Robot Audition and Drone Audition

Kazuhiro Nakadai

Principal Scientist, Honda Research Institute Japan Co. Ltd.
Specially-appointed Professor, Tokyo Institute of Technology

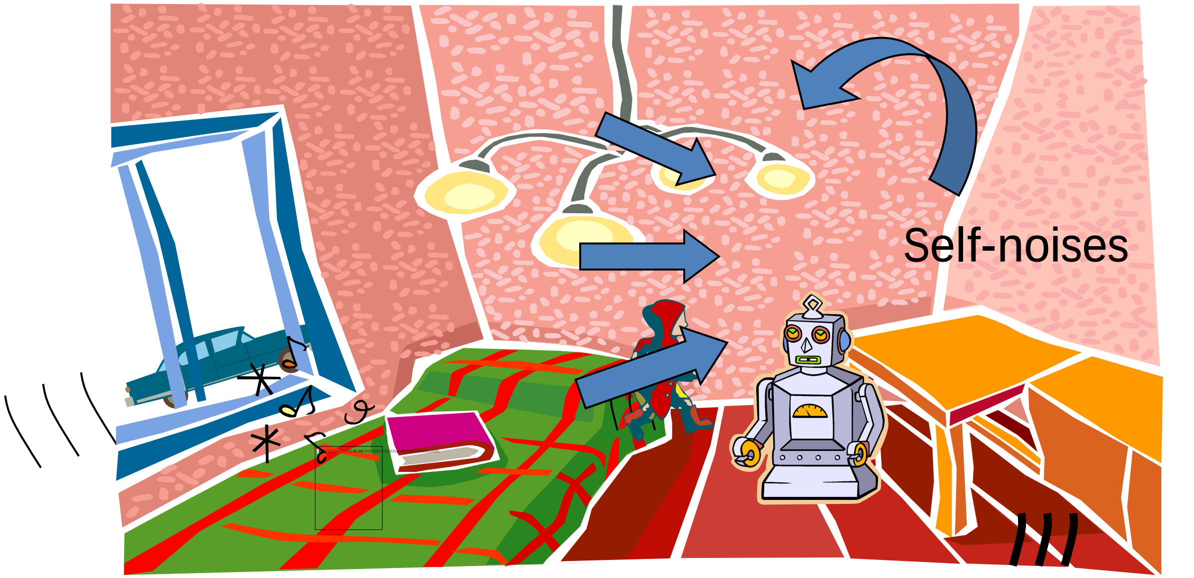
Hiroshi G. Okuno

Professor, Waseda University
Professor Emeritus, Kyoto University
Honorary Professor, Amity School of Engineering and Technology, India

Robot Audition [Nakadai & Okuno AAAI 2000]

- Not a headset microphone, but ***its own ears!***
  - **Noise-robustness**
    - Ego-noise (actuators, self-voice)
    - Environmental sounds
    - Simultaneous speech (barge-in)
  - **Cocktail Party Robot**
  - **Prince Shotoku Robot**

- ***Towards Auditory Scene Analysis***

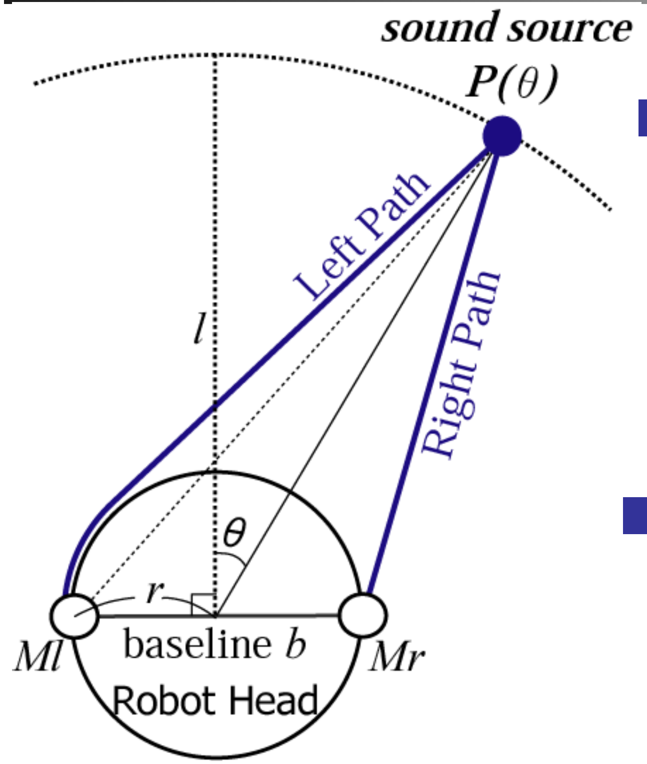


Primary Issues in Robot Audition

- **Sound Source Localization (SSL)**
  - **MUSIC based on Generalized Eigen/Singular-Value Decomposition (GEVD/GSVD-MUSIC) [Nakamura+, Okutani+, Ohata+ '09-'12]**
- **Sound Source Separation (SSS)**
  - **Geometric High-order Decorrelation based Source Separation with Adaptive Step-size Control (GHDSS-AS) [Nakajima+ '10, Takeda+ '12]**
- **Automatic Speech Recognition (ASR)**
  - **Missing feature theory based integration of separation and ASR [Yamamoto+ '07]**

- Published in multiple research fields; Robotics (ICRA, IROS), Acoustics & Speech (ICASSP, INTERSPEECH), AI (AAAI, IJCAI)
- Organized special sessions, workshop, tutorial, etc
- "Robot Audition" : an official keyword of RAS in 2014

Auditory Epipolar Geometry



- Head Related Transfer Function (HRTF)
  - Measured in anechoic room
  - prone to alter due to environmental change
- Auditory Epipolar Geometry
  - Method for horizontal localization
  - Estimate direction of a sound source from IPD mathematically

$$\Delta\varphi = \frac{2\pi f}{v} \times r\left(\theta + \sin\theta\right)$$

($\Delta\varphi$: *interaural phase difference (IPD)*)

Acoustic Model of Robot Head [IROS 2003]

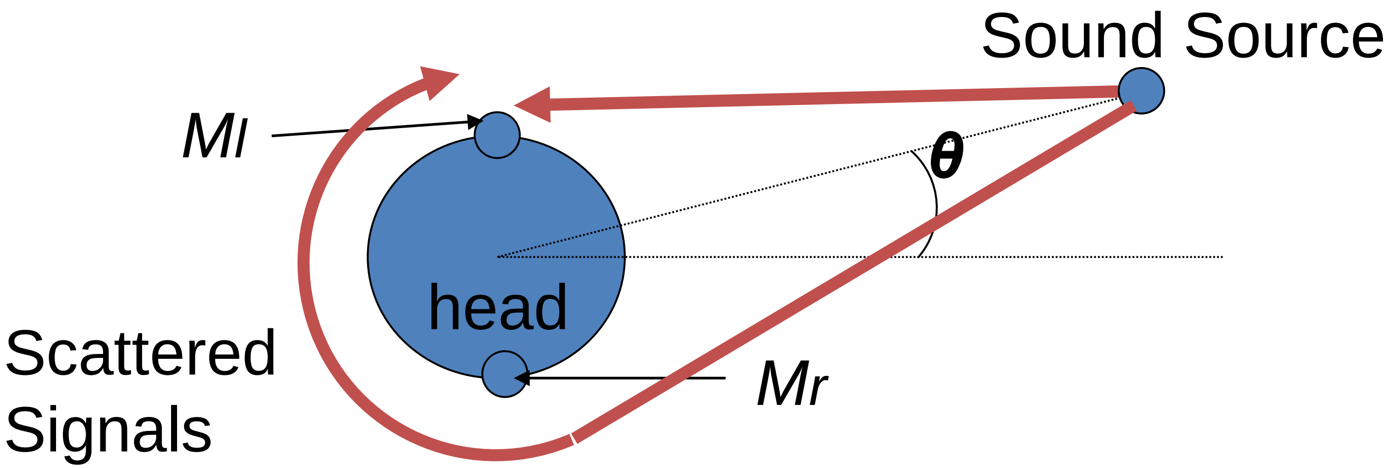
- Based on ***Scattering Theory***
- When spherical robot head is assumed, potential at a point on the surface is estimated by

$$S(\theta, f) = -\left(\frac{v}{2\pi a f}\right)^2 \sum_{n=0}^{\infty} (2n+1) P_n(\cos\theta) \frac{h_n^{(1)}\left(\frac{2\pi r_0}{v} f\right)}{h_n^{(1)\prime}\left(\frac{2\pi a}{v} f\right)}$$

