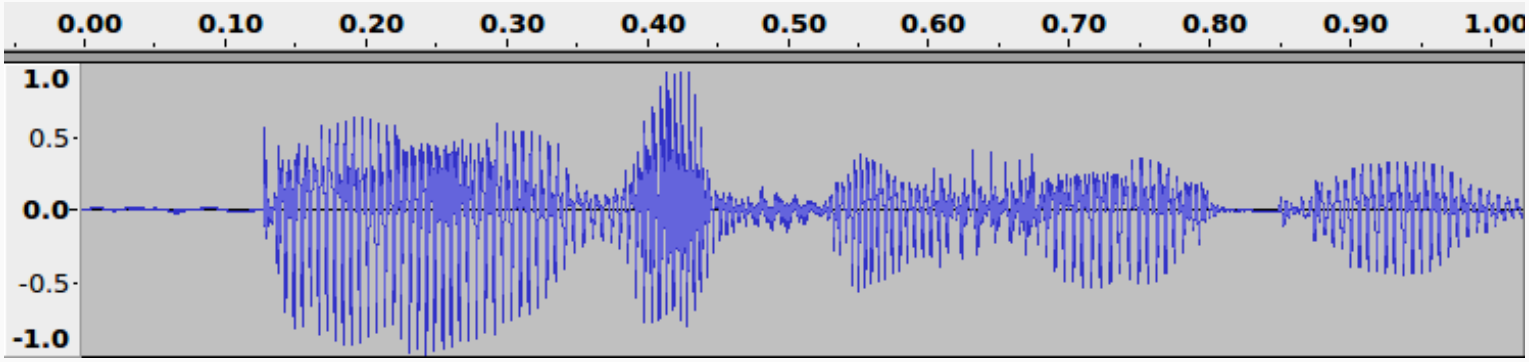
- Important for humans, stuttering makes recognition hard!
- Not that important for machines, as they can wait till they get the full signal

Signal Integrity

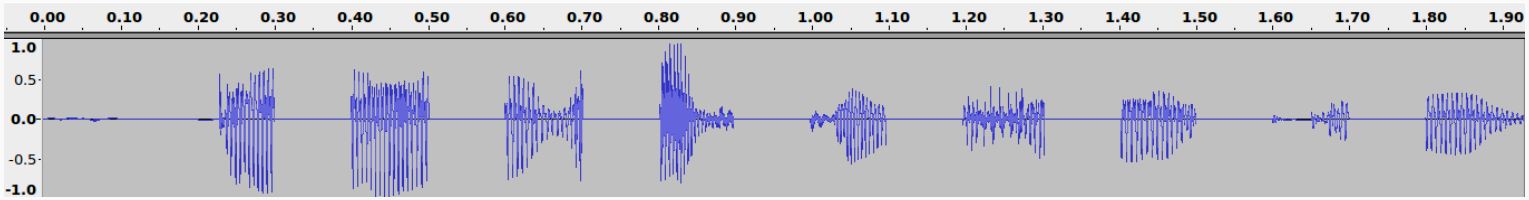
- Definitely recognizable by humans, but we can deal with it
- Important for machines, otherwise artifacts may occur!

