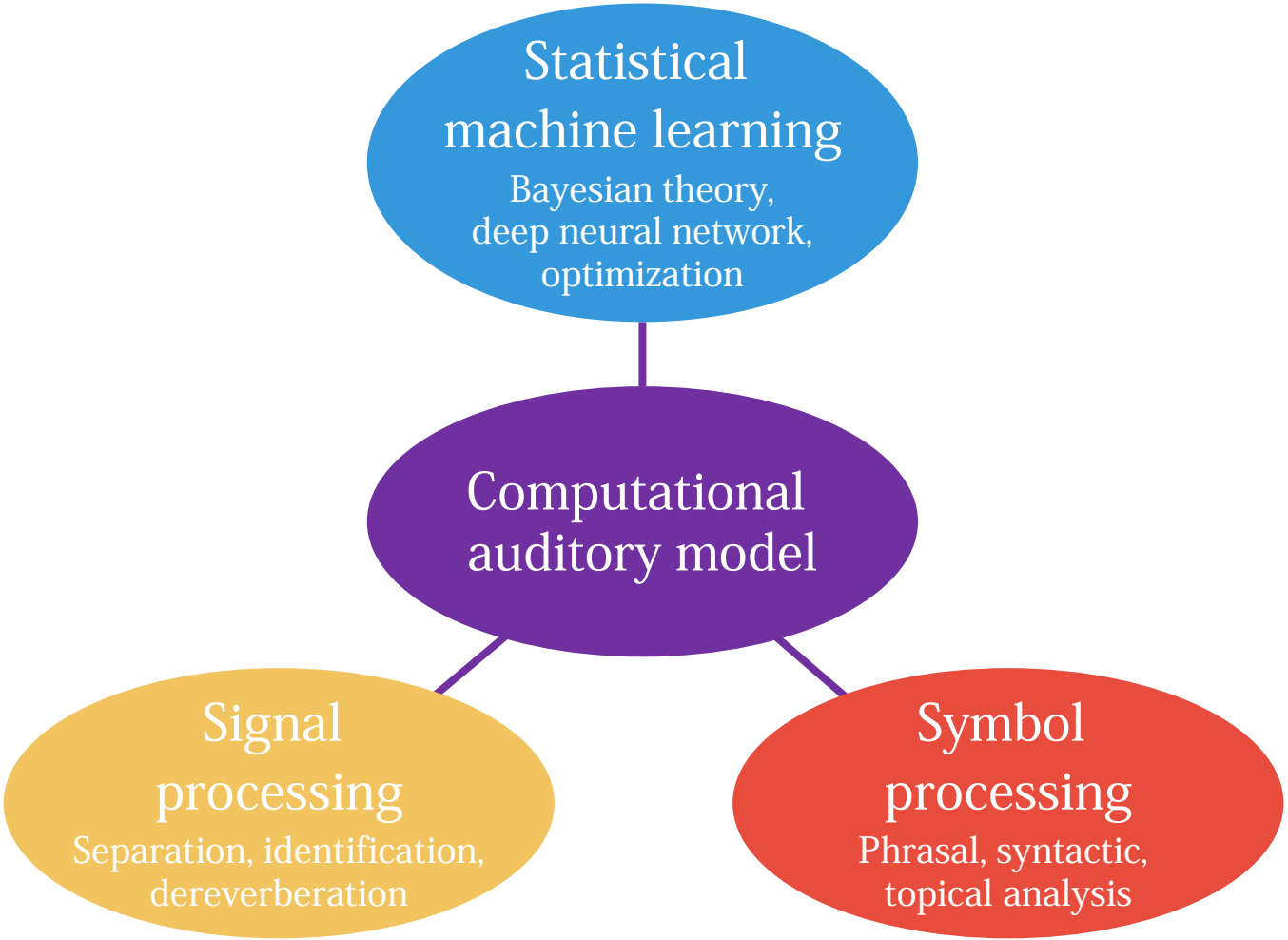
Audio Media Processing

Graduate School of Informatics
Kyoto University
Kazuyoshi Yoshii
yoshii@kuis.kyoto-u.ac.jp

Audio Media Processing



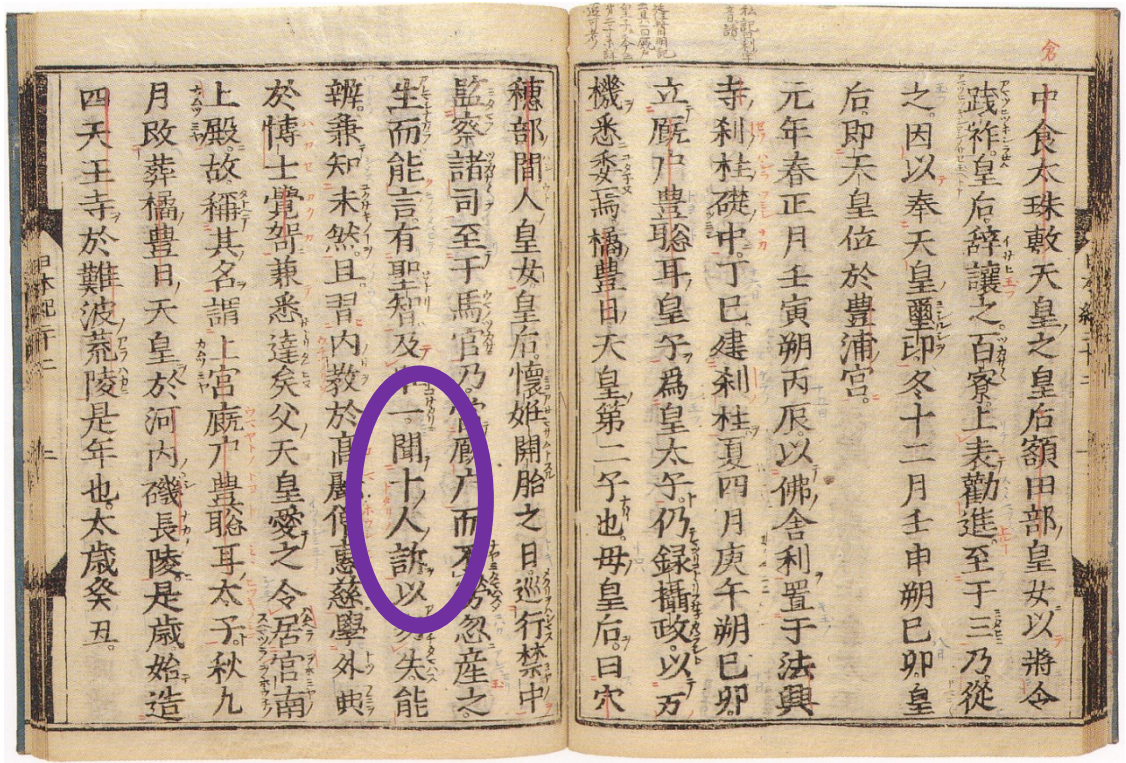
Keyword: Listen to Sound

- Listen to “speech”
  - Prince-Shotoku robot
    - Simultaneous speech recognition
  - Microphone-array processing
    - Sound source separation and dereverberation
- Listen to “music”
  - Music understanding and performance
    - Sound source separation and music transcription
    - Co-playing and accompaniment
- Listen to “environmental sounds”
  - Object detection in a disaster environment
  - Analysis of frog calling

Listen to Speech

Shotoku-Taishi (Prince Shotoku)