- Estimate IPD and IID ***mathematically***

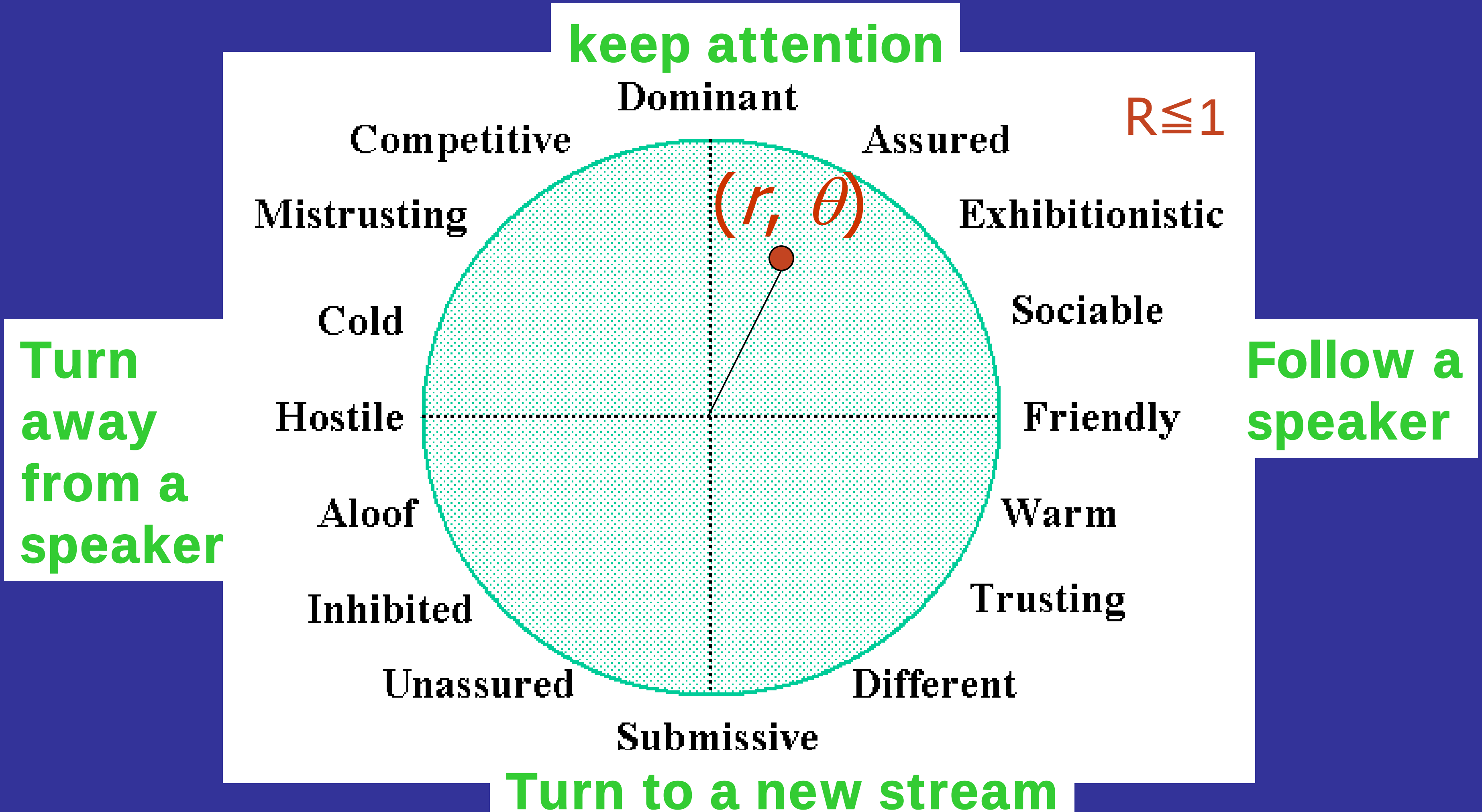
$$\text{IPD:} \quad \Delta\varphi_s(\theta, f) = \arg\left(S_l(\theta, f)\right) - \arg\left(S_r(\theta, f)\right)$$

$$\text{IID:} \quad \Delta\rho_s(\theta, f) = 20 \log_{10} \frac{|S_l(\theta, f)|}{|S_r(\theta, f)|}.$$



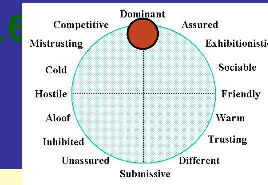
Personality in Focus-of-Attention [PRICAI 02]

- Personality is represented as a point (r, θ) in the Interpersonal Circumplex.