Artifacts

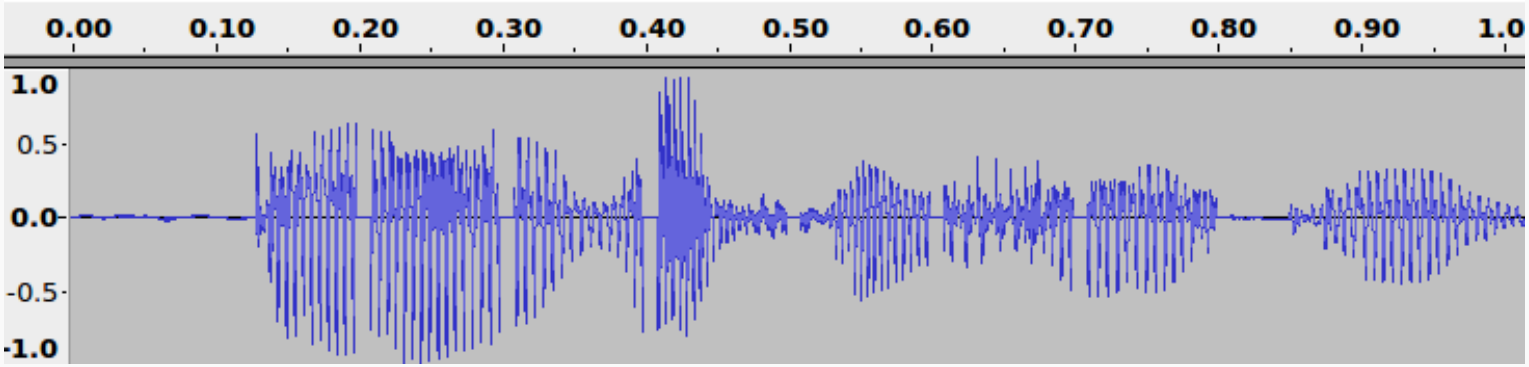
Play Normal Sound



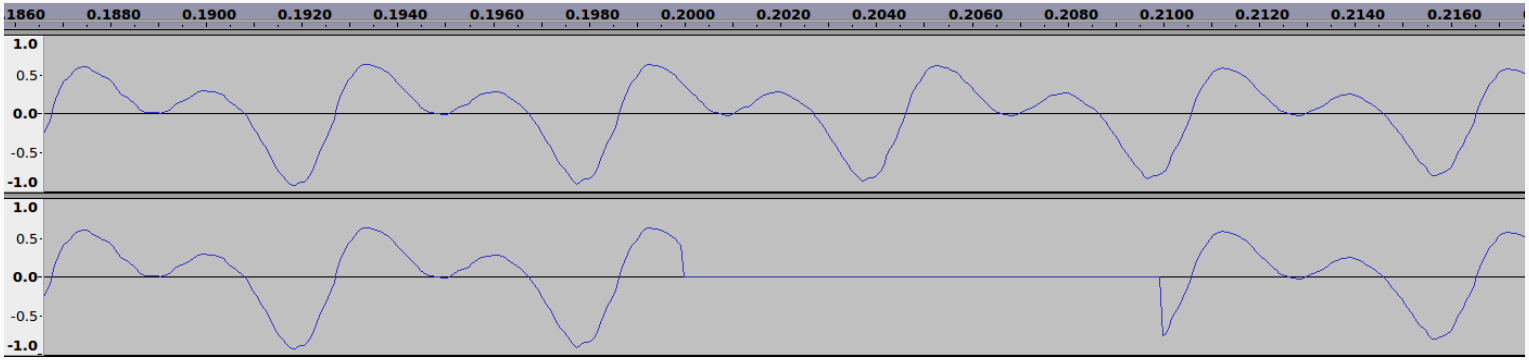
Play Sound with added silence



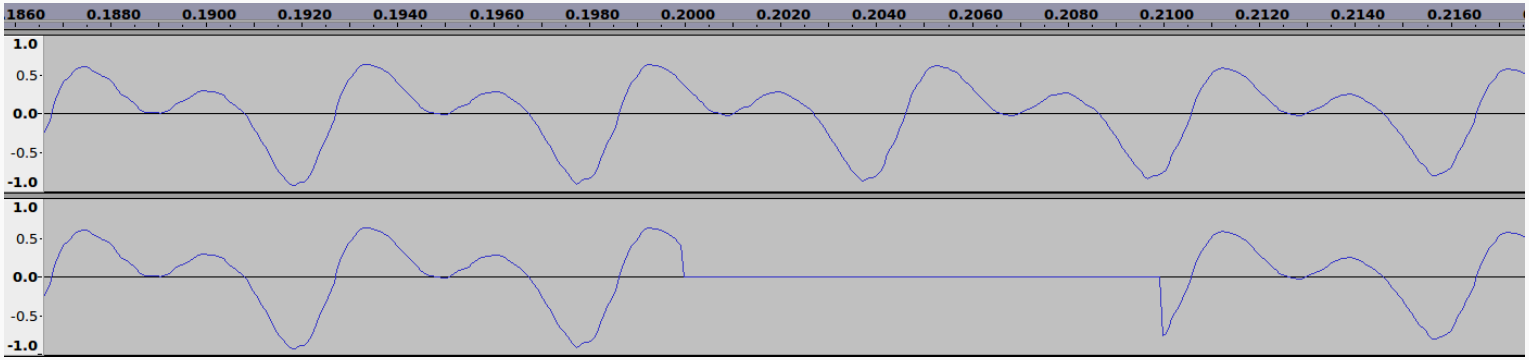
Play Soundwith removed parts



Artifacts II

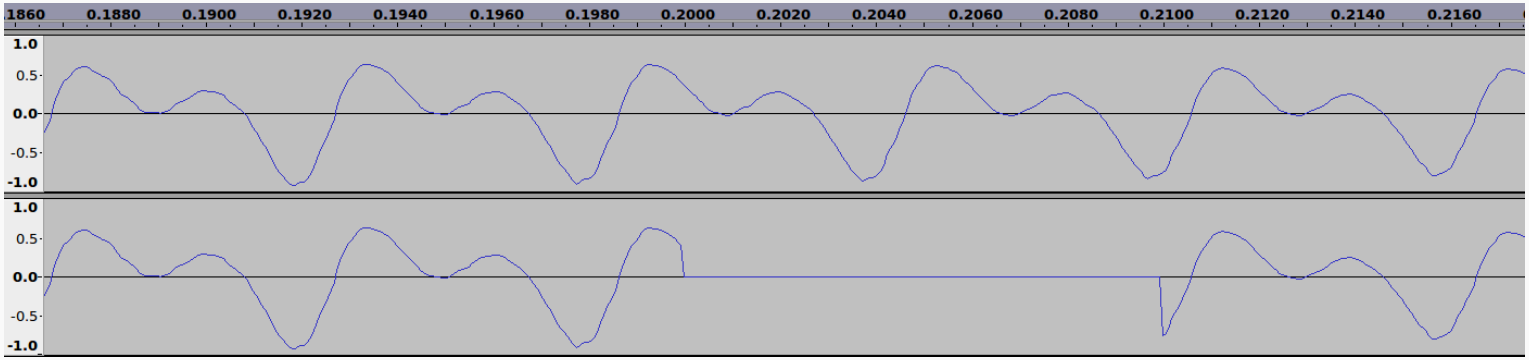


Artifacts II



Fast Fourier Transformations?

Artifacts II



Fast Fourier Transformations?

Good luck with that!

Evaluation of existing solutions

ALSA/ Pulseaudio

- Available everywhere
- Lack options for easy audio in- & output
- Timestamping/ adding custom meta information is basically impossible

Evaluation of existing solutions

ALSA/ Pulseaudio

- Available everywhere
- Lack options for easy audio in- & output
- Timestamping/ adding custom meta information is basically impossible

Jack Audio/ Gstreamer

- Strong focus on real time audio
- Timestamping/ adding custom meta information is basically impossible

Pipeline

Idea

Library

- Focus on transmitting audio
- No audio shall be lost!

Idea

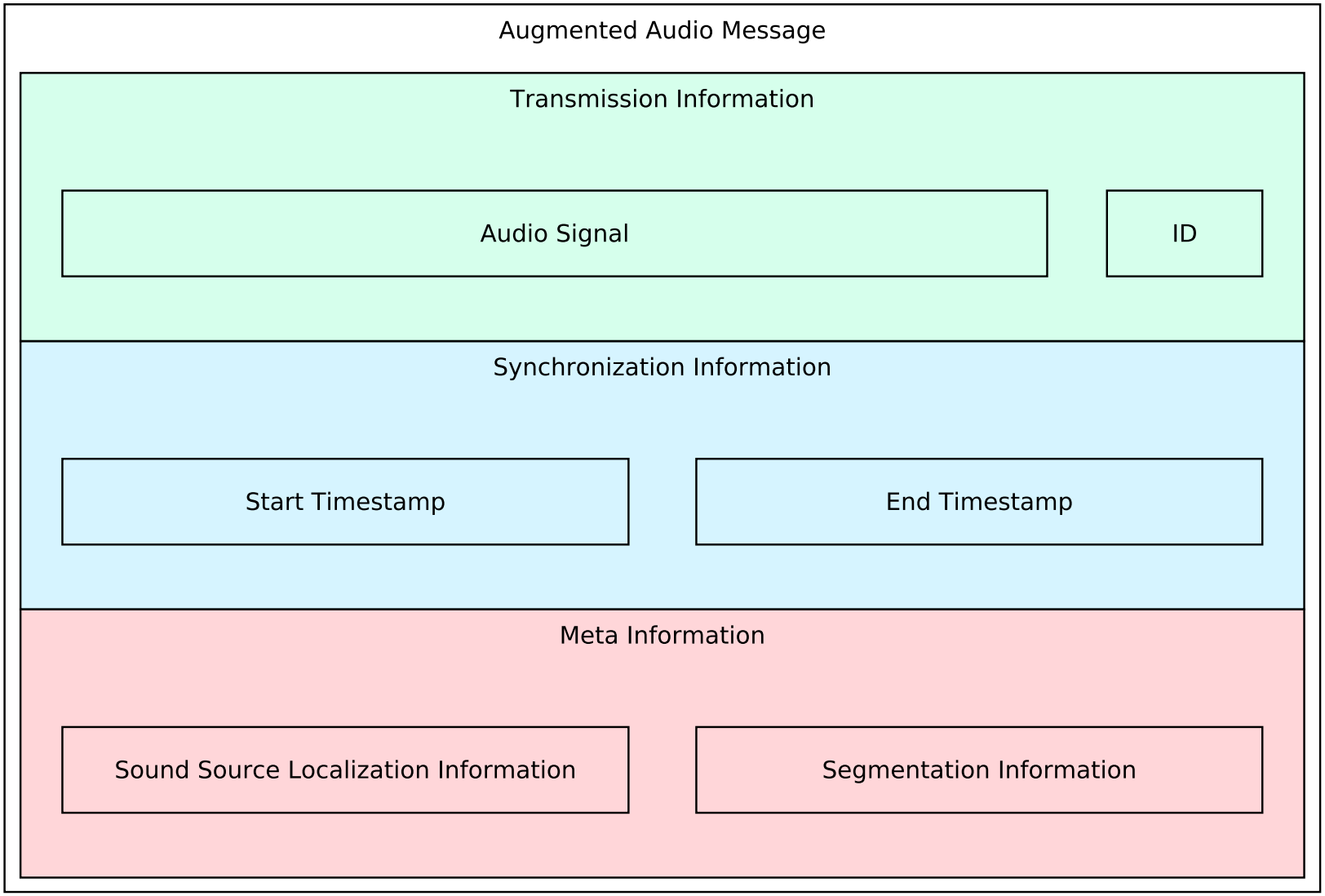
Library

- Focus on transmitting audio
- No audio shall be lost!