- Legendary person who can recognize simultaneous utterances of ten persons

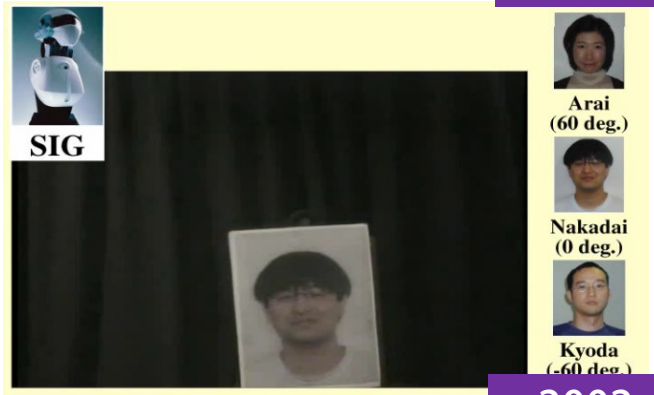


Simultaneous Speech Recognition

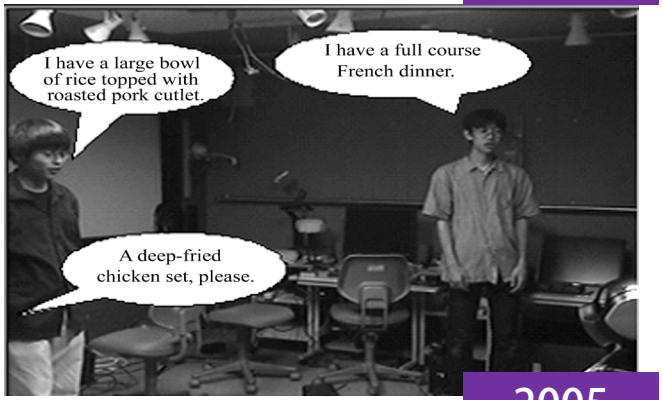
Closeness between speakers

2003

2006

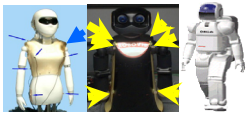


2002



2005

- Isolated word → Continuous speech
- Evaluation using three robots in a large room




料理の注文の
同時発話認識

Shotoku-Taishi Robot 2012

- Can many-ears robots go beyond humans?

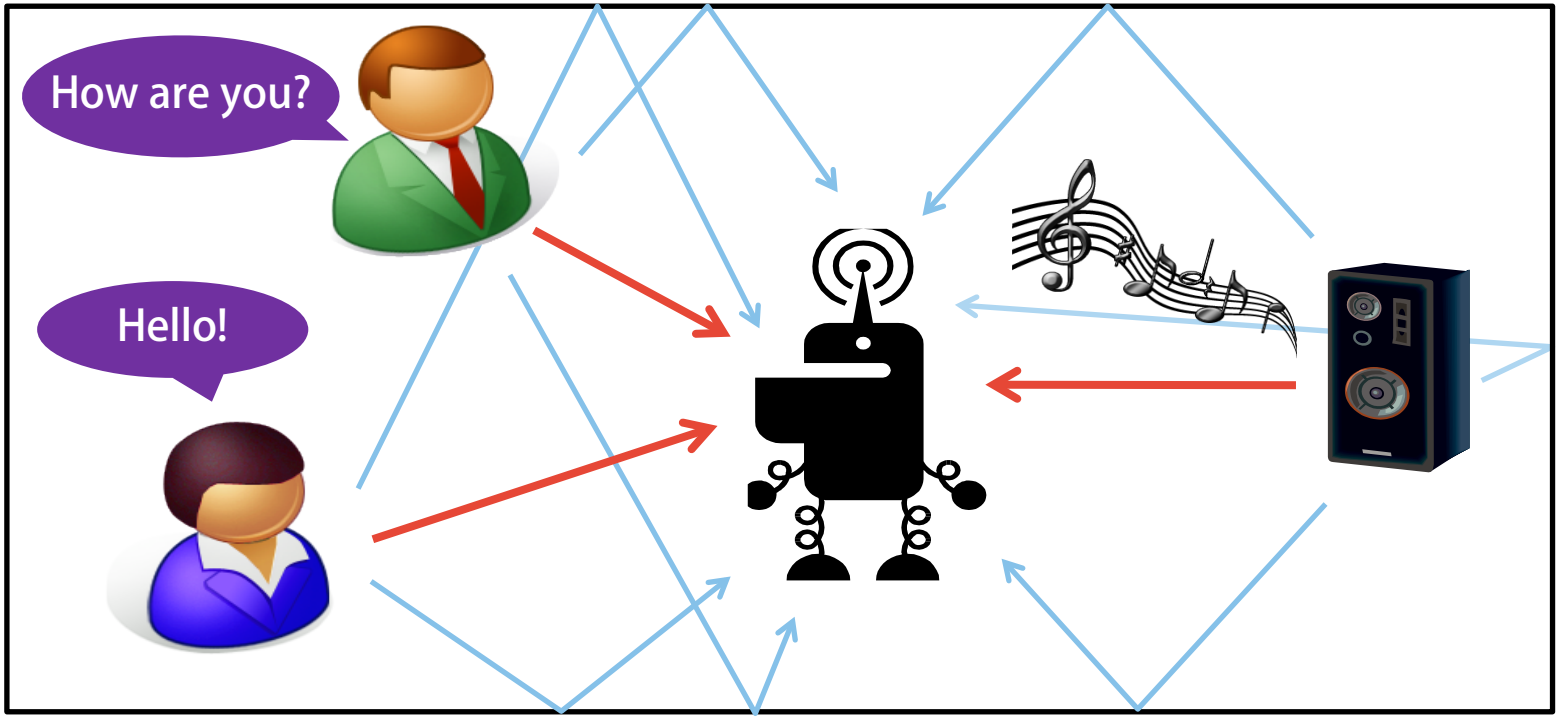
Simultaneous Speech Recognition

～ Meal Order Taking ～

16ch microphone-array processing • Sound directions are given

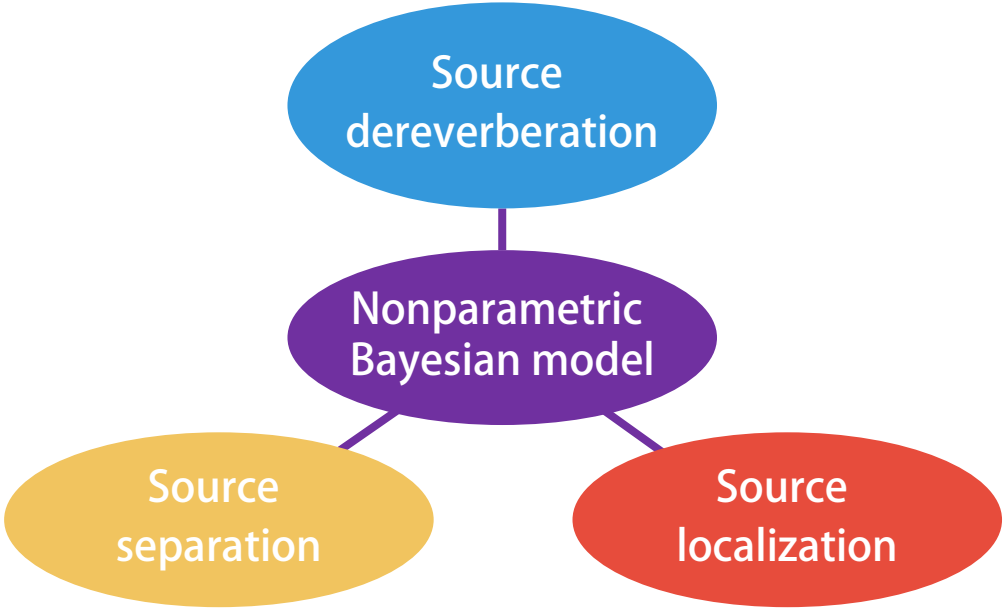
Microphone Array Processing

- Separate mixture signals into unknown number of sound sources with unknown reverberation time



Bayesian Unified Formulation

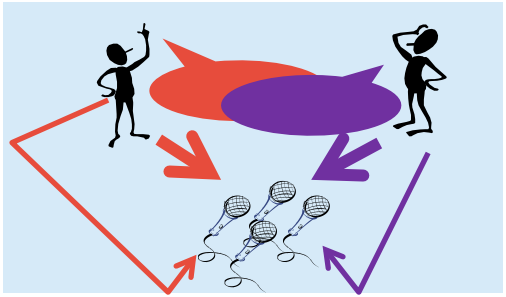
- A modern principled approach to the conventional egg-and-chicken problem
  - C.f. Errors are propagated in a cascade framework (localization → separation → dereverberation)