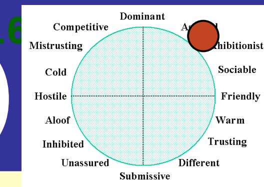
Demonstrations

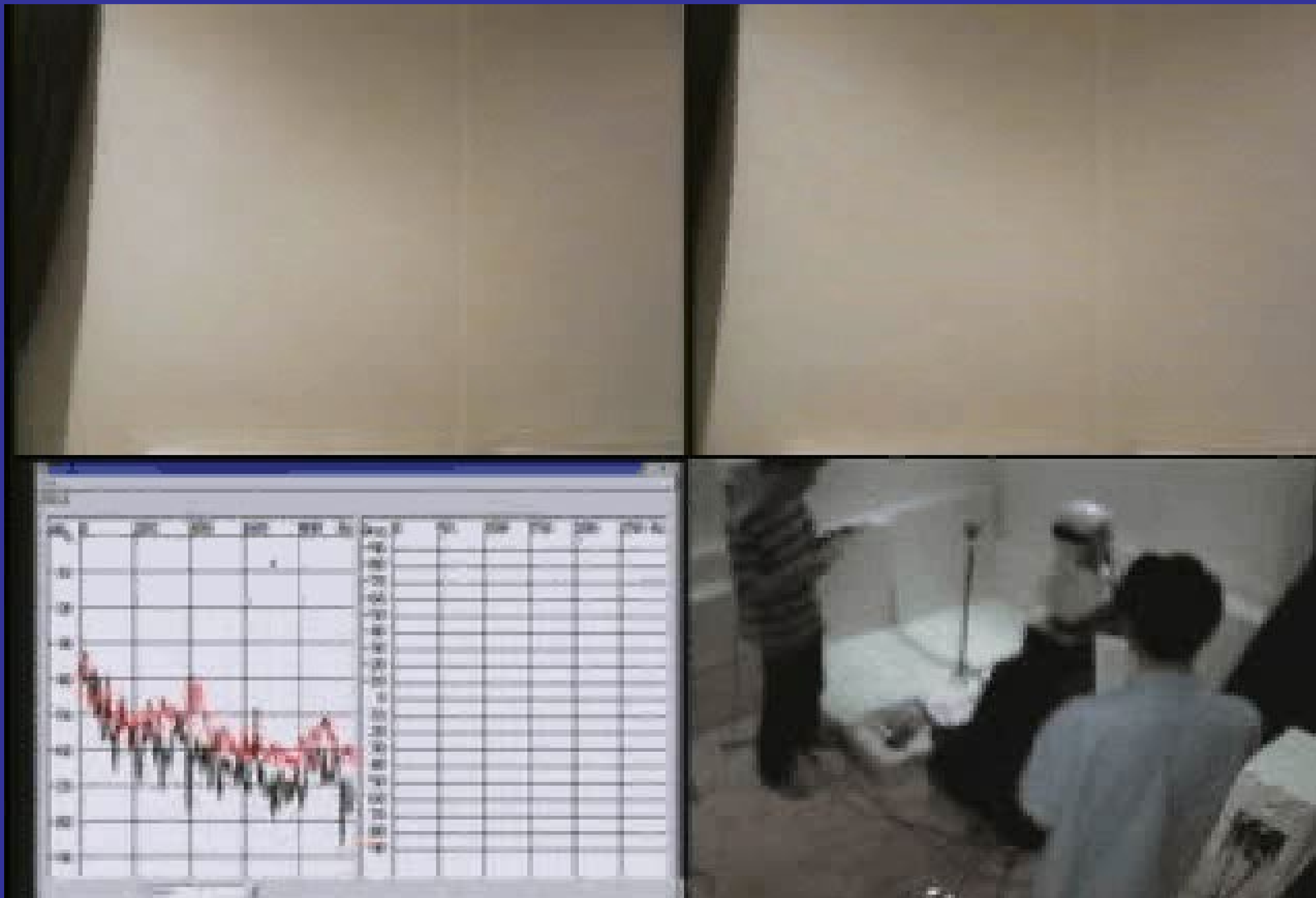
Demos	**Personality (r, θ)**
1. Follower	Dominant $(1, \pi/2)$
2. ~~Receptionist~~	Dominant $(1, \pi/2)$
3. ~~Stereo sound tracking~~	Dominant $(1, \pi/2)$
4. ~~Companion~~	Friendly $(1, 0)$
5. Listening to two people	Assured $(1, 3/8\pi)$
6. Hostile listener	Hostile $(1, \pi)$
7. ~~lose-interest listener~~	Friendly with large k



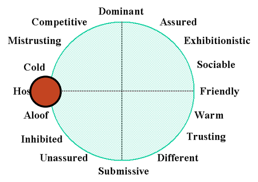
Follower (Dominant)

Listening to two people (Assured)





ICRA-03 (16



Hostile Listener (Hostile)

SIG as a physical non-verbal Eliza

1. A variety of non-verbal behaviors is attained by personality-based focus-of-attention control only with sound localization
2. The robot encouraged users to explore the robot's behaviors
   - Some people walk around talking with their hand covering *SIG* 's eyes in order to confirm the performance of auditory tracking.
   - Some people creep on the floor with talking in order to confirm the performance of auditory tracking.
   - Some people play hide-and-seek games with *SIG*.

Robot Quizmaster [Nishimuta+ 2012]

- **School-type interaction:** ask a right to answer
- **Auction-type interaction**: answer a quiz

A Robot Quizmater for the 'Fastest Voice First' Quiz Game

Izaya Nishimuta Naoki Hirayama
Katsutoshi Itoyama Kazuyoshi Yoshii
Hiroshi G. Okuno

Simultaneous Speech Recognition [Nakadai+ 2013]

Simultaneous Speech Recognition

～ Meal Order Taking ～

- Dealing with 11 directional sound sources, a diffuse noise source and ego-noise
- 16ch circular microphone array (speaker locations given).

Open Source Software for Robot Audition HARK [Nakadai+ 2008]

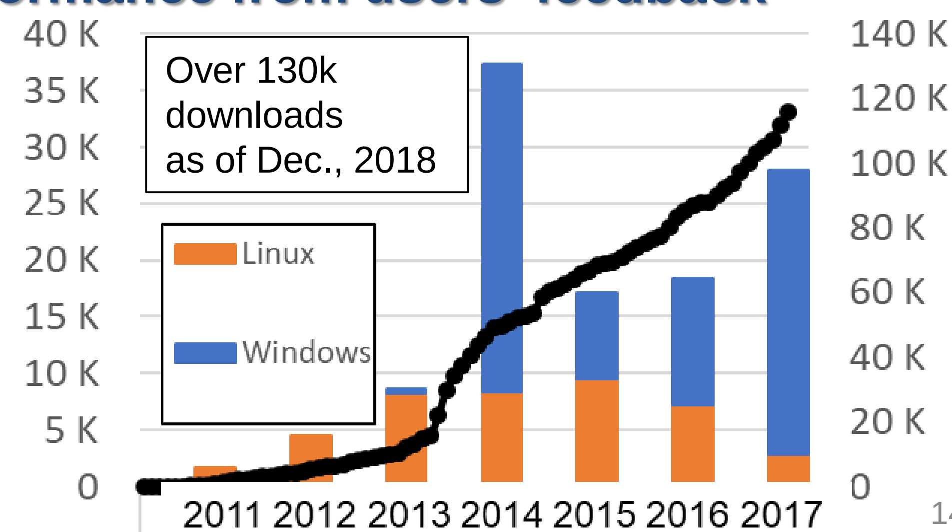
- **HRI-JP Audition for Robots with Kyoto University**
  (downloadable at http://www.hark.jp/)

HARK
Honda Research Institute Japan Audition for Robots with Kyoto University

hark = listen (Old English)

