人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-043

ＡＩチャレンジ研究会 (第４３回)

Proceedings of the 43th Meeting of Special Interest Group on AI Challenges

CONTENTS

日 時　2015 年 11 月 12 日　　場 所　慶応大学 日吉キャンパス 来往舎 シンポジウムスペース／中会議室
Keio University, Tokyo, Nov. 12, 2015

一般社団法人 人工知能学会
Japanese Society for Artificial Intelligence

人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-043-01 (11/12)

音声・画像・映像における Deep Learning を用いたパターン認識

Pattern Recognition using Deep Learning for Speech, Image and Video

篠田浩一
Koichi SHINODA
東京工業大学
Tokyo Institute of Technology
shinoda@cs.titech.ac.jp

Abstract

近年、マルチメディア分野では、Deep Learning(深層学習) が盛んに研究されている。特に、音声認識や画像における一般物体認識では、従来法から大幅にエラーを削減し、すでに標準的な技術として商用にも使われている。本稿では、まず、マルチメディア分野における深層学習のこれまでの研究を概観した上で、現段階における課題とそれに対するアプローチを解説する。研究の進展は急であり、そろそろできることとできないことがはっきりしてきた。最後に、今後、深層学習を用いたパターン認識の研究がどのような方向に進んでいくかを議論したい。

Empirical evidence: Summary

(Dahl, Yu, Deng, Acero 2012, Seide, Li, Yu 2011 + new result)

- Voice Search SER (24 hours training)

AM	Setup	Test
GMM-HMM	MPE	36.2%
DNN-HMM	5 layers x 2048	30.1% (-17%)

- Switch Board WER (309 hours training)

AM	Setup	Hub5'00-SWB	RT03S-FSH
GMM-HMM	BMMI (9K 40-mixture)	23.6%	27.4%
DNN-HMM	7 x 2048	15.8% (-33%)	18.5% (-33%)

- Switch Board WER (2000 hours training)

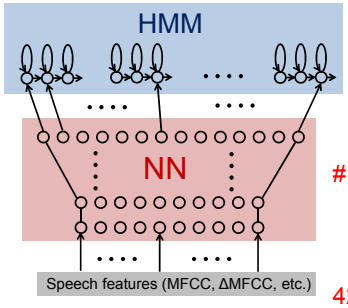
AM	Setup	Hub5'00-SWB	RT03S-FSH
GMM-HMM (A)	BMMI (18K 72-mixture)	21.7%	23.0%
GMM-HMM (B)	BMMI + fMPE	19.6%	20.5%
DNN-HMM	7 x 3076	14.4% (A: -34% B: -27%)	15.6% (A: -32% B: -24%)

(Dong Yu, 2012)

Neural network based speech recognition

1989: Time-Delay Neural Network (TDNN)
1994: Hybrid approach of NN and HMM
2000: Tandem connectionist features
2009: DNN phone recognition
2010: Recurrent NN (RNN) for language model
2011: DNN for LVCSR
(large vocabulary continuous speech recognition)
← The same as Hybrid approach (1994)

1994: Hybrid approach of NN and HMM



Use DNN outputs as HMM output prob.

$$p(x|c) = \frac{p(c|x)p(x)}{p(c)}$$

x: data, c: class

# of nodes:

132 x 50 x 5 (1994)

429 x 2048^5 x 761 (2011)

Bourlard and Morgan, "Connectionist Speech Recognition: A Hybrid Approach", The Springer International Series in Engineering and Computer Science, vol. 247, 1994

Replace GMM with DNN

- GMM (Gaussian Mixture Model) is mixture of experts (MoE), DNN is product of experts (PoE).
  - For GMM, it is difficult deal with multiple events in one window
  - GMM parameter estimation is easier to be parallelized
- DNN can get more info from multiple frames
  - GMM often use diagonal covariance and ignore correlation among them

Hinton et al., "Deep neural networks for acoustic modeling in speech recognition", IEEE Signal Processing Magazine, Nov. 2012.

Deep Learning (DL) in ICASSP2014

Already *de facto* standard

- 84 of 304 (28%) papers deals with DL
- Four sessions titled "DL" or "NN"
- DL penetrates into most speech sub-areas

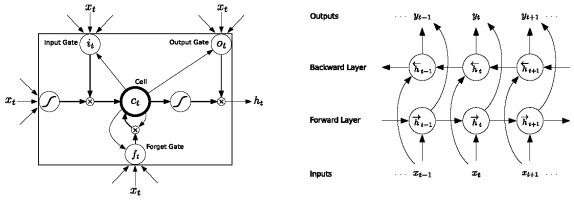
  Robustness (14), ASR systems (8), Features (7), Language model (5), Speaker recognition (5), Spoken term detection (3), Speech understanding (2), Emotion recognition (2)....

These trends continued in ICASSP2015

For high accuracy:

LSTM+Bi-directional RNN

Use LSTM (long-short-term memory) in RNN (Recurrent NN)
RNN: Effectively represents time sequence data
Bidirectional: Use info not only past but also future
LSTM: To use long contexts, make a cell which consists of 4 nodes



Graves et al., "Speech recognition with deep recurrent networks", ICASSP 2013.

For data sparsity:

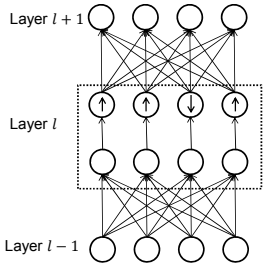
Speaker adaptation

To avoid overtraining, utilize prior knowledge about speakers

1. Regularization in parameter estimation (Bayesian approach)
2. Linear combination of speaker-cluster NNs
3. Add "speaker code" to NN inputs
4. Estimate activation function parameters

For data sparsity:

Estimate a new parameter of each node



Output of layer l

$$h^l = a(r^l) \circ \phi(W^{l\mathrm{T}} h^{l-1})$$

$\circ$: Element-wise multiplication

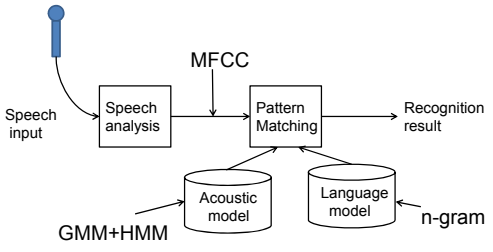
$a(r^l)$: Estimate for each speaker

free parameters ≃ # nodes

P. Swietojanski and S. Renals, "Learning hidden unit contribution for unsupervised speaker adaptation of neural network acoustic models", IEEE SLT 2014.

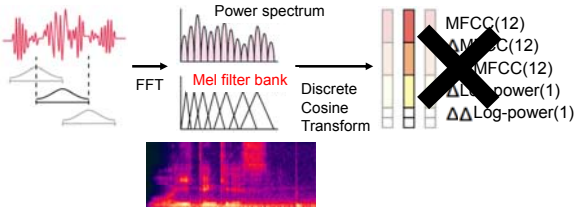
For end-to-end

Speech Recognition System - Before



For end-to-end

MFCC is no more needed



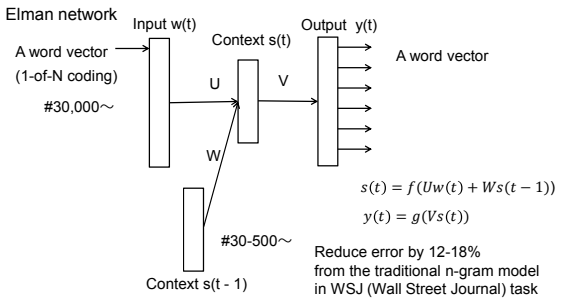
Mel filter bank features reduced 5-10% errors from MFCCs

- MFCC was used to de-correlate the Mel filter bank features
- In DNN, such de-correlation process is not needed

Mohamed et al. "Acoustic modeling using deep belief network", IEEE Trans. ASLP, vol. 20, no. 1, 2012.

For end-to-end

2010: Recurrent NN for language model



$$s(t) = f(Uw(t) + Ws(t-1))$$

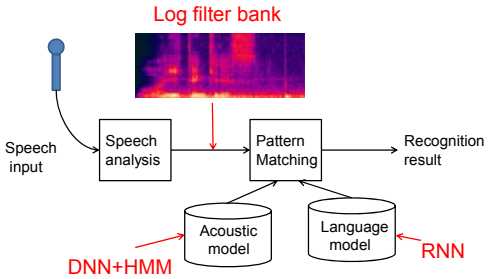
$$y(t) = g(Vs(t))$$

Reduce error by 12-18% from the traditional n-gram model in WSJ (Wall Street Journal) task

Mikolov et al. "Recurrent neural network based language model", INTERSPEECH2010

For end-to-end

Speech Recognition System - After

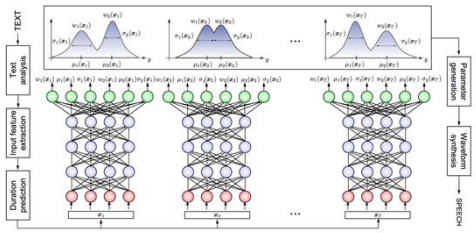


Mohamed et al. "Acoustic modeling using deep belief network", IEEE Trans. ASLP, vol. 20, no. 1, 2012.
Arisoy et al. "Deep neural network language models", NAACL-HLT 2012 workshop

Various applications

DNN for speech synthesis

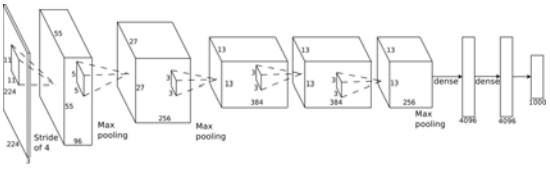
- Use DNN in reverse - input: label, output: data
- Output GMM parameters, mean and variance



Zen et al., Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis", ICASSP2014

ImageNet Challenge: ILSVRC 2012

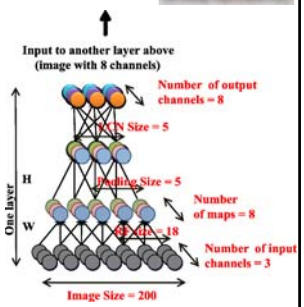
- Detect images of 1000 categories
- 1.2 million training samples
- Error 16% !



Krizhevsky et al. "ImageNet Classification with Deep Convolutional Neural Networks", NIPS2012.

Cat  Human face 

- Unsupervised learning
- 10 billion images from YouTube videos, each 200x200 pixels
- Sparse autoencoder with 9 layers, 1 billion nodes



Le et al. "Building high-level features using large scale unsupervised learning", ICML2012

TRECVID

(TREC Video Retrieval Evaluation)

Spinned out from Text REtrieval Conference (TREC) in 2001,
Organized by NIST(National Institute of Standard and Technology)
Aim : Promote research on video contents analysis and search
International, Competitive, Closed
Homepage: http://trecvid.nist.gov

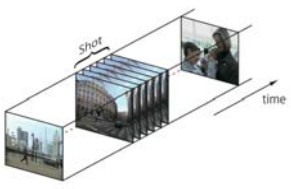
TokyoTech participated from 2006 (9 years)

2014 TRECVID task

- Semantic INdexing (SIN)
  Detect generic objects, scenes, actions
- Surveillance Event Detection (SED)
  Detect specific actions from surveillance video
- INstance Search (INS)
  Given a still image of an object, search video clips including it
- Multimedia Event Detection (MED)
  Detect complex "event"
- Multimedia Event Recounting (MER) (Pilot)
  Explain "event" detected

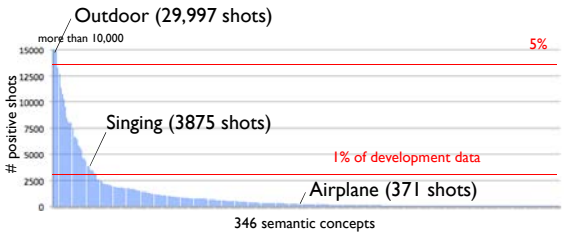
Semantic Indexing

Detect concepts from a set of video shots
Shot: The minimum unit of video
No. Concepts: 60
Training set: 549,434 shots, 800 hours
Test set: 332,751 shots, 400 hours



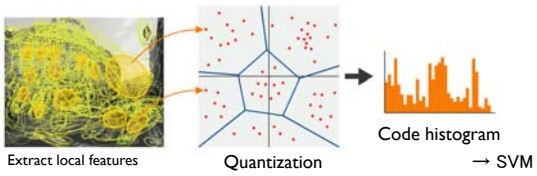
Frequency of Appearance (2011 task)

Number of positive samples in 264,673 training video shots



Bag of Visual Words

1. Quantize local features (e.g., SIFT) by using a codebook
   (Code word: Visual Word)
2. Use a code histogram as an input to SVM



Quantization Error!

Recent Trend

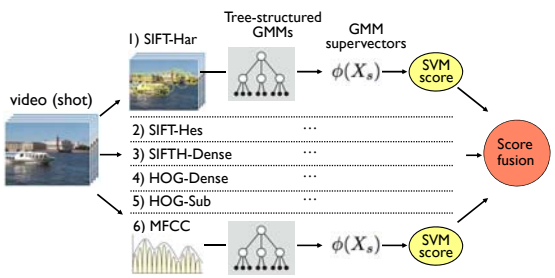
Tackle the data sparseness problem

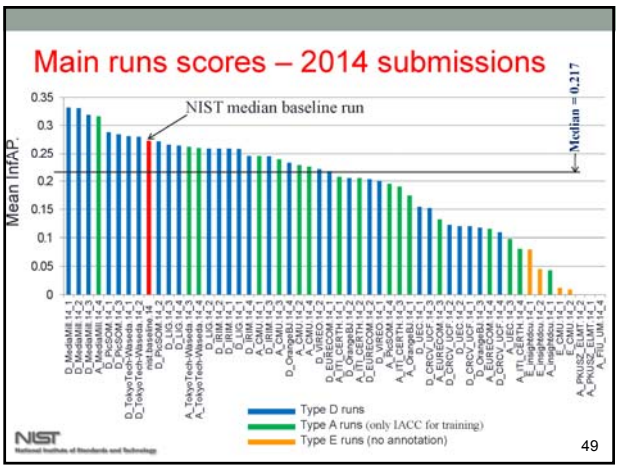
- More features
  SIFT, Color SIFT, SURF, HOG, GIST, Dense features
- Multi-modal
  Use Audio : Singing, Dance, Car, etc.
- Multi-frame
  Not only key frames
- Soft clustering
  Reduce quantization errors. GMM etc.

Less effective than expected

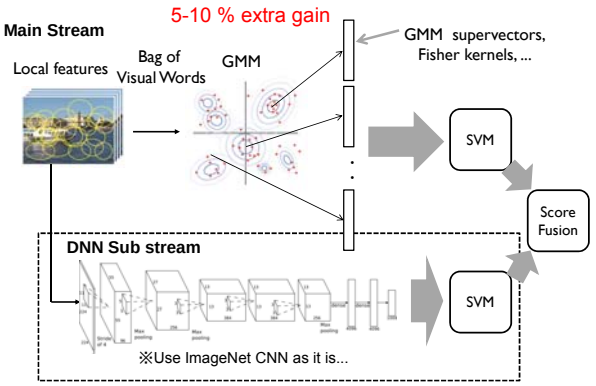
- Global features such as color histogram
  Local features are enough (no complementary info)
- Speech recognition, OCR
  Do not have performance high enough to contribute
- Object location
  Fail to detect. Many concepts do not have "location"
- Context between concepts
  Too Little data

TokyoTech Framework





Deep Learning



BoF is also deep learning!

Fisher Kernel based method is 5-layer DNN

stage	operation	type
SVM prediction	sign $f(X)$ $f(X) = \langle w, \phi(X) \rangle$	non-linear linear
per image vector, $\psi(X)$	square root, normalize (3) compute average of $\psi(x_i)$	non-linear linear
per descriptor vector, $\psi(x_i)$	multiply by γ_k in (1)/(2) bracket $(\cdot)$ in (1)/(2)	non-linear linear
preprocessing	L^2-normalization PCA projection	non-linear linear
SIFT	local pooling gradient filter	non-linear linear
image (as multiple overlapping regions)		

Table 1. Schematic description of a Fisher kernel SVM as a 5-layer feed-forward architecture (from bottom to top).

Sydorov et al., "Deep Fisher Kernels. End to End Learning of the Fisher Kernel GMM Parameters", CVPR2014

TRECVID Multimedia Event Detection (MED) task

- Extract "complex event" from many video clips (shot sequences)
  e.g. "Batting a run in", "Making a cake"
- Database : Home video 2000 hours
- Sponsored by IAPRA (The Intelligence Advanced Research Projects Activity)

Deep Learning at present

- Can be better than human in "well-defined" tasks with large data

MED task

- Multimedia
  Visual features, audio features, speech recognition, OCR
- Dynamic nature
- Training data for each event may be very small

Problems of Deep Learning

- How to deal with more complex problems such as MED?
- Only for "end-to-end" problems
  - Do we really need to solve them?
  - What is "semantics"?
- How to combine many modes in multimedia application
  - Combinatorial explosion
  - Time sequence

What we can do...

- Time Sequence
- Segmentation and Recognition
- Signal and symbol processing

Summary

- Deep learning is already de-facto in speech recognition
- Now, we are busy with replace "traditional" units by "DNN" units in a speech recognition system
  - What I explained today is only a small part of them
- Still ad-hoc, not enough theoretical background
  - How to optimize structures?
  - Why is Deep learning better?
  - How to combine acoustic and language models?

Speech is "lighter" compared with the other media.
Good test bed for exploring Deep learning!

Robust Dereverberation Adaptive to Speaker's Face Orientation

Randy Gomez, Keisuke Nakamura, Takeshi Mizumoto, and Kazuhiro Nakadai

Honda Research Institute Japan Co., Ltd.

Abstract

Reverberation poses a problem to the active robot audition system. The change in speaker's face orientation relative to the robot perturbs the room acoustics and alters the reverberation condition at runtime, which degrades the automatic speech recognition (ASR) performance. In this paper, we present a method to mitigate this problem in the context of the ASR. *First*, filter coefficients are derived to **correct the Room Transfer Function (RTF)** per change in face orientation. We treat the change in the face orientation as a filtering mechanism that captures the room acoustics. Then, joint dynamics between the filter and the observed reverberant speech is investigated in consideration with the ASR system. *Second*, we introduce a **gain correction** scheme to compensate the change in power as a function of the face orientation. This scheme is also linked to the ASR, in which gain parameters are derived via the Viterbi algorithm. Experimental results using Hidden Markov Model-Deep Neural Network (HMM-DNN) ASR in a reverberant robot environment, show that proposed method is robust to the change in face orientation and outperforms state-of-the-art dereverberation techniques.

Index Terms: Robust Robot Audition, Speech Enhancement, Dereverberation, Automatic Speech Recognition

1. Introduction

Reverberation is a phenomenon caused by the reflections of the speech signal in an enclosed environment. It smears the original speech due to the different time delays of arrival among the speech reflections. This phenomenon causes mismatch and degrades the ASR performance. To abate the effect of mismatch, the reverberant speech is enhanced, which is referred to as dereverberation. The problem concerning reverberation is further plagued when the room acoustics is perturbed as a result of the change in the speaker's face orientation. This event **alters the RTF resulting to another mismatch at runtime**. Consequently, the change in face orientation **affects the directivity pattern in which the speech is diffused, causing power issues**. There exists different types of dereverberation methods [1][2][13] but most of these have no mechanism in dealing with the acoustic perturbation due to the change in the speaker's face orientation.

In a human-robot communication scenario, the speaker may change its face orientation when communicating to the robot at any given time. Thus, the dereverberation mechanism should be able to cope with this mismatch as well. In this paper, we expand and improve our previous work [3] in mitigating the degradation of the ASR due to the change in the speaker's face orientation. The proposed method employs an **ASR-inspired RTF and gain correction** mechanisms to actively mitigate the changes in the room acoustics and the speech power due to the change in the face orientation. More importantly, the analysis and optimization employed in the proposed method is conducted jointly with the Hidden Markov Models (HMMs) for effective use in ASR application. These HMMs are used in the HMM-DNN ASR evaluation.

In our previous work [3], face direction compensation is achieved through **equalization**. The work in [3] is purely focused on the waveform compensation of the RTF and stops right there without any consideration of the HMMs [3]. Although [3] works well in enhancing the waveform, it has a very coarse treatment of the effect of dereverberation when applied to the HMM-DNN ASR. In contrast, the proposed method takes a HMM-centric approach, in both of the analysis and optimization procedures. In the proposed method, the change in the face orientation is hypothesized to impact the RTF as a filtering mechanism. **Filter coefficients are optimized in the context of the HMMs as per change in the speaker's face orientation**. This process ensures the link between the RTF and the HMMs. Next, we analyze the impact of the change in face the orientation to the power envelope of the speech signal. **Gain values are derived using the dual nature of the speech signal (i.e., acoustic waveform and the hypothesis) to characterize the change in power**. This mechanism links the power correction with the ASR system. Both the filter for RTF correction and the parameters for gain correction are used in the online dereverberation. Hence, the proposed method can adapt to the acoustic perturbation caused by the change in the speaker's face orientation. The derivation of these parameters are linked to the HMMs, a stark contrast from our previous work [3] which focuses purely on waveform enhancement only.

This paper is organized as follows; in Sec. 2, we show the background of the adopted dereverberation platform in our application. The schemes in extracting the filter coefficients, dereverberation parameter update and calculating gain parameters for power correction as per change in face orientation are discussed in Sec. 3. Experimental results and discussion are presented in Sec. 4, and we conclude the paper in Sec. 5.

2. Background

Microphone array processing based on beamforming and blind separation described in [9][17] is employed to convert the multi-microphone observed signals to a separated reverberant signal (single-channel). In our previous method [4][13], the smearing effect of reverberation is adopted from [15][5] and is solely dependent on the room transfer function (RTF) given as

$$\begin{aligned} r(\omega) &= A^E(\omega)c(\omega) + A^L(\omega)c(\omega) \\ &= \quad e(\omega) \quad + \quad l(\omega), \end{aligned} \tag{1}$$

where $r(\omega)$ is the separated reverberant speech w.r.t. ω frequency [9][17] and the right side of Eq. (1) is the reverberation model, where $c(\omega)$ is the clean speech, $A^E(\omega)$ and $A^L(\omega)$ are the early and late reflection components extracted from the full RTF $A(\omega)$. Both $A^E(\omega)$ and $A^L(\omega)$ are experimentally predetermined in [13]. $r(\omega)$ can be treated as the superposition

of $e(\omega)$ and $l(\omega)$, known as the early and late reflections, respectively. In this paper, we represent both $A^E(\omega)$ and $A^L(\omega)$ simply as the full RTF $A(\omega)$. We note that the measured $A(\omega)$ is matched with a speaker talking in front of the robot and hypothetically, **a change in the face orientation would require different sets of RTF measurements** which is a cumbersome process. Hence, we **propose a correction method that does not require any measurement**.

In [13] we treat $l(\omega)$ as long-period noise which is detrimental to the ASR, and dereverberation is defined as suppressing $l(\omega)$ while recovering $e(\omega)$ estimate. The latter is further processed with Cepstrum Mean Normalization (CMN) during ASR. Eq. (1) simplifies dereverberation into a denoising problem, and through spectral subtraction (SS) [10], the estimate $\hat{e}(\omega)$ in frame-wise manner j is given as

$$|e(\omega,j)|^2 = \begin{cases} |r(\omega,j)|^2 - |l(\omega,j)|^2 \\ \quad \text{if } |r(\omega,j)|^2 - |l(\omega,j)|^2 > 0 \\ \\ \beta|r(\omega,j)|^2 \quad \text{otherwise,} \end{cases} \tag{2}$$

where β is the flooring coefficient. In real condition, $l(\omega,j)$ is unavailable, precluding the power estimate $|l(\omega,j)|^2$. Therefore, the observed reverberant signal $r(\omega,j)$ is used instead of $l(\omega,j)$. This is made possible through a scheme in [13] serving as a workaround to this problem. The scheme introduces a multi-band suppression parameter δ_m optimized via the ASR likelihood criterion given as

$$\delta_m = \arg\max_{\boldsymbol{\delta}_{m,c\Delta}} P(\mathbf{y}^{\boldsymbol{\delta}_{m,c\Delta}}|\mathbf{w};\boldsymbol{\lambda}), \tag{3}$$

where $\boldsymbol{\lambda}$ and $\mathbf{w}$ are the speech acoustic and language models, respectively. $c\Delta$ is the discrete step in the search space while $\boldsymbol{\delta}_{m,c\Delta}$ are the suppression parameter values to be searched upon. For a given set of bands $\boldsymbol{Q} = \{Q_1,\ldots,Q_m,\ldots,,Q_M\}$, in the frequency ω, the dereverberation parameter δ_m dictates the extent of the suppression of the reverberant effects. The new estimate $\hat{e}(\omega,j)$ through the modified SS becomes

$$|e(\omega,j)|^2 = \begin{cases} |r(\omega,j)|^2 - \delta_m|r(\omega,j)|^2 \\ \quad \text{if } |r(\omega,j)|^2 - \delta_m|r(\omega,j)|^2 > 0 \\ \\ \beta|r(\omega,j)|^2 \quad \text{otherwise.} \end{cases} \tag{4}$$

It is obvious that the dereverberation platform in Eq. (4) is dependent on the dereverberation parameter δ_m. Consequently, δ_m **depends on the RTF** $A(\omega)$ as depicted in the model in Eq. (1) and needs to be corrected depending on the speaker's face orientation. Although Eq. (1) is effective for waveform enhancement, its formulation has no relation with HMM analysis. Thus, dereverberation performance is very limited to the original face orientation. In this paper, we will show the method of effectively correcting $A(\omega)$ as a function of the speaker's face orientation. The simplified block diagram of the proposed method is shown in Fig. 1. In the proposed method, the mechanism for RTF and power correction is implemented via an offline training scheme according to the change in the face orientation θ. The updated suppression parameters $\hat{\delta}_m^\theta$ resulting from RTF compensation with $\alpha^\theta A(\omega)$ and the gain parameters $G^\theta_{m_\tau}$ are stored for online dereverberation use. Details on Fig. 1 are discussed in the following section.

3. Methods

3.1. Microphone-array and Visual Processing

Sound source separation described in [9][17] is used to obtain the separated reverberant signal r^θ, where θ is the speaker's face orientation. It is defined by setting a straight line between the human and the robot (facing each other) as a reference axis. The change in speaker orientation is defined as the angular change θ from the reference axis from the human side. In our work we consider a deviation $-30 \leq \theta \leq 30$, where $\theta = 0$ is the reference angle in which the generic RTF is defined. The angle θ is estimated using the Kinect sensor.