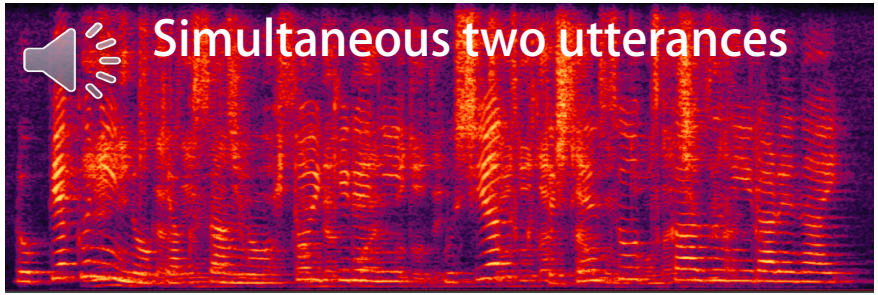


Signal processing + Machine learning + Application

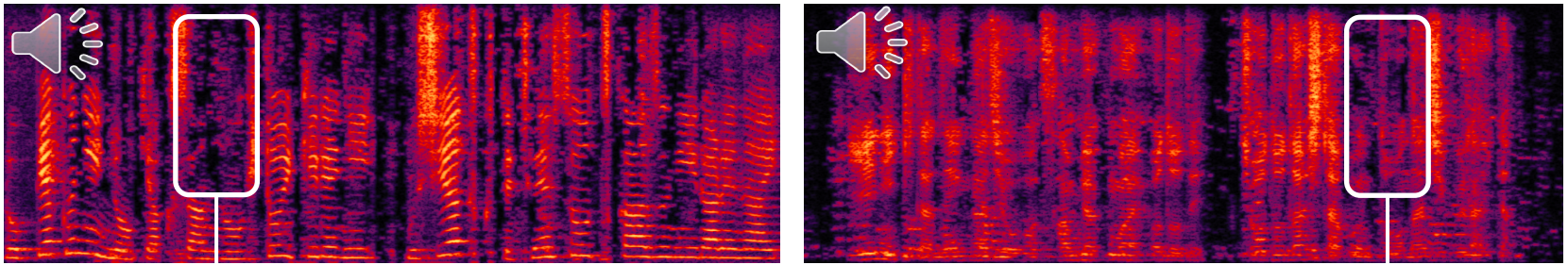
Hot topic!

Localization + Separation + Dereverberation

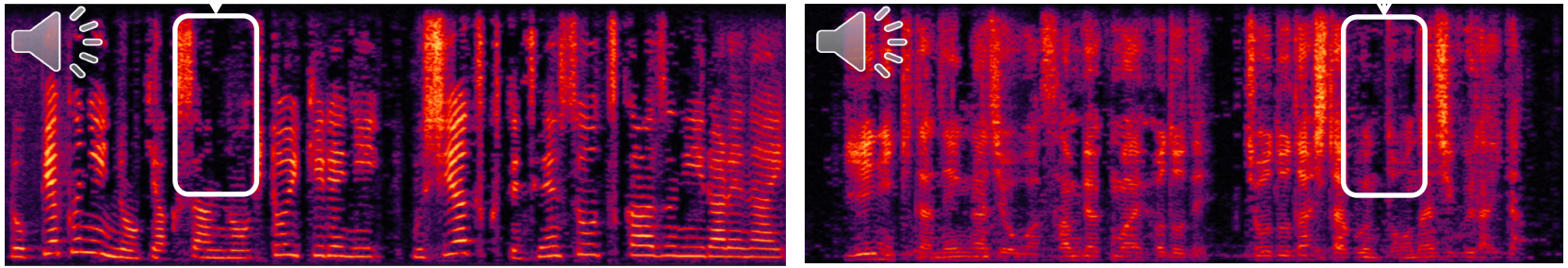




Without dereverberation



With dereverberation



Application to Real Environment

- Separate overlapping utterances in a noisy environment

Clock-tower international hall

Microphone array

Observed mixture signal

Separated signals

Background noise

Utterances

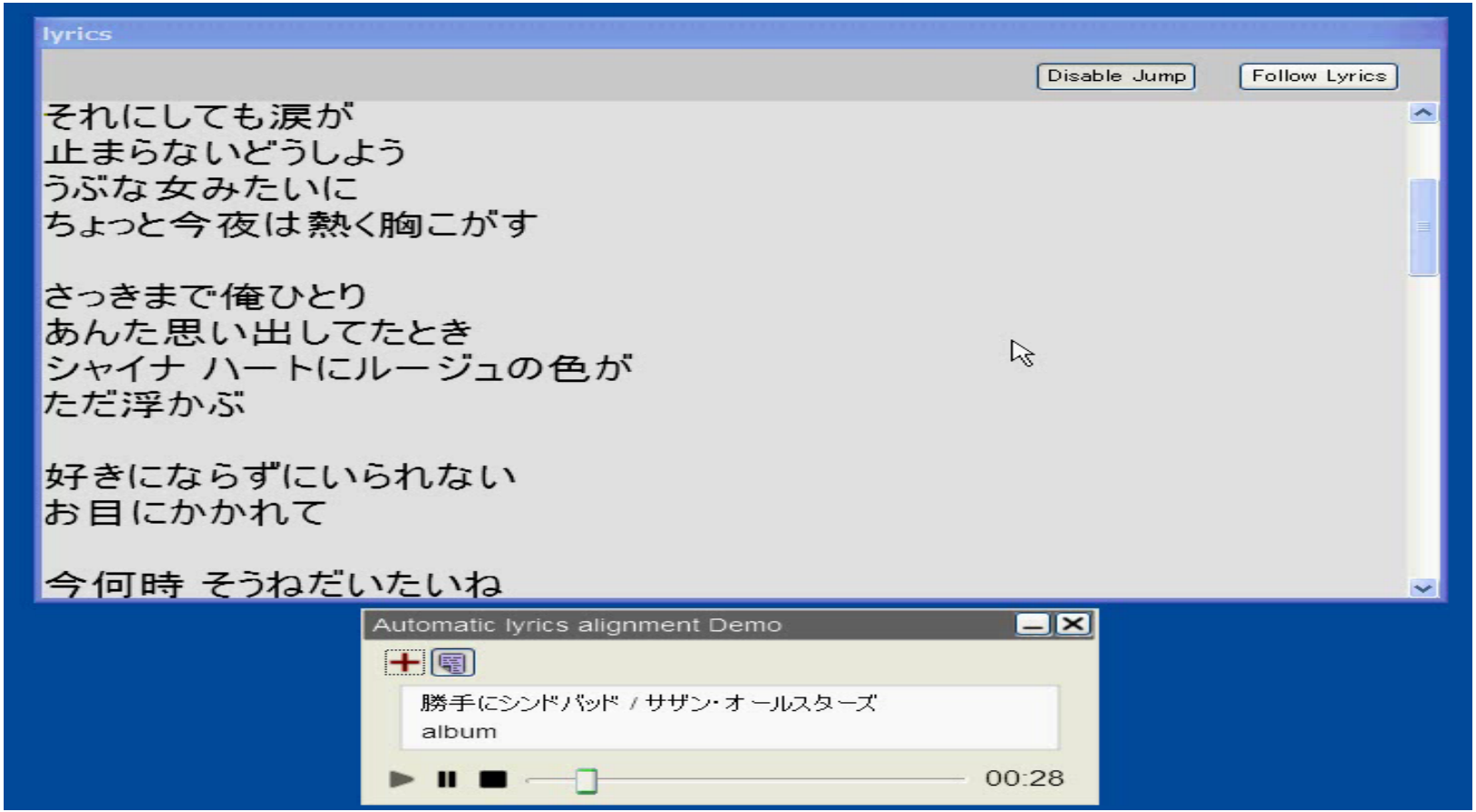
Listen to Music

Music Co-playing with Humans

- Real-time score-position tracking
  - Listen to partner's playing by using own ears