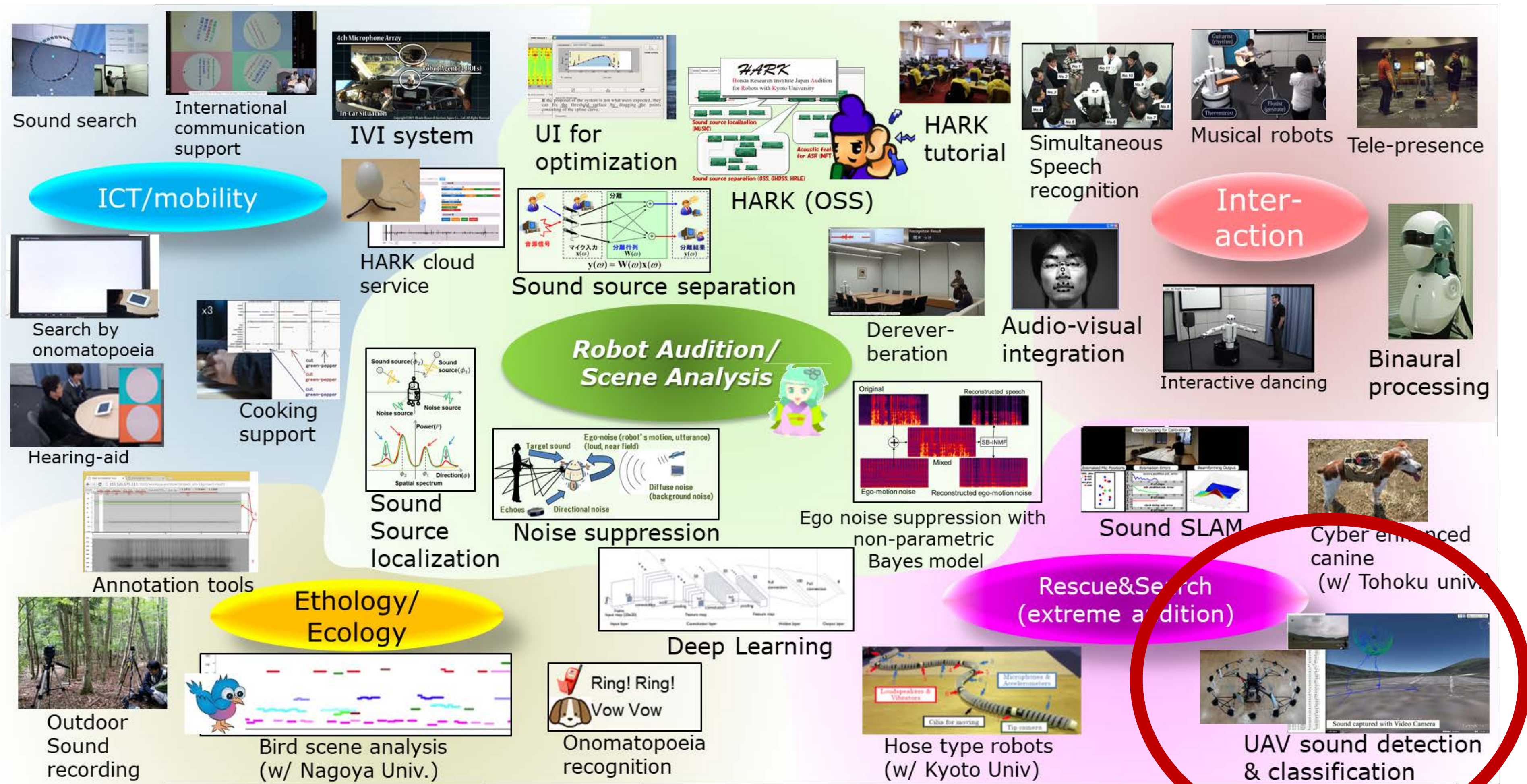
Research: Free
(Commercial: Licensing)

  - **Provides several algorithms for SSL, SSS & ASR with many utilities.**
- **Open sourced since Apr. 2008.**
  - **To accelerate robot audition related research**
  - **To provide a tool for collaboration (inter-field, intra-field)**
  - **To improve stability and performance from users' feedback**
- **Annually update with free tutorial and hackathon**
  - **15th Tutorial was held at IEEE/RSJ IROS 2018**



Research Map of Robot Audition

HONDA — Honda Research Institute JP



- From Robot Audition & Computational Auditory Scene Analysis to Many engineering & scientific fields
- Base technology: robotics, signal/speech processing, AI incl. DL,…

What is important in a disaster?

Great East Japan Earthquake '11/03

Kumamoto, Japan '16/04

Mexico City, Mexico '17/09

- Many disasters (e.g. earthquakes) happen in the world
  - Transportation paralyzed
  - Stacked emergency vehicles
- The Faster, The Better
  - Need to find people within three days to save their lives
    - **Golden 72 hours** in Japan or
    - **The rule of threes** in Western countries
  - One day faster claim = 6 month faster repair (Prof. R. Murphy)

Why microphone-array-embedded UAV?

- Realize a new rescue robot, i.e., an **Unmanned Aerial Vehicle (UAV) with a microphone array** to search for people in a disastrous situation
  - **Unmanned Aerial Vehicle (UAV)**
    - UAV ***moves quickly over a wide area*** even when traffic is cut off.
  - **Microphone array**
    - People can be ***detected from acoustic information***, even when people are occluded (a camera is not available).

UAV with a microphone array

Sound Source Localization: Multiple Signal Classification (MUSIC)

SEVD-MUSIC [Schmidt 1986]

– **S**tandard **E**igen **V**alue **D**ecomposition
MUltiple **SI**gnal **C**lassification

Observation: $\boldsymbol{x}(t)$

FFT

$X(\omega)$

Correlation Matrix
$R = XX^H$

R

Standard Eigen Value Decomposition
$R = E\Lambda E^{-1}$

$E = [e_1, \cdots, e_M]$

MUSIC spectrum

$$P(\psi) = \frac{|G(\psi)^H G(\psi)|}{\sum |G(\psi)^H e_m|}$$

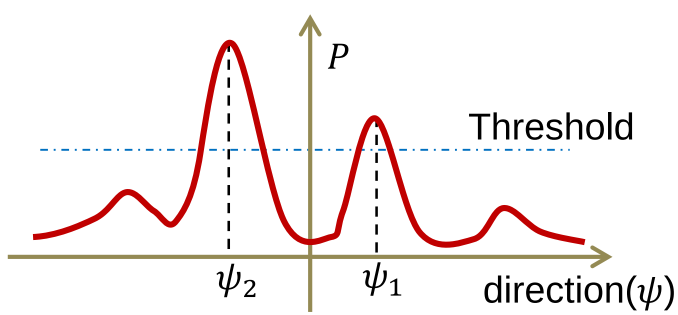
$P(\psi)$

eigenvector E : sound direction
eigenvalue Λ : sound power

Eigenvectors are *orthogonal* from each other.

Assumptions

- Large eigenvalues : target sound sources $[e_1, \cdots, e_L]$
- Small eigenvalues : noise sources $[e_{L+1}, \cdots, e_M]$



MUSIC spectrum

Performance degrades when the target sound source has smaller power than noise sources

$G(\psi)$: steering vector (= transfer function) for ψ
e_m : m-th eigenvectors of R for noise sources $(\#of\ mic. - \#of\ sound\ sources)$

Extension of MUSIC

Audio signal $x(t)$ → **F transpose** → $X(\omega)$ → **Correlation Matrix (CM)** $R = XX^H$ → R

Noise signal → **F transpose** → $N(\omega)$ → **Noise correlation matrix (noise CM)** $K = NN^H$ → K

Generalized Singular Value Decomposition

$$K^{-1}R = E_L \Lambda E_R$$

$E_L = [e_1, \cdots]$

Spatial spectrum

$$P(\psi) = \frac{|G(\psi)^H G(\psi)|}{\sum |G(\psi)^H e_m|}$$

→ $P(\psi)$

$G(\psi)$: Spatial transfer function

e_m : Singular vectors for noise sources

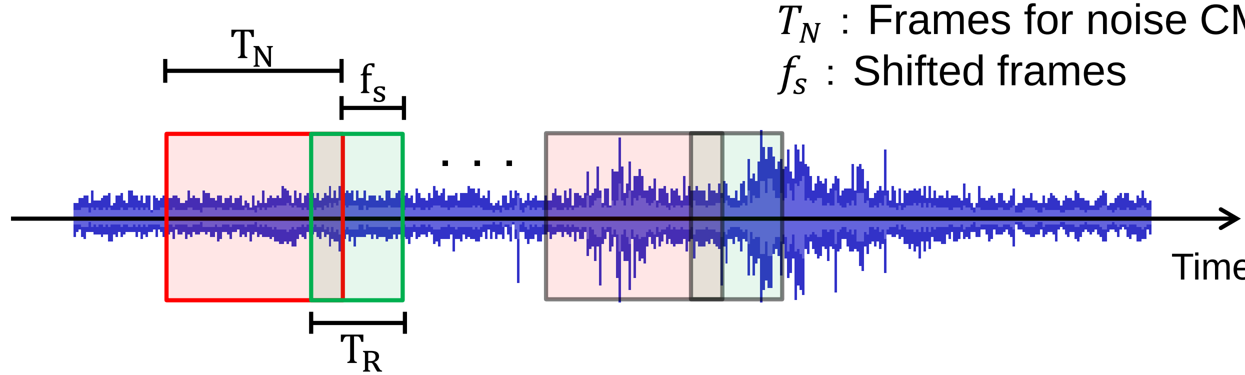
iGSVD-MUSIC [Ohata+'14]