3.2. Room Transfer Function Correction

Suppose that the observed reverberant speech at a particular face orientation θ when processed by a filter is given as

$$x^\theta[h] = \sum_{k=0}^{K-1} \alpha_k^\theta \;\; r^\theta[h-k], \tag{5}$$

where r^θ and α_k^θ are the observed reverberant speech and the filter coefficients, respectively. We note that the room acoustics information is captured in the observed reverberant speech via reflections on the enclosed space. We use the actual signal r^θ to analyze the reverberation condition as per change in face direction θ through the filter $\boldsymbol{\alpha}^\theta$. The filter of length K is given as

$$\boldsymbol{\alpha}^\theta = [\alpha_0^\theta, \alpha_1^\theta, ..., \alpha_{K-1}^\theta]^T. \tag{6}$$

The objective is to estimate $\boldsymbol{\alpha}^\theta$ in the context of the ASR. The resulting estimate captures the room acoustics at θ, and later used not just to correct the change in θ but making sure that the correction is more likely to improve the ASR performance. Since we are interested of the ASR's output (hypothesis), the actual signal x is immaterial. The hypothesis is expressed as

$$\hat{\boldsymbol{w}}^\theta = \underset{\boldsymbol{w}}{\text{argmax}} \log\, (P(f^{(x^\theta)}(\boldsymbol{\alpha}^\theta)|\boldsymbol{w})P(\boldsymbol{w}), \tag{7}$$

where $f^{(x^\theta)}(\boldsymbol{\alpha}^\theta)$ is the extracted feature vector from the utterance, $\boldsymbol{w}$ is the phoneme-based transcript, $P(f^{(x^\theta)}(\boldsymbol{\alpha}^\theta)|\boldsymbol{w})$ is the acoustic likelihood (i.e., using reverberant acoustic model) and $P(\boldsymbol{w})$ is due to the language (i.e., using language model). The latter can be ignored since phoneme-based transcript $\boldsymbol{w}$ is known, thus, argmax in Eq. (7) acts on $\boldsymbol{\alpha}^\theta$ and rewritten as

$$\hat{\boldsymbol{\alpha}}^\theta = \underset{\boldsymbol{\alpha}^\theta}{\text{argmax}} \log\, P(f^{(x^\theta)}(\boldsymbol{\alpha}^\theta)|\boldsymbol{w}). \tag{8}$$

In ASR, the total log likelihood in Eq. (8) when expanded [14] to include all possible state sequence is expressed as

$$\Gamma(\boldsymbol{\alpha}^\theta) = \sum_j \log\, P(f_j^{(x^\theta)}(\boldsymbol{\alpha}^\theta)|\hat{s}_j), \tag{9}$$

where s_j is the state at frame j. Eq. (9) heralds the formulation in the context of the HMMs via the state sequence. By using the ∇ operator, the total probability is maximized w.r.t the filter coefficient in Eq. (6), thus,

$$\nabla_{\boldsymbol{\alpha}^\theta}\, \Gamma(\boldsymbol{\alpha}^\theta) \;= \left\{ \frac{\partial\Gamma(\boldsymbol{\alpha}^\theta)}{\partial\alpha_0^\theta}, \frac{\partial\Gamma(\boldsymbol{\alpha}^\theta)}{\partial\alpha_1^\theta}, ..., \frac{\partial\Gamma(\boldsymbol{\alpha}^\theta)}{\partial\alpha_{K-1}^\theta} \right\}. \tag{10}$$

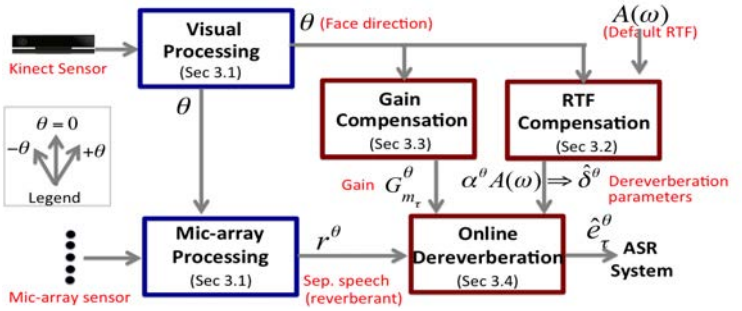


Figure 1: Overall System Structure.

Assuming a Gaussian mixture distribution with mean vector μ_{jv} and diagonal covariance matrix $\mathbf{\Sigma}_{jv}^{-1}$, respectively. Eq. (10) can be shown similar to that in [8] as

$$\nabla_{\boldsymbol{\alpha}^\theta} \Gamma(\boldsymbol{\alpha}^\theta) = -\sum_{j}\sum_{\mathrm{v}=1}^{\mathrm{V}} \gamma_{jv} \frac{\partial f_j^{(x^\theta)}(\boldsymbol{\alpha}^\theta)}{\partial \boldsymbol{\alpha}^\theta} \mathbf{\Sigma}_{jv}^{-1}(f_j^{(x^\theta)}(\boldsymbol{\alpha}^\theta) - \mu_{jv}), \tag{11}$$

where γ_{jv} is the posteriori of v-th mixture and j-th frame of the most likely HMM state. $\frac{\partial f_j{}^{(x^\theta)}(\boldsymbol{\alpha}^\theta)}{\partial \boldsymbol{\alpha}^\theta}$ is the Jacobian matrix of the reverberant feature vector. The filter coefficients are obtained using [11][12] based on Eq. (11). Correcting a generic RTF to the current face orientation θ of the speaker is given as

$$\hat{A}^\theta(\omega) = \alpha^\theta(\omega) A(\omega) \tag{12}$$

where $\alpha^\theta(\omega)$ is the face orientation-compensating filter in the frequency domain. It follows that a new dereverberation parameter can be extracted from the corrected RTF,

$$\hat{A}^\theta(\omega) \Rightarrow \hat{\delta}^\theta_m \tag{13}$$

The updated dereverberation parameters $\hat{\delta}^\theta_m$ are stored for online use in Sec 3.4.

3.3. Speech Power Compensation via Gain Correction

The change in face orientation does not only impact the RTF, but it also affects the power level of the separated signal r^θ. To mitigate the effect of the latter, we employed a power compensation scheme via gain correction. The process of deriving the gain is depicted in Fig. 2. Two sets of reverberant speech database are prepared, one is recorded facing directly the robot θ_A (s.t. $\theta = 0$), and the other set with face orientation θ_B (s.t. $\theta_B \neq 0$). θ_A is the reference face orientation in which θ_B is to be corrected to. The utterances are classified according to the time-duration referred to as template τ. Same duration utterances are grouped together (**time-duration classification**). We note that reverberation is characterized by the smearing phenomenon in which the power of the previous sound frames are carried over to the current frame. In this regard, the effect of reverberation is directly related to the duration of the speech utterance. Hence, it is noteworthy to analyze the impact of both the changes in the face orientation and speech duration, respectively. Consequently, the reverberant utterances are referred to as $r^{\theta_A}_\tau$ and $r^{\theta_B}_\tau$, respectively. Next, we analyze the change in power dynamics per change in face orientation θ_B relative to θ_A.

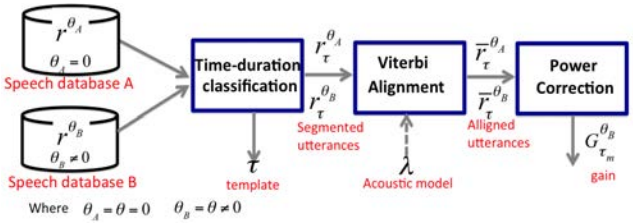


Figure 2: The offline training scheme used to calculate gain parameters for power gain correction.

To effectively establish the correspondence of the sound units (i.e. phonemes) between the two utterances in θ_B and θ_A, the utterances are **aligned via the Viterbi algorithm** using a known acoustic speech model λ. This is a very crucial step because we want to model the change in power similar to the concept of the reverberation phenomenon in which the energy of the current frame is affected by the previous frames. To achieve that, we need to have a correct association of the sound-frames between the speech database A and B. The alignment will guarantee that the particular sound of the current frame of interest in r^{θ_A} likely corresponds the same sound in r^{θ_B}, one-to-one correspondence is achieved. Moreover, the alignment scheme links the power analysis between the acoustic waveform and the hypothesis which are both used by the ASR system.

Frame-wise power spectral analysis is conducted to the aligned utterances $\bar{r}^{\theta_A}_\tau$ and $\bar{r}^{\theta_B}_\tau$ for face orientation θ and the template τ, respectively. The reverberant power of both are compared and analyzed. Then, band coefficients that minimizes the error between the two are extracted. The minimization of the error means minimizing the power mismatch between $\bar{r}^{\theta_A}_\tau$ and $\bar{r}^{\theta_B}_\tau$. For a total of O utterances indexed by o in a template τ, the error to be minimized is given as

$$E^{\theta_B}_\tau(j) = \frac{1}{\mathrm{O}} \sum_{\mathrm{o}} \sum_{\omega \in Q} |\bar{r}^{\theta_A}_\tau(\omega, \mathrm{o}, j) - G^{\theta_B}_{\tau_m}(\omega, \mathrm{o}, j) \bar{r}^{\theta_B}_\tau(\omega, \mathrm{o}, j)|^2, \tag{14}$$

where $G^{\theta_B}_{\tau_m}$ is the gain for the given set of bands $\boldsymbol{Q} = \{Q_1, \ldots, Q_m, \ldots, Q_M\}$ of template τ. $\bar{r}^{\theta_A}_\tau(\omega, \mathrm{o}, j)$ and $\bar{r}^{\theta_B}_\tau(\omega, \mathrm{o}, j)$ are the j-th frame viterbi-aligned utterance o from the speech database A and B, respectively. Since we are interested of the power dynamics for each frame in a given template τ, the summation in Eq. (14) is conducted on the

Table 1: Recognition performance in word accuracy (%)

Reverberation Time = 940 msec.@ Distance = 2.0 m	$\theta = -30$	$\theta = -15$	$\boldsymbol{\theta = 0}$	$\theta = +15$	$\theta = +30$
(A) No Enhancement	45.5 %	53.0 %	**64.7 %**	54.7 %	48.6 %
(B) Based on Feature Adaptation [16]	55.1 %	62.2 %	**70.0 %**	62.9 %	56.4 %
(C) Based on Wavelet Extrema [2]	57.3 %	63.7 %	**71.8 %**	63.2 %	57.1 %
(D) Based on LP Residuals [1]	59.7 %	65.4 %	**74.2 %**	66.1 %	59.3 %
(E) Based on Equalization (Previous work) [3]	68.1 %	75.9 %	**81.3 %**	76.5 %	69.3 %
(F-a) Proposed Method (RTF Comp. (Sec. 3.2))	74.9 %	77.4 %	**81.3 %**	78.1 %	75.7 %
(F-b) Proposed Method (RTF and gain Comp. (Sec. 3.2 & Sec. 3.3))	76.8 %	79.2 %	**81.3 %**	79.9 %	77.0 %
(G) Dereverberation with θ-matched RTF (Upperlimit) [13]	78.7 %	80.4 %	**81.3 %**	80.7 %	79.3 %
Reverberation Time = 940 msec. @ Distance = 3.0 m	$\theta = -30$	$\theta = -15$	$\boldsymbol{\theta = 0}$	$\theta = +15$	$\theta = +30$
(A) No Enhancement	30.7 %	37.2 %	**52.7 %**	40.5 %	32.1 %
(B) Based on Feature Adaptation [16]	37.0 %	43.4 %	**58.7 %**	44.7 %	36.8 %
(C) Based on Wavelet Extrema [2]	40.5 %	48.7 %	**62.4 %**	49.0 %	42.3 %
(D) Based on LP Residuals [1]	45.2 %	51.3 %	**66.1 %**	52.5 %	45.8 %
(E) Based on Equalization (Previous work) [3]	52.6 %	58.3 %	**73.9 %**	59.1 %	52.1 %
(F-a) Proposed Method (RTF Comp. (Sec. 3.2))	58.0 %	65.2 %	**73.9 %**	66.7 %	59.1 %
(F-b) Proposed Method (RTF and gain Comp. (Sec. 3.2 & Sec. 3.3))	63.8 %	67.3 %	**73.9 %**	68.8 %	64.9 %
(G) Dereverberation with θ-matched RTF (Upperlimit) [13]	65.8 %	69.2 %	**73.9 %**	70.4 %	66.7 %

same frame index across O. For a given template τ of j frames, we extract a sequence of multi band m gain values of $[\mathbf{G}^{\theta}_{\tau_m}(\omega,1),\ldots,\mathbf{G}^{\theta}_{\tau_m}(\omega,j),\ldots,\mathbf{G}^{\theta}_{\tau_m}(\omega,J)]$, for **power correction**. These values are then stored for online use in Sec 3.4.

3.4. Online Dereverberation

In the online mode (see Fig. 1), the visual processing scheme identifies the face orientation θ while the microphone array processing scheme converts the multichannel signal to a single channel separated reverberant signal r^{θ}. RTF and gain correction due to the change in face orientation θ as discussed in Sec 3.2-3.3 are used for dereverberation. Specifically, the adopted dereverberation platform based on spectral subtraction in Eq. (4) is rewritten as

$$|\hat{e}^{\theta}_{\tau}(\omega,j)|^2 = \begin{cases} |r^{\theta}_{\tau}(\omega,j)|^2 - \hat{\delta}^{\theta}_{m} G^{\theta}_{\tau_m}(\omega,j)|r^{\theta}_{\tau}(\omega,j)|^2 \\ \quad \text{if } |r^{\theta}_{\tau}(\omega,j)|^2 - \\ \quad \hat{\delta}^{\theta}_{m} G^{\theta}_{\tau_m}(\omega,j)|r^{\theta}_{\tau}(\omega,j)|^2 > 0 \\ \\ \beta |r^{\theta}_{\tau}(\omega,j)|^2 \quad \text{otherwise.} \end{cases} \tag{15}$$

Note that $\hat{\delta}^{\theta}_{m}$ and $G^{\theta}_{\tau_m}$ are the pre-stored values discussed in Sec 3.2-3.3 and are selected based on θ as identified through the visual processing scheme.

4. Experimental Results

4.1. Setup

We evaluate the proposed method in large vocabulary continuous speech recognition (LVCSR) based on a HMM-DNN framework. The training database is the Japanese Newspaper Article Sentence (JNAS) corpus with a total of approximately 60 hours of speech. The test set is composed of 200 sentences uttered by 50 speakers. The vocabulary size is 20K and the language model is a standard word trigram model. Speech is processed using 25ms-frame with 10 msec shift. The fBank features of 40 dimensions. The HMM-DNN has 6 layers with 2048 nodes. The reverberation time is approximately 940 msec., and testing is conducted at 2.0 m and 3.0 m distances, respectively. Speaker face orientation θ is defined in degree. The generic RTF matching that of the model training is at $\theta = 0$, in which the speaker is directly facing the robot. The test speakers' face orientation deviates at $\theta = -30, -15, +15, +30$, respectively. Key to evaluating the results of the different methods is the robustness of the recognition performance as θ deviates from $\theta = 0$ (matched condition) to $-30 \leq \theta \leq +30$ (mismatched conditions). The test data are recorded at $\theta = -30, -15, +15, +30$. This is done by re-playing the clean test database using a loudspeaker at angle θ and distances 2.0m and 3.0m, respectively. Hence, we use real reverberant speech.

4.2. ASR Performance

The ASR results are shown in Table 1. Method (A) is when no enhancement is employed while method (B) is the result based on feature adaptation by [16]. Instead of suppression, method [16], minimizes the reverberant mismatch through adaptation of the feature vector. The result in method (C) is based on wavelet extrema clustering [2], which operates in the wavelet domain to remove the effects of reverberation. Method (D) is based on the Linear Prediction residual approach [1]. By exploiting the characteristics of the vocal chord, it is able to remove the effects of reverberation. The method in (E) is based on our previous work [3] which employs an equalization technique to mitigate the change in face orientation. The proposed method (F-a) is evaluated when only the RTF compensation is in effect (Sec. 3.2); and (F-b) when both the RTF and gain compensation are employed (Sec. 3.2 and Sec 3.3), respectively. In method (G), the result of using a θ-matched RTF is shown; RTF are measured for each microphone and for each change in θ. The result in method (G) serves as the upperlimit for the adopted dereverberation platform. We note that methods (E)-(G) use the same dereverberation platform and differs only in the mitigation of the change in the face orientation. Therefore, methods (E)-(G) have the same performance at $\theta = 0$.

Table 1 shows that the proposed method outperforms the

existing methods and the previous work [3]. The recognition performance is robust to degradation when face orientation changes relative to the original condition $\theta = 0$. Moreover, it outperforms the previous work in method (E) [3]. This is because the proposed method is linked to the ASR system. The formulation to mitigate the change in the face orientation (i.e., RTF and gain corrections) evolves within the HMM construct. This hinged the optimization procedure to the ASR system itself. In contrast, the previous work and the rest of the methods are focused primarily on the waveform enhancement only.

5. CONCLUSION

In this paper, we have shown the method of analyzing the impact of the change in the face orientation through the alteration of both the RTF and power. These two creates a mismatch that degrades ASR performance when using the dereverberation framework. Moreover, we compensate its impact to the RTF by correcting it using optimized filter coefficients, specifically derived in the context of ASR. Also, the impact in power is corrected as per change in face orientation. Considerable amount of time is needed when measuring new RTFs. In the proposed method, the re-measurement of the RTF as a function of the face orientation can be avoided, this allows the robot to actively mitigate its impact online. We have compared our results with existing dereverberation methods, our previous work and the method when using a matched RTF.

Currently, our work is limited to the definition of the change in face orientation based on our experiment. In real world, the face orientation is more unpredictable resulting to unsymmetrical face orientation relative to the robot. In our future work, we will improve the current system to include random face directions. Although the proposed method involves the concept HMM in deriving the dereverberation and gain parameters, we did not consider actual model adaptation in this work. Hence, the latter will be part of our future work as well.

6. References

[1] B. Yegnanarayana and P. Satyaranyarana, "Enhancement of Reverberant Speech Using LP Residual Signals", *In Proceedings of IEEE Trans. on Audio, Speech and Lang. Proc.*, 2000.

[2] S. Griebel and M. Brandstein, "Wavelet Transform Extrema Clustering for Multi-channel Speech Dereverberation" *IEEE Workshop on Acoustic Echo and Noise Control*, 1999.

[3] R. Gomez, K. Nakamura, T. Mizumoto and K. Nakadai, "Dereverberation Robust to Speaker's Azimuthal Orientation in Multi-channel Human-Robot Communication" *In Proceedings IEEE Intelligent Robots and Systems IROS*, 2013.

[4] R. Gomez, K. Nakamura, and K. Nakadai, "Robustness to Speaker Position in Distant-Talking Automatic Speech Recognition" *In Proceedings IEEE Int. Conf. Acoust., Speech, Signal Proc. ICASSP*, 2013.

[5] P. Naylor and N. Gaubitch, "Speech Dereverberation" *In Proceedings IWAENC*, 2005

[6] Akinobu Lee, *Multipurpose Large VocabularyContinuous Speech Recognition Engine*, 2001.

[7] S. Vaseghi "Advanced Signal processing and Digital Noise reduction", Wiley and Teubner, 1996.

[8] M. Seltzer, "Speech-Recognizer-Based Optimization for Microphone Array Processing *IEEE Signal Processing Letters*, 2003.

[9] H. Nakajima, K. Nakadai, Y. Hasegawa and H. Tsujino, "Adaptive Step-size Parameter Control for real World Blind Source Separation" *In Proceedings IEEE Int. Conf. Acoust., Speech, Signal Proc. ICASSP*, 2008.

[10] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction" *In Proceedings IEEE Int. Conf. Acoust., Speech, Signal Proc. ICASSP* ,1979.

[11] , "On numerical analysis of conjugate gradient method" *Japan Journal of Industrial and Applied Mathematics*, 1993.

[12] , W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, "Numerical Recipes in C: The Art of Scientific Computing" *Cambridge University Press*, 1988 .

[13] R. Gomez and T. Kawahara, "Robust Speech Recognition based on Dereverberation Parameter Optimization using Acoustic Model Likelihood" *In Proceedings IEEE Transactions Speech and Acoustics Processing*, 2010.

[14] "The HTK documentation http://htk.eng.cam.ac.uk/docs/docs.shtml"

[15] E. Habets, "Single and Multi-microphone Speech Dereverberation Using Spectral Enhancement" *Ph.D. Thesis*, June 2007.

[16] H.-G. Hirsch and H. Finster, "A new approach for the adaptation of HMMs to reverberation and background noise" *Speech Communication*, pp 244-263, 2008.

[17] "http://winnie.kuis.kyoto-u.ac.jp/HARK/"

音環境知能技術を活用した聴覚支援システムのプロトタイプの開発

Developing a prototype of hearing support system using sound environment intelligence

石井カルロス[1]，劉超然[1]，Jani Even[1]
Carlos ISHI, Chaoran LIU, Jani EVEN
国際電気通信基礎技術研究所
[1]石黒浩特別研究所
[1]ATR/HIL

carlos@atr.jp, chaoran.liu@irl.sys.es.osaka-u.ac.jp, even@atr.jp

Abstract

難聴者に対して従来の補聴器が持つ問題点を解決するため、提案者らがこれまで培ってきた音環境知能（音の時空間的構造化）技術を発展させ、利用者と利用環境に適応して、聞き取るべき音（対話相手の声、呼びかけ、アラームなど）とその妨げとなる不要・不快な音（ドア、エアコン、対話相手以外の声など）を取捨選択でき、更に選択された音に対する空間的感覚を再構築できる聴覚支援システムの実現を目的とする。本稿では、聴覚支援システムのプロトタイプの開発について進捗を報告する。

1 はじめに

世界各国で共通して、その国における人口の 1 割～2 割程度が難聴・聴覚障害を持っているといわれている。2009 年の日本補聴器販売店協会による「補聴器供給システムの在り方に関する研究」報告書の中で、日本の難聴者人口は 15.7% (1944 万人)と報告されている。そのうち、自覚のない難聴者(7.2%)、自覚がある難聴者(4.5%)、ほとんど使用しない補聴器所有者(1.0%)、常時または随時使用の補聴器所有者(2.7%)に分かれる。高齢者の難聴は、神経細胞などの老化現象としての老人性難聴で、65 歳以上では 25～40%、75 歳以上では 40～66%の割合で見られる。高齢化に伴い、難聴者数は更に増加すると予想される。

日本で補聴器を使っている人は 400 万人程度であり、難聴者のうち 5 人に 1 人しか補聴器を使っていないことになる。補聴器を途中で使わなくなる難聴者も多く、その理由として以下が記載されている：

「会話中、周りの音も大きくて、肝心な言葉が聞き取れない。」
「テレビのセリフが聞こえない。」
「コップをテーブルに置いた音、ドアの音などが大きくてびっくりする。」
「水音、新聞をめくる音などが気になる。」
「ピーピー音（ハウリング）が鳴る。」
「玄関チャイムが聞こえない。」
「自分の声が最も大きく聞こえる。」
「自分の声が変に聞こえて気持ち悪い。」
「声や音が聞こえても、どこから鳴ったのかが分からない。」

一般の補聴器は、マイクが補聴器に埋め込まれているため、周囲の雑音も増幅されてしまうという根本的な問題がある。ハウリング（ピーピー音）も起きやすく利用者に苦痛を感じさせる。最近の補聴器は、デジタル処理の導入により、周波数帯域ごとの音量調整や騒音抑制などの機能が埋め込まれ、性能は上がっている。ハウリング防止の信号処理も施しているものがあるが、その分、音量を抑える必要があり、重度難聴には十分な音量が出力できない。

補聴器コンサルタントによると、補聴器を止める原因は多くの場合、利用者に合った補聴器を選べていない、または設定が難しく誤った設定で使用しているためとされているが、それらが適切であっても補聴器単体による快適さ（聞こえやすさ）には限界がある。

ピンマイクやペン型などの遠隔マイクにより、FM 経由で遠隔の声を送受信する機能を持つ補聴器もあるが、遠隔のマイク周辺の雑音も増幅する問題や、音の方向を感知するための空間的情報も保たれない問題が残る。

空間的情報の伝達においては、マイク埋め込みの補聴器を両耳にかけることにより、ある程度解決されるが、自分の声も大きく聞こえる問題は残る。

遠隔センサによる空間的情報の伝達における問題点は、センサと音源の相対的角度が利用者と音源の相対的角度と異なることが原因で、音の方向情報を取得できる多チャンネルの場合でも生じる。聴覚支援を目的に多チャンネルのマイクロホンアレイ技術を活用した研究は国内外多数あるが、ほとんどが一つの音源を強調させ、モノラル信号を出力する仕組みで、空間的情報が失われる。

以上、従来の補聴器の問題点は、次の(1)～(3)にまとめられる。

(1) 利用者に必要な音と不要な音を選択することができない。
(2) 音の空間的情報が失われる。
(3) 設定が複雑で使いにくい。

提案者らは、これまで環境内に設置した複数のマイクロホンアレイと人位置検出システムを組み合わせて、いつ誰がどこで発話したのかを検出できる音環境知能の基盤技術の研究開発を進めてきた。本提案では、環境センサネットワークによる音環境知能技術を発展させ、上述の従来の補聴器の問題点を解

決することにより、利用者が快適な日常生活を可能とする聴覚支援システムの実現を目的とする。

まず問題点(1)に対し、環境内の個々の音を分離することにより、これまで補聴器単体では出来なかった、利用者に対して必要な音と不要な音を取捨選択的に制御可能な聴覚支援システムを提案する。環境センサの利用により、対象音の強調と不要音の抑圧に加え、ハウリングの問題および自分の声が大きく聞こえる問題も解決できる。これにより、従来の補聴器より音量を上げることができ、対象となる音や声が聞きやすくなる。

問題点(2)に対処するために、環境センサにより分解された個々の音源に対し、センサと利用者の相対的な位置や向きに応じた音像（音の空間的情報の感覚）の再構築手法を提案する。これにより、どの方向から音が鳴ったのか、といった空間的情報の知覚を可能にする。

問題点(3)に対して、時と場と利用者の好みに合わせて、環境センサにより、利用者の注意対象および利用者向けの発話対象をシステムが自動的に学習する手法を提案し、利用者の負担を最小限にする対象音選択インタフェースを追究する。スマートホンやタブレットを用いたものや利用者の頭部動作を用いたジェスチャ入力など、複数の利用者層を想定した数種類のインタフェースを提案する。

図１に提案する聴覚支援システムの利用場面のイメージ図を示す。老人ホームや介護施設などの供用空間で複数の利用者が環境センサを共用して、ドアの音や足音、食器の音など、不要・不快な音を抑圧し、利用者が注意している対話相手の声やテレビの音（利用者指向の注意対象）と利用者に背後から話しかけられた声（利用者向けの発話対象）を強調し、利用者に応じてその場で聞くべき音のみを提供するようなシステムの実現を目指す。

本論文では、上記の問題点(1)と(2)を解決するための基本的機能を備えた聴覚支援システムの概要を紹介し、プロトタイプの実現に向けた進捗を報告する。

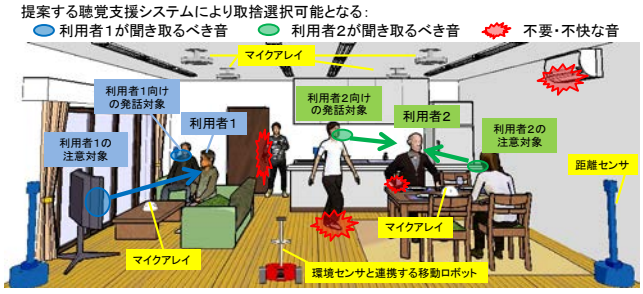


図１．提案する聴覚支援システムの利用場面の例。

2 関連研究

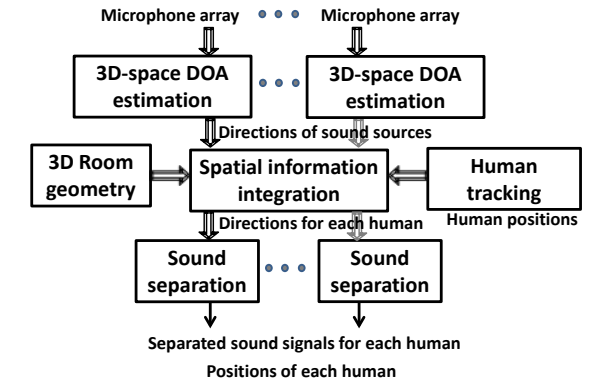
補聴器への応用においては、バイノーラル処理（両耳に装着した補聴器のマイクを利用した信号処理）が、国内外で多く研究されている。例えば、猿渡らは，バイノーラル信号を用いてブラインド信号処理とポストフィルタリングを中心に，両耳補聴器に適用した研究を進めてきた[高藤 2008]. 鵜木らは、「聞き耳」型補聴システムの研究開発が実施し[鵜木 2013]、中藤らも、高齢者の聴覚機能の低下に向けた聴覚支援システムに関する研究を進めている[中藤 2014]。

海外でも、補聴器への応用として、アレイ処理や多チャンネル Wiener フィルタなどの信号処理を導入した研究が多い（[Desloge 1997],[Bogaert 2008], [Cornelis 2012]など）。しかし、その殆どは利用者が装着した補聴器のバイノーラル処理を施したものであり、本研究のように環境センサを利用したものはあまり存在しない。

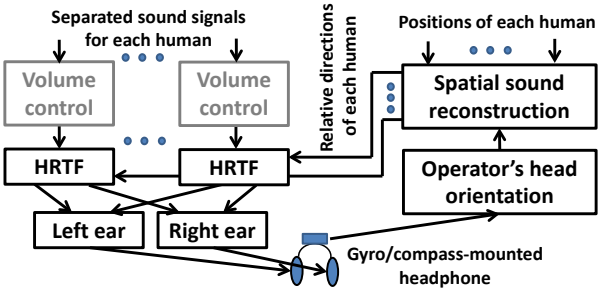
3 提案する聴覚支援システム

図２に提案システムのブロック図を示す。提案システムは二つの部分から構成される。一つは環境センサネットワーク側の音源位置推定・トラッキングと複数人の音源分離であり，もう一つは利用者側の頭部回転トラッキングと空間的感覚の合成である。