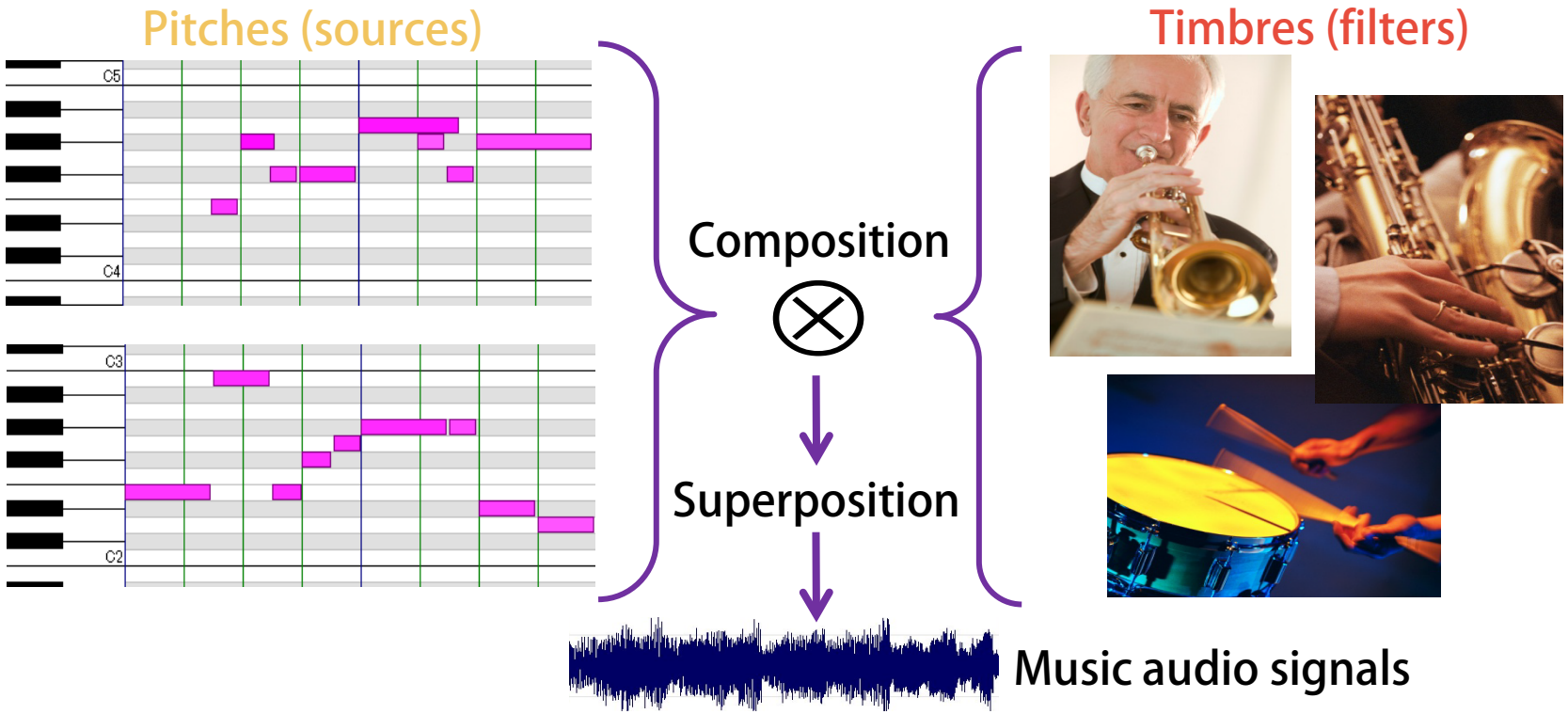
Lyric-to-Audio Synchronization

- Efficient navigation to a section of interest



Music Understanding

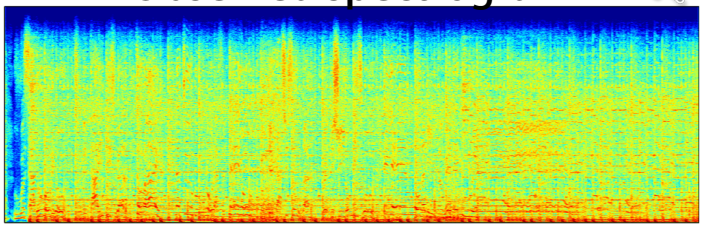
- Parts-based representation of music
  - Combinations of "pitches" and "timbres"