- Whitening noise sources by estimating a noise correlation matrix(CM) incrementally
- To estimate a noise CM, a signal observed f_s frames before is used by assuming that no target signal is included in the signal.
- (green box) for CM, (red box) for noise CM

T_R : Frames for CM

T_N : Frames for noise CM

f_s : Shifted frames

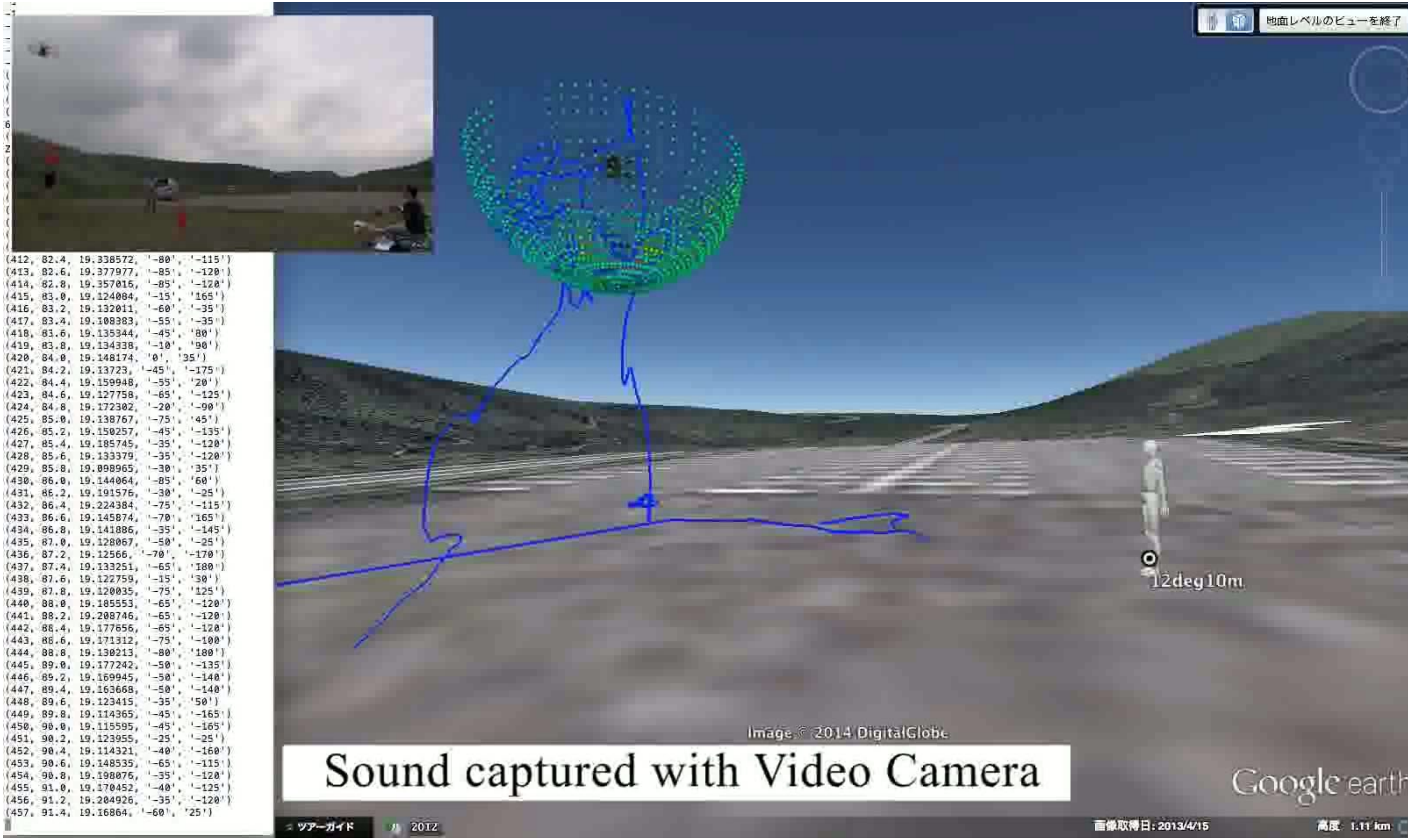


Incremental estimation of correlation matrix for iGSVD-MUSIC

A target sound source with smaller power can be localized.

Offline Sound Source Localization with iGSVD-MUSIC [Ohata+14, Nagamine+14]

HONDA Honda Research Institute JP

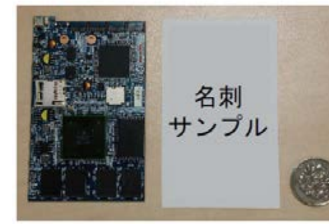


Extremely noisy sound sources (-15dB) were successfully localized.

Issues and approaches for online demo

1. Real-time processing, communication reduction

- Achieved real-time processing and reduction of communication to 1/100 developing an embedded microphone array system



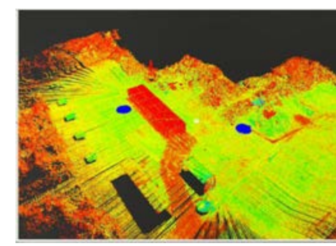
RASP-MX
(SiF, Inc)

2. 3D sound source localization

- Estimate 3D position by triangulation with weighted LMS [Washizaki+ IROS 2016]

3. Robustness for online outdoor demo

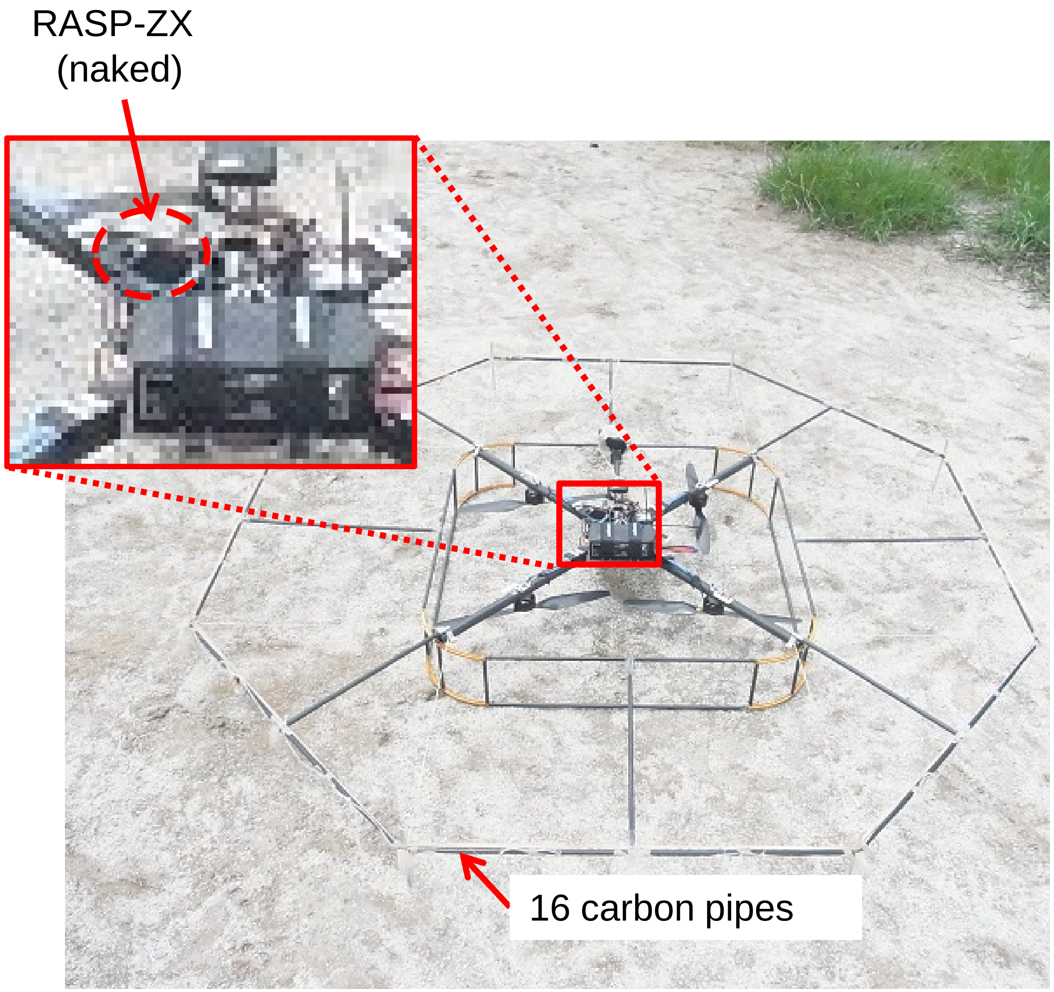
- Take "less time to get ready" with "less human errors" in "all weather"
  - Intuitive UI
    - 3D visualization on a point cloud map
  - Microphone array structure
    - 16ch spherical array
  - All weather
    - Water-resistant microphone array