Orchestrator

- Focus on fusion of meta-data/ synchronizing audio
- Handle configuration of components (under the hood)

Augmented Audio



Augmented Audio

AugmentedAudio.msg

```
1  int8[] signal
2  RecordingTimeStamps time
3  bool segmentation_ended
4  int32 id
5  SSLDir[] directions
```

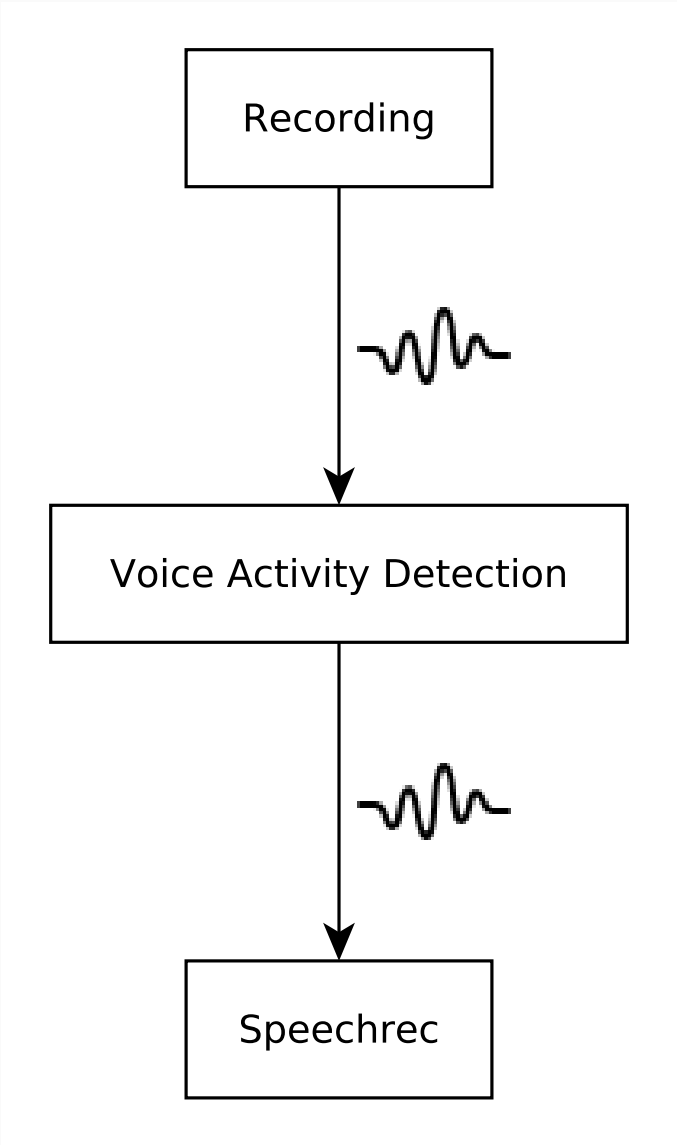
RecordingTimeStamps.msg

```
1  time start
2  time finish
```

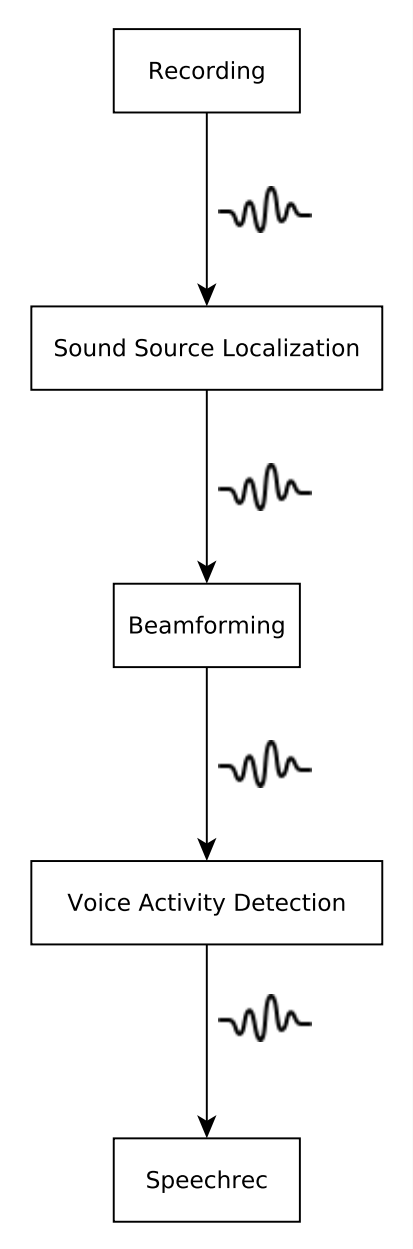
SSLDir.msg

```
1  string sourceId
2  float32 angleVertical
3  float32 angleHorizontal
```