本システムの構成は、著者らが先行研究[Liu 2015]で提案した遠隔操作ロボットシステムにおいて音響臨場感を操作者に伝達する手法と類似している。その違いとして、遠隔操作システムでは操作者は遠隔地にいるが、本研究で提案する聴覚支援システムの場合は、利用者は環境センサと同じ場にいる。また、先行研究で報告したシステムに対し、本研究では主に音源分離のリアルタイム実装およびアルゴリズムの改善を進めた。



(a) 環境センサ側の処理



(b) 利用者側の処理

図２．提案する聴覚支援システムの概要

環境センサネットワーク側の処理では，まず，各マイクロホンアレイによって音の３次元到来方向（DOA）を推定する。環境とアレイの位置関係と各音源の DOA を統合することにより，３次元上での人位置（厳密には口元の位置）情報が得られる。人位置情報は，ヒューマントラッキングシステムにより，非発声時にも常時追跡されている。次に，推定した人位置情報に基づいて各人の音声を分離し，位置情報と合わせて利用者側のシステムに送信する。

利用者側の処理では，まず，人位置情報と利用者の顔の向きによって，左右のチャンネルに対応した最適な頭部伝達関数（HRTF: Head-Related Transfer Functions [Cheng 2001]）をデータベースから選択する。次に，分離した音声に畳み込み演算を行い，ステレオヘッドフォンに再生する。利用者の頭部回転トラッキングには，ヘッドホンの上部に取り付けたジャイロセンサーとコンパスを用いた。また，分離した各音源のボリュームは，独立して調節することができるユーザインタフェースを開発した。

3.1 ３次元音源定位

音源定位に関して，まず，各マイクロホンアレイで DOA 推定を行う。複数のアレイによる DOA 情報と人位置情報を統合することで，音源の３次元空間内の位置を推定する。

実環境での音の DOA 推定は広く研究されてきた。MUSIC 法は，複数のソースを高い分解能で定位できる最も有効な手法の一つである。この手法を使うには事前に音源数が必要であるため，本研究では[Ishi 2009]で提案した解決法を用いる。音源数を固定した数値に仮定し，閾値を超えた MUSIC スペクトルのピークを音源として認識する。この研究で使用した MUSIC 法の実装は 100 ms ごとに 1 度の分解能を有しており，2 GHz のシングルコア CPU でリアルタイムに探索することができる。

聴覚支援システムにおいて，利用者にとって最も重要な音源は人の音声である。本研究では人の声を抽出するために，複数の 2D-LRF（Laser Range Finder）で構成したヒューマントラッキングシステムを使用した[Glas 2007]。複数のマイクロホンアレイからの DOA 推定出力と LRF のトラッキング結果が同じ位置で交差すれば，そこに音源がある可能性が高い[Ishi 2013]。本システムでは 2D の LRF を用いているため、人位置情報は 2D に限られる。ここでは、検出された音源の位置が口元の高さの範囲内にあるかの制限もかけている（z = 1 ~ 1.6m）[石井 2014] [Ishi 2015]。無音区間や音源方向推定が不十分な区間では、最後に推定された口元の高さと最新の 2D 位置情報を用いて，音源分離を行う。

3.2 音源分離

音源分離では，選択された複数の人物を並列に分離する。図３に処理の流れを示す。

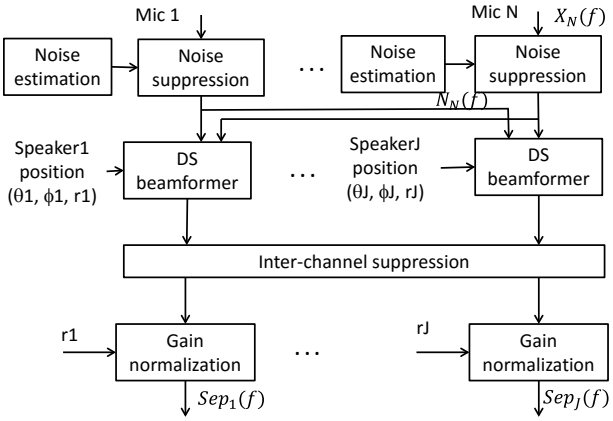


図 3. 音源分離の処理の流れ

まず、分離の第 1 ステップとして、エアコンなどの定常雑音抑圧（noise suppression）をチャンネル毎に行う。定常雑音抑圧手法として式(1)に示すように Wiener filter を用いる。

$$H_{WFi}(f) = \frac{1}{1+\frac{N_i(f)}{X_i(f)}} \quad (1)$$

定常雑音（$N_i(f)$）は、対象となる人の声が存在しない区間での平均スペクトルとして推定する。

定常雑音抑圧処理は、ポストフィルタとして、ビームフォーマを施した後に行うことも可能であるが、ここでは、musical ノイズの発生を抑えるため、ビームフォーマの前に施す。

次に、音源定位部から得られる方向（方位角、仰角）と距離情報を基に、ビームフォーマを施す。ここでは計算量が少なく且つロバストな DS ビームフォーマ（Delay-Sum Beamformer）を用いて、対象となる人の声を強調する。フレーム長は 32 ms で，シフト長は 10 ms である。

本研究で使用した 16 チャンネルのマイクロホンアレイ（半球 30cm にマイクを配置した形状）の DS ビームフォーマのレスポンスの特徴として，低周波領域の分解能が低いことが挙げられる。そのため，無指向性雑音の低周波成分が分離音に多く混在してしまい、臨場感の伝達に悪影響を与える可能性がある。

空間に指向性音源 S と無指向性雑音源 N が存在すると仮定した場合，DS ビームフォーマの出力は以下の形になる：

$$Y_{DS}(f) = \boldsymbol{w}_{Sdir}(f) \cdot S(f) + \int_0^{2\pi} (\boldsymbol{w}_{\theta}(f) \cdot N(f)) d\theta \quad (2)$$

$Y_{DS}(f)$は周波数 f に対応したビームフォーマの出力で，S_{dir} は信号の方向，$\boldsymbol{w}_{Sdir}$ は S_{dir} 方向のビームフォーマレスポンスを指す。式の二つ目の項目は，分離音声に混在する雑音を表している。この雑音成分を低減させるために，各周波数に以下のようなウェイトを掛けた。

$$w_{norm}(f) = \frac{1}{\int_0^{2\pi} \boldsymbol{w}_\theta(f)\, d\theta} \quad (3)$$

$$Y_i = \sum_f w_{norm}(f) \cdot Y_{DS}(f) \quad (4)$$

Y_i はウェイト掛けした後のビームフォーマ出力である。

また、DS ビームフォーマのみでは、十分な音源分離が出来ず、チャンネル間の信号（妨害音）の漏れを抑えるための処理（inter-channel suppression）を行う。妨害音抑圧処理には、式(5)に示すように Wiener filtering を用いる。

$$H_{WFi}(f) = \frac{1}{1+\frac{I_i(f)}{Y_i(f)}} \quad (5)$$

$$I_i(f) = \max_{j \neq i}\{Y_j(f)\} \quad (6)$$

$I_i(f)$ は式(6)に示すように、分離された対象音以外の音源の中で、最も強い周波数成分を表す。上述の妨害音抑圧処理の一つの問題点として、同じ方向に対象音と妨害音が存在する場合、対象音に歪みが生じる可能性が高い。そこで、ここでは対象音と妨害音の差が 5 度以内であれば、抑圧処理を行わない制約を設けた。

$$I_i(f) = \frac{|dir_1 - dir_2|}{5} I_i(f), \text{ if } |dir_1 - dir_2| < 5 \quad (7)$$

最後に、音源とマイクロホンアレイの距離によって、観測される音圧が異なるため、距離による振幅の正規化（gain normalization）を施す。

$$g_j = \frac{1}{r_i} \quad (8)$$

3.3 音の空間的情報の再構築

環境センサ側から提供される分離音を受信し、利用者と対象音源の相対的位置関係を考慮して、音の空間的感覚を再構築する。処理としては、複数音源に対する音量調整と、頭部伝達関数（HRTF）を用いた音像の合成となる。

まず、音量調整に関しては、各音源とアレイの間の距離による違いを補正するため，分離された各音源に対して、それぞれの距離によって以下のように正規化を行う。

$$g_i = \frac{\sum_{n=1}^{N} dist_n - dist_i}{(N-1)\cdot\sum_n^N dist_n} \quad (9)$$

$$Y_i = g_i \cdot Y_{PF,i} \quad (10)$$

ここで，N は音源の数で，$dist_n$ は n 番目の音源とアレイの距離を表す。g_i は i 番目の音源に掛ける正規化ファクタで，Y_i は i 番目の音源の分離結果を示している。

音像の合成においては、一つの音源を特定の方向から聞こえるようにするため，その方向に対応した HRTF によってフィルタリングするステレオ化方法が一般的である。本研究では，一般公開されている KEMAR (Knowles Electronics Manikin for Acoustic Research) ダミーヘッドの HRTF データベースを利用した[Gardner 1995]。KEMAR は HRTF 研究のために一般的な頭部サイズを使って作られたダミーヘッドで，データベースには空間からのインパルス信号に対するダミーヘッドの左右耳のレスポンスとして，仰角-40 度から 90 度までの総計 710 方向のインパルス応答が含まれている。各インパルス応答の長さは 512 サンプルで，サンプリング周波数は 44.1 kHz である。

前述のように, HRTF を用いて動的に音像を合成するには，頭部の向きのリアルタイム検出が必要である。このため，本研究ではヘッドホンの上部にジャイロセンサーとコンパスを取り付け，頭部回転のトラッキングを行った。角度情報はシリアルおよびブルートゥース経由のいずれかでシステムに送られる。音場の合成に使う方向は音源方向から頭部角度を引いたもので，この方向に対応した左右チャンネルのインパルス応答がデータベースから選出され，分離音と畳み込み演算を行った音声が利用者の両耳に再生される。

4 予備的評価

現段階では、開発したシステムの定性的な評価に留まっている。まず、研究室内での予備的評価により、wiener filter のパラメータは、$\alpha = 1$, $\beta = 0.001$ とした。式(8)の振幅の正規化に関しては、距離が大きくなり過ぎると、背景雑音も増幅されてしまうため、距離による正規化は 2m までと制限した。

著者らの研究所のオープンハウス（2015 年 10 月）で開発したシステムのデモを行った。デモシステムとして、LRF 2 個で人位置推定を行い、ポスター前のテーブル上にマイクアレイ 1 個を設置して、訪問者にヘッドホンをかけてもらい、ポスターの周りにいる人のうち、強調したい人をマウスの左クリックで選択し、抑圧したい人を右マウスで選択する機能を設けたインタフェースを開発した。取捨選択型の機能を体験していただいた方々には、高評価の感想をいただいた。一つの大きな課題として、処理後の音声が再生される遅延が大きすぎることが挙げられる。現在は遅延が 300ms 程度で、対話相手が目の前で発話している状況では、口の動きや頷きなどのタイミングが音声とずれて見えるため、違和感があるという意見が多かった。この遅延は、処理時間に加え、再生用のバファリングも大きな原因となっているが、ハードウェアの開発により、短くすることは可能である。その他、訪問した一般の高齢者の方も数人体験していただき、使いたいので早く実用化していただけないかとの意見もいただいた。

分離音の音質においては、研究室で予備評価を行った際、図３に表示したすべての処理を用いるのが最も聞きやすかった。しかし、オープンハウス会場では、入力の noise suppression を用いない方が分離音の音質が良かった。研究室では空調音が最も強い背景雑音源であるが、ポスター会場の雑音はカクテルパーティ効果のようなバブル雑音が大きかったため、システムを起動した際に推定した背景雑音のレベルが大きく、定常雑音の wiener filter 処理を施すと強い歪みが生じてしまうことが原因と考えられる。定常雑音の推定については、今後改善する予定である。また、システム全体の詳細な評価についても今後進める予定である。

謝辞

本研究は、総務省 SCOPE の委託研究によるものである。

参考文献

[高藤 2008] 高藤、森、猿渡、鹿野 (2008). SIMO モデルに基づく ICA と頭部伝達関数の影響を受けないバイナリマスク処理を組み合わせた両耳聴覚補助システム、電子情報通信学会技術研究報告. EA, 応用音響 108(143), 25-30, 2008.

[鵜木 2013] 鵜木祐史. 「聞き耳」 型補聴システムの研究開発. 「戦略的情報通信研究開発推進事業 SCOPE ）」 平成 25 年度新規採択課題 http://www.soumu.go.jp/main_content/000242634.pdf

[中藤 2014] 高齢者の聴覚機能の低下に向けた聴覚支援システムに関する研究、文部科学省科学研究費基盤研究(C) 、2014 年 04 月 ～ 2017 年 03 月

[Desloge 1997] J.G. Desloge, W.M. Rabinowitz, and P.M. Zurek, Microphone-Array Hearing Aids with Binaural Output- Part I: Fixed-Processing Systems, IEEE Trans. Speech Audio Processing, vol. 5, no. 6, pp. 529□542, Nov. 1997.

[Bogaert 2008] Bogaert, T.V., Doclo, S., Wouters, J., Moonen, M. The effect of multimicrophone noise reduction systems on sound source localization by users of binaural hearing aids, J. Acoust. Soc. Am. 124 (1), 484-497, July 2008

[Cornelis 2012] Cornelis B., Moonen, M., Wouters, J. Speech intelligibility improvements with hearing aids using bilateral and binaural adaptive multichannel Wiener filtering based noise reduction. J Acoust Soc Am. 2012 Jun;131(6):4743-4755.

[Liu 2015] Liu, C., Ishi, C., Ishiguro, H., Bringing the Scene Back to the Tele-operator: Auditory Scene Manipulation for Tele-presence Systems, Proc. ACM/IEEE International Conference on Human Robot Interaction (HRI 2015), USA. 279-286, March, 2015.

[Cheng 2001] Cheng, C. I., Wakefield, G. H. Introduction to head-related transfer functions (hrtfs): Representations of hrtfs in time, frequency, and space. *J. Acoust. Soc. Am*, 49(4):231-249, April 2001.

[Ishi 2009] Ishi, C. T., Chatot, O., Ishiguro, H., Hagita, N. Evaluation of a MUSIC-based real-time sound localization of multiple sound sources in real noisy environments. Proceedings of the *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 09)*. 2027‑2032. 2009.

[Glas 2007] Glas, D.F. et al, 2007. Laser tracking of human body motion using adaptive shape modeling. In Proceedings of the *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2007)*, 602-608. 2007.

[Ishi 2013] Ishi, C., Even, J., Hagita, N. (2013). Using multiple microphone arrays and reflections for 3D localization of sound sources. In Proc. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2013)*, 3937-3942, Nov., 2013.

[石井 2014] 石井カルロス寿憲, Jani EVEN, 萩田紀博, (2014) "複数のマイクロホンアレイと人位置情報を組み合わせた音声アクティビティの記録システムの改善", 第３２回日本ロボット学会学術講演会, Sep. 2014.

[Ishi 2015] Ishi, C., Even, J., Hagita, N. (2015). “Speech activity detection and face orientation estimation using multiple microphone arrays and human position information,” *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2015)*, pp. 5574-5579, Sep., 2015.

[Gardner 1995] Gardner, W. G., Martin, K. D. HRTF measurements of a KEMAR. *J. Acoust. Soc. Am*. 97(6):3907-3908, Jun. 1995.

Coarse-to-Fine チューニングを用いた HARK の音源定位パラメータの最適化

杉山　治[1], **小島 諒介**[1], **中臺 一博**[1,2]

Osamu SUGIYAMA[1], **Ryosuke KOJIMA**[1], **Kazuhiro NAKADAI**[1,2]

1. 東京工業大学 大学院 情報理工学研究科,

2.（株）ホンダ・リサーチ・インスティチュート・ジャパン

1. Graduate School of Information Science and Engineering, Tokyo Institute of Technology,

2. Honda Research Institute Japan Co., Ltd.

{sugiyama.o, kojima, nakadai}@cyb.mei.titech.ac.jp

Abstract

本稿ではオープンソースロボット聴覚ソフトウェア HARK (Honda Research Institute Japan Audition for Robots with Kyoto University) の音源定位におけるパラメータ最適化のためのインタフェースを提案する．HARK でパラメータ調整用のインタフェースは存在するものの，HARK に熟練していてもそのパラメータの最適化には時間を要する．本稿で提案するインタフェースは，HARK のパラメータ最適化における課題を，可視化，操作，最適化における課題に分類し，それぞれを解決する機能を設計・実装した．そして，ユーザ評価において，可視化性・設定の柔軟さの点で，従来のインタフェースを上回るという結果を得た．

1　はじめに

本稿では，オープンソースロボット聴覚ソフトウェア HARK (Honda Research Institute Japan Audition for Robots with Kyoto University) のパラメータ最適化を効率的におこなうことができるよう，HARK の音源定位機能に焦点をあて，インタラクティブなインタフェースを提案する．

HARK は，2008 年にロボット音響における OpenCV を目指しリリースされたオープンソースソフトウェアである [Nakadai 10]．複数のマイクロホンからなるマイクロホンアレイを用いた処理に対応し，音源定位 [Nakamura 09, Ohata 13]，音源分離 [Nakajima 08]，音声認識といった機能を，HARKDesigner と呼ばれるグラフィカルユーザインタフェースを用いて組み合わせることで柔軟なロボット聴覚ソフトウェアを作成することができる．HARK を用いることで，例えば４人のユーザが同時に発話するような状況においても，個々の発話を音声認識するロボットアプリケーションを容易に作成することが可能になる．

HARK の最新版であるバージョン 2.2 でもパラメータを調整するためのインタフェースは存在するが，熟練した作業者がパラメータの最適化を行った場合でも数日を要することもあり，ソフトウェアが安定して使えるようになるまでのオーバヘッドが高い．本研究では，これらの音源定位のパラメータ最適化における課題を，可視化・操作・最適化の３つの観点から整理し，それぞれの課題を解決するためのインタラクティブなインタフェースを設計・開発する．提案するインタフェースでは，音源定位の処理過程を可視化し，マウスジェスチャによる直感的に変数の変更を可能にした．さらに，システムが変数の最適値の予想を示し（Coarse チューニング），それを元にユーザがより正確に変数を最適化する（Fine チューニング）手順を踏む Coarse-to-Fine チューニング [Fujii 11] を取り入れた．これらのインタフェースの機能を利用することで，ユーザは従来のインタフェースより直感的に音源定位のパラメータを設定・最適化することができる．また，ユーザによる定性評価を実施し，提案インタフェースの有効性を検証した．

2　課題とアプローチ

図 1 に HARK における音源定位のプロセスを示す．まず，マイクアレイから多チャンネル音声信号を取得し，短時間フーリエ変換 (Short-Time Fourier Transform, STFT) にかけて周波数スペクトラムへと変換する．その後，MUltiple SIgnal Classification (MUSIC) 法 [Schmidt 86] を用いることで，横軸が時間・縦軸が方位角，色がパワーを示す MUSIC スペクトログラムを得る．最後に，音源追跡により，MUSIC スペクトログラムから音源の位置情報を抽出する．この過程で，ユーザは，以下のパラメータを設定する必要がある．

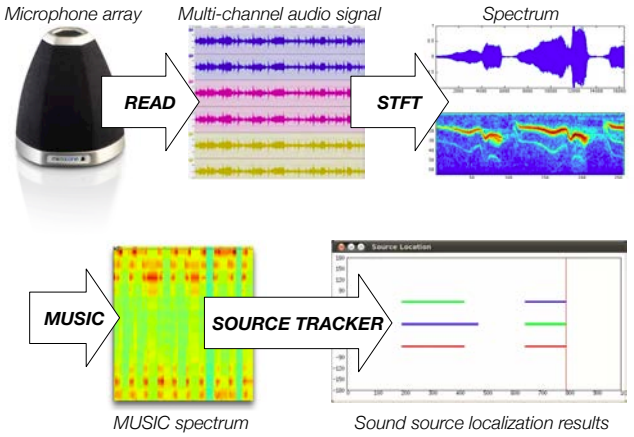


図 1: 音源定位のプロセス

- **num sources**: 音源数
- **thresh:** 音源と雑音を分けるパワーの閾値
- **pause length**: 音源の前区間長
- **preroll length**: 音源の後区間長

これらのパラメータを個々に最適化するには時間がかかり，実験環境における HARK の即時セットアップの障害となっている．本研究では，この問題を以下の 3 つの課題に分類し，それぞれを解決するインタラクティブなインタフェースを設計・開発する．

- **可視化の課題**: 音源定位の途中のプロセスを可視化できていないため，経過を見ながらパラメータの調整ができない
- **操作の課題**: 閾値などのパラメータを直接数値で調整することは非直感的であり，またその結果が即時に反映されない
- **最適化の課題**: システムによる最適化支援機能がない．ユーザは一からパラメータを調整しなければならない

以降の節では，提案するインタフェースがこれらの課題をどのように解決するのかを詳細に述べる．

3 音源定位のためのインタラクティブインタフェースの提案

本稿で提案するインタフェースは，先に述べた可視化・操作・最適化における 3 つの課題の解決を図り，HARK における音源定位のパラメータ調整の時間を短縮することを目的とする．一般に，HARK を用いて音源定位のパラメータを調整する時，ユーザは，1) 適当なパラメータセットを選択し，それを用いて音源定位を行い，定位結果と MUSIC スペクトログラムを得る．2) 得られた音源定位結果と MUSIC スペクトログラムを比較し，雑音部と音源部を推測する．3) 推測した通りにそれらの雑音部と音源部を分けるパラメータセットを導き出す．

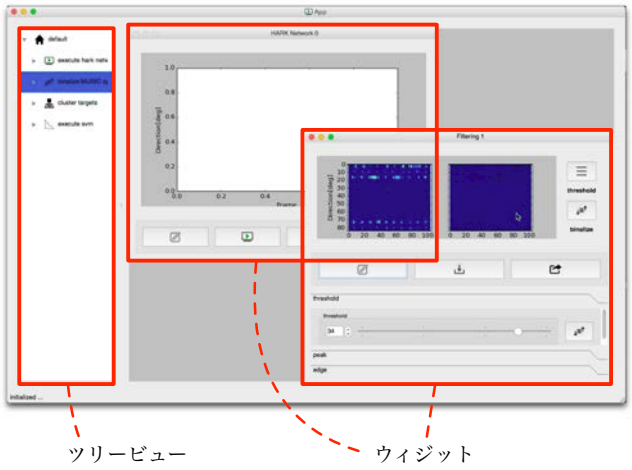


図 2: 提案するインタフェースの概観

3.1 音源定位プロセスの可視化

本稿で提案するインタフェースの概要を図 2 に示す．提案するインタフェースは上記の 3 プロセスを実行する以下の 3 つのウィジットを持つ．

a) 音源定位実行ウィジット

b) 音源のラベル付けウィジット

c) 動的閾値最適化ウィジット

音源定位のパラメータ調整に必要な処理を複数のウィジットに分けることで，ユーザはそれらのプロセスを同時に確認しながら，多面的にパラメータの調整をすることができる．

図 3 に 3 つのウィジットの概観を示す．それぞれのウィジットは共通してチャートボックスとコントロールボックスを持ち，チャートボックスでは，各音源定位過程の可視化を，コントロールボックスでは各定位過程の実行とパラメータ調整を行う．

音源定位の実行　音源定位の実行は音源定位実行ウィジットで行う（図 3a)）．このウィジットでは，HARK による音源定位を実行することができ，コントロールボックスで，解析する多チャネル音声ファイル，チャネル数，伝達関数，音源数，音源時間長をパラメータとして指定することができる．チャートボックスは，音源定位のプロセスの途中で得られる MUSIC スペクトログラムが表示され，得られた定位結果と MUSIC スペクトログラムを音源のラベル付けウィジットに出力することができる．

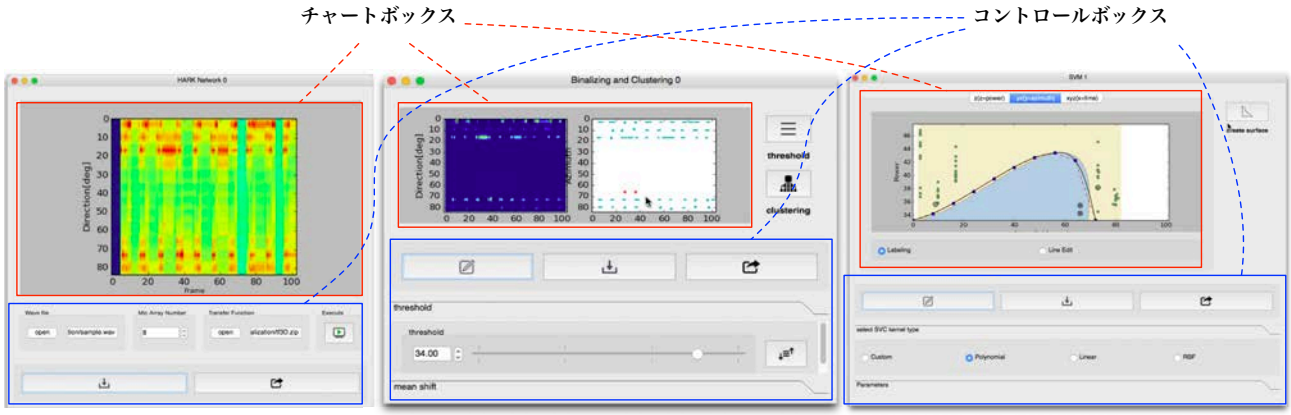


図 3: 音源定位のパラメータ調整のためのウィジット

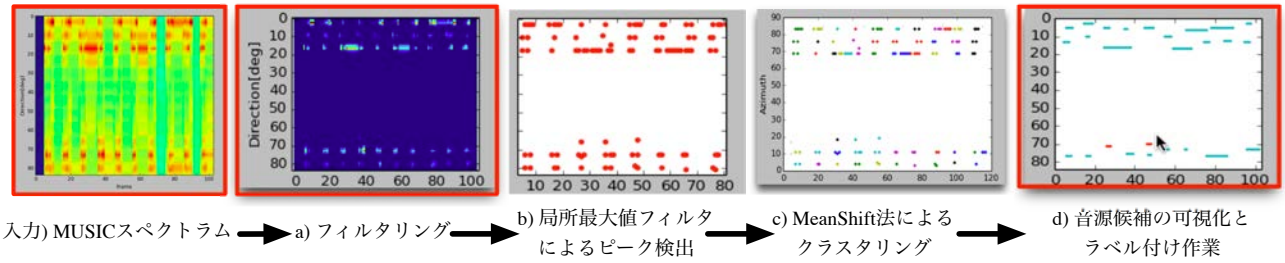


図 4: ラベル付けウィジットのバックグラウンド処理

音源のラベリング　音源ラベル付けウィジットでは，MUSIC スペクトログラムの表示に対して直接，音源のラベル付けを行うことができる（図 3b)）．チャートボックスには 2 つのチャートが表示され，一方には MUSIC スペクトログラムが，もう一方には音源候補が図示される．ユーザは最初のチャートを用いて雑音部を除去することで音源部を，次のチャートで音源の候補を確認，その候補が音源なのか，雑音なのかをラベリングする．これらの操作を実行するため，音源ラベル付けウィジットはバックグラウンドで以下の処理を実行する．

a) 閾値によるフィルタリング

b) 局所最大値フィルタによるピーク検出

c) 検出されたピークをクラスタリングすることによる音源候補の抽出

d) 音源候補の可視化とラベリング

図 4 は，上記のバックグラウンドプロセスの過程を図示したものである．図 4 中，赤い枠線を持つものは処理結果が可視化される処理を表し，それ以外のものはチャート上には図示されずバックグラウンドで処理される．

閾値によるフィルタリングでは，1) ピーク検出にむけて MUSIC スペクトログラムの低パワー部を除去する．2) 局所最大値フィルタ [Nishiguchi 04] によってピーク検出を行う．3) この処理によって得られたピーク群を，Mean-Shift 法 [Okada 08] を用いてクラスタリングし，得られたクラスタを音源候補とする．4) それぞれのクラスタをチャートボックスの右のチャートにレンダリングする．この際，クラスタを構成するピーク時間軸の最大値と最小値の差をそのクラスタ長とする．またこの間の方向軸の平均が縦軸の値としてプロットされる．