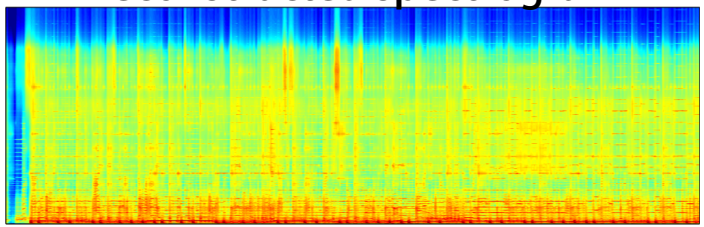
Music Signal Decomposition

- Timbre-based audio source separation

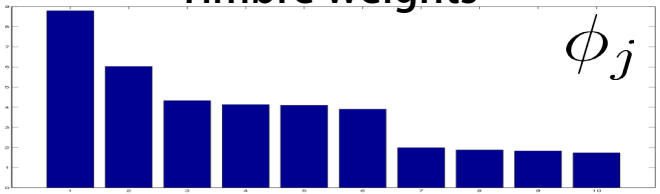
Observed spectrogram



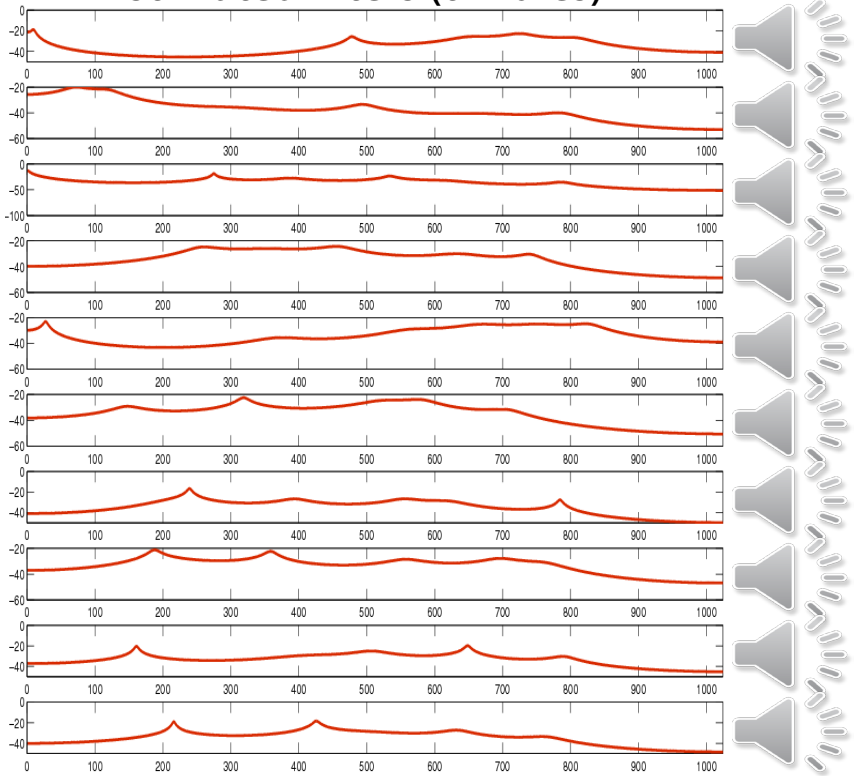
Reconstructed spectrogram



Timbre weights

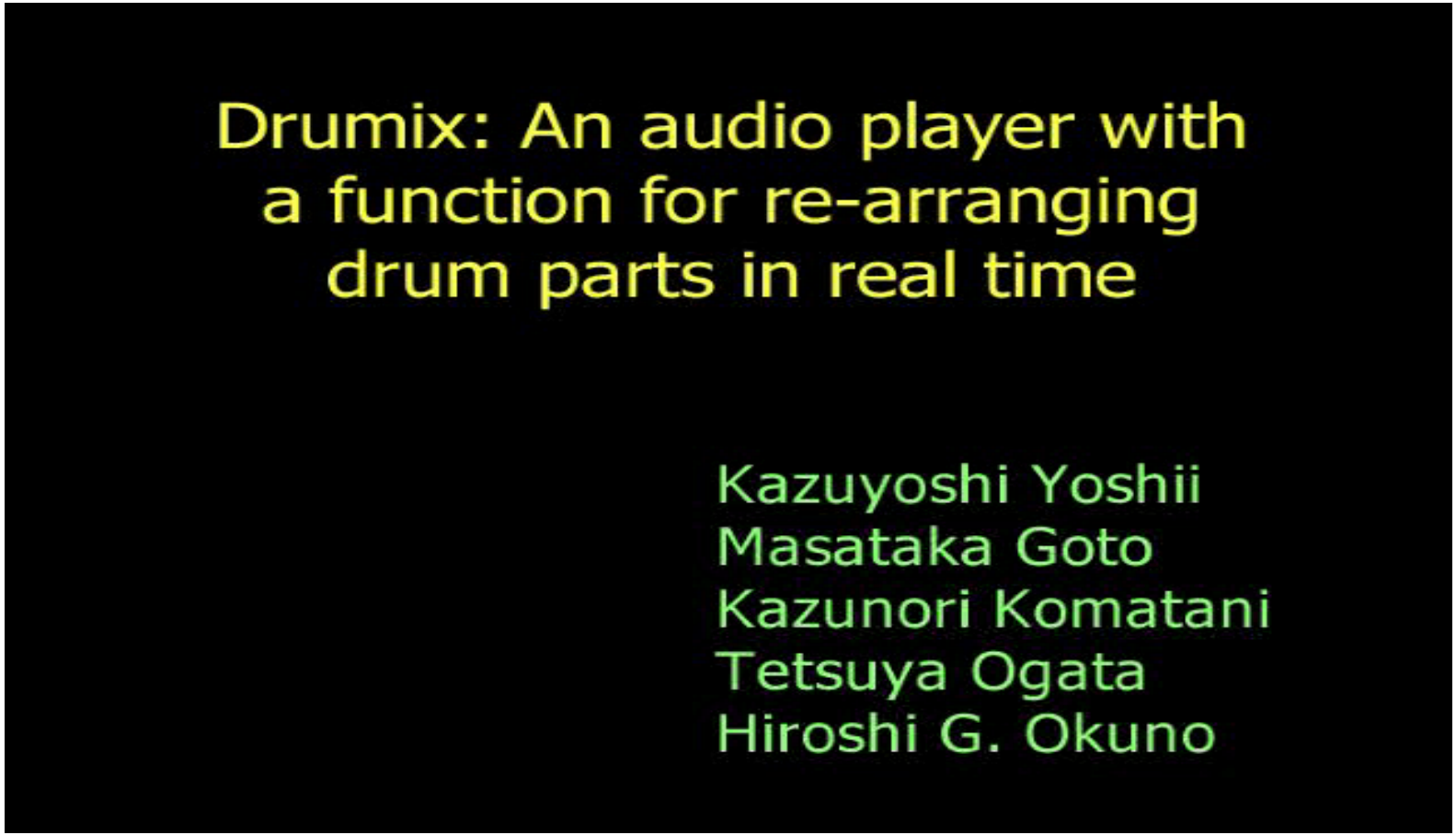


Estimated filters (timbres)



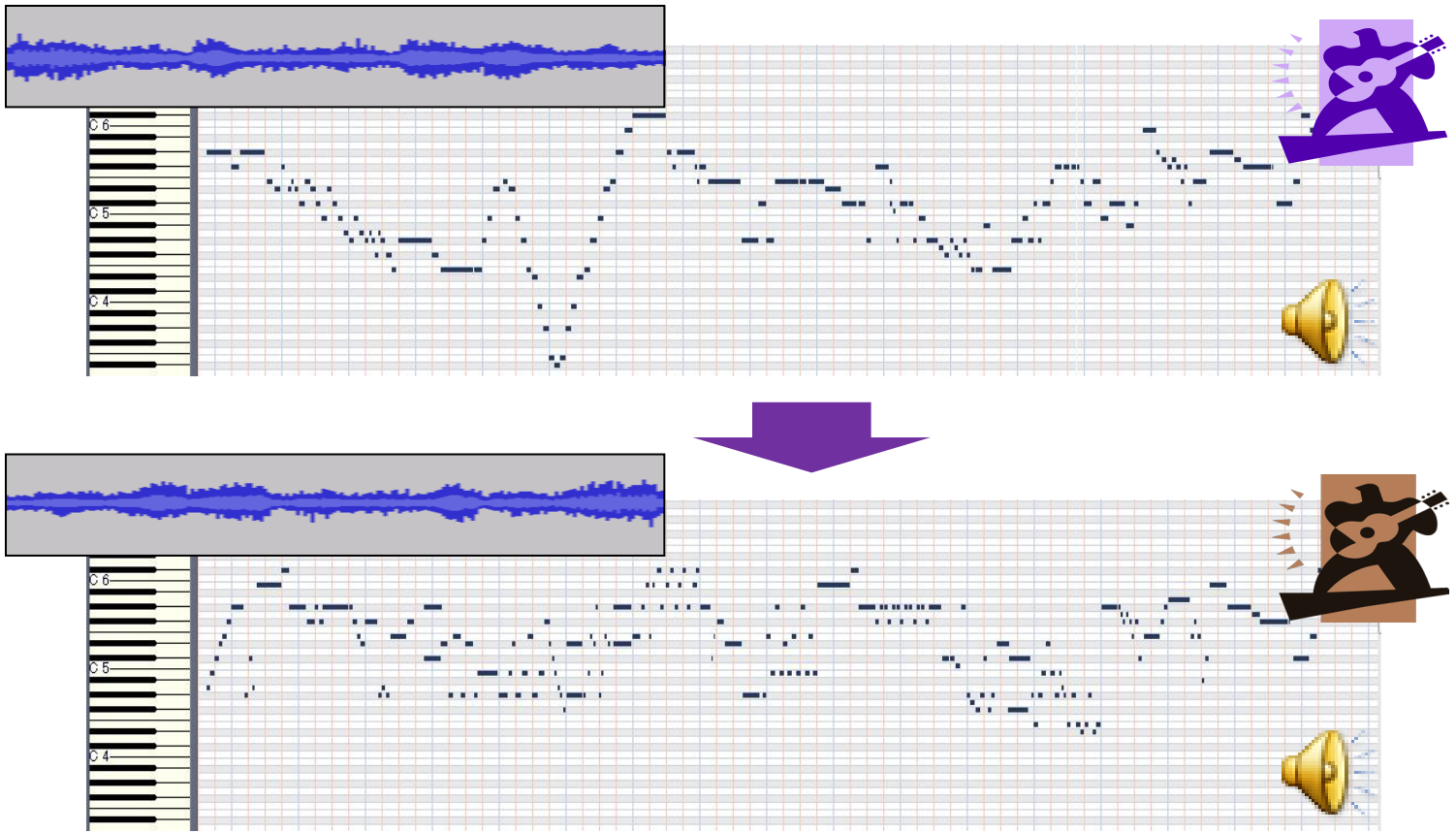
Replacement of Drum Parts

- Edit only drum parts in mixture signals



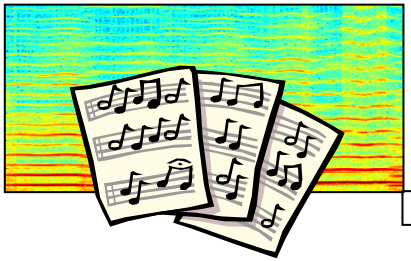
Replacement of Guitar Solo

- Edit only a guitar part while preserving original timbres



Timbre and Effect Estimation

- Preserve timbers and reverberation of original guitar solo



- Audio signal
- Music score

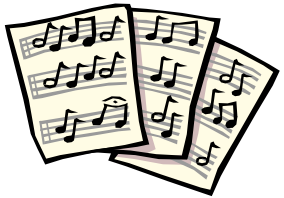
Part separation and effect estimation

Guitar part (phrase)

Accompanying parts

Reverb.