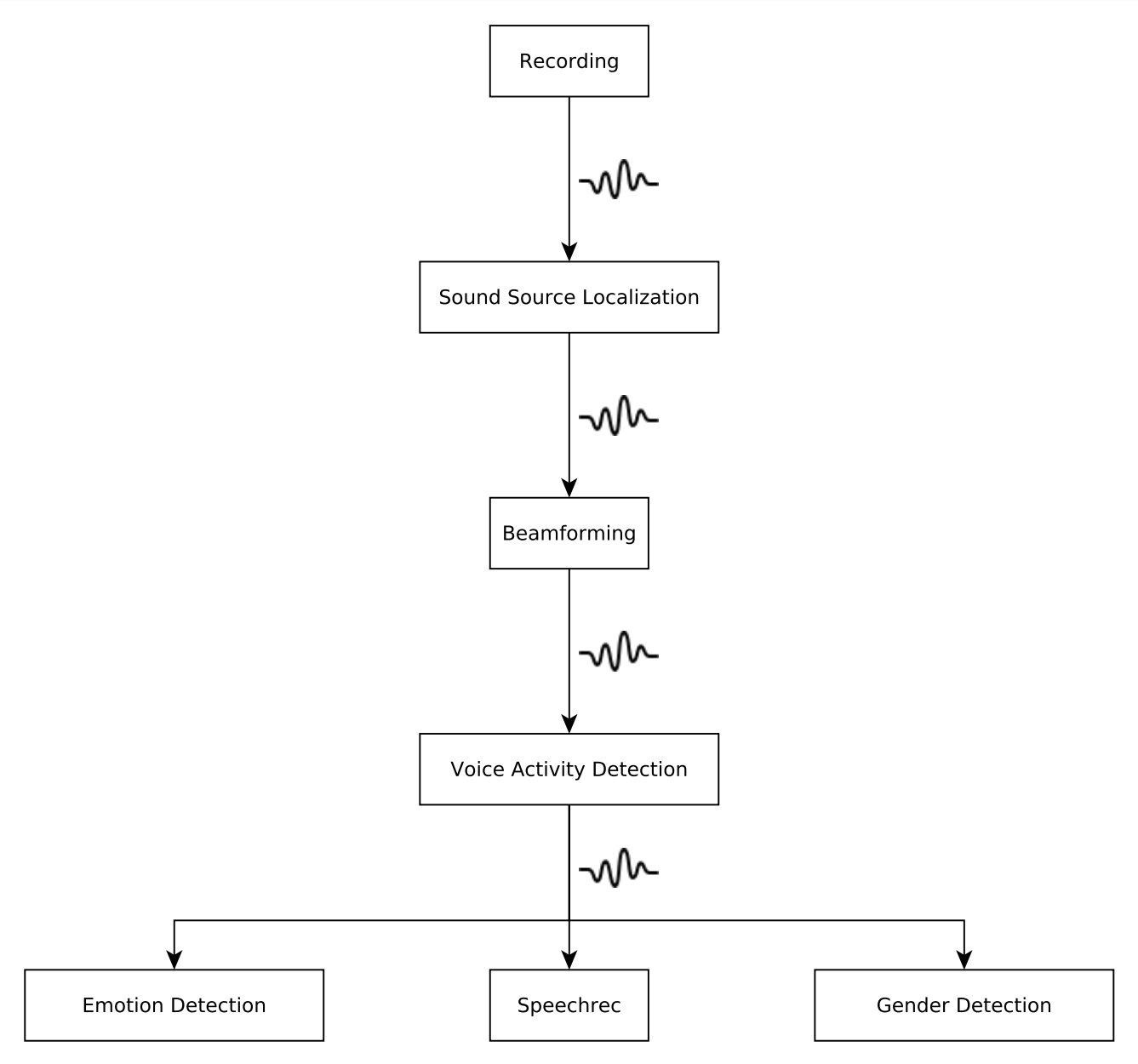
Architecture (Sound Flow)



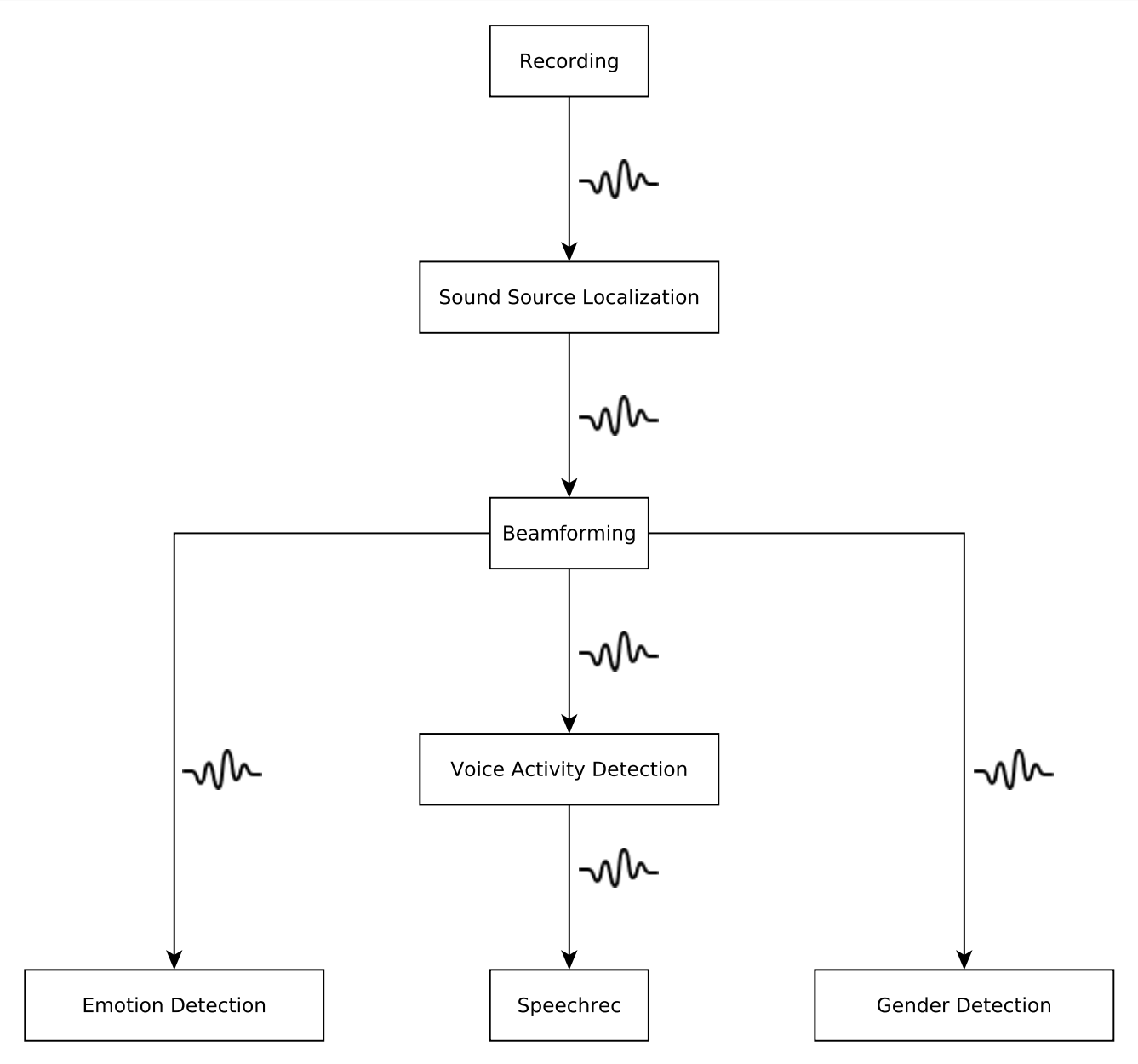
Architecture (Sound Flow)