Real-time visualization
Blue points are sound locations

Water-resistant test (passed 12h submersion test)

First UAV system for online demo



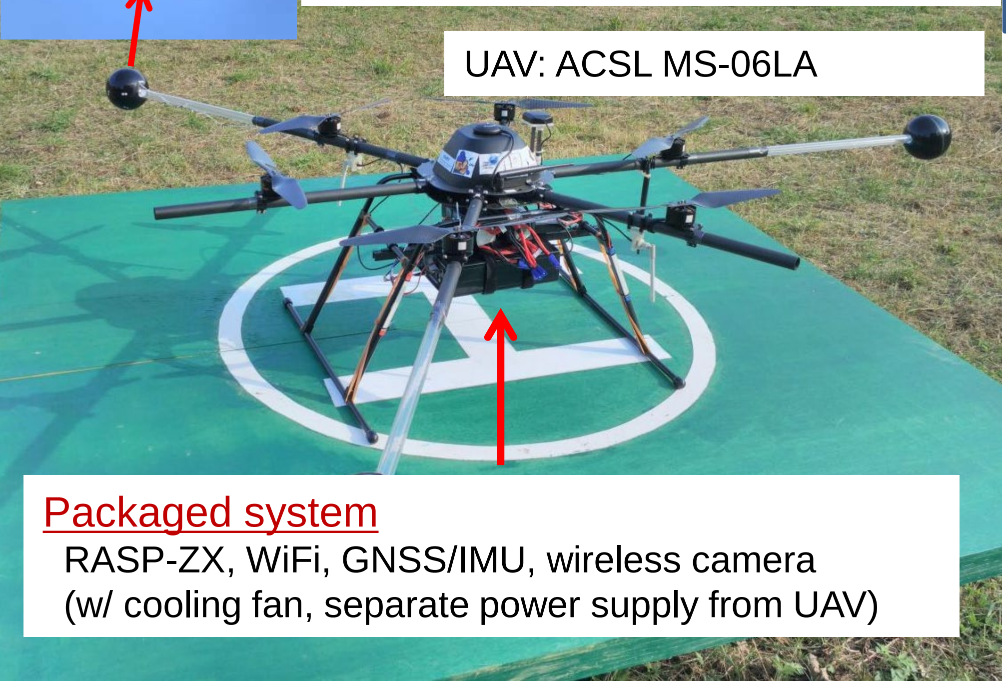
<u>Prototype: enRoute UAV</u>

- 16 pipes and many cables
- Octagonal layout of microphones
- Naked multichannel audio device RASP-ZX for cooling purposes

- *2-hour setup time* necessary
- Many *disconnections* and *contact failures*
- *Not water-resistant*

Revised UAV system for online demo
[Nakadai+ IROS 2017]

16ch spherical microphone array
- Water-resistant
- Only a single cable connection
- Free from frame assembly



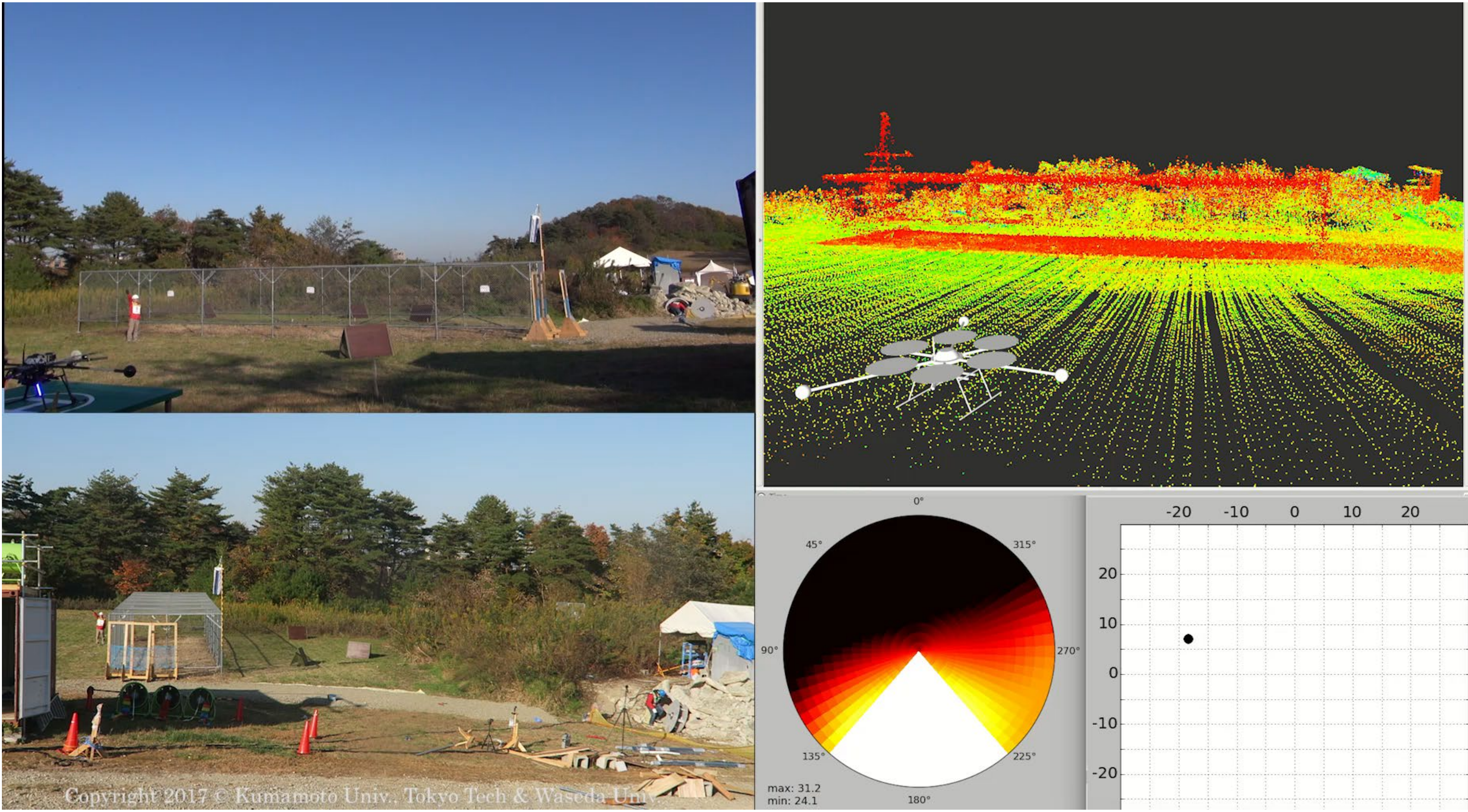
Packaged system
RASP-ZX, WiFi, GNSS/IMU, wireless camera
(w/ cooling fan, separate power supply from UAV)