表 1: 最適化処理に用いるパラメータ

ウィジット	アルゴリズム	変数名	型	初期値
ラベリング	フィルタリング	power	float	32.0
ラベリング	局所最大値フィルタ	x	int	1
		y	int	2
ラベリング	Mean Shift	kernel_size	float	0.02

これらの過程で必要な局所最大値フィルタのフィルタサイズや Mean-Shift 法のカーネルサイズなどの各パラメータはコントロールボックスのスライダーで調整することができる．また，その値を数値としても確認することができる．各パラメータの初期値を表 1 にまとめる．

動的閾値の最適化　動的閾値最適化ウィジットでは，音源と雑音を分けるパワー閾値を動的に設定することができる（図 3c)）．閾値を複数の視点から設定できるようにす

るためチャートボックスはマルチタブ構成になっており，それぞれのタブでは以下に示す複数の次元で音源候補をプロットする．

1D 縦軸を各音源候補のパワーの平均とし，それぞれの音源候補のパワーを降順に並べたもの

2D 縦軸を各音源候補の各方向軸ごとのパワーの平均とし，横軸を方向として音源候補をプロットしたもの

3D 縦軸を各音源候補のパワーとし，横軸を時間フレーム，奥行きを方向として音源候補をプロットしたもの

音源候補は，ラベル付けウィジットで事前にラベル付けされており，音源とラベル付けされたものは青く，雑音とラベル付けされたものは赤くプロットされる．ユーザはこれらの音源と雑音を切り分ける境界を，サポートベクタマシン (Support Vector Machine, SVM) によってラフに求め (Coarse チューニング)，マウスジェスチャによって閾値を詳細に設定することができる（Fine チューニング）．これらの挙動については，3.3 節で詳しく述べる．一方，コントロールボックスでは，SVM のカーネルの選択，それぞれのパラメータを調整することができる．ユーザは，これらのインタフェースを用いることで直感的に音源と雑音を分ける閾値を設定し，音源定位に反映することができる．図 6 は多項式カーネルを用いた場合の閾値の設定例である．

3.2 ジェスチャ操作によるインタラクティブなインタフェース

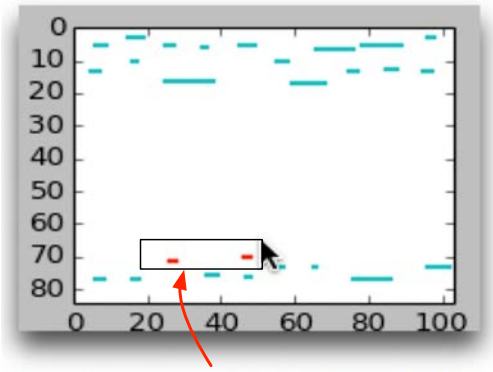
提案インタフェースのジェスチャ操作について述べる．既存の HARK のインタフェースでは，音源定位のパラメータを数値で指定するため，その値がどのように結果に反映されるのかわかりにくいという課題があった．本稿では，この課題を解決するために 2 つの機能をインタフェースに実装した．

3.2.1 マウスジェスチャによる音源候補の選択

図 3b), c) のチャートボックスでは，マウスジェスチャによる音源候補のラベリングをすることができる．ユーザはラベル付けしたい音源候補の周辺の矩型領域を，マウスのドラッグ&リリースジェスチャで指定することでラベル付けを行うことができる（図 5）．この情報は，動的閾値最適化ウィジットで，音源と雑音をわける閾値を設定するときに使われる．

3.2.2 パラメータ変更の即時反映

提案インタフェースのすべてのチャートボックスは，パラメータの変更やマウスジェスチャの結果が即時に反映される．ユーザは，自身のパラメータの変更がどのように音源定位結果の各プロセスに影響を与えるのかをチャートボックスから直感的に読み取ることができるため，それぞれの過程で反映される結果を見ながらパラメータの最適化作業をインタラクティブにすることが可能となる．



ドラッグ&リリースで囲った矩型領域の候補をラベリングする

図 5: ラベル付けのためのマウスジェスチャ

3.3 Coarse-to-Fine チューニング

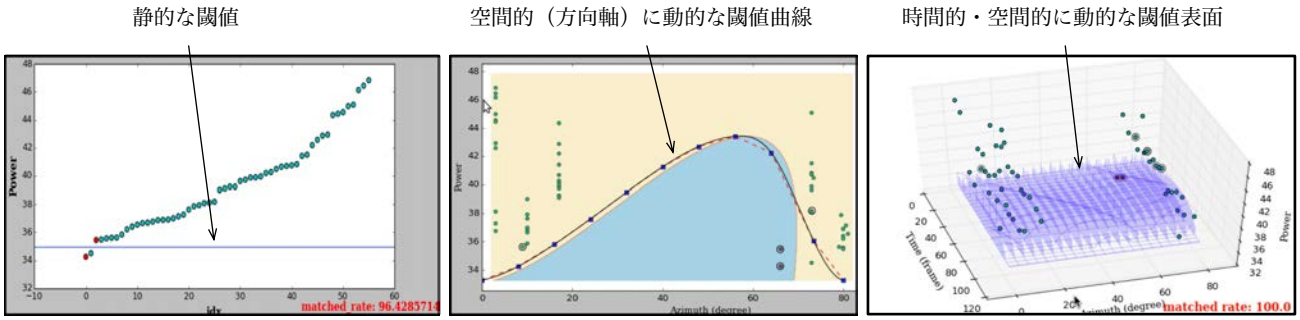
Coarse-to-Fine メカニズムとは，人間の視覚はまず全体を見てから，細部を詳細に見るという動きをするというメカニズムのことである [Menz 03]．このメカニズムは，画像処理における物体認識などに応用されており，本稿では，このメカニズムを組み込んだシステムと人の協調作業の方法を提案する．

環境や状況依存で最適な値が変わってしまうため，機械学習技術を用いても音源定位パラメータの完全な最適化を行うことは困難である．本稿では，機械学習のマシンループにユーザのアドバイスを加えることで短時間で詳細なパラメータチューニングを行うことを目指し，そのためのインタフェースを開発する．

Coarse-to-Fine チューニングの最適化対象は，前述の 3 つのパラメータのうち，音源と雑音を分離する際のパワーの閾値である．HARK の既存のインターフェースでは，この閾値は時間的，空間的に静的にしか設定できなかった．しかし，音源や方向性雑音のパワーに違いがある場合や，ある一定期間，高いパワーのノイズがのってしまった場合には，静的な閾値では対応できないことがある．本稿では，この閾値を空間・時間軸で動的に設定できるようにし，その最適化を Coarse-to-Fine チューニングで行う．

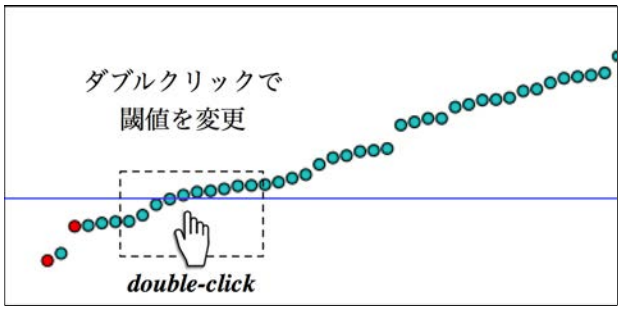
3.3.1 Coarse チューニング

Coarse チューニングでは，システムがラフにパラメータの最適値をユーザに提示する．具体的には，音源と雑音を分けるパワー閾値の動的な変化に対応できるように空間・時間方向に対する閾値曲線（面）として表す．この閾値曲線（面）は SVM を用いて推定する．動的閾値最適化



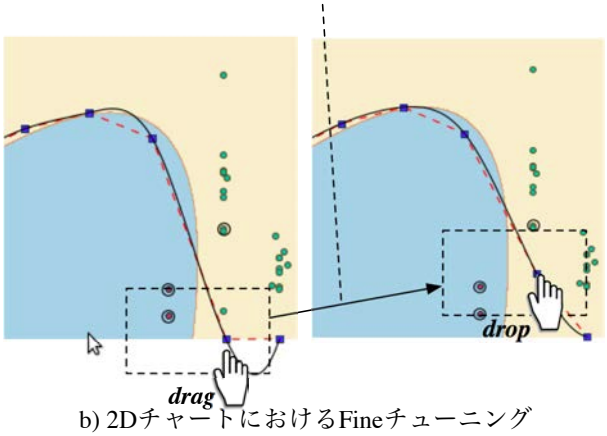
a) 1Dチャート　　b) 2Dチャート　　c) 3Dチャート

図 6: 動的閾値の調整画面



a) 1Dチャートにおける Fine チューニング

ドラッグ&ドロップで閾値曲線を構成するノードを移動



b) 2Dチャートにおける Fine チューニング

図 7: Coarse-to-Fine チューニング

ウィジットは，閾値を設定する 3 つの異なるチャート画面を持つ（図 6）．図 6a) のチャートでは，設定する閾値は静的で既存の HARK と変わらないが，音源フィルタリングウィンドウでラベル付けした音源候補を抽出する閾値を求め，ユーザに提示する．図 6b) のチャートでは空間（方向）軸に沿って，MUSIC スペクトログラム上のパワーの強い領域のピーク座標がプロットで表示される．同時に，ユーザがラベル付けした音源を定位するために最適な閾値の境界曲線を多項式カーネルを用いて求め，提示する．図 6c) のチャートでは，3 次元（時間，空間（方向），パワー軸）空間上に MUSIC スペクトログラムのパワーの強い領域のピーク座標がプロットされており，ユーザがラベル付けした音源を定位するために最適な閾値の境界面を提示する．ユーザはこれらの提示される 3 つの静的閾値，閾値の境界曲線，境界面の中からその状況に最もあったものを選択し，Fine チューニングを行う．

3.3.2 Fine チューニング

Fine チューニングでは，Coarse チューニングで提示されたパラメータの値に基づき，ユーザが詳細にパラメータの最適化を行う．図 7 は，動的閾値最適化ウィンドウにおけるマウスジェスチャ操作を示す．動的閾値最適化ウィジットでは，SVM に基づいてシステムが閾値候補を提示し，その後にユーザが最適値を調整する．その際，図 7a) の 1D チャートでは各音源候補のパワーの平均値が降順にプロットされており，音源を示す青いプロットと雑音を示す赤いプロットをうまく切り分けるように閾値を設定する．閾値の設定は画面をダブルクリックすることで行い，ダブルクリックされた y 軸の値を閾値として採用する．図 7b) の 2D チャートでは境界線をノードをマウスでドラッグ & リリースすることで閾値を自由に変更することができる．Coarse チューニングでシステムから提案された境界曲線は，境界曲線上のノード群とそれらを補完する spline 曲線として，ユーザに提示される．ユーザは提示されたノード群の位置をマウスのドラッグ & ドロップジェスチャで任意の位置に変更することができる．これらのマウスオペレーションは即座にシステムに伝達され，変更された結果が図に反映されるため，ユーザは反映結果を見ながらインタラクティブに閾値の調整を行うことができる．

4 システム評価

提案システムの有効性を評価するために，評価実験を行った．実験では，ロボット実験で収集した多チャンネル音声信号を提案インタフェースでパラメータ調整する様子と，HARK の既存インタフェースで調整する様子をビデオで撮影し，その様子を 8 名の大学院生に見せ，その印象を

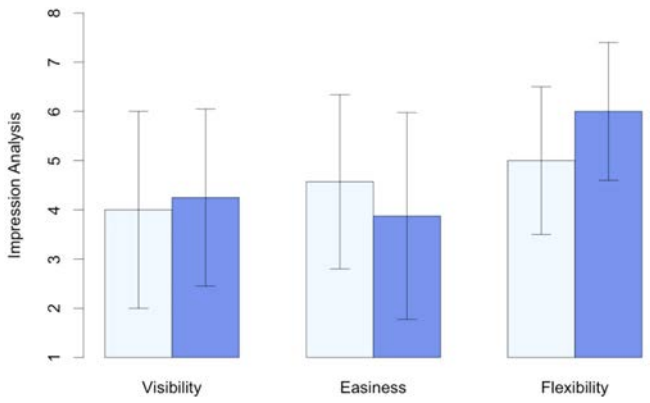


図 8: 定性評価

a) 可視化性，b) 操作性，c) 設定の柔軟性の観点から 7 段階で評価してもらった．なお，実験前に学生はそれぞれのインタフェースの使い方に関するレクチャを 10 分間受けており，その使い方，操作の意味を理解してもらった．

4.1 実験結果

実験結果を 図 8 に示す．図 8 からわかるように，提案インタフェースは，可視化性，設定の柔軟性の 2 つの観点で既存の HARK のインタフェースの評価を上回ることが示された．対して，操作性に関しては既存の HARK インタフェースが上回るという結果になった．

可視化性と設定の柔軟性で既存の HARK インタフェースより良い評価を得たことは本稿の提案するインタフェースが設計の意図通りにユーザの負荷を軽減できていることを示していると考えられる．一方，操作性に関しては，良い評価が得られなかった．実験アンケート後，被験者に実施したインタビューでは，複数の被験者から提案インタフェースは設定する項目が多く，便利だと思われる反面，いろいろと覚えるべきことが多いのではないかという指摘を受けた．これらの懸念が，設定項目が少なく操作できる既存の HARK インタフェースの評価が提案インタフェースより高くなった原因であると考えられる．本稿では，これらのユーザの評価から，それぞれのウィジットでショートカット機能を実装することでシステムによるユーザの補助機能を追加し，操作性においても既存インタフェースを上回る機能を実装する予定である．これらの設計・実装と評価は将来課題である．

5 結論

本稿では，HARK における音源定位のパラメータ最適化のため，インタラクティブなインタフェースを設計・開発した．提案インタフェースは，可視化・操作・最適化における定位パラメータ調整の課題を解決することで，直感的な最適化を行うことができる．そして，ビデオによる評価実験を通じて，可視化性と操作の柔軟性において既存の HARK インタフェースよりも高く評価されることを示した．

謝辞

科研費 24220006 および，JST ImPACT タフロボティクスチャレンジの支援を受けた．

参考文献

[Nakadai 10] K. Nakadai *et al.*: “Design and Implementation of Robot Audition System “HARK”,” Advanced Robotics, Vol.24, pp.739-761, VSP and RSJ, 2010.

[Nakamura 09] K. Nakamura *et al.*, “Intelligent sound source localization for dynamic environments,” IROS 2009, pp. 664-669.

[Ohata 13] 大畑 他, “クワドロコプタを用いた屋外環境音源探索，” SICE SI2013, pp. 360-363.

[Nakajima 08] H. Nakajima *et al.*, “Adaptive step-size parameter control for real-world blind source separation,” IEEE ICASSP 2008, pp. 149-152.

[Fujii 11] 藤井 他, “ロボット聴覚ソフトウェア HARK における音源定位パラメータチューニングの検討,” SICE SI-2011, pp. 202-205.

[Menz 03] M.D. Menz *et al.*, “Stereoscopic depth processing in the visual cortex: a coarse-to-fine mechanism,” Nature neuroscience Vol.6, No.1, pp. 59-65, 2003.

[Schmidt 86] R.O. Schmidt, “Multiple emitter location and signal parameter estimation,” IEEE Trans. on Antennas and Propagation, Vol.34, No.3, pp. 276-280, 1986.

[Carle 04] C. Carle, *et al.* “Code reusability tools for programming mobile robots,” IEEE/RSJ IROS 2004, pp.1820-1825.

[Nishiguchi 04] 西口 他, “スターセンサ画像の暗い星検出への繰り返し型最大値フィルタの応用,” 計測自動制御学会論文集, Vol.40, No.5, pp.573-581, 2004

[Okada 08] 岡田, “ミーンシフトの原理と応用,” 信学技報, Vol. 107, No. 539, PRMU2007-308, pp. 308-346, 2008.

身体的拘束に基づく音声駆動体幹動作生成システム

Speech Driven Trunk Motion Generating System Based on Physical Constraint

○境　くりま [*1,2], 港　隆史 [*1], 石井　カルロス寿憲 [*1], 石黒　浩 [*1,2]
Kurima SAKAI[*1,2], Takashi MINATO[*1], Carlos Toshinori ISHI[*1], Hiroshi ISHIGURO[*1,2]
ATR[*1], 大阪大学大学院 基礎工学研究科 [*2]
sakai.kurima@irl.sys.es.osaka-u.ac.jp, carlos@atr.jp, minato@atr.jp, ishiguro@sys.es.osaka-u.ac.jp

Abstract

近年，様々なヒューマノイドロボットが開発されてきており，人の代わりとなり社会的な役割を果たすことが期待されている．ヒューマノイドロボットが人間らしい動きをすることで，我々はロボットに対し親密感を覚える．特に，人々に受け入れられる対話ロボットを実現するためには，発話に伴う動作が必要となる．本論文では，人と対話するヒューマノイドロボットの頭部，腰部動作に着目し，ヒューマノイドロボットの発話に合わせて，人らしい頭部，腰部動作をリアルタイムで生成するシステムを構築する．

1　はじめに

近年通信技術やセンサ技術の発達によりロボットがより身近なものになってきた．特にヒューマノイドロボットは，遠隔操作することで場の共有感や身体動作といった非言語情報を伝達することができるため，電話やビデオチャット以上に遠隔地の人と直接対面しているような対話が実現できる [1]．特に，人手不足が深刻な高齢者介護の現場では，高齢者と遠隔地の人をつなぐことで役立っている [2]．また，自律ヒューマノイドロボットによるイベント会場の案内役 [3]，デパートでの販売員 [4]，病院での陪席者 [5]，や受付 [6] など社会的役割を人の代わりに果たそうという試みも行われている．以上のようにヒューマノイドロボットには，人の代わりとなり社会的な役割を果たすことが期待される．

ここで問題となるのは，人々に受け入れられるためのロボットの振る舞いのデザインである．人はエージェントの外見からその振る舞いを予測し，人間らしい見た目には人間らしい振る舞いを期待する傾向にある（適応ギャップ）[7]．特に，人間に外見が酷似したアンドロイド（図 1）に対して，それに応じた人間らしい動きを期待する．アンドロイドが期待される振る舞いを行わなければ，悪い印象を与えることとなる．実世界で動くアンドロイドでは，アクチュエータの自由度などのハードウェア的な制約があり，人間と同一の動きが実現できないため，人がどのような動きに人間らしさを感じるのか，その要素を明らかにして動きをデザインする必要がある．また，人間らしい動きは，外見にかかわらず人型エージェントに対する親密度を向上させることが報告されており [8]，人間らしい動きを感じさせる要因を明らかにすることは，人型エージェント全般において意義がある．

図 1: Android ERICA

従って，人々に受け入れられる対話ロボットを実現するためには，発話に伴ってどのような動作を表出すべきかが課題となる．対話ロボットにおいて，人らしさの要因として最も重要な点は，ロボット自身が発話しているという印象である．その印象を与えるための基本的な動作は，発声のための運動である．発声のための動き（口唇動作だけでなく，首，胸，腹の動き）が，発声と同期して表出されれば，ロボット自身が発話しているという印象を強める．人の発話と動きの関係をモデル化し，発話情報から動作を自動生成すれば，最も基本的な発話時の人らしい振る舞いとなる．本研究では，人と対話するヒューマノイドロボットの頭部，腰部動作に着目し，ヒューマノイドロボッ

トの発話に合わせて，人らしい頭部，腰部動作をリアルタイムで生成するシステムを構築する．

2　関連研究

コンピューターグラフィックスの研究分野では，エージェントの発話に合わせ頭部動作を自動生成する手法がいくつか提案されている．Le et.al. は発話音声のパワー，ピッチと頭部の 3 自由度の動きを Gaussian Mixture Model を用いてモデル化し，リアルタイムで頭部動作を生成するシステムを提案している [9]．また，隠れマルコフモデルを用いた同様のモデル化も行われている [10, 11, 12]．しかし機械学習を用いた自動生成システムでは，学習に使われているモーションデータが収録された状況に合った動作しか生成できない．特に，対話相手との関係性により話し方が変化するため，すべての状況での動作を収録することは困難である．また，これら手法は収録されたデータを復元することを目的にしているため，異なる状況で使用するための動きの変調や他の動きと複合することができない．エージェントの動作は対話状況に応じて複数の動作をミキシングすることが重要になり，様々なミキシングの手法が提案されている [13, 14, 15]．そのため，エージェントの発話する動作のみに着目したシステムが必要となる．

本論文では日本語の発話に合わせた動作生成を扱うのに対し，上記の研究は主に英語を母国語とする動作生成手法である．日本語に対する動作生成もいくつか提案されている．Watanabe et.al. は，発話の on/off 情報から頷きのタイミングを推定する手法を提案している [16]．しかし，頷き生成のタイミングを生成するだけで，どのような関節の動きが人間らしさを生むかまでわかっておらず，実際のアンドロイドで使用するには不十分である．Ishi et.al. は，発話の意味に対する動作のマッピング方法を提案している [17, 18]．発話の意味を推定するためには，韻律特徴のみならず言語特徴も利用する必要があるため [19]，リアルタイムシステムを構築することが困難である．

一方で，解剖学の知見から，口の開閉動作に伴い頭部が動くことも報告されている [20]．この知見から頭部の発話動作も社会的状況の要素以外の身体的拘束をもとに生成できる可能性がある．

本論文では，社会的状況に依存せず，純粋に発話のための動作を，人間の身体的拘束を利用し発話情報に基づいてリアルタイムで生成することを目的とする．また，機械学習で構築したモデルでは，発話と動作のどのような特徴が人間らしさに関わっているのか，解析するのは容易ではない．本研究では，動作の要因が直感的に分かりやすい動作生成モデルの構築を目指す．特に，目線をそらす動作は対話のコンテキストに依存し [21]，そのパターンは個性に依存する [22] ことから，本論文では発話に合わせた首と腰の縦方向の動きに着目する．

3　韻律と頭部動作の関係見つける実験

本節では人間らしい発話動作を自動生成するためのルールを見つけるための実験を説明する．人間が発声する際頭部動作などが音声に同期することが報告されており，特にパワーとピッチの変化と動作の変化が同期することが知られている [23]．しかし，日本語ではパワー，ピッチの韻律特徴と頭部動作の相関は高くないことも報告されている [24]．また，解剖学の知見から，口の開閉動作に伴い頭部が動くことも報告されている [20]．そのため，従来の音声のパワー・ピッチに加え，口の開き度合の 3 要素が社会的なインタラクションを含まない状況でも動きと相関があるのかを明らかにする．

3.1　実験設定

口の開閉が母音を発音する際に大きく変化するため，実験参加者に「あ・い・う・え・お」を 3 秒間発声してもらい，その発声に伴う首の動きの変化を計測する．母音の発声はそれぞれを高音・中音・低音で発音する条件 (Voice Pitch Condition) と，発声しやすい声の高さで大声で発声する条件 (Mouth Openness Condition) を設けた．被験者には，各発声ごとに正面を一旦向くよう指示を出し，姿勢をリセットした．予備実験より，被験者は母音を発声する際に 2 要因（高音で大きな声など）を混同させると発声しづらかったため，本実験では，2 要因を分けて頭部動作の変化を計測した．また，小さな声で発声すると頭部が動かないことも予備実験にて確認されていたため，Mouth Openness Condition では，大きな声のみ発声させた．

頭部動作は被験者の頭頂に取り付けた Inertial Measurement Unit(IMU) で計測した．被験者には口の形をはっきり作るように教示することで，母音に対する口の開き具合を統制した．

3.2　実験手順

各条件ごとに被験者には 2 回試行させた．1 回目は実験室での発声に馴化するために行った．また，身体動作を正しく計測できているかの確認も行った．すべての発声後に，発音する際に意識した姿勢がどのようなものかアンケートにて調査した．

3.3　実験結果

実験被験者は 11 人（男：6 人，女：5 人，平均年齢 22.0，標準分散 0.54）であった．そのうち男性被験者 1 人が正しく声の高さを発声できていなかったため解析から除いた．

Voice Pitch Condition の計測結果を図 2 に示す．縦軸は発声定常状態での首の角度を示す．高音，中音，低音を発声する際の首の角度を分散分析にかけたところ，有意差が認められた ($F(2, 18) = 12.843, p < 0.01$)．さらに，多重比較したところ，高音を発声する際に首の角度が最も上がり ($p < 0.05$)，低音を発声する際に最も下がることが

明らかとなった ($p < 0.05$)．すなわち，高音を発声する際は頭部をそらし，低音を発声する際は頭部を下げる傾向が認められた．

Mouth Openness Condition の計測結果を図 3 に示す．縦軸は発声に伴う首の角度の変化量を示す．この変化量は，発話開始前と発声定常状態での首の角度の差の絶対値で定義した．口を開いて発声する「あ」「え」「お」群と口を閉じて発声する「い」「う」群に分け，発声に伴う首の角度の変化量の大きさを比較したところ，口の開きを伴う発声条件のほうが有意に首を大きく動かすことが認められた (ウィルコクソンの順位和検定, $p < 0.05$)．

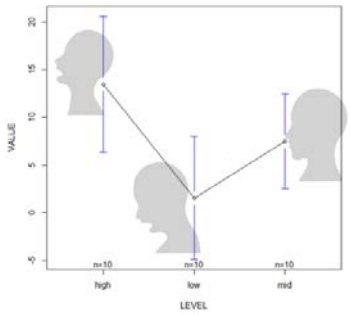


図 2: Head position according to pitch

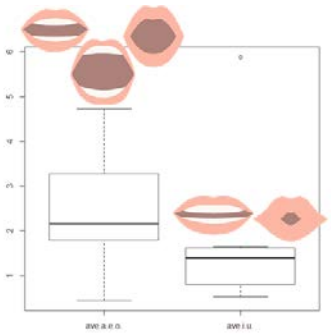


図 3: Head displacement according to mouth openness

以下にアンケートによる発声しやすい姿勢についての自由記述結果を示す．この記述からも，高音を発声する際は頭部をそらし，低音を発声する際は頭部を下げる傾向が認められた．

- 声の高低を意識して使い分けることが難しく感じ、高く出そうと思えば背筋が伸び顎が上がりました。低く出そうと思えば、背筋を少しだけ丸め顎を引き、なるべく口の中に篭るように発声しました。
- 高い音を出す際は上を向き、低い音を出す際には下を向く
- 口を大きく開ける あといは上から声を出し、うえおは下からあげるイメージで声を出す 体の中心に力を集めるイメージ
- 高い音は背筋が伸びる感じでした。低い音になるほど下を向いていたと思います。
- 高い音を出すときは顔を上向きに、逆に低い音を出すときは下向きにすると出しやすかった

4　身体的拘束に基づく発話動作生成システム

以上の知見をもとに音声特徴から頭部動作を生成するアルゴリズムを以下に説明する．人間らしい動作には滑らかな関節制御が重要である [25, 8]．そのため，音声特徴という間欠的な情報から連続的に滑らかな動作を生成する必要がある．また，二次遅れ系のダイナミクスに基づいて生成される動作が人間らしさ印象を与えることが報告されている [26]．そこで，本論文ではばねダンパ系を用いた運動モデルを利用することで，音声特徴という間欠的な情報から常時滑らかな動作を生成する（図 4, 式 1）．また，筋肉のモデル化をばねダンパ系を用いた運動モデルを用いた試みもあるため [27, 28]，この動作生成モデルの動作パラメータは筋肉の硬さに比例したパラメータとなっている．筋肉の硬さは発話の緊張度合・感情状態によって変化すると考えられ，発話時の感情や緊張度合といった人間が理解できるパラメータから動作パターンを調節することが期待される．

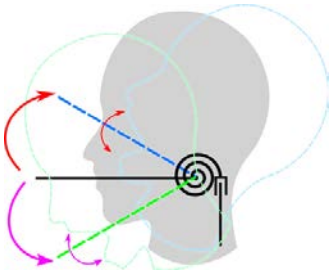


図 4: Classification of generating motion

$$J\ddot{\theta}_{base} + D\dot{\theta}_{base} + K\theta_{base} = T(t)Dir(t) \qquad (1)$$

4.1　ばねダンパ系による頭部動作生成

式 1 に対する外力を音声特徴をもとに定義することで，音声から頭部動作を自動生成する（式 2）．節 3 の実験結果から，口を大きく開けると首も大きく動くことから，式 4 のように，口の開く大きさによる外力を定義する．口の開きが大きくまたは均一である場合は，外力は口の開きの大きさに比例するようにする．口の開きが小さくなる場合は，首に与える外力をなくすことで運動モデルのばねの力により基準位置へ滑らかに戻る．口の開きが小さく

なる場合も口の開き度合をそのまま外力として与えてしまうと首の戻りが遅くなりリアルタイムで動作生成することが困難となる．また，予備実験から大きな声を出さないと首が顕著に動かなかったことから，声の大きさに比例した外力を式 3 のように外力を定義する．口の開き度合同様に，声のパワーが増えるまたは均一である場合は，外力は声の大きさに比例するようにする．声が小さくなる場合は，首に与える外力をなくすことで運動モデルのばねの力により基準位置へ滑らかに戻る．声が小さくなる場合も声のパワーをそのまま外力として与えてしまうと首の戻りが遅くなりリアルタイムで動作生成することが困難となる．V と L は声の大きさと口の開き度合という異なるスケールの外力を合わせるための定数である．

$$T(t) = VP(t) + LH(t) \tag{2}$$

$$P(t) = \begin{cases} Power(t) & (Power(t) >= Power(t-1)) \\ 0 & (otherwise) \end{cases} \tag{3}$$

$$H(t) = \begin{cases} LipHeight(t) & (LipHeight(t) >= LipHeight(t-1)) \\ 0 & (otherwise) \end{cases} \tag{4}$$

節 3 の実験結果から，首の動く方向は声の高さで決定されるため，式 1 の外力の運動モデルに対する方向を式 5 のように定義した．式 5 は，高音域を発声する場合は頭部をそらし，低音域を発声する場合は頷く方向に首を動かし，中音域では首を動かさないことを表す．

$$Dir(t) = \begin{cases} 1(Headup) & (HighTone) \\ -1(Headdown) & (LowTone) \\ 0(Nomovement) & (MiddleTone) \end{cases} \tag{5}$$

また，口の開閉度合は Ishi et.al. のフォルマント抽出に基づく口唇動作推定の手法を用いる [29].