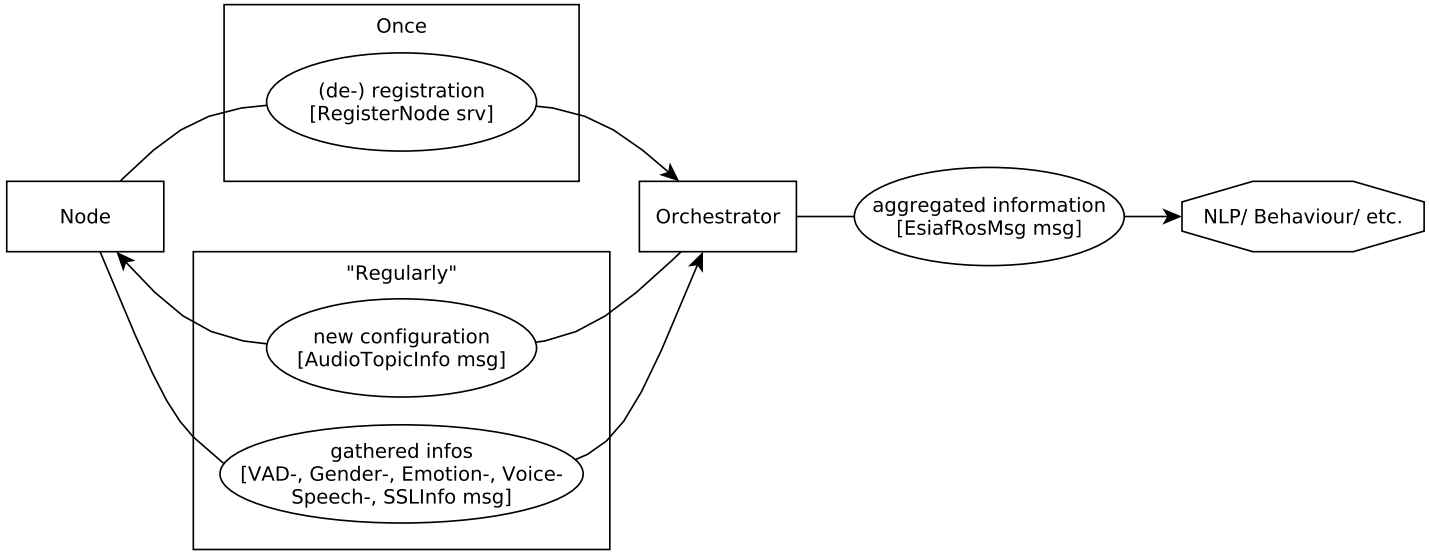
Architecture (Sound Flow)



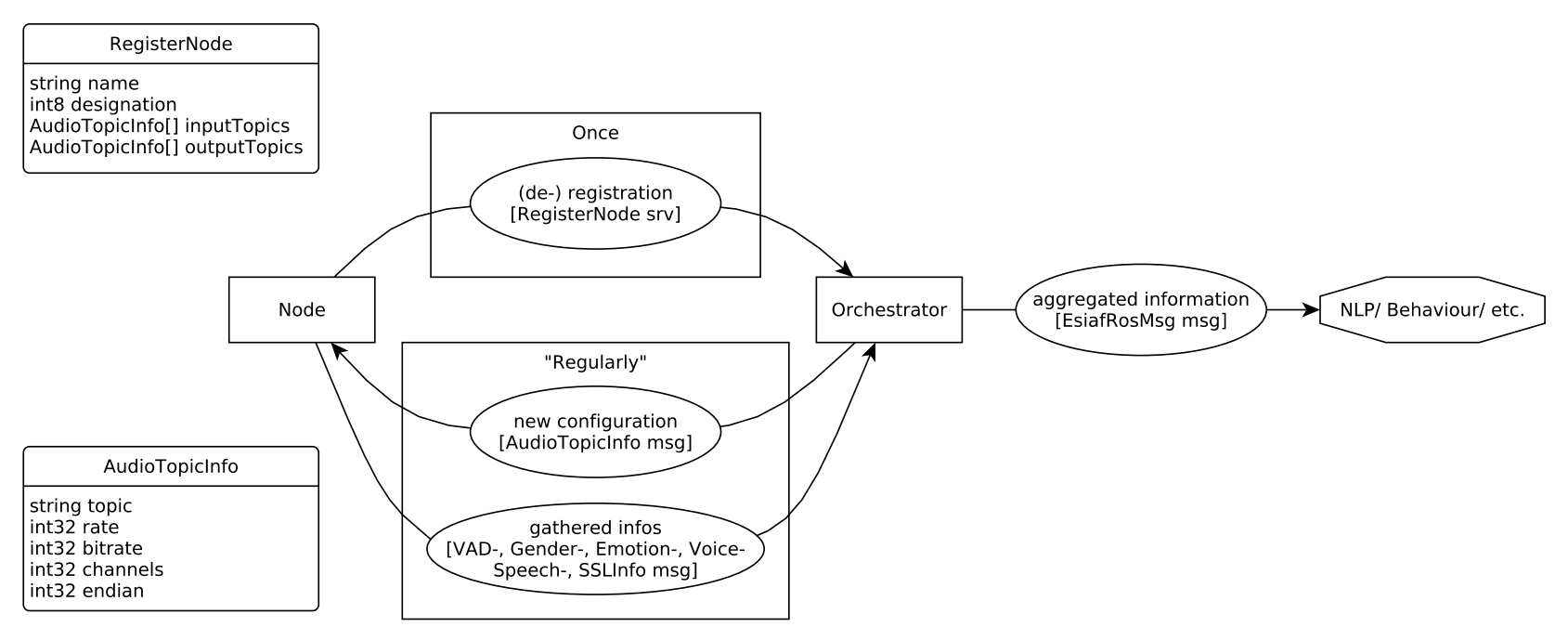
Architecture (Sound Flow)



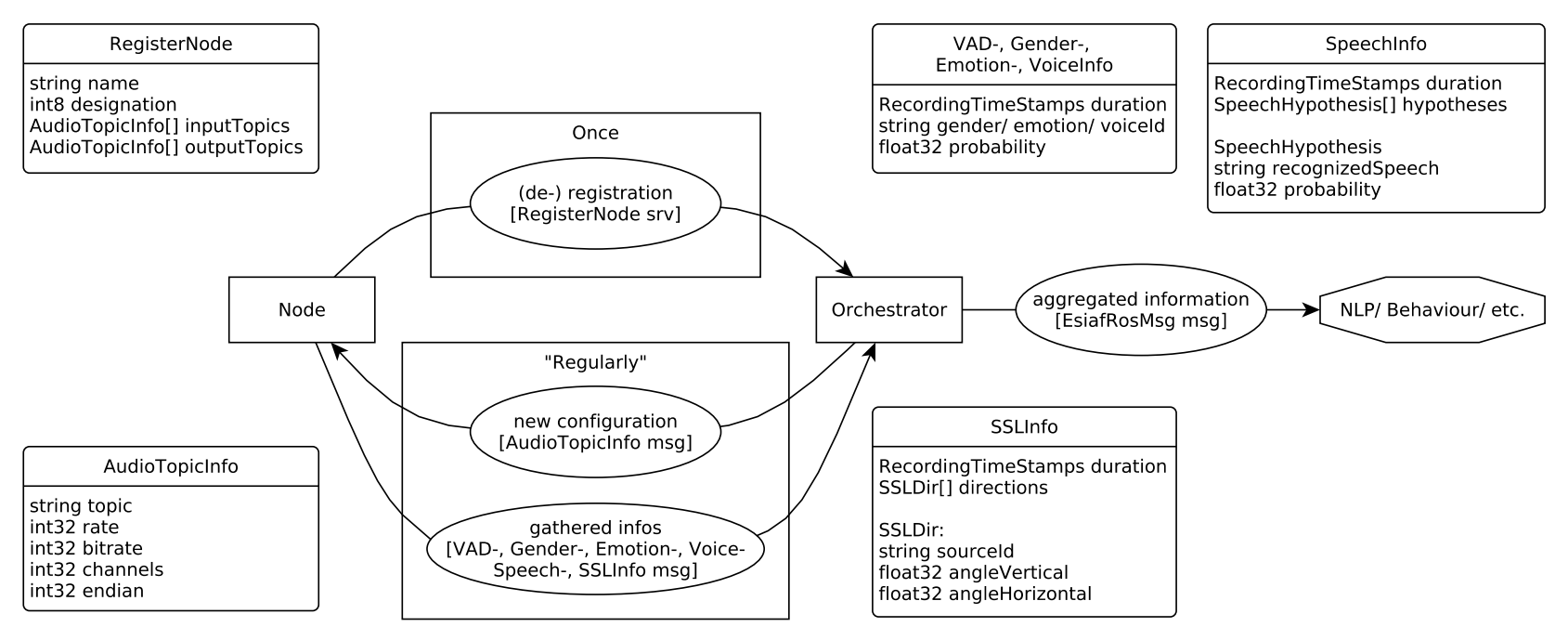
Architecture (Information Flow)