Parallel distributed processing with HARK and ROS

Displayed information for participants

Preparation Time: reduced **2 hours → 40 min**
Just 2min necessary to take off after the UAV is switched on.
Human errors/disconnections: **not observed**
Weather conditions: operates even in **rainy conditions**

Online Demo@Nov. 11, 2017 (ImPACT TRC field evaluation)

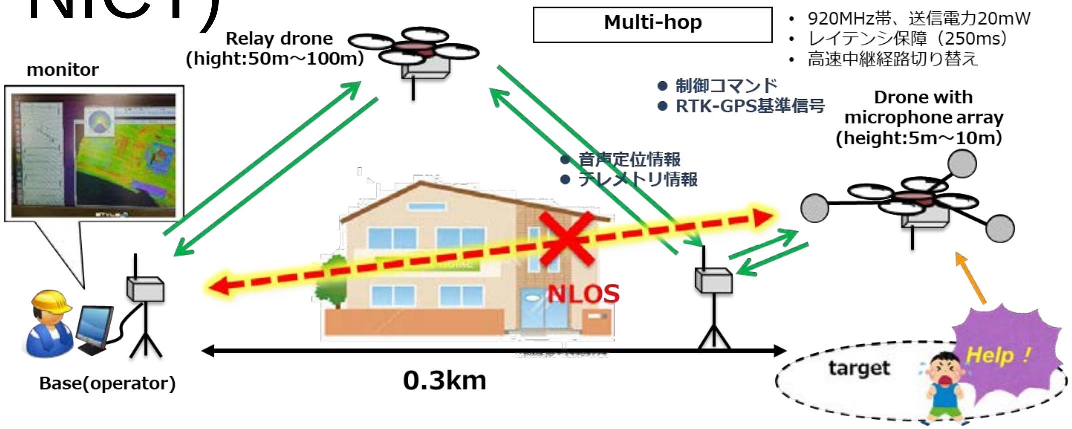


Detected sound locations, and displayed them on 3D map

For more practical tasks

■ Non Line of Sight (NLOS)

– Use multi-hop wireless communication (collaboration with NICT)



■ High Mobility

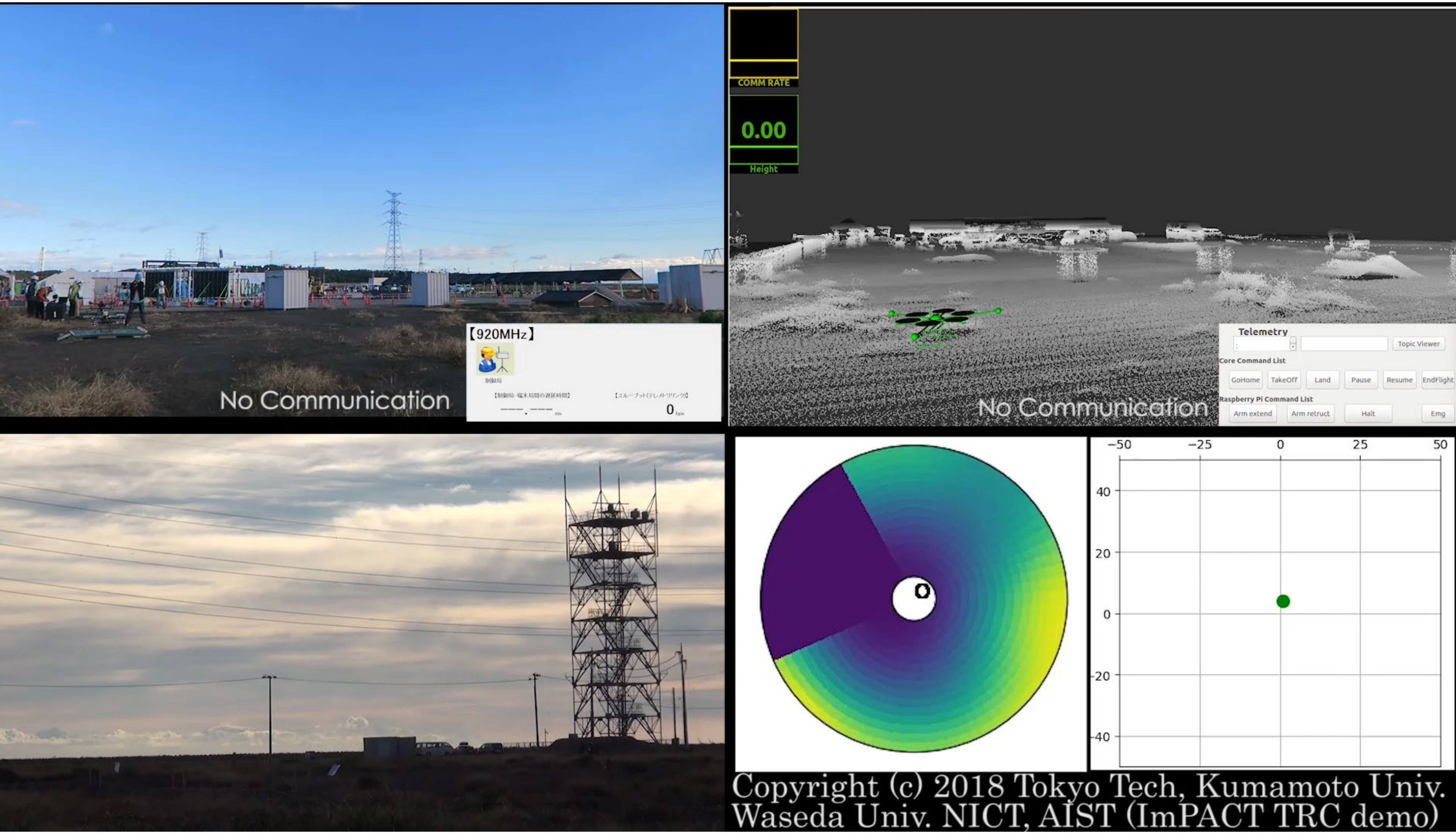
– Slow speed with arms (microphone array) spread

– An open & close mechanism installed

Motion mode (arm closed)

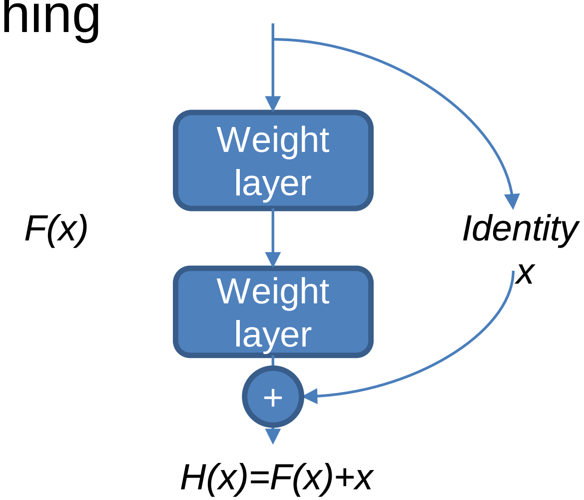
search mode (arm open)