4.2 韻律情報の抽出

F0 の値の抽出には，32 ms のフレーム幅で 10 ms 毎に LPC(Lear Predictive Coding) 逆フィルタによる残差波形の自己相関関数の最大ピークに基づいた処理を行う．さらに，人間のイントネーションの知覚特性と一致するよう，F0 の値を対数スケールに変換した．

$$F0[\text{semitone}] = 12 \times \log_2(F0[\text{Hz}]) \tag{6}$$

次に，音節内で F0 の変化量を表す $\Delta F0$(人間の音調の知覚に基づくパラメータ [30]) を抽出した．$F0move$ は音節の後半の F0 の近似直線上の音節末の F0($F0tgt2b$) と前半部の F0 平均値 ($F0avg2a$) との差分を用いて計算する（式 7）．そして，音節の音調は式 8 に応じて，上昇調，下降調，平坦調に分類した．

$$\Delta F0 = F0tgt2b - F0avg2a \tag{7}$$

$$tone = \begin{cases} rising\ (Rs) & (\Delta F0 > 1 \text{ semitone}) \\ falling\ (Fa) & (\Delta F0 < -2 \text{ semitones}) \\ flat\ (Ft) & (\text{otherwise}) \end{cases} \tag{8}$$

4.3 首と腰の協調動作

頭部が動く際には上下方向だけではなく，前後方向にも動くことが判っている [31]．このことから，首の 1 自由度の回転だけではなく，腰も連動させることでより人間らしい動きが実現できると考えられる．また，口と首の動き出すタイミングは異なり，口のほうがやや早いく動くことが報告されていることから [32]，動かす関節により位相差があることが考えらる．そこで，式 9 の変換式を用いて図 5 のような協調動作を実装する．

$$\theta_{act}(t) = \alpha_{act}\theta_{base}(t + \beta_{act}) \tag{9}$$

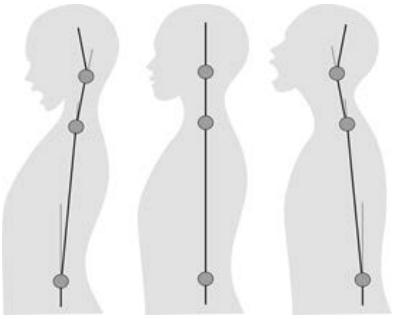


図 5: Multi-Joint Control

5 展望

節 4 で提案したモデルは，人間の身体的拘束に基づき，ばねダンパ系を用いた筋肉のダイナミクスを利用している．そのため，この動作生成モデルの動作パラメータは筋肉の硬さに比例したパラメータとなっている．筋肉の硬さは発話の緊張度合・感情状態によって変化すると考えられ，発話時の感情や緊張度合といった人間が理解できるパラメータから動作パターンを調節することが期待される．今後は，発話時の緊張・感情状態にあった動作を生成することができるかの検証や直感的に動作パラメータを決定できるかのユーザビリティの面から提案手法を評価する．

6 謝辞

本研究は，JST 戦略的創造研究推進事業 (ERATO） 石黒共生ヒューマンロボットインタラクションプロジェクトの一環として行われたものです．

参考文献

[1] Daisuke. Sakamoto, Takayuki Kanda, Tetsuo Ono, Hiroshi Ishiguro, and Norihiro Hagita. Android as a telecommunication medium with a human-like presence. In *Human-Robot Interaction*, pp. 193–200, 2007.

[2] 海光桑村, 竜二山崎, 修一西尾. テレノイドによる高齢者支援-特別養護老人ホームへの導入の経過報告-. 電子情報通信学会技術研究報告, Vol. 113, No. 272, pp. 23–28, 2013.

[3] Yutaka Kondo, Kentaro Takemura, Jun Takamatsu, and Tsukasa Ogasawara. A gesture-centric android system for multi-party human-robot interaction. *Journal of Human-Robot Interaction*, Vol. 2, No. 1, pp. 133–151, 2013.

[4] Miki Watanabe, Kohei Ogawa, and Hiroshi Ishiguro. Can Androids Be Salespeople in the Real World? In *ACM Conference Extended Abstracts on Human Factors in Computing Systems*, pp. 781–788, 2015.

[5] Masahiro Yoshikawa, Yoshio Matsumoto, Masahiko Sumitani, and Hiroshi Ishiguro. Development of an android robot for psychological support in medical and welfare fields. In *Robotics and Biomimetics*, pp. 2378–2383, 2011.

[6] Takuya Hashimoto and Hiroshi Kobayashi. Study on natural head motion in waiting state with receptionist robot SAYA that has human-like appearance. In *Robotic Intelligence in Informationally Structured Space*, pp. 93–98, 2009.

[7] Takanori Komatsu and Seiji Yamada. Adaptation gap hypothesis: How differences between users' expected and perceived agent functions affect their subjective impression. *Journal of Systemics, Cybernetics and Informatics*, Vol. 9, No. 1, pp. 67–74, 2011.

[8] Lukasz Piwek, Lawrie S McKay, and Frank E Pollick. Empirical evaluation of the uncanny valley hypothesis fails to confirm the predicted effect of motion. *Cognition*, Vol. 130, No. 3, pp. 271–277, mar 2014.

[9] Binh Huy Le, Xiaohan Ma, and Zhigang Deng. Live Speech Driven Head-and-Eye Motion Generators. *Visualization and Computer Graphics*, Vol. 18, No. 11, pp. 1902–1914, 2012.

[10] Mehmet Emre Sargin, Yucel Yemez, Engin Erzin, and Ahmet Murat Tekalp. Analysis of head gesture and prosody patterns for prosody-driven head-gesture animation. In *Pattern Analysis and Machine Intelligence*, Vol. 30, pp. 1330–1345. Department of Electrical and Computer Engineering, University of California-Santa Barbara, Santa Barbara, CA 93106-9560, USA. msargin@ece.ucsb.edu, 2008.

[11] Carlos Busso, Zhigang Deng, Ulrich Neumann, and Shrikanth Narayanan. Natural head motion synthesis driven by acoustic prosodic features. *Computer Animation And Virtual Worlds*, Vol. 16, No. 3-4, pp. 283–290, 2005.

[12] Mary Ellen Foster and Jon Oberlander. Corpus-based generation of head and eyebrow motion for an embodied conversational agent. *Language Resources and Evaluation*, Vol. 41, No. 3-4, pp. 305–323, 2007.

[13] Jelle Saldien, Bram Vanderborght, Kristof Goris, Michael Van Damme, and Dirk Lefeber. A motion system for social and animated robots. *International Journal of Advanced Robotic Systems*, Vol. 11, No. 1, pp. 1–13, 2014.

[14] Andrew G Brooks and Ronald C. Arkin. Behavioral overlays for non-verbal communication expression on a humanoid robot. *Autonomous Robots*, Vol. 22, No. 1, pp. 55–74, 2007.

[15] Miles L Patterson. 非言語コミュニケーションの統合モデルに向けて. 対人社会心理学研究, 第 7 巻, pp. 67–74, 2007.

[16] Tomio Watanabe, Masashi Okubo, Mutsuhiro Nakashige, and Ryusei Danbara. InterActor: Speech-Driven Embodied Interactive Actor. *International Journal of Human-Computer Interaction*, Vol. 17, No. 1, pp. 43–60, 2004.

[17] Chaoran Liu, Carlos Toshinori Ishi, H Ishiguro, and N Hagita. Generation of nodding, head tilting and eye gazing for human-robot dialogue interaction. In *Human-Robot Interaction*, pp. 285–292, 2012.

[18] Carlos Toshinori Ishi, ChaoRan Liu ChaoRan Liu, H Ishiguro, and N Hagita. Head motion during dialogue speech and nod timing control in humanoid

robots. In *Human-Robot Interaction*, pp. 293–300. Ieee, 2010.

[19] Kurima Sakai, Carlos Toshinori Ishi, Takashi Minato, and Hiroshi Ishiguro. Online speech-driven head motion generating system and evaluation on a tele-operated robot. In *Robot and Human Interactive Communicationtion*, pp. 529–534, 2015.

[20] Per-Olof Eriksson, Hamayun Zafar, and Erik Nordh. Concomitant mandibular and head-neck movements during jaw opening-closing in man. *Journal of oral rehabilitation*, Vol. 25, No. 11, pp. 859–870, 1998.

[21] Sean Andrist, Xiang Zhi Tan, Michael Gleicher, and Bilge Mutlu. Conversational Gaze Aversion for Humanlike Robots. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-robot Interaction*, pp. 25–32, 2014.

[22] Randy J Larsen and Todd K Shackelford. Gaze avoidance: Personality and social judgments of people who avoid direct face-to-face contact. *Personality and Individual Differences*, Vol. 21, No. 6, pp. 907–917, 1996.

[23] Dwight Bolinger. *Intonation and Its Parts: Melody in Spoken English*. 1985.

[24] Hani C. Yehia, Takaaki Kuratate, and Eric Vatikiotis-Bateson. Linking facial animation, head motion and speech acoustics. *Journal of Phonetics*, Vol. 30, No. 3, pp. 555–568, 2002.

[25] Michihiro Shimada and Hiroshi Ishiguro. Motion Behavior and its Influence on Human-likeness in an Android Robot. In *Annual meeting of the Cognitive Science Society*, pp. 2468–2473, 2008.

[26] 正幸中沢, 卓也西本, 茂樹嵯峨山. 力学モデル駆動による音声対話エージェントの動作生成. In *Human-Agent Interaction Symposium*, pp. 2C–1, 2009.

[27] Cho-chung Liang and Chi-feng Chiang. A study on biodynamic models of seated human subjects exposed to vertical vibration. *International Journal of Industrial Ergonomics*, Vol. 36, pp. 869–890, 2006.

[28] Astrid Linder. A new mathematical neck model for a low-velocity rear-end impact dummy: Evaluation of components influencing head kinematics. *Accident Analysis and Prevention*, Vol. 32, pp. 261–269, 2000.

[29] Carlos Toshinori Ishi, Chaoran Liu, Hiroshi Ishiguro, Norihiro Hagita, Intelligent Robotics, and Communication Labs. Evaluation of formant-based lip motion generation in tele-operated humanoid robots. *IROS2012*, pp. 2377 – 2382, 2012.

[30] Carlos Toshinori Ishi. Perceptually-Related F0 Parameters for Automatic Classification of Phrase Final Tones. *IEICE transactions on information and systems*, Vol. 88, No. 3, pp. 481–488, March 2005.

[31] Hamayun Zafar, Erik Nordh, and Per-Olof Eriksson. Spatiotemporal consistency of human mandibular and head-neck movement trajectories during jaw opening-closing tasks. *Experimental Brain Research*, Vol. 146, No. 1, pp. 70–76, 2002.

[32] Hamayun Zafar, Erik Nordh, and Per-Olof Eriksson. Temporal coordination between mandibular and head-neck movements during jaw opening-closing tasks in man. *Archives of Oral Biology*, Vol. 45, No. 8, pp. 675–682, 2000.

Using Sensor Network for Android gaze control*

Jani Even, Carlos Toshinori Ishi, Hiroshi Ishiguro
Hiroshi Ishiguro Laboratories, Advanced Telecommunications Research Institute International, Japan.
even@atr.jp *

Abstract

This paper presents the approach developed for controlling the gaze of an android robot. A sensor network composed of RGB-D cameras and microphone arrays is in charge of tracking the person interacting with the android and determining the speech activity. The information provided by the sensor network makes it possible for the robot to establish eye contact with the person. A subjective evaluation of the performance is made by subjects that were interacting with the android robot.

1 INTRODUCTION

The eyes of a human convey a considerable amount of information during interaction. For this reason, it is important to implement a human like gaze behavior in robots that communicate with humans. In [1], the authors followed the example of the human visual system to develop the gaze of their humanoid robot. Their robot, called Kismet, controls his eyes and neck to look at target detected by four cameras located in the eyes and on the face.

The ability to perform eye contact is important but gaze also plays an important role in the mutual attention [1, 2] (and reference herein) and pointing [3]. In [4], a reactive gaze implementation for mutual attention and eye contact is presented for a humanoid robot in an explanation setting. A motion capture system is used to get the head orientation of the human. Only the robot's head is actuated. Another example of robot head with human-style gaze ability is the system presented in [5].

This paper presents the gaze control of the android robot developed for the ERATO Ishiguro Symbiotic Human-Robot Interaction Project [6]. This android is called Erica which stands for ERATO Intelligent Conversational Android. Erica was designed to have a realistic human like appearance, see Fig.1.

Figure 1: Close-up of Erica the android robot of the ERATO Ishiguro Symbiotic Human-Robot Interaction Project.

The goal of this paper, is to investigate the ability of Erica to look at a given direction in the environment. This is done by using a sensor network for finding and tracking the point of interest and controlling Erica to look at this point.

Erica is sitting on a chair, but contrary to the robots in [5, 1, 7], Erica has a complete body. Consequently, the gaze implementation presented in this paper actuates not only the eyes and the neck but also the waist of Erica.

2 SENSOR NETWORK

Before describing the gaze control, let us present how the system determine the point of interest. The basic idea is that a sensor network provides information on the context around Erica and depending of the intended interaction, a point of interest is determined.

In the current state, the sensor network main role is to track human [8, 9, 10] and determine who is talking [11, 12]. For this purpose a human tracking system is combined with a sound localization system. Figure 2

*Research supported partly by the JST ERATO Ishiguro Symbiotic Human-Robot Interaction Project and partly by the Ministry of Internal Affairs and Communications of Japan under the Strategic Information and Communications R&D Promotion Programme (SCOPE).

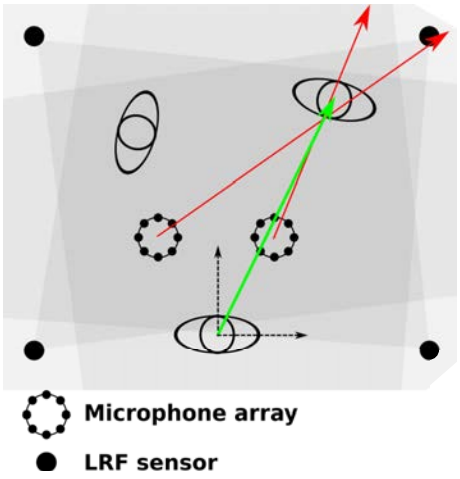


Figure 2: Example of possible sensor network configuration.

shows one example of configuration with four laser range finders (LRFs) for tracking humans and two microphone arrays for performing sound localization. During the experiments, the human tracker system was not using LRFs but RGB-D cameras attached to the ceiling of the room [13]. Using the sound localization (the red arrows in Fig.2) it is possible to determine who is talking. Then the goal is to have Erica pays attention to that person (the green arrow). Namely, the sensor network gives a point of attention that can vary. This point of attention is referred to as the *focus point* in the remainder.

3 KINEMATIC CHAIN

In this section, we describe the kinematic chain to be controlled for setting the gaze of Erica at a given focus point.

Figure 3 shows the joints involved in the gaze control. The kinematic chain controlling the eyes direction has 7 degrees of freedom (DOF):

- yaw and pitch for the eyes,
- yaw pitch and roll for the neck,
- yaw and pitch for the waist.

However, the current implementation does not use the neck roll.

Pneumatic actuators are used to move the joints. These actuators are controlled by on board PID controllers. The commands are sent to the robot at a frequency of 20 Hz. The robot provides a feedback measured by potentiometers also at the frequency of 20 Hz. The on board PID are tuned to favor smoother movements which results in a lesser control accuracy. Consequently, it is necessary to rely on the feedback to get the achieved positioning.

Using the specifications of Erica, a computer model of the kinematic chain was implemented. The posture of the model is updated when the feedback from the actuators is received. Namely, the model provides an estimate of the current posture of Erica.

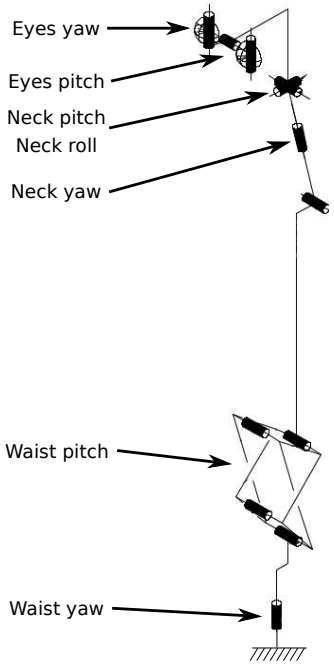


Figure 3: Kinematic chain for the gaze.

4 GAZE CONTROL

The kinematic model presented in the previous section provides the current gaze direction of Erica's eyes.

This is illustrated in Fig.5 that depicts a part of the graphical user interface (GUI). In the left view, the current gazing direction of Erica is shown by the pink line. The green line shows the direction of the focus point (the red box). At this moment Erica is not requested to look at the focus point. When asked to look at the focus point, the look direction (pink) is aligned to the focus point direction (green) as in the right part of Fig.5. The goal of the gaze control is to send command to move the joints of Erica in order to perform this alignment.

Figure 4 shows the flow chart of the algorithm. The control sequence is as follows:

1. the position of the focus point $f(k)$ is given to the gaze control,
2. the gaze control requests the current gaze direction (i.e. the current orientation of the joints $\widetilde{\theta(k)}$) to the kinematic model,
3. if the gaze direction is close enough to the focus point direction go to 9 otherwise go to 4,

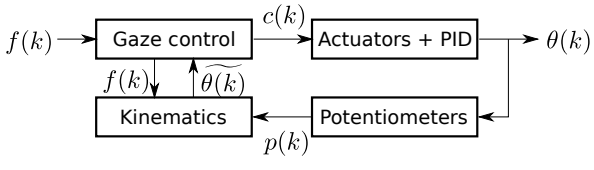


Figure 4: Flowchart showing the different blocks of the gaze control.

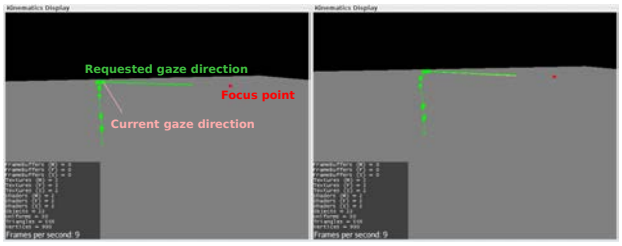


Figure 5: Visualization of the kinematic chain and the focus point.

4. the gaze controller determines the commands $c(k)$ to send to the joints,
5. the actuators move the joints,
6. the potentiometers give the feedback $p(k)$,
7. the kinematic model is updated,
8. loop to 2,
9. gaze control completed.

The step 4 is the most important ones. Given the direction of the focus point and the current direction of the gaze, the controller has to determine the commands to send to the different joints.

Figure 6 illustrates the procedure for the yaw command of the waist, the neck and the eyes. In Fig.6-a, Erica has a posture determined by the waist yaw, neck yaw and eye yaw and is requested to look in the set gaze direction. All these directions are represented by the colored arrows. The controller determines the desired angles for the joints. These angles are represented by the dash arrows in Fig.6-b. For the waist (red) and neck (orange), the desired angles are converted in absolute commands defined as a fraction of the set angles. For the eyes (yellow), the desired angle is converted to a command relative to the current position of the eyes which is represented by the black arrow.

Let us denotes the set angle by $\theta(k)$ then the waist and neck angles are

$$\begin{aligned} \theta_{\text{waist}}(k) &= \alpha_{\text{waist}}\theta(k) \\ \theta_{\text{neck}}(k) &= \alpha_{\text{neck}}\theta(k) \end{aligned} \tag{1}$$

where α_{waist} and α_{neck} control the amount of rotation distributed to the waist and neck.

For the eye angle, the relative value is

$$\theta_{\text{eyes}}(k) = \theta(k) - \widetilde{\theta_{\text{eyes}}(k)} \tag{2}$$

where $\widetilde{\theta_{\text{eyes}}(k)}$ is the estimated eye angle given by the kinematic model.

Only the eyes are controlled in a closed loop because the accuracy on the eye movement is greater than on the waist and neck.

When the joints have started to move, as in Fig.6-c, the absolute angles for the waist and neck do not change whereas the relative one for the eye is updated.

Figure 6-d shows an example of gaze control completion. In this case, a small error still exists for the neck that did not reach the absolute angle (the double black arrow). However, the gaze direction is reached as the relative angle computed for the eyes compensated the residual error on the neck.

In practice, for all the joints, the angles are converted in command values that are in the range $[0, 255]$ before sending them to the robot. The feedback values received are also in the range $[0, 255]$. The conversion is a simple linear mapping. For example for the eyes

$$c_{\text{eyes}}(k) = \theta_{\text{eyes}}(k) * \frac{255}{\theta_{\text{eyes, max}} - \theta_{\text{eyes, min}}}$$

$$\widetilde{\theta_{\text{eyes}}}(k) = \theta_{\text{eyes, min}} + p_{\text{eyes}}(k) * \frac{\theta_{\text{eyes, max}} - \theta_{\text{eyes, min}}}{255}$$

where $\theta_{\text{eyes, min}}$ and $\theta_{\text{eyes, max}}$ are the angles corresponding to the command or the potentiometer values 0 and 255.

5 EXPERIMENTAL RESULTS

5.1 objective evaluation

In this experiment, the focus point was set to subject tracked by the sensor network. This subject was walking in front of Erica for four minutes. The direction of the subject (the focus point) and the estimated gaze direction given by the kinematic model were recorded. The command and potentiometer values were also recorded. The goal of this experiment is to check if Erica is able to track a moving focus point using the proposed gaze control approach.

The top of Fig.7 shows the yaw of the focus direction (solid line) and the yaw of the gaze direction given by the kinematic model (dashed line). The three other graphs are showing the command values (solid lines) and the potentiometer values (dashed lines) for the control of the waist, neck and eyes yaw. The focus direction is well tracked by the gaze direction except for the period between the two vertical red dashed lines.

Figures 8 and 9 respectively show a good tracking period (the green vertical dashed lines in Fig. 7) and the bad tracking region. The top graph of Fig.8 clearly shows that the gaze direction closely follows the focus direction. We can note a slight delay, which is expected, and some overshoots. However, the graph for the neck control shows some large errors and the one for the waist some small errors. These two graphs are by construction scaled version of the focus angle, see Eq.(1). Then, we can see on the graph for the eyes that the command is different and it compensated for the error as expected.

The tracking error that appears in Fig.9 is explained by the fact that the large error on the neck angle could not be corrected by the eyes because they saturated (the command reached 0). This is due to the fact that the subject was at a large focus angle.

Figure 10 shows the cumulative density functions (CDFs) for the errors on the yaw (left) and the pitch (right). The horizontal black dashed lines indicate the 90% quantiles. For the yaw, 90% of the errors are smaller

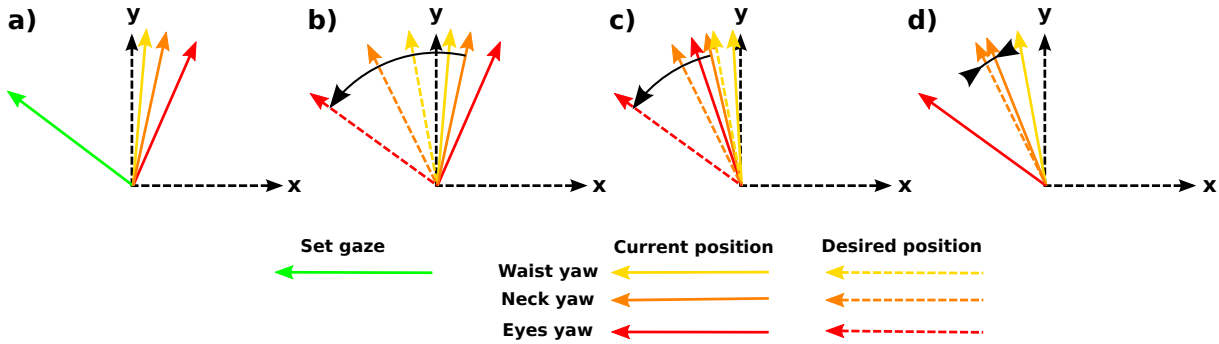


Figure 6: Ratio for the different body parts during gaze setting. Note the remaining error on the neck in d.

than 11 degrees and for the pitch smaller than 5 degrees. The larger error on the yaw is due to the fact that while the person was moving in front of Erica, the pitch did not vary much whereas the yaw presented large variations. The error showed in Fig.9 created the small bump around 45 degrees in the CDF of the yaw. Note that these errors are computed while tracking a moving person. Then the small tracking delays contribute to the error for the lower values of the CDFs.

Figure 11 shows the cumulative density functions (CDFs) for the errors on the yaw command for the waist (left), the neck (center) and the eyes (right). As expected, the 90% quantile is significantly higher for the neck.

This experiment showed that the proposed approach is able to accurately track a moving focus point. The performance was measured on the feedback given by the potentiometers. This means that some bias may be present if the calibration is not done properly. Namely, the measured focus direction and the true focus direction may differ.

A finding is also that most of the error comes from the neck. In particular, for some large angles, Erica could not look at the desired directions because of the error on the neck positioning. These situations correspond to cases where the human would also turn on themselves to look. This is due to friction forces that prevent the neck actuator to achieve the desired positioning while moving smoothly. To solve this problem, a low level controller that is aware of the friction will be implemented.

5.2 Subjective evaluation

The subjective evaluation of the gaze control is performed by setting a focus point and asking a subject to position herself/himself where she/he feels Erica is making eye contact with her/him. This is done for several focus points in front of Erica, see Fig.12. Then for each of the focus points, the position where eye contact is felt the best is recorded using the human tracker. The height of the subject eyes is measured to set the height of the focus points. For the focus points, the yaw angle θ is computed and for the corresponding position of perceived eye contact, the yaw angle $\widehat{\theta}$ is also computed. For the selected focus point, in green in Fig.12, the yaw angle θ is represented by the green arrow and the yaw angle $\widehat{\theta}$ of the perceived eye contact is represented by a red arrow.