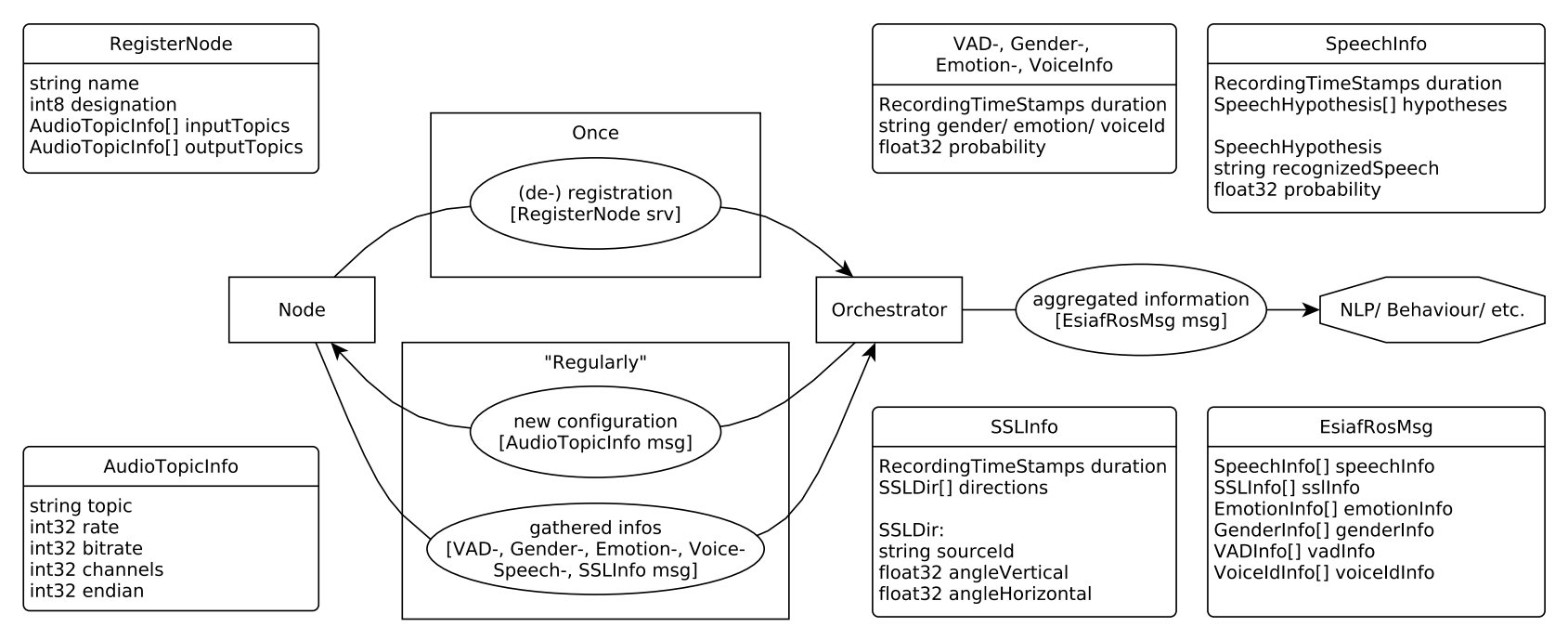
Architecture (Information Flow)



Architecture (Information Flow)



Architecture (Information Flow)



Planned Evaluation

RoboCup Speechrec Task

RoboCup Speechrec Task

Pros:

- Is widely used to test robots speech recognition capabilities
- Score can be used to compare against dozens of other robots
- Fusion could result in higher score

RoboCup Speechrec Task

Pros:

- Is widely used to test robots speech recognition capabilities
- Score can be used to compare against dozens of other robots
- Fusion could result in higher score

Cons:

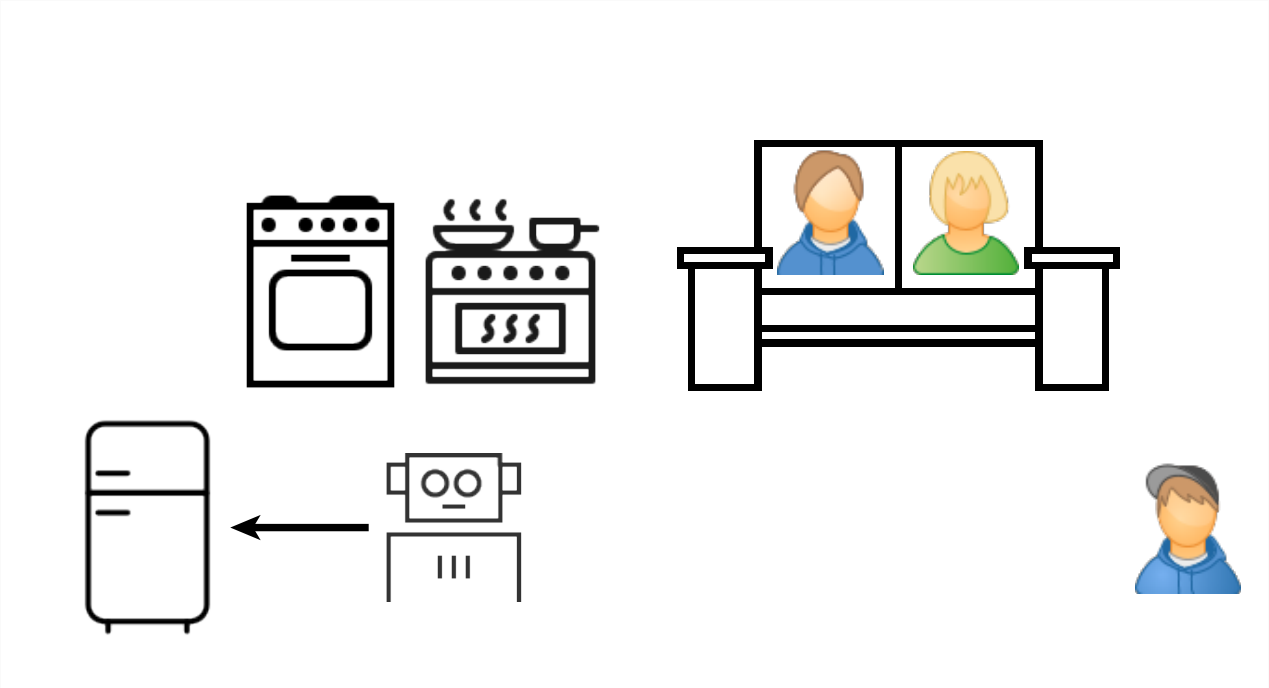
- Setup (noise) is a bit tricky, but can be adequately simulated using several speakers in a big echoing hall
- Is somewhat random, because some questions may be more difficult than others (however, difficulty may depend on components)
- Encompasses a bit of visual person recognition as well as "NLP", but both are irrelevant for us