Timbre

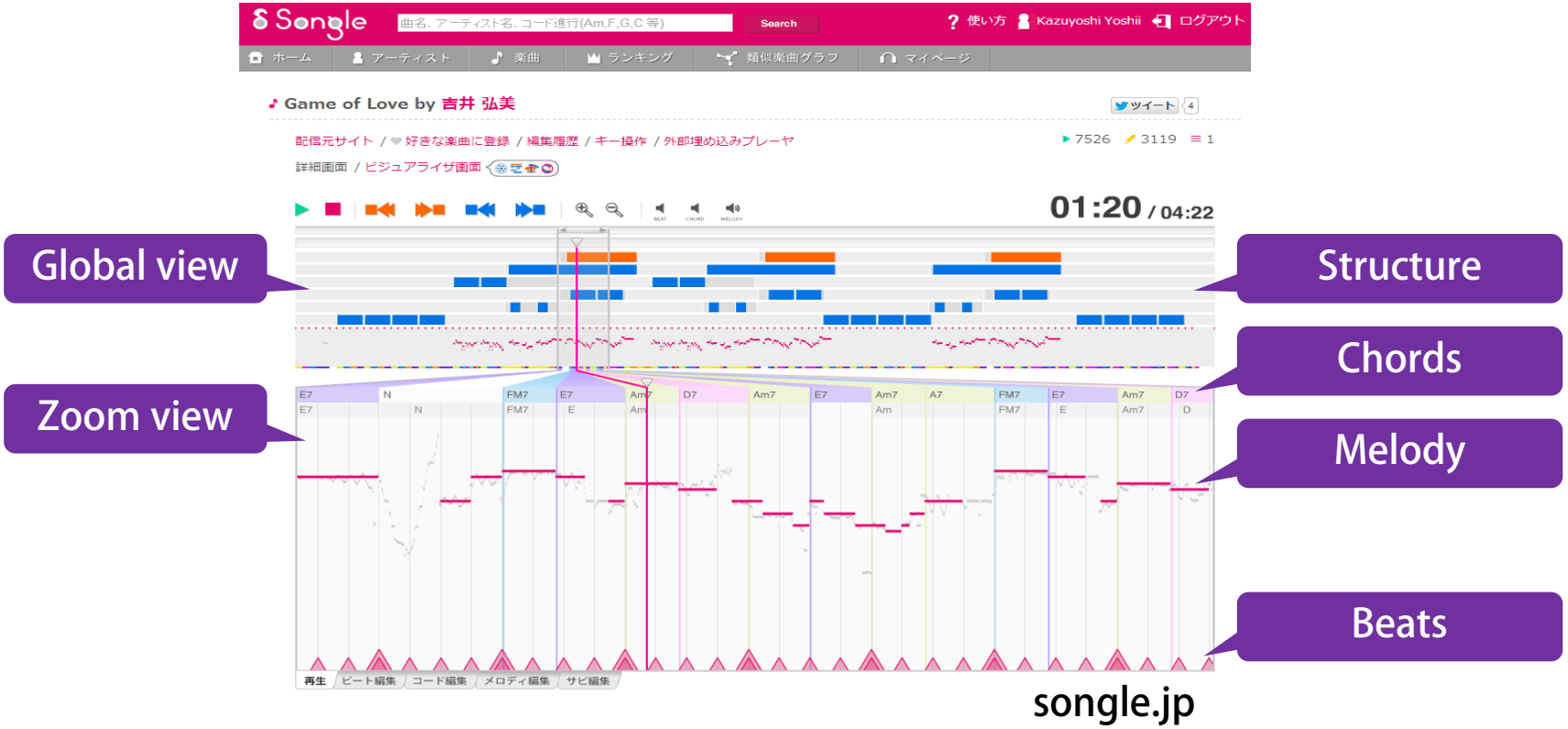


Score of new guitar solo

Synthesized guitar solo

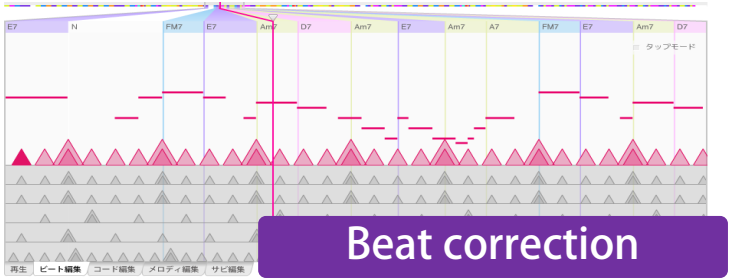
Songle: Active Music listening Service

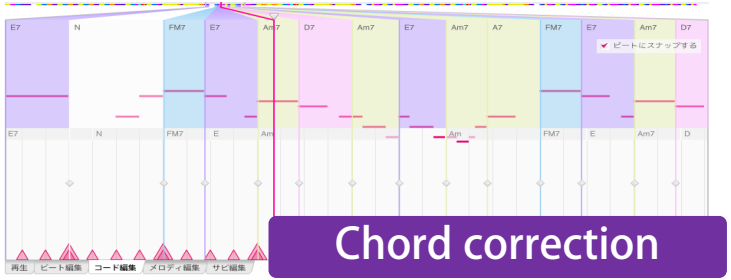
- We can enjoy automatically estimated, visualized, and sonificated musical elements of songs on the Web

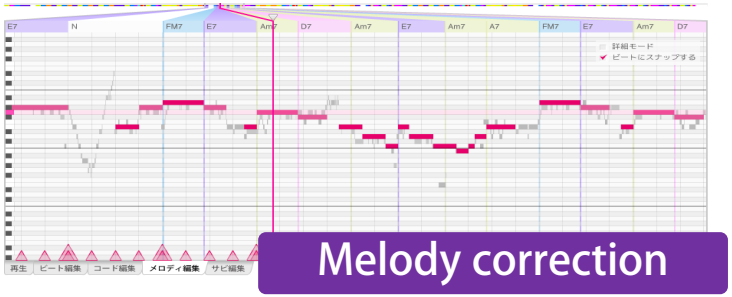


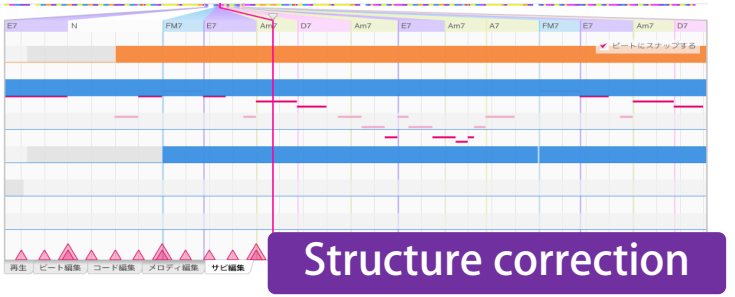
A New Form of Outreach Activity

- We can amplify user contributions by using machine-learning techniques
  - Corrections by some users → Retraining → Accuracy improvement → Reward to all users