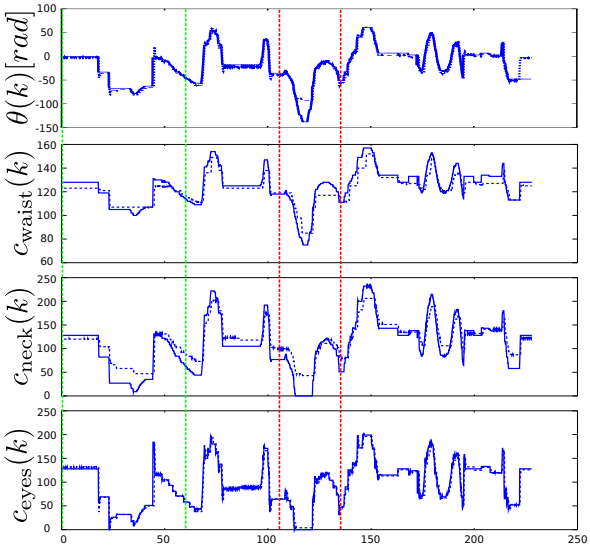


Figure 7: Axis command (dashed) and potentiometer feedback (solid) for the yaw.

Figure 13 is a plot of the perceived angles versus the focus point angles. The data points for two different subjects are plotted (circles and crosses). The black line is $\widehat{\theta} = \theta$ and the red line is the linear fit:

$$\widehat{\theta} = 0.79\ \theta + 7.92 \tag{3}$$

the RMSE is 5.27.

The angle $\widehat{\theta}$ of the perceived eye contact does not correspond to the set angle θ. Meaning that the subjects did not feel the eye contact at the exact set position.

However, a linear fit of the data is possible. The bias of 7.92 degrees and the scaling error of 0.79 could be explained by calibration errors. The ranges $\theta_{\text{XXX, max}}$ and $\theta_{\text{XXX, min}}$ (where XXX is for waist, neck or eyes) have to be adjusted.

Without re-calibration of the ranges, the linear fit could be used to select the set angle to look at a position given by the human tracker:

$$\theta = 1.21\ \widehat{\theta} - 9.47 \tag{4}$$

the RMSE is 6.54. Figure 14 shows this linear fit.

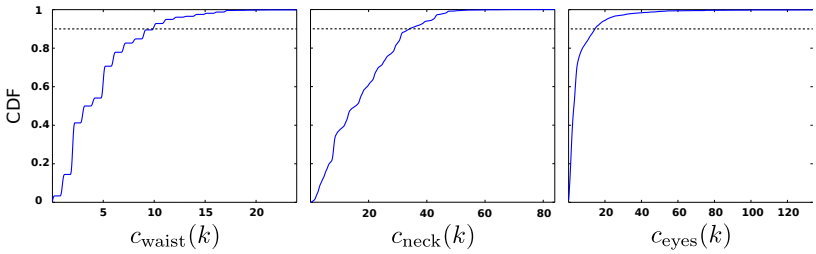


Figure 11: Cumulative density functions for the command errors for the waist (left), the neck (center) and the eyes (right).

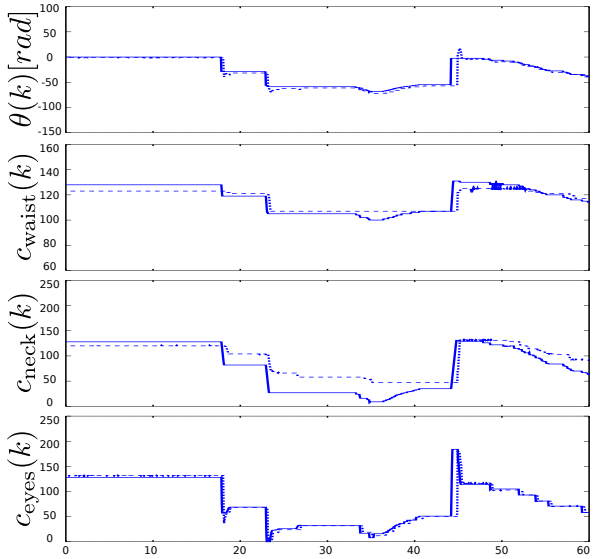


Figure 8: Close-up of the axis command (dashed) and potentiometer feedback (solid) for the yaw.

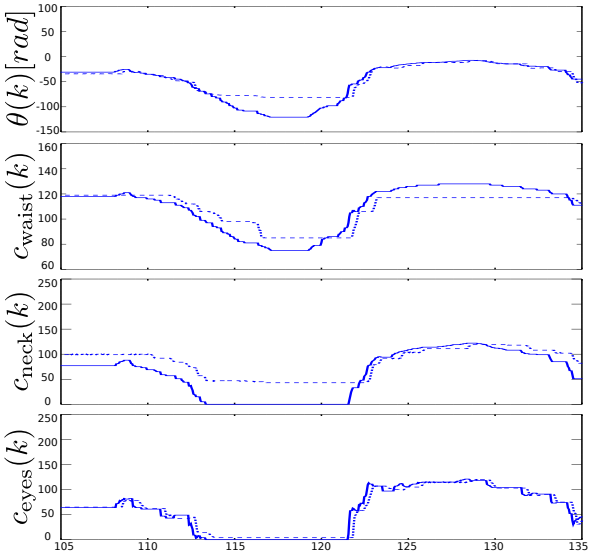


Figure 9: Close-up of the axis command (dashed) and potentiometer feedback (solid) for the yaw.

6 CONCLUSIONS

This paper presented the low level gaze function of Erica. The objective experiment showed that the gaze control is behaving as expected. The system is able to compensate the measured error. However, the subjective evaluation suggests that there is still a calibration to be done in order to obtain eye contact. An alternative way would be to use the linear mapping between perceived gaze angle and set angle.

In addition to the ability to look at a given point, a humanoid robot should also reproduce a human like behavior [14, 15]. Human like features of the gaze are implemented at a higher level in Erica's control architecture. The integration of these higher level features with the low level control will be the focus of future research.

References

[1] C. Breazeal, A. Edsinger, P. Fitzpatrick, and B. Scassellati, "Active vision for sociable robots," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, vol. 31, no. 5, pp. 443–453, 2001.

[2] B. Scassellati, "Investigating models of social development using a humanoid robot," in *Neural Networks, 2003. Proceedings of the International Joint Conference on*, 2003, vol. 4, pp. 2704–2709 vol.4.

[3] Sotaro Kita, *Interplay of gaze, hand, torso orientation and language in pointing, in Pointing: Where Language, Culture, and Cognition Meet*, Lawrence Erlbaum Associates, 2003.

[4] Y. Mohammad and T. Nishida, "Reactive gaze control for natural human-robot interactions,"

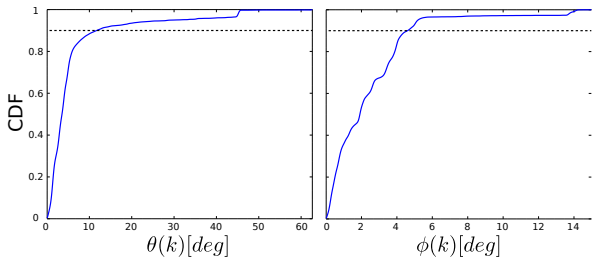


Figure 10: Cumulative density functions for the angular errors for the yaw (left) and the pitch (right).

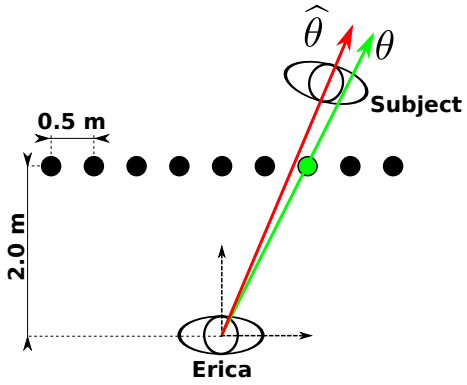


Figure 12: Settings for the subjective test showing the focus points.

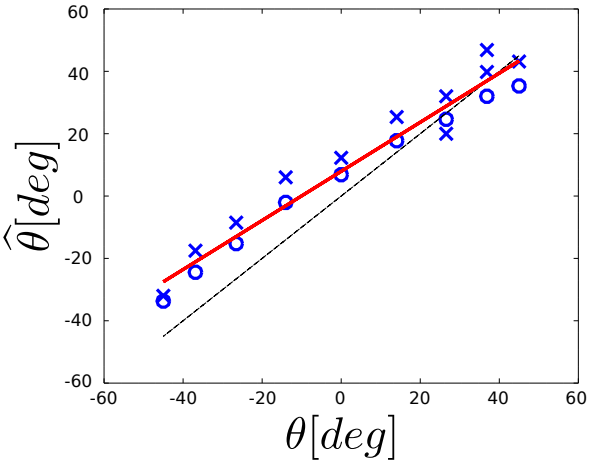


Figure 13: Subjective angle $\widehat{\theta}$ versus set angle θ. The red line is a linear fit of the data points.

in *Robotics, Automation and Mechatronics, 2008 IEEE Conference on*, 2008, pp. 47–54.

[5] A. Takanishi, H. Takanobu, I. Kato, and T. Umetsu, "Development of the anthropomorphic head-eye robot we-3rii with an autonomous facial expression mechanism," in *Robotics and Automation, 1999. Proceedings. 1999 IEEE International Conference on*, 1999, vol. 4, pp. 3255–3260 vol.4.

[6] Hiroshi Ishiguro et al., "Erato ishiguro symbiotic human-robot interaction project," `http://www.jst.go.jp/erato/ishiguro/en/index.html`, 2015.

[7] D. Hanson and The University of Texas at Dallas, *Humanizing Interfaces: An Integrative Analysis of the Aesthetics of Humanlike Robots*, University of Texas at Dallas, 2007.

[8] Jae Hoon Lee, T Tsubouchi, K Yamamoto, and S Egawa, "People tracking using a robot in motion with laser range finder," 2006, pp. 2936–2942, Ieee.

[9] D.F. Glas et al., "Laser tracking of human body motion using adaptive shape modeling," *Proceedings of 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 602–608, 2007.

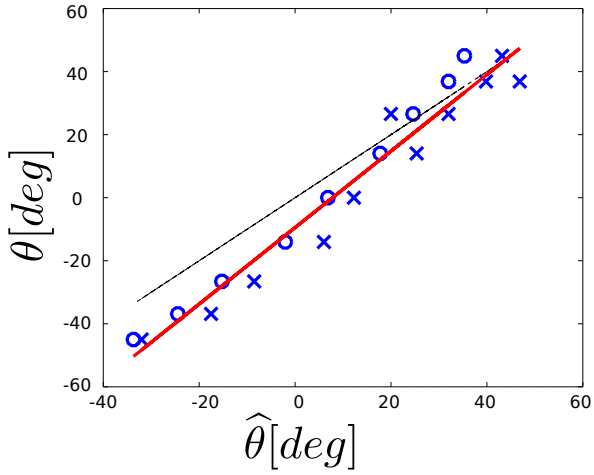


Figure 14: Set angle θ versus subjective angle $\widehat{\theta}$. The red line is a linear fit of the data points.

[10] L. Spinello and K. O. Arras, "People detection in rgb-d data.," in *Proc. of The International Conference on Intelligent Robots and Systems (IROS)*, 2011.

[11] C.T. Ishi et al., "Evaluation of a music-based real-time sound localization of multiple sound sources in real noisy environments," *Proceedings of 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2027–2032, 2009.

[12] C.T. Ishi, J. Even, and N. Hagita, "Using multiple microphone arrays and reflections for 3d localization of sound sources," *Proceedings of 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3937–3942, 2013.

[13] D. Brscic, T. Kanda, T. Ikeda, and T. Miyashita, "Person tracking in large public spaces using 3-d range sensors," *Human-Machine Systems, IEEE Transactions on*, vol. 43, no. 6, pp. 522–534, 2013.

[14] J.M. Wolfe, "Guided search 2.0: A revised model of visual search," *Psychonomic Bulletin & Review*, vol. 1, no. 2, pp. 202–238, 1994.

[15] R Weidner, J Krummenacher, B Reimann, H Muller, and G Fink, "Sources of top-down control in visual search," *Cognitive Neuroscience, Journal of*, vol. 21, no. 11, pp. 2100–2113, 2009.

小型クアドロコプタの群を用いたコンセンサスに基づく音源定位

Sound Source Localization Based on Consensus using a Swarm of Micro-Quadrocopters

中村圭佑 [1], シナパヤ ラナ [2], 中臺一博 [1], 高橋秀幸 [2], 木下哲男 [2]

Keisuke NAKAMURA, Lana SINAPAYEN, Kazuhiro NAKADAI, Hideyuki TAKAHASHI, Tetsuo KINOSHITA

1.（株）ホンダ・リサーチ・インスティチュート・ジャパン, 2. 東北大学

1. Honda Research Institute Japan Co., Ltd., 2. Tohoku University

{keisuke,nakadai}@jp.honda-ri.com, lana@sacral.c.u-tokyo.ac.jp,
{hideyuki,kino}@riec.tohoku.ac.jp

Abstract

本稿では，単独マイクを搭載した複数の小型クアドロコプタを用いた音源の検出および定位について述べる．群ロボットによるロボット聴覚機能の実現には，1) 各個体が環境中の音源を用いて自己位置を推定する機能，2) 各個体が未知音源を定位する際に推定状態の不確かさを考慮した群としての情報統合の枠組が必要である．それぞれの問題を解決するため，UKF を用いた自己位置推定手法，および，コンセンサスの概念を導入した UKCF による群ロボットによる未知音源定位手法を提案する．各手法を実環境で収録したデータを用いて有効性の確認を行った．

1 序論

本稿では，大きさが 0.1m を下回る小型クアドロコプタの群を用いた屋内環境下音源定位を提案する．小型クアドロコプタはペイロードが小さいため，内蔵されたセンサに加えて２つ以上のマイクロホンを搭載することが困難である．また，小型クアドロコプタ上でのロボット聴覚機能の実現には，マイクロホンに近接した大きなパワーのプロペラ雑音を持つこと，内蔵された CPU の計算速度や性能に限界があること，カメラ等のマルチモーダル情報を付加するに十分なペイロードがないこと等の問題がある．これらの問題に対し，我々は小型クアドロコプタを複数用いて群を形成することで解決を図る．群の中から定位対象である環境中の音源に近いクアドロコプタを積極的に用いることで信号対プロペラ雑音比を改善し，各個体に計算を分散化することで各個体に搭載された低性能の CPU でも実現可能な音源定位を提案する．各個体の分散処理および群としての情報統合を用いた音源定位を実現するためには，1) 環境中の音源をランドマークとして各個体が独立に自己位置を推定しながら未知音源の位置を同時に推定する機能，2) 各個体で独立に推定された位置情報を各個体の推定誤差を考慮して効果的に統合する枠組が必要である．これらの要件を満たすため，本稿は GPS や モーションキャプチャ等を使用せず，クアドロコプタ内蔵センサと搭載マイクのみを用いて屋内環境でも適用可能な二つの手法を提案する．1) については，環境中の音源ランドマークを用いた Uncented Kalman Filter (UKF) ベースの自己位置と未知音源位置推定を提案する．2) については，各個体で推定された音源位置情報の Uncented Kalman Consensus Filter (UKCF) を用いたコンセンサスに基づく統合を提案する．評価では二つの提案法の有効性をシミュレーションと実機を用いたデータを用いて確認した．

2 関連研究

クアドロコプタを含めた飛行ロボットは広大な空間中を短時間で探索でき，がれきや段差，水たまりなどの地形によらず移動できる．また，群を形成して [2] 屋内を移動することも可能である [1]．以上の点から，飛行ロボットは災害時探索に適しており，探索における音情報は暗闇や煙，がれきの中から被害者を見つけるのに鍵となる情報の一つである．Basiri らは翼を持つ飛行体の群を用いて，各個体にマイクロホンアレイを搭載することで，音を用いた自己位置推定と，人が地上から鳴らした笛の音源定位を実現した [3; 4]．しかし，翼を持つ飛行体は高度が高く，飛行に大きな空間を要するため，屋内環境には適していない．一方，クアドロコプタなどの回転翼を持つ飛行体は安定した姿勢を保つことができるため，屋内環境でも使用可能であるが，プロペラ雑音のパワーが大きいために環境中の目的音信号対雑音比が低くなってしまう．プロペラ雑音に対し，クアドロコプタにマイクロホンアレイを搭載してプロペラ雑音を白色化しつつ環境中の音源を定位する手法が提案されている [5; 6] が，マイクロホンアレイを搭載するには少なくとも数十〜数百グラムのペイロードが必要である．このために

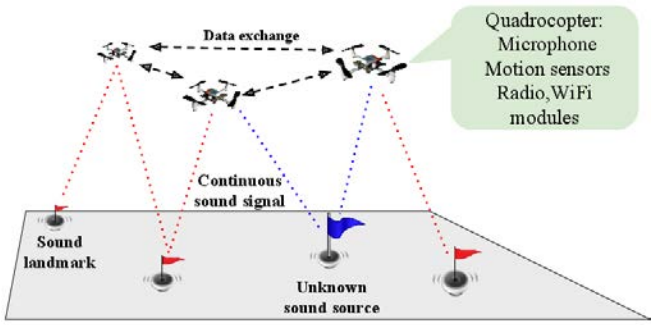


Figure 1: Considered Environment in SSL using a Swarm

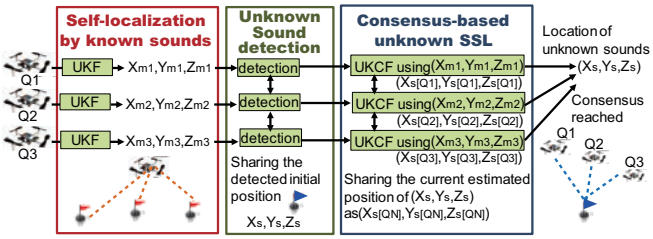


Figure 2: Process Flow in SSL using a Swarm

は，大きな機体が不可欠となるため，屋内環境適用には不向きである．屋内環境用の小型クアドロコプタを用いることで，プロペラ雑音を小さくし，より環境中の音源との信号対雑音比を向上できると考えられる．しかし，ペイロードが数グラム〜十数グラムに限定されるため，マイクロホンアレイの搭載が難しく，搭載された CPU は負荷の高いマイクロホンアレイ処理には不向きである．そこで本稿では，Figure 1 のように各クアドロコプタに対して 1 個のマイクロホンを搭載し，各個体で音を分散処理し，処理された情報を群として後段で統合する手法を提案する．提案法では，Figure 2 のように，まず，既知の音源ランドマークを用いて UKF に基づいて自己位置と未知の音源位置を推定し（3.1 節），小型クアドロコプタの群によって未知音源を定位するための UKCF を用いたコンセンサスに基づく推定音源位置情報の統合を行う．UKCF では，Kalman Consensus Filter (KCF) [9] による線形分散システムに対する最適状態推定の考え方を，UKF [8] によって非線形拡張した非線形分散システムの最適状態推定を行う（3.2 節）．

Table 1: Notation of Variables for UKF

Quad. coordinates	x_q, y_q, z_q
Velocity	$\dot{x}_q, \dot{y}_q, \dot{z}_q$
Acceleration	$\ddot{x}_q, \ddot{y}_q, \ddot{z}_q$
Iteration	k
State	$x_k = (x_q, y_q, z_q, \dot{x}_q, \dot{y}_q, \dot{z}_q)^T$
Time step	t
Landmark intensity (1m)	I
l-th landmark coordinates	x_l, y_l, z_l
Number of landmarks	L
Initial sigma weight	$e = 0.9$

Table 2: Model of UKF

Time update function f

$$f_x(\alpha_q) = \alpha_q + t \times \dot{\alpha}_q + \frac{t * \ddot{\alpha}_q^2}{2} \text{ for } \alpha \text{ in } \{x, y, z\}$$

$$f_{\dot{x}}(\dot{\alpha}_q) = \dot{\alpha}_q + t \times \ddot{\alpha}_q \text{ for } \alpha \text{ in } \{x, y, z\}$$

Output function h

$$i_{k,l} = \frac{I}{(x_l - x_q)^2 + (y_l - y_q)^2 + (z_l - z_q)^2}$$

$$h(x_k) = \begin{pmatrix} i_{k,1} & \dots & i_{k,L} \end{pmatrix}$$

3 提案手法

3.1 音源ランドマークを用いた自己位置の推定

群中のそれぞれのクアドロコプタは内蔵された 9 軸のモーションセンサ（3 軸加速度センサ，3 軸角速度センサ，3 軸地磁気センサ）に加えて，クアドロコプタのコア部分に取り付けた単独マイクロホンを用いて，自己位置と音源位置推定を行う．9 軸モーションセンサは自身の観測情報を用いて Dead Reckoning により自己位置をある程度推定することができるが，累積誤差が大きくなり精度良く自己位置を推定することが難しい．そこで，本稿では搭載されたマイクロホンを用いて，環境中の既知の位置に固定された音源ランドマークから発せられた音源（スピーカー）の強度を観測することで自己位置の推定誤差を軽減する．プロペラ雑音や環境雑音が大きいため，音源強度にも観測誤差を生じるが，ランドマークの絶対位置を使用できるため，Dead Reckoning よりも累積誤差の少ない位置推定が期待できる．

モーションセンサから得られる加速度情報とマイクロホンから得られる音源強度情報を統合したモデルに基づいて UKF [8]を用いて位置推定を行う．ここで，本稿では，自己位置推定に使う音源ランドマーク位置は既知であること（未知音源の位置は未知）と，音源ランドマークは複数個存在し，その音は定常で指向性はなく独立した周波数を持つことを仮定する．従って，各クアドロコプタは，各音源強度を周波数独立に観測できることになる．また，クアドロコプタの初期位置は既知とし，初期速度はないものとする．以上の仮定より，UKF での推定対象はクアドロコプタの自己位置となることから，状態遷移モデルの状態は，Table 1 のようにクアドロコプタの状態のみで表され，状態はモーションセンサから得られる加速度情報を用いて更新される．また，音源強度は音源からマイクロホンまでの距離の二乗に反比例することが知られているため，観測モデルは，推定された位置情報から期待され

Table 3: Notation of Variables for UKCF

Source coordinates	x_s, y_s, z_s
Quad. coordinates	x_q, y_q, z_q
State	$x = (x_s, y_s, z_s)^T$
Sate dimension	$n = 3$
Iteration	k
Source intensity (at 1m)	I
Sigma points	$X_k = \left\{(x_k^j, w^j) \mid j = 0 \dots 2n\right\}$
Initial sigma weight	$w^0 = 0.009$
Predicted state	x_k^f
Predicted error	P_k^f
Corrected state	x_k
Corrected error	P_k
Predicted measurement	z_k^f
Process noise	Q_k
Measurement noise	R_k
Kalman gain	K_k
Consensus gain	C_k
Consensus order	$\varepsilon = 0.01$
Frobenius norm	$\|.\|_F$

Table 4: Model of UKCF

Model

Output function h

$i_k = \frac{I}{(x_s-x_q)^2+(y_s-y_q)^2+(z_s-z_q)^2}$

$h(x_k) = \left(i_k\right)$

Prediction step

(for each individual quadrocopter)*Sigma point generation*

$x_{k-1}^0 = x_{k-1}\ x_{k-1}^i = x_{k-1} + \left(\sqrt{\frac{n}{1-w^0}P_{k-1}}\right)_i$ for $i = 1 \dots n$

$x_{k-1}^{i+n} = x_{k-1} - \left(\sqrt{\frac{n}{1-w^0}P_{k-1}}\right)_i$ for $i = 1 \dots n$

$w^j = \frac{1-w^0}{2n}$ for all $j = 1 \dots 2n$

State Transition

$x_k^f = x_{k-1}$

Mean and covariance computation

$x_k^f = \sum_{j=0}^{2n} w^j x_k^{f,j}$

$P_k^f = \sum_{j=0}^{2n} w^j \left(x_k^{f,j} - x_k^f\right)\left(x_k^{f,j} - x_k^f\right)^T + Q_{k-1}$

Predicted measurement computation

$z_{k-1}^{f,j} = h(x_{k-1}^j)$

$z_{k-1}^f = \sum_{j=0}^{2n} w^j z_{k-1}^{f,j}$

Kalman Gain computation

$Cov(z_{k-1}^f) = \sum_{j=0}^{2n} w^j \left(z_{k-1}^{f,j} - z_{k-1}^f\right)\left(z_{k-1}^{f,j} - z_{k-1}^f\right)^T + R_k$

$Cov(x_k^f, z_{k-1}^f) = \sum_{j=0}^{2n} w^j \left(x_k^{f,j} - x_k^f\right)\left(z_{k-1}^{f,j} - z_{k-1}^f\right)^T$

$K_k = Cov(x_k^f, z_{k-1}^f) Cov^{-1}(z_{k-1}^f)$

Correction step (for quad. q in a swarm of size M)

Consensus Gain computation

$C_k = \varepsilon \frac{P_k^f}{1+\|P_k^f\|_F}$

State and error correction

$x_k^q = x_k^{f,q} + K_k(z_k - z_{k-1}^f) + C_k \sum_{m=0}^{M}(x_k^{f,m} - x_k^{f,q})$

$P_k = P_k^f - K_k Cov(z_{k-1}^f) K_k^T$

る音源強度として，Table 2 のようにモデル化した．以上の状態遷移モデルと観測モデルを用いて，UKF では，観測周期ごとに，予測ステップにおいてモーションセンサから得られる加速度情報からクアドロコプタの位置と期待される音源強度を予測し，更新ステップでは観測された音源強度と予測音源強度の誤差を用いて状態を更新することを繰り返す．

我々はこれまでにも，複数のマイクが環境中に設置された状況で，移動する拍手音を用いて，拍手位置とマイク位置を推定する手法を提案してきた [7]．本稿の UKF では，移動するものがマイクとなり，固定されるものが音源であるという意味で，[7]の逆問題として類似している．しかし，拍手音であればマイクまでの到達時間差が陽に使えるため，距離を容易に求めることができるが，本稿の問題では音源ランドマークから発せられる複数の音の同期などを仮定できないため，強度情報のみしか用いることができないという意味で発展的な問題であるといえる．

3.2 UKCF を用いたコンセンサスに基づく音源定位

前節の UKF によってクアドロコプタが自身の位置を定位できている状況において，本節では，各クアドロコプタが未知の音源を検知した時にそれを定位しつつ，他個体で推定された未知音源位置と情報統合する手法について述べる．未知音源の検出について，我々はこれまで，各クアドロコプタをランダムに移動させ，未知音源に近づいてから離れた時に観測される音源強度時系列データのピークを検出した手法を提案した [10]．また，音源強度がピークとなった時の値と，その時刻のクアドロコプタの位置を用いて，未知音源位置のおよその初期位置を計算した．本稿では未知音源の検出については，この手法を用いることとし，説明を省略する．詳細は [10] を参照されたい．本稿では検出後の位置推定について述べる．[10] で計算された初期位置はピークの音源強度とその時刻のクアドロコプタの位置のみによるため，誤差が大きい．本稿では，UKCF によって分散した非線形システムの状態推定を行いつつ，各分散システムの推定結果を誤差の収束性を保証しつつ統合する手法を提案する．これまでの分散システムに対する誤差の収束を保証した状態推定として KCF が知られているが，線形システムにしか適用できなかった．

本稿の音源強度を用いた音源位置推定のモデルは，音源からクアドロコプタまでの距離を用いて記述されるため，非線形システムの状態推定となり，直接 KCF を適用することができない．UKCF では Uncented 変換を用いることで，コンセンサスに基づく推定を未知音源位置推定のための非線形分散システムに適用することができる．本手法では，前節と同様に，未知音源は環境中に固定され，その音は定常で指向性はなく独立した周波数を持つことを仮定する．したがって，各クアドロコプタは，各未知音源の音源強度を周波数独立に観測できることになる．また，前節の UKF によってクアドロコプタの自己位置は推定できているため，UKCF による推定対象は未知音源位置のみとなり，状態遷移モデルの状態は Figure 3 のように未知音源位置のみを用い，モデルは固定音源として記述される．観測モデルは前節と同様に音源強度を Figure 4 のように用いる．

Figure 4 に推定ステップを示す．Q_k と R_k は共分散行列であり，観測雑音をガウス白色雑音で無相関と仮定することで対角行列とした．UKF のように予測ステップでは，未知音源位置の状態と，状態に対する誤差共分散を推定するため，状態空間の中からシグマ点を予測し，それらの点の重み平均を算出する．予測された状態を用いて音源強度を非線形な観測モデルに従って予測する．更新ステップでは，状態を Figure 4 のように以下の式で更新する．

$$x_k^q = x_k^{f,q} + K_k(z_k - z_{k-1}^f) + C_k \sum_{m=0}^{M} (x_k^{f,m} - x_k^{f,q}) \qquad (1)$$