Dataset Evaluation

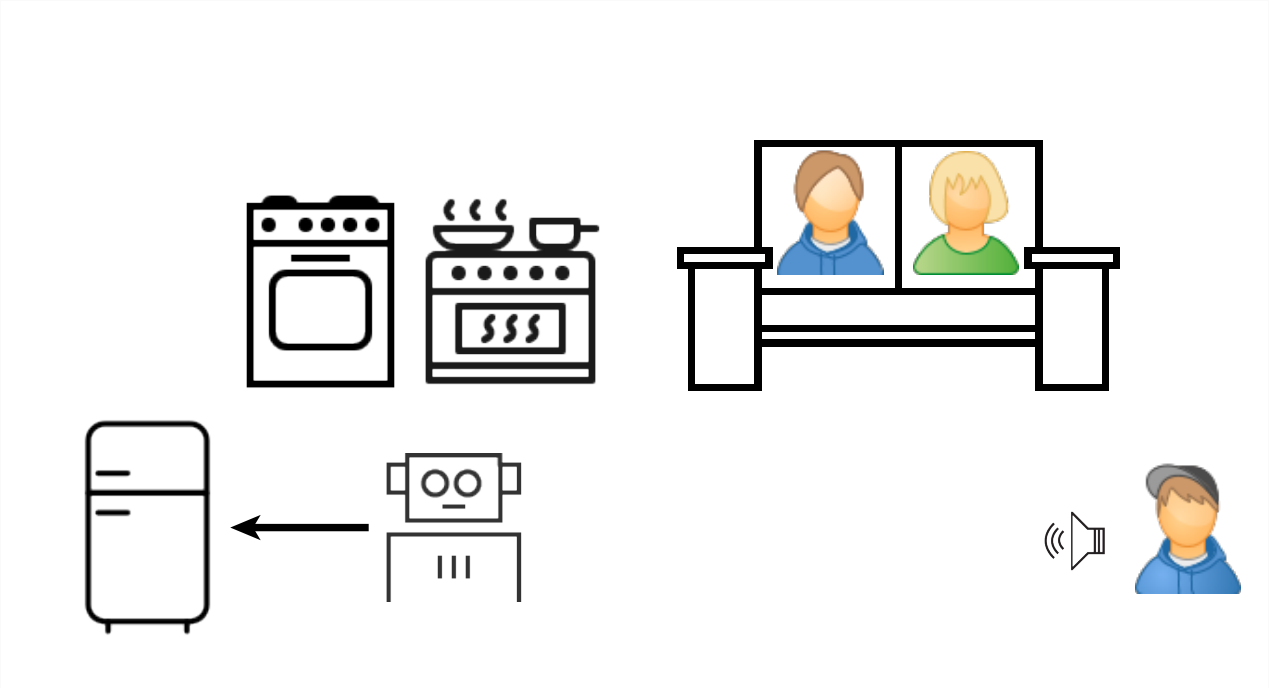
Everything is modular, so even the input can be switched out, so just read wav files instead of microphone input.
Compare WER and CPU/ overall time between...

- Bare minimum pipeline
- "Longer" pipeline (more preprocessing)
- "Wider" pipeline (more parallel components)
- All encompassing pipeline
- Other pipelines (e.g. apartment)

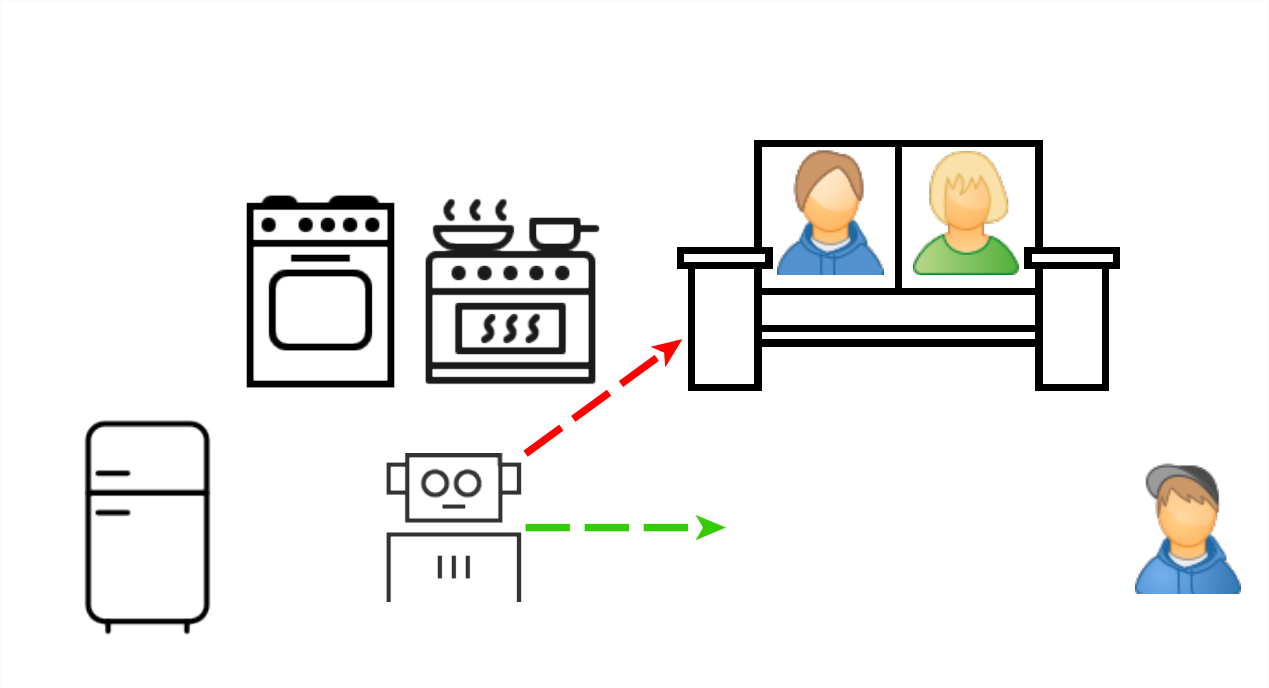
Fusion Test



Fusion Test



Fusion Test



Fusion Test

What to evaluate with this test?