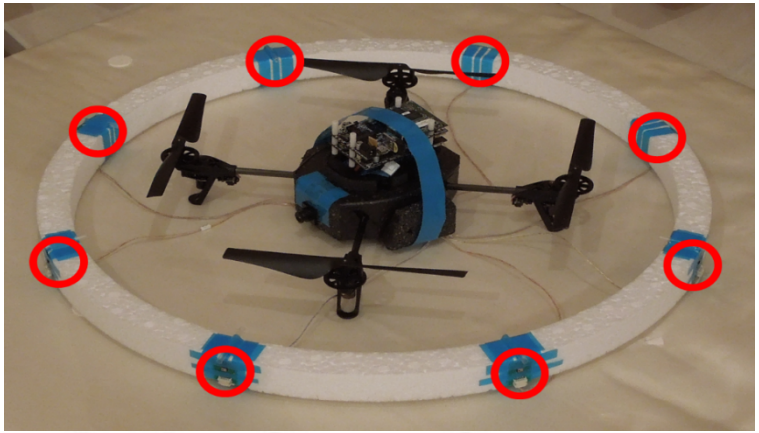




Listen to Environmental Sounds

Give Ears to Flying/Snake Robots

- Robot audition could help in disaster



Active Scope Camera

日経新聞2013年3月24日

Audition in Flying Robots

- Use a microphone array for localization

Visualization of Frog Calling

- Discriminate calling of two kinds of frogs

Visualization of Bird Singing

- Separation and localization in a park

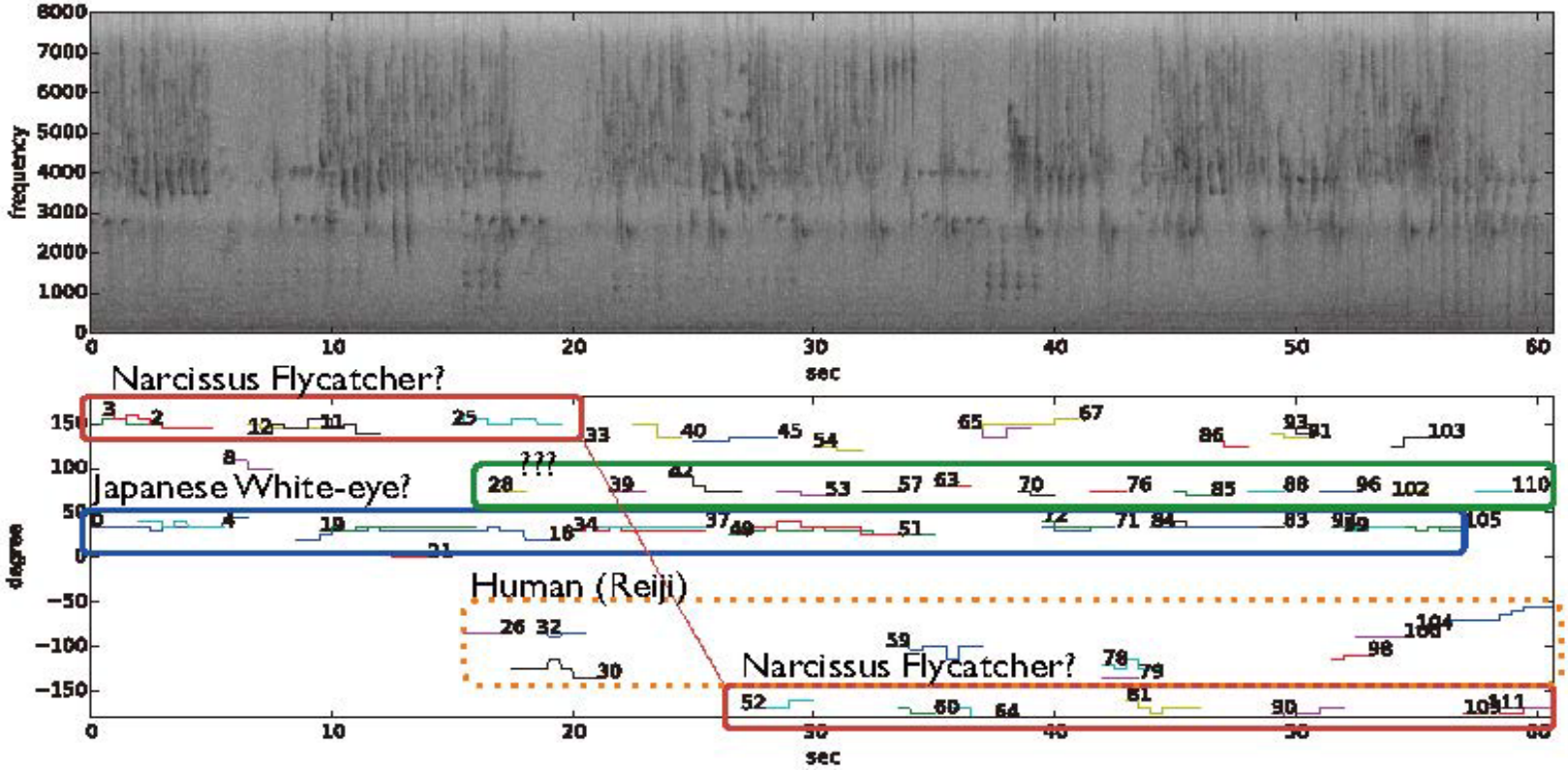


Figure 5: HARK を用いた野鳥の歌の音源定位の例.

Robot Audition

Why Robot Audition?

Conventional problem:
We need to speak around microphones

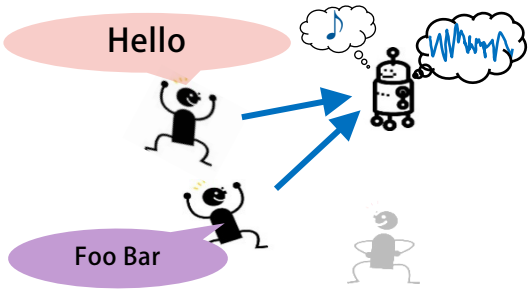
Why?

The microphones inevitably catch noise sound with target utterances

Our approach:
We aim to separate and recognize sounds

Sound Source Localization (SSL)	Sound Source Separation (SSS)

Computational Auditory Scene Analysis (CASA)

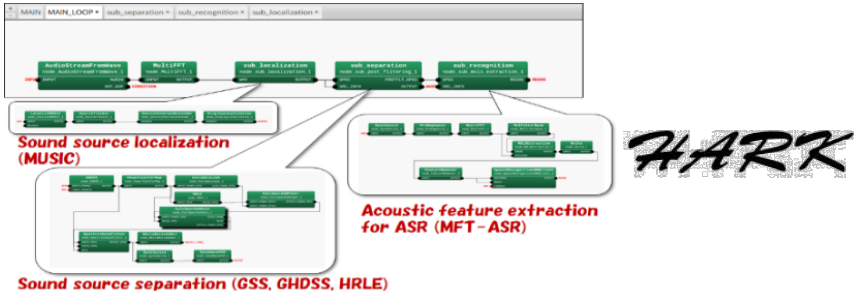


Robot Audition for MC Robots

- Robot MCs need to interact with multiple people
  - Use an open-source robot audition software “HARK” developed by Honda Research Institute Japan and Kyoto University
  - HATTACK 25: Speech-based quiz game

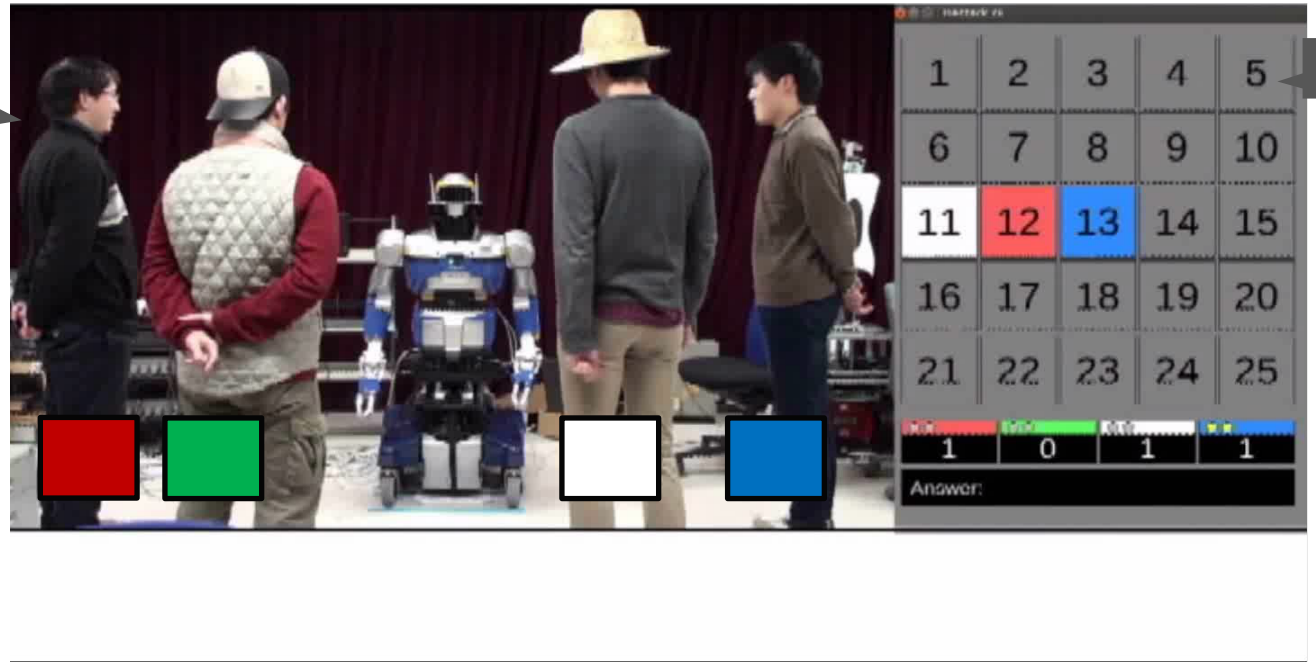
Inspired by the well-known quiz game in Japan