UKF では状態は，カルマンゲイン K_k を持つ項である式 (1) の右辺第一項，第二項のみで更新される．コンセンサスを考慮するため，本稿では，式 (1) の第三項の導入を提案する．$x_k^{f,m}$ と $x_k^{f,q}$ はそれぞれ，他個体が推定している未知音源位置と自分が推定している未知音源位置を表すため，$x_k^{f,m} - x_k^{f,q}$ は自分が推定している状態が他と離れているほど大きくなる項となる．$x_k^{f,m} - x_k^{f,q}$ を小さくするようにそれぞれのクアドロコプタが状態を更新すれば，有限時間でコンセンサスを達成できるというアイデアに基づいている．また，UKCF では制御ゲインであるコンセンサスゲイン C_k を各クアドロコプタが推定している誤差の分散値に従って変化させる．KCF [9] では，コンセンサスゲインを

$$C_k = \varepsilon \frac{P_k^f}{1 + \|P_k^f\|_F} \qquad (2)$$

と設計することで，平衡点 $x_k^{f,1} = x_k^{f,2} = \ldots = x_k^{f,M}$ が漸近安定となることを保証しており，本稿でもこれを用いることとした．UKCF を用いることで，全てのクアドロコプタが短時間で精度良く音源位置を推定することが期待される．

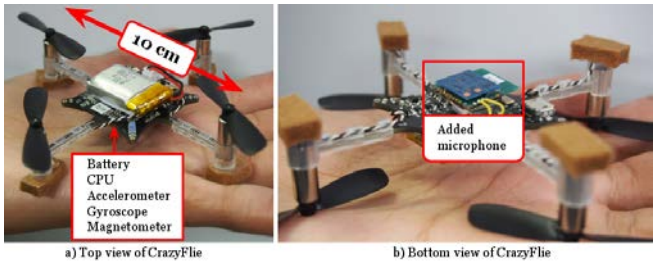


Figure 3: Micro-quadrocopter with a Sinle Microphone

4 評価実験

本章では以下の 3 つの評価を行い，提案法の有効性を検証する．

- 音源ランドマークを用いた UKF ベースの自己位置推定の有効性検証のための実環境下の音源とクアドロコプタ間の距離推定精度（一次元定位推定精度．4.1 節）
- 音源ランドマークを用いた UKF ベースの自己位置推定の有効性検証のための実環境下のクアドロコプタの二次元位置推定性能（4.2 節）
- 提案する UKCF と既存の UKF を用いた時のシミュレーション環境下の未知音源定位性能比較（4.2 節）

Figure 3 に使用した小型クアドロコプタを示す．小型クアドロコプタには加速度計，角速度計，地磁気計が全て搭載された Bitcraze 社の CrazyFlie を用いた．マイクは一つで小型クアドロコプタ中央下部に設置した．録音は 16kHz，16 bit で行い，音源強度計算のためのフレーム長は 512 とした．環境中の音源には指向性のないスピーカーを用い，音源毎に定常で周波数の異なるサイン波を流した．残響による性能劣化がないよう，実験には 3m × 4m の無響室を用いた．自己位置推定では，無響室にモーションキャプチャを敷設し，小型クアドロコプタにマーカーをつけて正解位置を計測して誤差を評価した．

4.1 UKF を用いた一次元自己位置推定性能

本節では自己位置推定の最も単純な場合である距離（一次元）推定の評価を行う．実験では，一台の小型クアドロコプタと一つの音源ランドマークを用いて，小型クアドロコプタを音源から 1 m の円周上を回るように移動させて距離推定性能を評価した．小型クアドロコプタを円周上に飛ばすことが困難であったことと，音源強度ベースの手法が実環境で正しく動作するかを確かめるため，小型クアドロコプタのプロペラを動かさない状態（プロペラ雑音がない状態）で手で円周上に動かして評価した．Figure 4 に自己位置推定結果を示す．図のように，提案法では，平均誤差 0.06 m 程度で累積誤差なく 1 m の距離を推定できていることがわかる．一方，9 軸モーションセンサ情報の

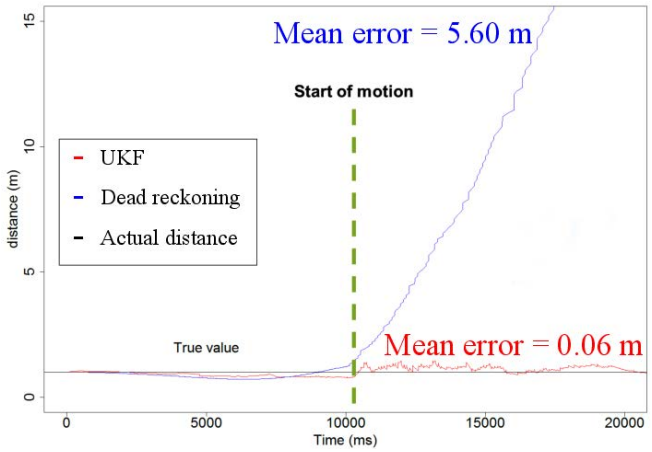


Figure 4: Error on distance estimation results

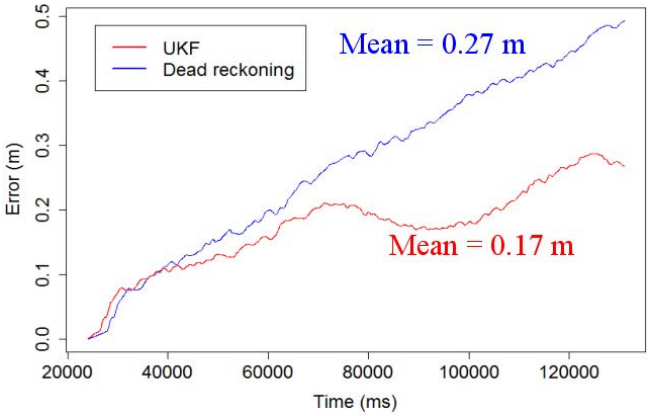


Figure 5: Error on 2D self-localization results

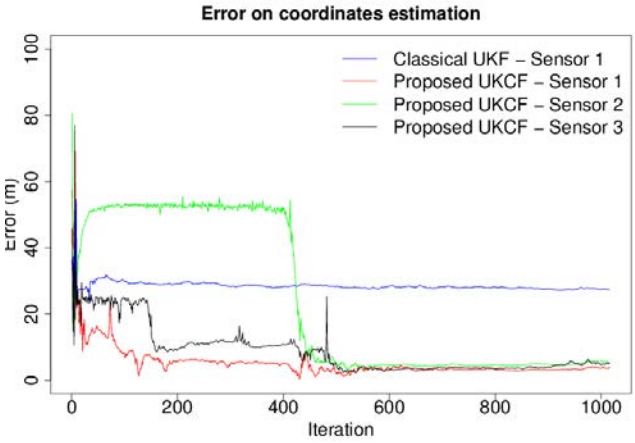


Figure 6: SSL Performance Comparison between UKF and UKCF

みを用いた Dead Reckoning による自己位置推定では，累積誤差が大きくなり，平均誤差 5.6 m となって発散していることがわかる．以上のことから，音源ランドマークの絶対位置情報と，モーションセンサの情報を両方使用した UKF によって累積誤差を軽減できていることの有効性が確認できた．

ただし，本実験では，クアドロコプタのモーターを動作しておらずプロペラ雑音がなかったこと，一つの音源ランドマークしか無かったため定常で周波数独立だという音源に対する仮定でも動作しやすい環境であったこと，距離のみの推定しかできなかったことなどから，まだ実環境ロバストとは言い難い．次節ではこれらを考慮した評価を行う．

4.2 UKF を用いた二次元自己位置推定性能

前節の評価を発展させ，音源ランドマークを 5 個に増やして二次元の自己位置推定性能を評価した．5 個の音源ランドマークは無響室の床面にランダムに配置し，音源の位置はモーションキャプチャで計測された正解位置を用いた．本実験では，小型クアドロコプタを以下の仮定のもとで実際に飛行させて評価した．具体的には，本実験は二分間程度小型クアドロコプタを飛行させ収録したデータを用いて評価はオフラインで行った．ただし，モーションセンサから得られる加速度の観測雑音が非常に大きかったため，モーションキャプチャの時系列データから加速度を計算して用いた．また，使用したスピーカは水平面上のみ無指向性を保証しており，高さ方向の推定が困難であったため，水平面上の二次元自己位置推定性能の評価とした．最後に，UKF での音源位置の初期値計算に必要な 1m の距離での各音源ランドマークに対する音源強度は未知であるため，音源毎にあらかじめ計測したものを用いた．

Figure 5 に二次元自己位置推定性能の比較を示す．前節と同様に Dead Reckoning との比較を示している．UKF を用いた場合は累積誤差が大きくなるものの，0.17 m と小さな誤差で定位を達成できているのに対し，Dead Reckoning による平均誤差は 0.27 m であり，音源ランドマークを用いた UKF による自己位置推定性能の向上が複数ランドマークを使った場合でも確認できた．

UKF による手法で累積誤差が大きくなったことには二つの原因が考えられる．第一に，完全な無指向性スピーカーを使用できず，音源強度が定常かつ距離の二乗に反比例するという仮定が水平面上のみでしか成立しなかったことである．実際には小型クアドロコプタは三次元空間上を飛行していたため，モデルとの相違があった．第二に，単一周波数の定常音に対する 1 m の音源強度が不均一になってしまったことである．単一周波数であったことから信号対雑音比も劣化してしまう場合が見受けられた．調波構造を持つ音など，観測しやすい音源の選択は今後検討の余地があると考えられる．

4.3 UKCF を用いた未知音源定位性能

最後に，未知音源定位の評価をシミュレーション環境下で行った．本実験では，一つの未知音源を三つの小型クアドロコプタを用いて位置推定することを想定した．音源位

置と小型クアドロコプタ位置の初期値は誤差つきで与え，UKCF によって各クアドロコプタが音源位置の推定状態を共有して音源位置を推定できるかを検証した．情報を共有しない場合は UKF に相当するため，UKF との比較を行った．小型クアドロコプタ位置の初期誤差は 70.7 m とし，UKCF と UKF で同じ初期値を用いた．UKCF と UKF 共に観測誤差である音源強度に対するガウス雑音の分散は 20 m とした．

Figure 6 に比較結果を示す．図のように，三台全てのクアドロコプタに対して UKCF では 4.5 m まで誤差が収束していることがわかる．一方で UKF では，30 m と大きな誤差で収束していることが確認できる．よって，提案法の UKCF が既存法に対して，大きな初期誤差を持つ上に高雑音下の状況であっても精度良く音源定位できることが確認できた．さらに，初期誤差を大きくし，観測音源強度に対する雑音を大きく設定して，シミュレーションしたことで，三台のクアドロコプタは最初音源定位のコンセンサスを取れていなかったが，状態更新するに従って，およそ 500 回のイタレーション（30.75 秒）でコンセンサスを達成できた．このことから，UKCF が平衡点で安定していることを数値計算でも確認できた．

複数の小型クアドロコプタが自由に飛行していれば，遠方の未知音源や音響的オクルージョンを扱う必要があるため，各クアドロコプタに対する未知音源の観測強度は必ずしも信頼できない．例えば，Figure 6 の Sensor 2 は最初の 400 フレームで推定誤差が増加していることから観測強度に誤差があったことが考えられる．そのような不確かな観測であっても，他の信頼できる観測を持つ個体の情報を利用することで最終的に Sensor 2 も小さな推定誤差を達成できたことから，コンセンサスの有効性を分散システム全体から見ても確認することができた．

5 結論

本稿では，小型クアドロコプタの群を用いた自己位置推定と未知音源位置推定について述べた．小型クアドロコプタではペイロードが小さいことからマイクが一つしか搭載できない場合を考え，各クアドロコプタが分散して自己位置と未知音源位置を推定しつつ，全個体が共通して推定している未知音源位置をコンセンサスの概念を導入することで情報統合する手法を提案した．自己位置推定には 環境中にある音源ランドマークを用いた UKF を，コンセンサスに基づく未知音源推定には UKCF を提案した．評価では，様々な仮定を置いたものの，自己位置推定と未知音源位置推定それぞれにおいて，実環境下もしくはシミュレーション環境下において有効性を確認することができた．本稿の手法はモデルに多くの仮定がある上，限られた環境での評価に留まっている．実践的に使えるような技術にするには，音源ランドマークが移動する，定常でない，指向性を持つ，周波数的に独立でない場合等の環境に対する仮定を緩和すること，および，三次元でも推定できること，残響環境下でも推定できるなどのモデルに対する仮定を緩和することなど，多くの課題を抱えている．これらの仮定や課題を解決しつつ，さらには群の形成法や移動法，個体が音源定位するための最適な運動計画など，群を積極的に利用した技術革新も今後の課題である．

参考文献

[1] F. Wang *et al.*, “A mono-camera and scanning laser range finder based UAV indoor navigation system”, in *Proc. of IEEE ICUAS*, pp. 694–701, 2013.

[2] A. Kushleyev *et al.*, “Towards a swarm of agile micro quadrotors”, in *Autonomous Robots*, vol. 35, no. 4, pp. 287–300, 2013.

[3] M. Basiri *et al.*, “Robust acoustic source localization of emergency signals from micro air vehicles”, in *Proc. of IEEE/RSJ IROS*, pp. 4737–4742, 2012.

[4] M. Basiri *et al.*, “Audio-based Positioning and Target Localization for Swarms of Micro Aerial Vehicles”, in *Proc. of IEEE ICRA*, pp. 4729–4734, 2014.

[5] K. Okutani *et al.*, “Outdoor auditory scene analysis using a moving microphone array embedded in a quadrocopter”, in *Proc. of IEEE/RSJ IROS*, pp. 3288–3293, 2012.

[6] T. Ohata *et al.*, “Improvement in outdoor sound source detection using a quadrotor-embedded microphone array”, in *Proc. of IEEE/RSJ IROS*, pp. 1902–1907, 2014.

[7] H. Miura *et al.*, “SLAM-based Online Calibration for Asynchronous Microphone Array”, in Advanced Robotics, vol. 26, pp. 1941–1965.

[8] E. A. Wan *et al.*, “The unscented Kalman filter for nonlinear estimation”, in *Proc. of IEEE AS-SPCC*, pp. 153–158, 2000.

[9] R. Olfati-Saber, “Kalman-consensus filter: Optimality, stability, and performance”, in *Proc. of IEEE CDC/CCC*, pp. 7036–7042, 2009.

[10] L. Sinapayen *et al.*, “Sound Source Localization with an Autonomous Swarm of Quadrocopters”, in *Proc. of IEEE/RSJ IROS Workshop on Modular and Swarm Systems*, 2014.

複数移動ロボットによる協調音源分離のための分離精度予測を用いた配置最適化

Layout Optimization of Multiple Mobile Robots for Cooperative Sound Source Separation by Predicting Source Separation Performance

関口航平, 坂東昭宜, 糸山 克寿, 吉井 和佳

Kouhei Sekiguchi, Yoshiaki Bando, Katsutoshi Itoyama, Kazuyoshi Yoshii

京都大学 大学院情報学研究科

Graduate School of Informatics, Kyoto University

sekiguch@kuis.kyoto-u.ac.jp

Abstract

本稿では，複数音源が存在する状況において，注目したい音源を高精度に分離することを目的として，マイクロホンアレイを搭載した複数移動ロボットの配置を最適化する手法について述べる．音源分離はマイクロホンアレイを搭載した全てのロボットを 1 つのマイクロホンアレイとみなして行う．音源分離の精度はロボットと音源の位置関係に依存するため，ロボットを最適な配置に移動させることで音源分離性能の向上を行うことができる．しかし，音源分離に最適な複数台のロボットの配置は自明ではない．本研究では，ロボット配置の分離精度を事前予測して，複数台のロボットの配置を最適化する．分離精度の予測値は，瞬時混合モデル上での混合行列と分離行列から計算する．実験では，提案法によりランダムな場合に比べて SDR が最大 8.6 dB 向上することを確認した．さらに，各ロボットで独立に分離音を生成してから統合する場合よりも提案法での分離精度が高くなることを確認した．

1 はじめに

近年の通信技術の発達に伴い遠隔地とのコミュニケーションを行う様々な手段が開発されている．その一つがテレプレゼンスロボットである．テレプレゼンスロボットとは，移動機構にカメラやマイクロホンを搭載したロボットで，遠隔地にいる操縦者がまるで現地にいるかのようにコミュニケーションを行うことを可能にする．例えば，在宅勤務者が自宅から社内の人とコミュニケーションをとるなどの目的でテレプレゼンスロボットが使用されている [Ng 15, Berri 14, Yan 13].

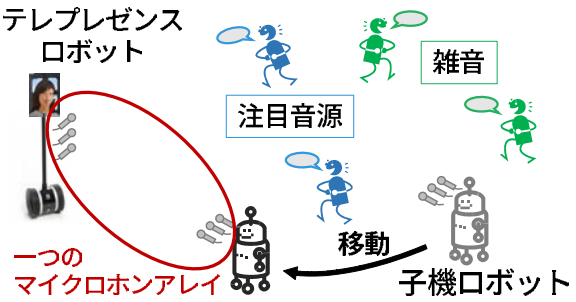


図 1: 複数ロボットの配置最適化の一例

テレプレゼンスロボットを用いて遠隔地とのコミュニケーションを円滑に行うためには，目的音以外の雑音への対策が不可欠となる．一般に，実際の環境では他人の話し声や音楽，空調機などの様々な雑音が存在しており，操縦者が実際に聞く音は複数の音を含む混合音となり，目的音の認識が困難になる．このような状況に対処するため，マイクロホンアレイ処理を用いて混合音を各音源信号に分離する研究が行われている [Makino 05, Lee 07, Nakajima 10]．水本らは音源分離を用いて，操縦者が指定した方向の音だけを聞くテレプレゼンスロボットを開発した [Mizumoto 11].

音源分離の精度はマイクロホンと音源の位置関係に依存し，音源とロボットの位置関係によっては分離が困難となる問題がある [Nakadai 02]．例えば，ロボットから見て複数の音源が同一方向に存在する場合や，音源間の距離差が大きい場合などである．注目音源が一つならば音源に近づくことで聞きやすくなるが，目的音源が複数存在する場合には，最適なロボットの配置は自明ではない．

本研究では，複数の子機ロボットを用いた音源分離支援システムの開発を行う (図 1)．操縦者が聞きたい音源を指定することで，その音源の配置に応じてマイクロホンアレイを搭載した子機ロボットが適切な位置に移動し，分離精度を向上させる．このとき，複数のロボットに搭載されたマイクロホンアレイ全体を一つの大きなマイクロホンアレイとみなし，すべてのマイクロホンでの観測音を

用いて音源分離を行う．ロボットの最適配置は音源分離精度を予測して決定する．各ロボット配置での実際の音源分離精度は，もとの音源信号が未知であるため計算することができない．そこで，瞬時混合モデルを仮定し，音源とロボットの位置関係を用いて音源分離をシミュレーションすることで音源分離精度の予測を行う．実験では，ランダムな配置と提案法による最適配置での分離精度の比較と，各ロボットで独立に分離音を生成して統合した場合と，複数のロボット全体を一つのマイクロホンアレイとみなして分離音を生成した場合の比較を行った．

2　音源分離に最適なロボット配置の探索

音源が複数存在する環境において，複数ロボットの配置を最適化することで，目的音源を高精度に分離する手法について述べる．本研究の課題は，音源分離に最適な複数ロボットの配置が自明ではないことである．ある配置で実際に音を録音して音源分離を行っても，元信号がないため分離音から分離精度を計算できず，最適配置を探索することができない．したがって，複数ロボットの最適配置探索には，実際に音源分離を行わず各ロボット配置での音源分離精度を予測することが必要となる．本研究では，音源分離に GICA や GHDSS[Nakajima 10] などの幾何制約付きブラインド音源分離手法を使用する．この手法は分離性能や環境適応性が高く計算量も少ないため，実時間での動作が望まれるロボット聴覚に適した手法である．一方，この手法では音源分離精度の予測が困難であるという問題がある．そこで，GICA や GHDSS と分離精度について相関のある遅延和ビームフォーミング（DSBF[Johnson 92]）の利得 (図 2) を用いて音源分離精度の推定を行う．利得とは分離音中に含まれる目的音と雑音の比率であり，分離音と音源信号の関係から計算することができる．この関係は音の混合過程と分離過程を推定することによって求めることができる．注目音源についての利得を用いた評価関数により遺伝的アルゴリズムで最適配置を決定する．

本稿で扱う配置最適化問題を以下のように定める．

入力	$\boldsymbol{X}_t = [\boldsymbol{x}_{t1}, \cdots, \boldsymbol{x}_{tM}]^T \in \mathbb{C}^{M\times F}$ N 個の音源が混合した M チャンネル観測音
出力	(1) $\boldsymbol{Y}_t = [\boldsymbol{y}_{t1}, \cdots, \boldsymbol{y}_{tN'}]^T \in \mathbb{C}^{N'\times F}$ 注目している N' 個の音源の分離音 (2) $\boldsymbol{B}^* = [\boldsymbol{b}_1^*, \cdots, \boldsymbol{b}_R^*] \in \mathbb{R}^{R\times 2}$ R 台のロボットの最適配置の座標
仮定	(1) 各マイクロホンはすべて同期済み (2) N 個の音源座標 $\boldsymbol{C}=[\boldsymbol{c}_1, \cdots, \boldsymbol{c}_N] \in \mathbb{R}^{N\times 2}$ は音源定位と三角測量により既知 [Sasaki 06]

ここで，音源の総数を N とし，そのうち注目する音源の数を N' と定める．$\boldsymbol{X}_t, \boldsymbol{Y}_t$ はそれぞれ，録音した音響信号，分離音の t フレーム目を短時間フーリエ変換して得る．F は周波数ビンの数を表し，$\boldsymbol{x}_{tm} = [x_{tm1}, \cdots, x_{tmF}]$，$\boldsymbol{y}_{tm} = [y_{tm1}, \cdots, y_{tmF}]$ である．

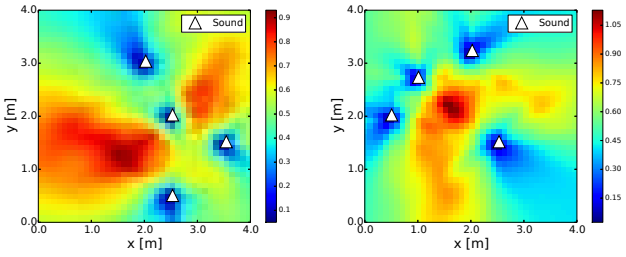


図 2: 1 台のロボットを部屋の各点に配置した場合の利得の一例．三角は音源位置を表し，値が大きい位置ほど分離精度が高くなると予想される．

マイクロホンアレイの配置最適化の関連研究には Martinson らの手法 [Martinson 11] と佐々木らの手法 [Sasaki 11] がある．前者は 1 チャンネルマイクロホンを搭載した複数のロボットを用いる．音源の配置から音源定位に最適なロボット配置を幾何的に決定する．後者はマイクロホンアレイを搭載した 1 台のロボットを用いる．DSBF の利得を用いて，すべての方向に対して高い分離精度をもつ，音源配置によらないマイクロホンアレイの最適配置を探索する．

2.1　音の混合過程

音源信号 $\boldsymbol{S}_t = [\boldsymbol{s}_{t1}, \cdots, \boldsymbol{s}_{tN}] \in \mathbb{C}^{N\times F}$ とマイクロホンによる観測音 $\boldsymbol{X}_t$ の関係について述べる．ここで，$\boldsymbol{s}_{ti}$ は音源 i の t フレーム目の音源信号の短時間フーリエ変換を表す．音の伝搬を線形時不変システムと仮定すると，音源信号と観測音の関係は以下のように表される．

$$\boldsymbol{x}_{t\cdot f} = \boldsymbol{H}_f \boldsymbol{s}_{t\cdot f} \tag{1}$$

ここで，$\boldsymbol{x}_{t\cdot f} = [x_{t1f}, \cdots, x_{tMf}]^T \in \mathbb{C}^M$，$\boldsymbol{s}_{t\cdot f} = [s_{t1f}, \cdots, s_{tNf}]^T \in \mathbb{C}^N$ であり，$\boldsymbol{H}_f \in \mathbb{C}^{M\times N}$ は混合行列である．雑音と残響を考慮せず，音の距離減衰と到達時間差のみを考慮した場合，x_{tmf} と s_{tnf} の関係は次のように表される．

$$x_{tmf} = \sum_{n=1}^{N} \frac{1}{d_{nm}} s_{tnf} \exp(-j2\pi f \tau_{nm}) \tag{2}$$

ここで，d_{nm} は音源 n とマイクロホン m の間の距離を表し，τ_{nm} は音源 n のマイクロホン m への到達時間を表し，$\tau_{nm} = d_{nm}/c$（c は音速）で計算される．音の振幅は距離に反比例するため，$1/d_{nm}$ の項は距離減衰を表す．式 (1) と式 (2) を比較すると，混合行列 $\boldsymbol{H}_f$ の (m,n) 成分 h_{mnf} は以下のように表される．

$$h_{mnf} = \frac{1}{d_{nm}} \exp(-j2\pi f \tau_{nm}) \tag{3}$$

2.2　音源分離

マイクロホンでの観測音 $\boldsymbol{x}(t)$ と分離音 $\boldsymbol{y}(t)$ の関係について述べる．音の混合過程と同様に，音源分離が線形時不変

システムで表されると仮定すると，観測音と分離音の関係は以下の式で表される．

$$\boldsymbol{y}_{t\cdot f} = \boldsymbol{W}_f \boldsymbol{x}_{t\cdot f} \tag{4}$$

ここで，$\boldsymbol{y}_{t\cdot f} = [y_{t1f}, \cdots, y_{tNf}]^T \in \mathbb{C}^N$ であり，，$\boldsymbol{W}_f \in \mathbb{C}^{N\times M}$ は分離行列を表す．式 (1) と式 (4) から，$\boldsymbol{W}_f = \boldsymbol{H}_f^{-1}$ のとき，$\boldsymbol{y}_{t\cdot f} = \boldsymbol{W}_f \boldsymbol{x}_{t\cdot f} = \boldsymbol{W}_f \boldsymbol{H}_f \boldsymbol{S}_{t\cdot f} = \boldsymbol{S}_{t\cdot f}$ となり，分離音は音源信号と等しくなる．

GICA は ICA を基にした手法であり，音源信号の独立性を過程して，分離音が独立となるような分離行列 $\boldsymbol{W}$ を推定する．分離行列 $\boldsymbol{W}$ を推定するために，以下の二つのコスト関数を用いる．

$$\boldsymbol{J}_{\mathrm{ICA}}(\boldsymbol{W}) = \int p(\boldsymbol{y}) \log \frac{p(\boldsymbol{y})}{\prod_k p(y_k)} \tag{5}$$

$$\boldsymbol{J}_{\mathrm{GC}}(\boldsymbol{W}) = \|\boldsymbol{W}\boldsymbol{H} - \boldsymbol{I}\|^2 \tag{6}$$

ただし，$p(\boldsymbol{y}) = p(y_1, \cdots, y_N)$ である．$\boldsymbol{J}_{\mathrm{ICA}}(\boldsymbol{W})$ は $p(\boldsymbol{y})$ と $\prod_k p(y_k)$ の KL-divergence であり，独立性の尺度となっている．$\boldsymbol{J}_{\mathrm{GC}}$ は幾何制約を表す．実際の環境での混合行列 $\boldsymbol{H}$ は未知であるため，ここで与える $\boldsymbol{H}$ はあらかじめ録音したインパルス応答や幾何的に計算したインパルス応答から作成する．本研究ではリアルタイムで音源分離を行うために以下の更新式を用いて，逐次的に分離行列 $\boldsymbol{W}$ を推定する．

$$\boldsymbol{W}_{t+1} = \boldsymbol{W}_t - \alpha \boldsymbol{J}'_{\mathrm{ICA}} - \beta \boldsymbol{J}'_{\mathrm{GC}} \tag{7}$$