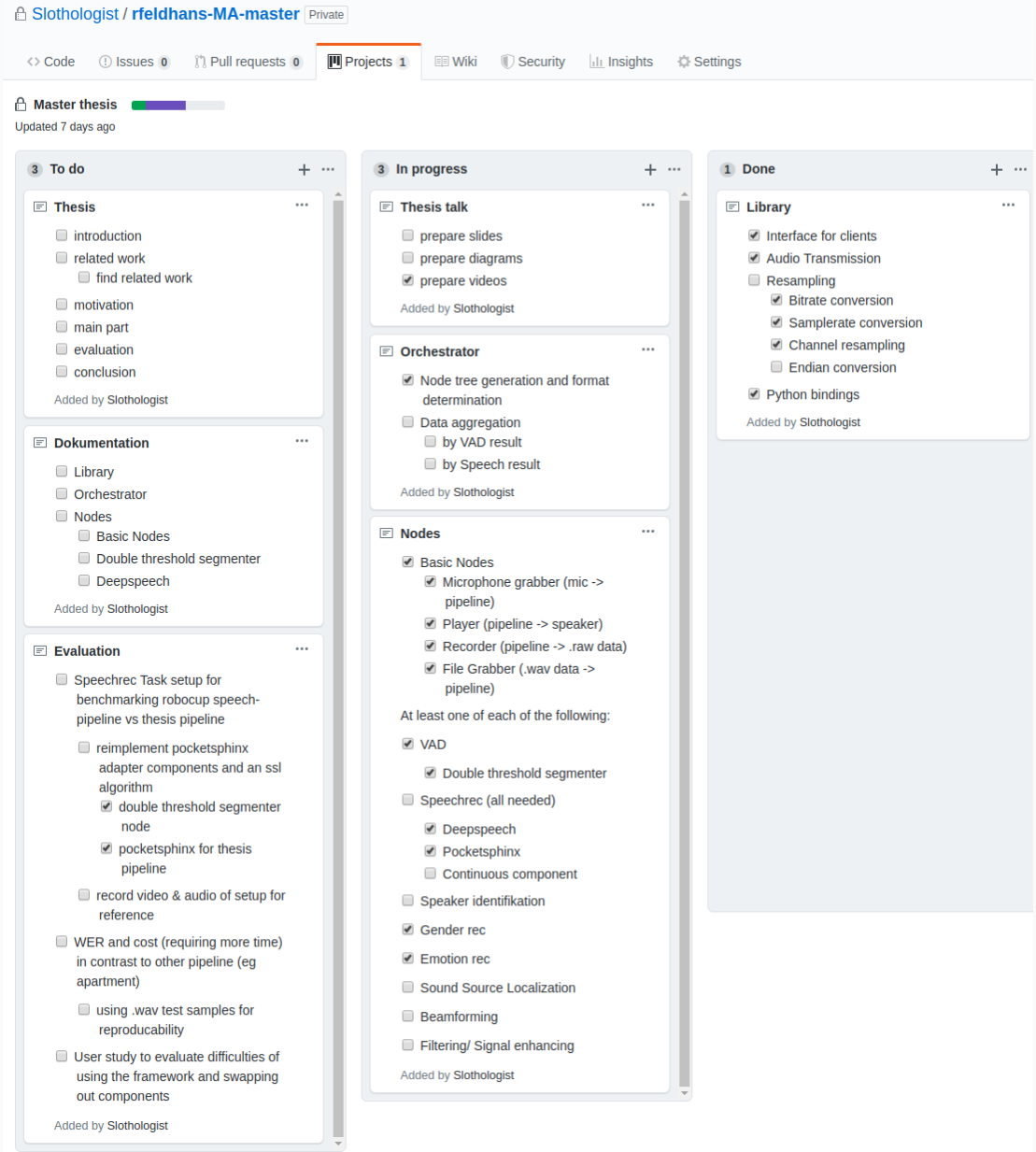
Fusion Test

What to evaluate with this test?

- Is the robot capable to robustly determine who is talking to it?
- Robustness tested with either manual jerking of the robot or with the help of Peppers autonomous life
- Qualitative analysis

Current State

Progress



Slothologist / **rfeldhans-MA-master** Private

Code | Issues 0 | Pull requests 0 | Projects 1 | Wiki | Security | Insights | Settings

Master thesis

Updated 7 days ago

3 To do

Thesis

- [ ] introduction
- [ ] related work
  - [ ] find related work
- [ ] motivation
- [ ] main part
- [ ] evaluation
- [ ] conclusion

Added by Slothologist

Dokumentation

- [ ] Library
- [ ] Orchestrator
- [ ] Nodes
  - [ ] Basic Nodes
  - [ ] Double threshold segmenter
  - [ ] Deepspeech

Added by Slothologist

Evaluation

- [ ] Speechrec Task setup for benchmarking robocup speech-pipeline vs thesis pipeline
  - [ ] reimplement pocketsphinx adapter components and an ssl algorithm
    - [x] double threshold segmenter node
    - [x] pocketsphinx for thesis pipeline
  - [ ] record video & audio of setup for reference
- [ ] WER and cost (requiring more time) in contrast to other pipeline (eg apartment)
  - [ ] using .wav test samples for reproducability
- [ ] User study to evaluate difficulties of using the framework and swapping out components

Added by Slothologist

3 In progress

Thesis talk

- [ ] prepare slides
- [ ] prepare diagrams
- [x] prepare videos

Added by Slothologist

Orchestrator

- [x] Node tree generation and format determination
- [ ] Data aggregation
  - [ ] by VAD result
  - [ ] by Speech result

Added by Slothologist

Nodes

- [x] Basic Nodes
  - [x] Microphone grabber (mic -> pipeline)
  - [x] Player (pipeline -> speaker)
  - [x] Recorder (pipeline -> .raw data)
  - [x] File Grabber (.wav data -> pipeline)

At least one of each of the following:

- [x] VAD
  - [x] Double threshold segmenter
- [ ] Speechrec (all needed)
  - [x] Deepspeech
  - [x] Pocketsphinx
  - [ ] Continuous component
- [ ] Speaker identifikation
- [x] Gender rec
- [x] Emotion rec
- [ ] Sound Source Localization
- [ ] Beamforming
- [ ] Filtering/ Signal enhancing

Added by Slothologist

1 Done

Library

- [x] Interface for clients
- [x] Audio Transmission
- [ ] Resampling
  - [x] Bitrate conversion
  - [x] Samplerate conversion
  - [x] Channel resampling
  - [ ] Endian conversion
- [x] Python bindings

Added by Slothologist

Thanks for your Attention!

Discussion