A player position can be identified by his or her voice



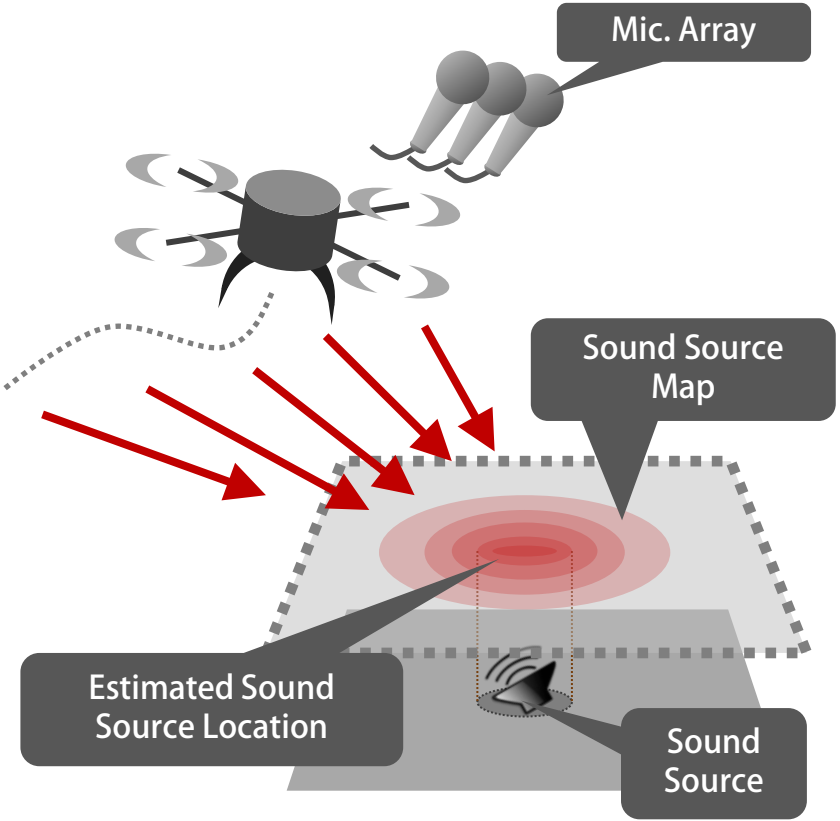
Demo : HATTACK25

- Players can barge in when the robot is talking about questions
  - To answer, players have to say “yes!” first
  - Impossible for standard dialogue systems



Robot Audition for Flying Robots

- Localize source sources on the ground by using flying robots with microphones



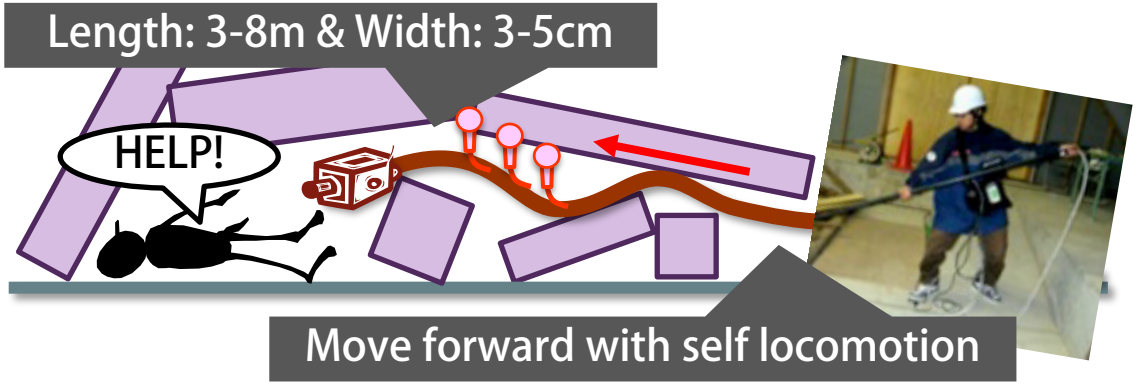
Robot audition is disturbed by self-generating noise

Video of the flying robot

1. Learn self-generating noise (with Gaussian Process)
2. Suppress noise from input

Robot Audition for Rescue Robot

- Estimate robot shapes and detect sound sources in collapsed buildings



Moving hose-shaped robot

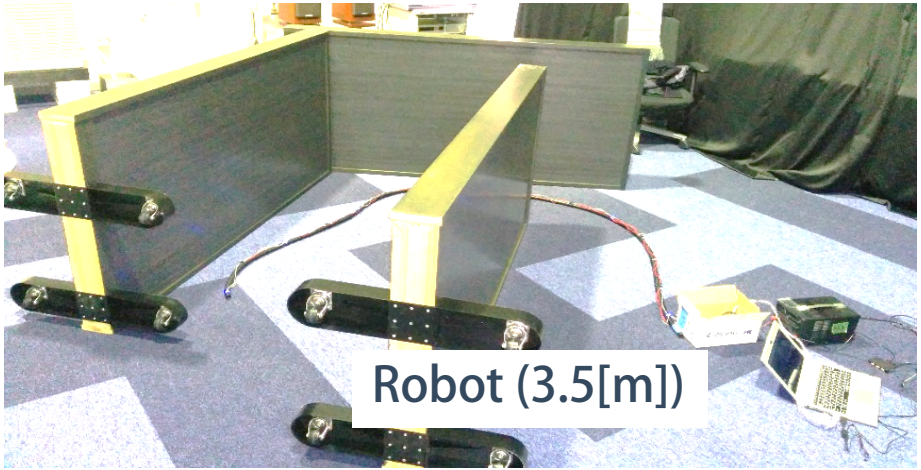
How to estimate microphone positions on the robot?
→ Statistical signal processing techniques

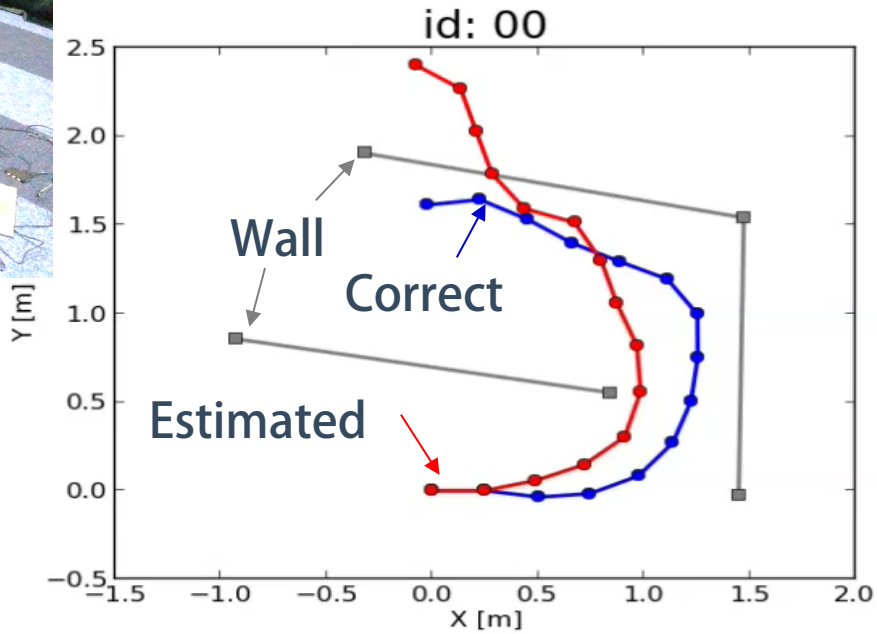
We designed a state space model of robot posture
and estimated the posture by measuring TDOA of sounds

Time Difference of Arrival

Posture Estimation for a Hose-shaped Robot

- Accurately estimate shapes even when obstacles exist





Keyword: Listen to Sound

- Listen to “speech”
  - Prince-Shotoku robot
    - Simultaneous speech recognition
  - Microphone-array processing
    - Sound source separation and dereverberation
- Listen to “music”
  - Music understanding and performance
    - Sound source separation and music transcription
    - Co-playing and accompaniment
- Listen to “environmental sounds”
  - Object detection in a disaster environment
  - Analysis of frog calling