Final Demo@Nov. 2, 2018 (ImPACT TRC field evaluation)



Sound Source Localization Based on Deep Learning [Nelson+ 2016]

- **Deep Residual Network [He+ 2015]**
  - Winner algorithm for ILSVRC 2015 based on CNN
  - Making optimization easier by learning residue
  - Easy to increase the number of layers by avoiding vanishing



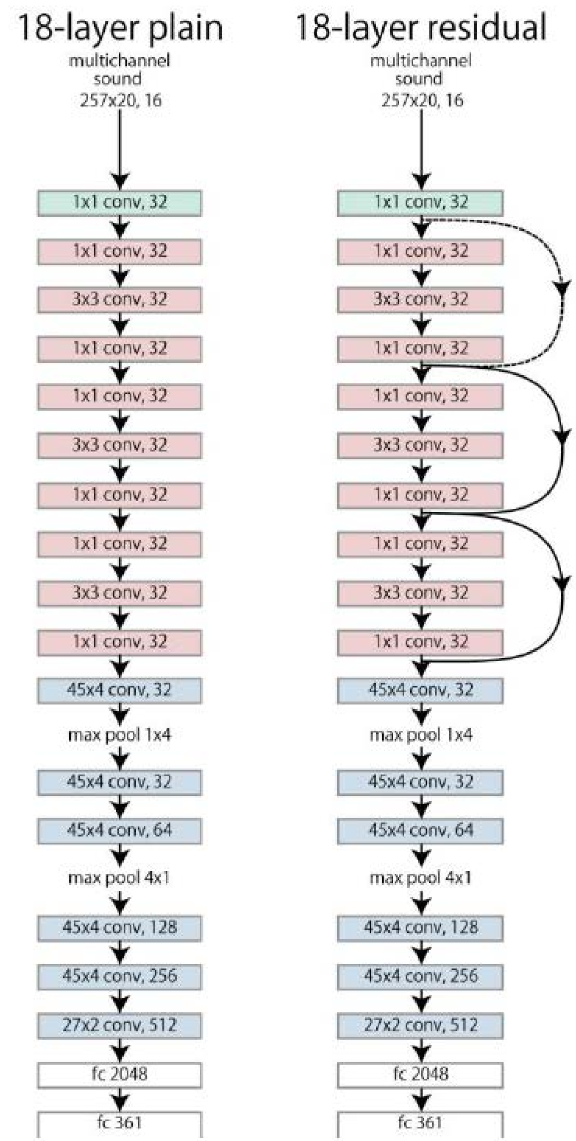
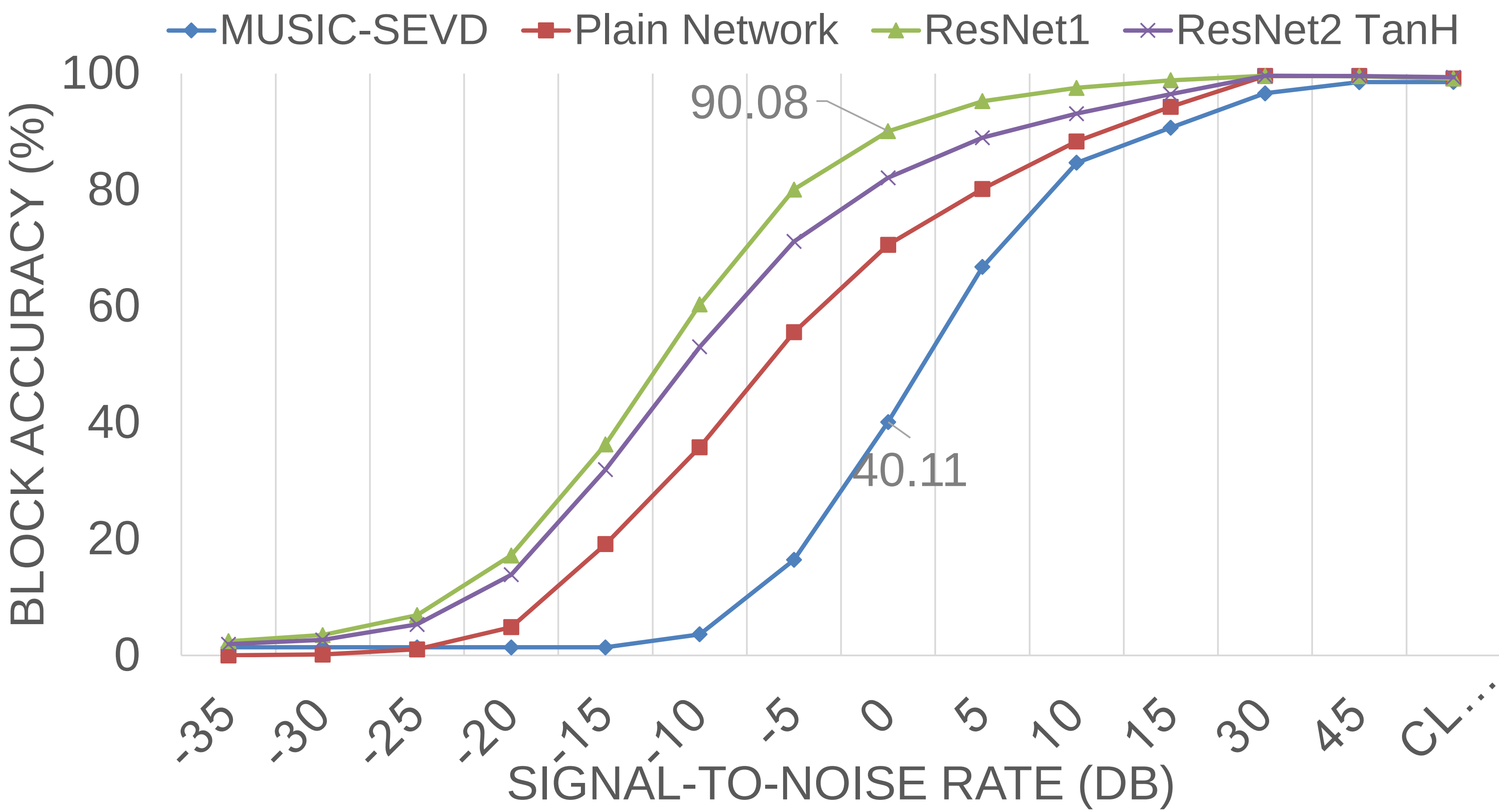


Fig. 4 Network Architecture
Left: Plain Network Right: Residual Network. The dotted line shortcut represents a 1x1 convolutional layer

Performance of Sound Source Localization



- Training data: JNAS (3,350,000 files),
  - Add robot's motor noise to training data(SNR clean～-35dB)
  - Input: 257-dim STFT feature (raw)
  - Output: 36-dim localization result (10 deg resolution)
- Test data: 7,200 files (200 per direction)

Amplitude input only, and no phase information

Summary

- **Introduced robot audition and drone audition**

- **Sound source detection for a UAV with a microphone array**
  1. Real-time processing, communication reduction
  2. 3D sound source localization
  3. Robustness for online outdoor demo
  4. NLOS wireless communication
  5. Open&close microphone array for high mobility

- **Our activities in ImPACT (finished Mar, 2019) were published a book Disaster Robotics 2019**

- **Future work**
  1. Frame-based sound source localization
     - Multiple microphone arrays [Yamada+ 2019] workshop poster
  2. Sound source classification [Morito+16, Uemura+17, Nakadai+18]
  3. Multimodal scene understanding: visual processing, thermal camera, etc
  4. Deep sound source localization for dynamic environments