ただし，α, β はステップサイズパラメータであり，$\boldsymbol{J}'_{\mathrm{ICA}} = \nabla_{\boldsymbol{W}^*} \boldsymbol{J}_{\mathrm{ICA}}$, $\boldsymbol{J}'_{\mathrm{GC}} = \nabla_{\boldsymbol{W}^*} \boldsymbol{J}_{\mathrm{GC}}$ である．$\{\}^*$ は複素共役を，∇ は微分作用素を表す．GHDSS は GICA と類似した手法であり，分離行列 $\boldsymbol{W}$ を推定するために，$\boldsymbol{J}_{\mathrm{ICA}}$ の代わりに以下で定める $\boldsymbol{J}_{\mathrm{HDSS}}$ を用いる．

$$\boldsymbol{J}_{\mathrm{HDSS}} = \|E[\mathbf{E}_\phi]\|^2 \tag{8}$$

ただし，

$$\mathbf{E}_\phi = \phi(\mathbf{y})\mathbf{y}^H - \mathrm{diag}[\phi(\mathbf{y})\mathbf{y}^H] \tag{9}$$

$$\phi(y_i) = -\frac{\partial}{\partial y_i} \log p(y_i) \tag{10}$$

GICA や GHDSS では観測音などに応じて分離行列 $\boldsymbol{W}$ が異なるため，事前に分離行列を推定することが困難である．したがって，これらの手法を用いた場合，利得を計算することができないという問題がある．

本研究では遅延和ビームフォーミング (DSBF) という手法に注目する．DSBF と GICA や GHDSS の分離精度には高い正の相関があるため，本研究では，この相関を利用して DSBF の利得を使って GICA や GHDSS での分離精度の予測を行う．音源分離手法として DSBF を用いた場合，分離行列 $\boldsymbol{W}_f$ の要素はマイクロホンと音源の位置関係のみから決定される．DSBF とは注目音源の座標から各マイクへの到達時間差を推定し，観測信号を到達時間差だけ時間シフトして足し合わせることにより注目音を強調する音源分離手法である．本研究では，各マイクロホンと音源の距離を考慮し，音源に近いマイクロホンの観測音の比率を高く，音源と遠いマイクロホンの観測音の比率を低くして足し合わせる．したがって，分離音と観測音の関係は周波数領域では次のように表される．

$$y_{tnf} = \sum_m \frac{1}{d_{nm}} x_{tmf} \exp(j2\pi f \tau_{nm}) \tag{11}$$

式 (4) と式 (11) から，分離行列 $\boldsymbol{W}_f$ の (n, m) 成分 w_{nmf} は以下の式で表される．

$$w_{nmf} = \frac{1}{d_{nm}} \exp(j2\pi f \tau_{nm}) \tag{12}$$

2.3 目的関数

複数ロボットの配置最適化における目的関数を DSBF の利得の調和平均として定める．ロボット配置 $\boldsymbol{B} = [\boldsymbol{b}_1, \cdots, \boldsymbol{b}_R]$ における目的関数の値を $f(\boldsymbol{B})$ とすると，

$$f(\boldsymbol{B}) = \frac{N'}{\sum_{n\in D} \frac{1}{g_n(\boldsymbol{B})}} \tag{13}$$

ここで，D は注目音源の集合を表す．$g_n(\boldsymbol{B})$ は音源 n の利得を表し，音源 n の分離音中の音源 n と雑音の比率として定める．利得の調和平均を目的関数としたのは，本研究ではすべての注目音源の高精度な分離を目的とするためである．もし一つでも分離精度が悪い音源が存在する場合，目的関数の値は大きく低下する．

利得は分離音と音源信号の関係式から計算することが可能である．式 (1) と式 (4) から，分離音と音源信号の関係は周波数領域で以下のように表される．

$$\boldsymbol{y}_{t\cdot f} = \boldsymbol{A}_f \boldsymbol{s}_{t\cdot f} \tag{14}$$

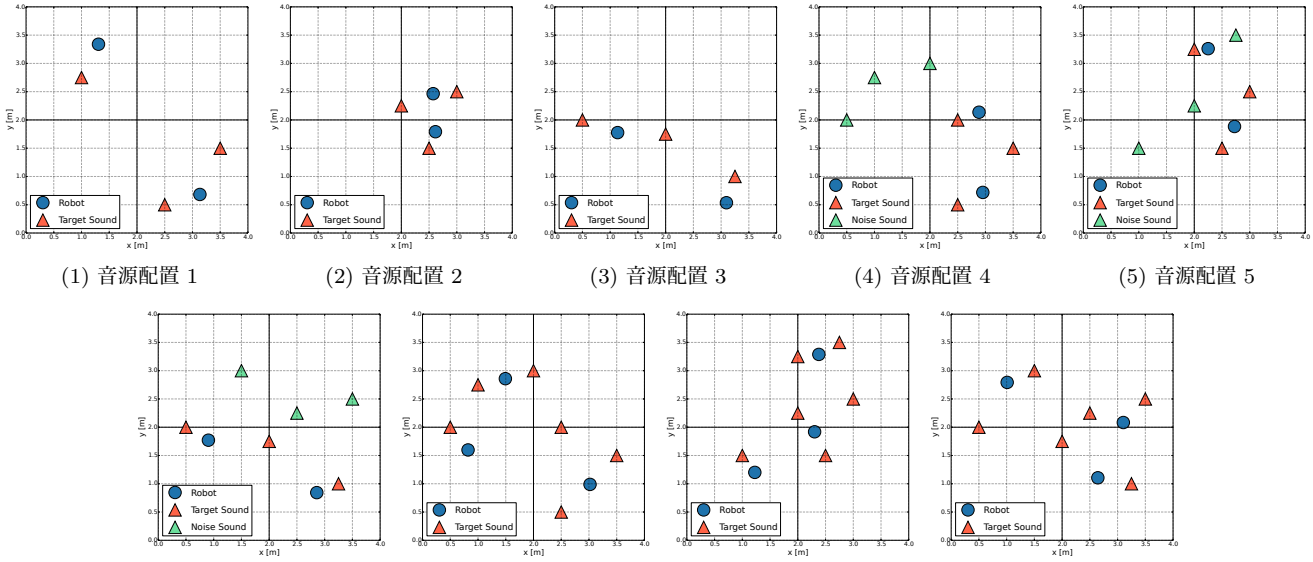
ここで，$\boldsymbol{A}_f \in \mathbb{C}^{N\times N}$ は利得行列であり，$\boldsymbol{A}_f = \boldsymbol{W}_f \boldsymbol{H}_f$ として定める．利得行列 $\boldsymbol{A}_f$ の対角成分は分離音中に含まれる目的音源の比率を，非対角成分は雑音の比率を表している．したがって，音源 n の利得 $g_n(\boldsymbol{B})$ は以下のようになる．

$$g_n(\boldsymbol{B}) = \frac{\sum_f a_{nnf}}{\sum_{n\neq k} \sum_f a_{nkf}} \tag{15}$$

ここで，a_{nkf} は利得行列 $\boldsymbol{A}_f$ の (n, k) 成分であり，音源 n の分離音に含まれる音源 k の割合を示す．

DSBF の利得を用いた場合には，式 (11) と式 (2) から a_{nkf} は以下のようになる．

$$a_{nkf} = \left| \sum_{m=1}^{M} \frac{1}{d_{nm} d_{km}} \exp(j2\pi f(\tau_{nm} - \tau_{km})) \right| \tag{16}$$



(1) 音源配置 1　(2) 音源配置 2　(3) 音源配置 3　(4) 音源配置 4　(5) 音源配置 5

(6) 音源配置 6　(7) 音源配置 7　(8) 音源配置 8　(9) 音源配置 9

図 3: 音源配置と提案法で求めた複数ロボットの最適配置．青丸がロボット，赤三角が注目音源，緑三角が雑音を表す．

2.4　最適配置探索

本研究では遺伝的アルゴリズムを用いて最適配置探索を行う．これは，グリッドサーチによる全探索を用いた場合，ロボットの台数に指数的に比例して計算量が増加してしまい，また，勾配法などを用いた場合には局所最適解に陥ってしまう可能性があるためである．複数ロボットの座標と方向の組を 1 つの個体とみなし，個体の組み替えは現在位置の近傍へ移動することで行い，突然変異によりランダムに移動することで局所最適解に陥ることを防ぐ．ロボットの向きはランダムに与える．ただし，ロボット間の距離が離れすぎた場合，1 つの時間フレーム内に含まれる音源信号の区間がロボット間で大きく異なってしまうため，分離精度が低下してしまう問題がある．そのため，ロボット間の距離が一定距離以内に収まるように制約を設ける．選択はエリート選択とルーレット選択を併用する．目的関数は複数台のロボットに搭載した全てのマイクを 1 つのマイクロホンアレイとみなして計算を行う．世代交代を一定回数行った後，評価関数の値が最大となる個体を最適配置とする．

3　評価実験

ランダムな配置と提案法による最適配置の比較と，複数ロボット全体を一つのマイクロホンアレイとみなして音源分離する場合と各ロボットで独立に分離音を生成して統合した場合を比較するために，シミュレーション混合を用いた評価実験を行った．

3.1　実験条件

一辺 4 m の正方形の部屋に音源 N 個，ロボット R 台がある場合を想定する．N 個の音源の中で，N' 個の音源を注目したい音源，$N-N'$ 個の音源を雑音とみなす．以下の 3 つの条件それぞれについて 3 種類の音源配置を用意し実験を行った (図 3).

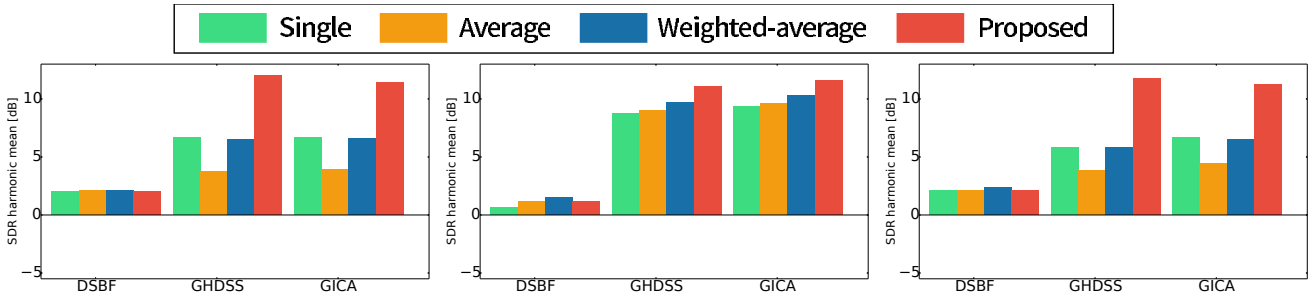
1. 注目音源 3 つ，ロボット 2 台 ($N = N' = 3, R = 2$)
2. 注目音源 3 つ，雑音 3 つ，ロボット 2 台 ($N = 6, N' = 3, R = 2$)
3. 注目音源 6 つ，ロボット 3 台 ($N = N' = 6, R = 3$)

各ロボットは 8 チャンネルマイクロホンアレイを搭載し，ロボットの配置を提案法を用いて最適化する．最適配置での観測音は幾何学的に計算したインパルス応答を使ったシミュレーション混合を用いて生成する．音源信号は JNAS 音素バランス文 (量子化 24 bit, サンプリング周波数 16 kHz)[Sagisaka 92] を用いた．

ランダムな配置と，提案法での最適配置で，複数のロボット全体を一つのマイクロホンアレイとみなして音源分離を行った場合と，最適配置で各ロボットで独立に分離音を生成して統合した場合の比較を行った．音源分離手法には DSBF, GHDSS, GICA の 3 つを用いた．GICA と GHDSS を用いて各ロボットで独立に分離音を生成した場合，各分離音の位相がわからないという問題がある．そのため，分離音をそのまま足しただけでは，位相が合わず分離精度が悪くなってしまう可能性がある．この問題を考慮して，分離音の統合方法として以下の 3 つの手法を用いた．

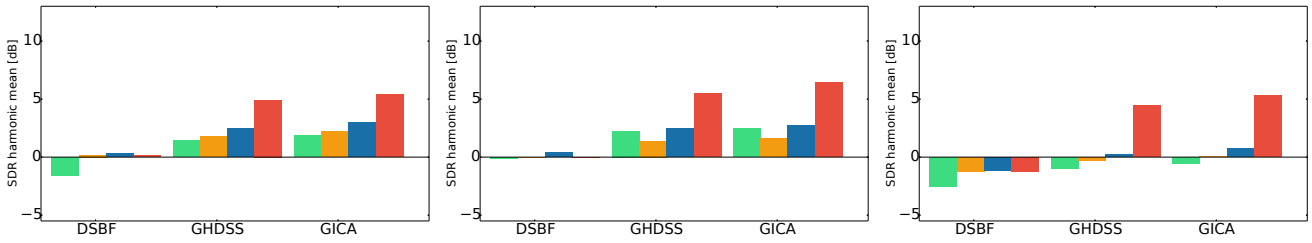
(a) Single：各音源に最も近いロボットでの観測音のみを用いて分離音を生成する．

(b) Average：ロボット R 台の場合，$R-1$ 個の分離音をそれぞれ $0.5n$ サンプルずらして加算し，分離精度を計算する ($n = 0, 1, \cdots, 20$)．n^{R-1} の組み合わせの中で最も分離精度が高くなったものを，分離音として出力する．



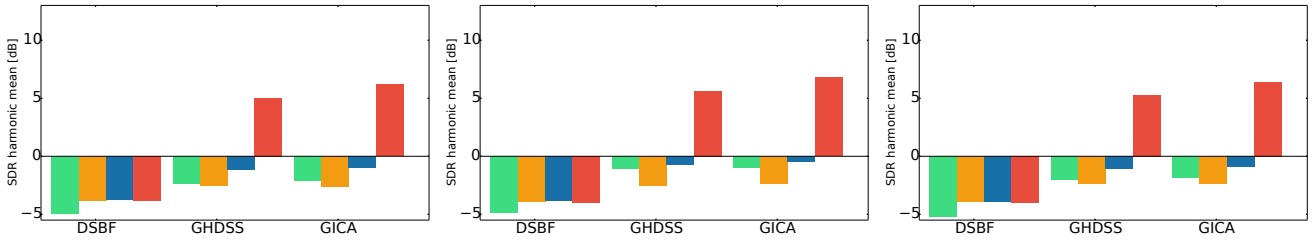
(1) 音源配置 1　(2) 音源配置 2　(3) 音源配置 3

図 4: ロボット数：2，注目音源数：3，雑音数：0 の場合の SDR の調和平均



(4) 音源配置 4　(5) 音源配置 5　(6) 音源配置 6

図 5: ロボット数：2，注目音源数：3，雑音数：3 の場合の SDR の調和平均



(7) 音源配置 7　(8) 音源配置 8　(9) 音源配置 9

図 6: ロボット数：3，注目音源数：6，雑音数：0 の場合の SDR の調和平均

(c) Weighted-average：音源とロボットの距離で重み付けして (b) を行う．

音源分離精度の指標には sound-to-distortion ration(SDR)[Vincent 06, Raffel 14] の注目音源についての調和平均を用いた．調和平均を使用したのは，目的関数で利得の調和平均を用いたのと同様，本研究では全ての注目音源を高精度に分離することを目的としており，調和平均は一つでも分離精度が低い音源が存在すると値が大きく下がるためである．最適配置探索は遺伝的アルゴリズムを用いることによりランダム性を持つため，各音源配置について提案法による最適化を 30 回行い，各ロボット配置について注目音源の SDR の調和平均を求め，その平均を計算した．

3.2 実験結果

ランダムな配置と提案法による最適配置での分離精度を表 1 に示す．すべての場合において，提案法による最適配置がランダムな配置を上回っている．提案法による最適配置で，複数ロボットを 1 つのマイクロホンアレイと見なす場合と，各ロボットで独立に分離を行う場合の実験結果を図 4，図 5，図 6 に示す．全ての条件で，DSBF を用いた場合よりも GHDSS や GICA を用いた場合の分離精度が上回っている．これは，DSBF では雑音によらず注目音源の位相に合わせて信号をずらして足し合わせるだけであるが，GHDSS や GICA では，雑音方向に Null ビームを形成することにより，雑音を消すことができるためである．

分離	配置	1	2	3	4	5	6	7	8	9
DSBF	ランダム	-2.7	-2.5	-2.5	-6.3	-6.6	-6.3	-8.7	-8.8	-9.5
	最適位置	2.1	1.2	2.1	0.2	-0.1	-1.3	-3.8	-4.0	-4.0
GHDSS	ランダム	4.5	2.4	4.4	-3.0	-3.8	-3.5	-3.5	-5.5	-4.8
	最適位置	12.0	11.0	11.8	4.9	5.5	4.4	5.0	5.6	5.2
GICA	ランダム	5.0	4.4	5.5	-1.8	-2.5	-2.4	-1.2	-2.6	-1.8
	最適位置	11.4	11.5	11.2	5.4	6.4	5.3	6.2	6.8	6.4

表 1: ランダムな配置と提案法による最適配置での分離精度 (SDR [dB])

全てのマイクロホンを同時に用いる提案法と，各ロボットごとに分離音を生成する Single, Average,Weighted-average を比較する．DSBF を用いた場合には，提案法と Average が同一となり，提案法と比べて Single は平均で 0.78 dB 低く，Weighted-average は平均で 0.16 dB 高くなり，ほとんど差がなかった．提案法と Average が同一となったのは，Average が行っているロボットごとに平均を取り，さらにロボット間で平均を取る操作は，全てのロボットで一度に平均を取る操作と同じためである．一方，GHDSS や GICA では，提案法が他の手法を平均で 5.2 dB

以上上回る結果となった．これは，GICA や GHDSS では，目的音源をほとんど含まない観測音も Null ビームを形成するために使うことができるためである．各ロボットで独立に分離音を生成した場合，目的音源をほとんど含まない観測音を有効に使うことができず，また，この観測音から生成される分離音は分離精度が悪くなるため，その音を足すことで全体の分離精度も下がってしまう場合があると考えられる．

今回試した手法以外にも複数の音源を統合する手法が考えられるが，GHDSS と GICA では平均で 5.2 dB 以上他の手法を上回り，また，実際には各ロボットで分離音を生成する場合，統合の際に位相のずれの問題があるため分離精度が更に低下することが予想される．これらのことから，提案法の有効性が確認できたと言える．

3.3 まとめ

本稿では，複数の音源が存在する状況において，注目したい音源に応じて複数ロボットの配置を最適化することで，音源分離精度の向上を行う手法を開発した．音源分離は複数のロボット全体を一つの大きなマイクロホンアレイとみなして行った．複数ロボットの最適配置はロボットと音源の位置関係から分離精度を予測することで決定した．実験では，提案法によりランダムな場合に比べて SDR が最大 8.6 dB 向上することを確認した．さらに，各ロボットで独立に分離音を生成してから統合する場合よりも提案法での分離精度が高くなることを確認した．今後は，実環境で複数ロボットを用いて音源分離を行う際に問題となるロボット間の同期に取り組む予定である．

謝辞 本研究の一部は，科研費 24220006，および ImPACT「タフ・ロボティクス・チャレンジ」の支援を受けた．

参考文献

[Berri 14] Berri, R., et al.: Telepresence Robot with Image-based Face Tracking and 3D Perception with Human Gesture Interface using Kinect Sensor, in *JCRIS*, pp. 205–210 (2014)

[Johnson 92] Johnson, D. H. and Dudgeon, D. E.: *Array Signal Precessing: Concepts and Techniques*, Prentice Hall (1992)

[Lee 07] Lee, I., et al.: Fast Fixed-point Independent Vector Analysis Algorithms for Convolutive Blind Source Separation, *J. Signal Processing*, Vol. 87, No. 8, pp. 1859–1871 (2007)

[Makino 05] Makino, S., et al.: Blind Source Separation of Convolutive Mixtures of Speech in Frequency Domain, *IEEE Trans. Fundamentals of Electronics, Communications and Computer Sciences*, Vol. 88, pp. 1640–1655 (2005)

[Martinson 11] Martinson, E., et al.: Optimizing a Reconfigurable Robotic Microphone Array, in *IEEE/RSJ IROS*, pp. 125–130 (2011)

[Mizumoto 11] Mizumoto, T., et al.: Design and Implementation of Selectable Sound Source Separation on the Texai Telepresence System using HARK, in *IEEE ICRA*, pp. 2130–2137 (2011)

[Nakadai 02] Nakadai, K., et al.: Real-Time Sound Source Localization and Separation for Robot Audition, in *ICSLP*, pp. 193–196 (2002)

[Nakajima 10] Nakajima, H., et al.: Blind Source Separation With Parameter-Free Adaptive Step-Size Method for Robot Audition, *IEEE Trans. Audio, Speech and Language Processing*, Vol. 18, No. 6, pp. 1476–1485 (2010)

[Ng 15] Ng, M. K., et al.: A cloud robotics system for telepresence enabling mobility impaired people to enjoy the whole museum experience, in *IEEE DTIS*, pp. 1–6 (2015)

[Raffel 14] Raffel, C., et al.: mir_eval: A Transparent Implementation of Common MIR Metrics, in *ISMIR*, pp. 367–372 (2014)

[Sagisaka 92] Sagisaka, Y. and Uratani, N.: ATR Spoken Language Database, *J. The Acoustic Society of Japan*, Vol. 48, No. 12, pp. 878–882 (1992)

[Sasaki 06] Sasaki, Y., et al.: Multiple Sound Source Mapping for a Mobile Robot by Self-motion Triangulation, in *IEEE/RSJ IROS*, pp. 380–385 (2006)

[Sasaki 11] Sasaki, Y., et al.: 32-Channel Omni-Directional Microphone Array Design and Implementation, *J. Robotics and Mechatronics*, Vol. 23, No. 3, pp. 378–385 (2011)

[Vincent 06] Vincent, E., et al.: Performance Measurement in Blind Audio Source Separation, *IEEE Trans. Audio, Speech and Language Processing*, Vol. 14, No. 4, pp. 1462–1469 (2006)

[Yan 13] Yan, R., et al.: An Attention-Directed Robot for Social Telepresence, in *HAI*, pp. III–1–2 (2013)

ビッグデータ解析とクラウドソーシング
Big data analysis and crowdsourcing

鹿島 久嗣
Hisashi KASHIMA
京都大学
Kyoto University

kashima@i.kyoto-u.ac.jp

Abstract

機械学習をはじめとするデータ解析技術の進歩が実世界において様々なブレークスルーを起こしている一方で、ビッグデータの解析や処理のプロセスはいまだ極めて労働集約的であり、これらを行う人手をいかに調達するかが重要な課題である。この人的ボトルネックの問題を解消するための有望なアプローチの一つとしてクラウドソーシングの考え方が注目されている。クラウドソーシングを利用して人間による判断や処理をプロセスに組み込むことによって、機械だけでは解決できない、いわゆる「データの外側」 を人間の知識や判断によって補うことが可能となる。本講演ではビッグデータ解析・処理をクラウドソーシングで実現するための要素技術となる品質保証技術、クラウドソーシングで収集したデータからの機械学習、クラウドソーシングを利用したデータモデリング事例などを紹介するとともに、セキュリティやプライバシ、人間と機械の協働問題解決といった今後の課題についても述べる。

1 ビッグデータ解析のボトルネック：人材不足

近年、機械学習をはじめとするデータ解析技術は様々な分野における差別化のカギとして認識されつつある。しかしながらデータ解析研究においてしばしば中心的に捉えられるこれらの自動解析技術はデータ解析のプロセス全体からみるとごく一部にすぎない。データの収集や洗浄・結果の解釈などを含むデータ解析プロセスの多くの部分がデータを解析する人間に依存する極めて属人的で労働集約的なものであり、急速に高まるデータ解析需要に反して、データ解析において主導的な役割を果たすいわゆる「データサイエンティスト」の不足が各所で指摘されている。

2 クラウドソーシングの台頭

米国政府が 2012 年に打ち出した「ビッグデータ研究開発イニシアティブ」の中で、注力すべき情報技術分野として「機械学習」「クラウドコンピューティング」とともに挙げている技術が、インターネットを介して不特定多数の人に仕事や作業を依頼する「クラウドソーシング」である。2005 年に登場した米 Amazon 社の提供するクラウドソーシング市場 Mechanical Turk はクラウドソーシングの利用を広く浸透させる契機となったが、国内においても同様の商用サービスが多数登場しており、発注側にとってはオンデマンドで労働力を調達する手段として、働き手にとっては場所や時間にとらわれない新しい働き方として注目されている。クラウドソーシングの対象範囲は、マイクロタスク（特別なスキルを要しない比較的単純な労働）から、より高度で専門的な業務を行うものへと拡大しつつあり、データ解析業務はその最たるものであるといえる。

計算機科学分野においても HCI、メディア処理など様々な分野でその利用が拡大しており、従来の計算機を中心としたパラダイムに変革を起こしつつある。

3. クラウドソーシングによるビッグデータ解析

データ解析のプロセスには比較的誰にでも実行可能なデータ収集や電子化のステップ、多少の専門知識やドメイン知識を要するデータクレンジングやキュレーションのステップ、そしてデータ解析手法の高い専門技能を要するモデル化・視覚化のステップへと続く。 最終的に得られた結果の評価や解釈には対象ドメインの深い知識が必要であり、また、そもそもの課題立案にはビジネス的な洞察も必要となる。このようにデータ解析のプロセスの各々のステップが要する様々な種類・レベルの専門性や適性を少人数でカバーすることは極めて困難であり、クラウドソーシングによってこれらの人材をオンデマンドで調達し、並列・協調的にプロセスを実行することが、この人的資源のボトルネックの解消へ向けた極めて有望なアプローチとなるだろう。その実現のためにはプラットフォーム技術・品質保証技術・インセンティブ設計・セキュリティ/プライバシ保護技術など様々な観点からの技術開発が必要である[鹿島 14, 鹿島 16]。

参考文献

[鹿島 14] 鹿島久嗣, 馬場雪乃: ヒューマンコンピュテーション概説. 人工知能学会誌, 29(1) (2014).

[鹿島 16] 鹿島久嗣, 小山聡, 馬場雪乃: ヒューマンコンピュテーションとクラウドソーシング. 講談社サイエンティフィク (2016). [刊行予定]

凧型無人航空機を用いた音源探査

公文 誠, 田嶋 脩一, 永吉 駿人
Makoto KUMON, Shuichi TAJIMA and Hayato NAGAYOSHI
熊本大学
Kumamoto University
kumon@gpo.kumamoto-u.ac.jp

Abstract

本論文では，ゆっくりと飛行可能な凧型の主翼を有する無人航空機を用いて地上の音源を探査する方法を考察する．この航空機は推力を生むプロペラを有し，飛行速度を制御することで飛行高度を操作できる飛行特性があるが，一方でプロペラの駆動音は大きく，音信号の計測におけるエゴノイズの主要因である．そこで，飛行高度を著しく乱さない範囲で，周期的にプロペラの駆動を停止することで音信号の観測を実現する方法を提案する．加えて，観測された音信号から推定された音源方向に飛行し，音源位置を推定する飛行経路計画についても考察する．これらの方法は数値シミュレーションを通じてその有効性を検証したので，あわせてこの結果を報告する．

1 はじめに

無人航空機は飛行しながら広範囲を効率的に探査可能なため，捜索や救助といったタスクでの活用が期待されている．これらのタスクを実現する上で対象を検出することが重要で，無人航空機にはカメラなどの種々のセンサが搭載されている．実際の捜索においては，単に無人航空機が探索をするだけでなく，要救助者が笛を吹く，大声を上げるなどで助けを求めることが考えられる．このことから，音信号も捜索における重要なモダリティの一つと言え，マイクロホンを搭載した無人航空機による音源探査について研究がなされている（例えば　[1, 2, 3] がある.）

このような無人航空機から音源探査を実現するには以下を考慮する必要がある.

1. 音源と航空機の距離が離れている（10m-100m）
2. 無人航空機自身の発するエゴノイズ
3. 受聴可能範囲内に複数の音源の存在

本論文では，これらのうち特に 1 と 2 について考察する.

無人航空機について考えると，定点ホバリング飛行が可能であることから，昨今マルチロータヘリコプタのような回転翼機の活用が期待されているものの，回転翼機が飛行し続けるにはロータを常に回転させる必要があり，この駆動音が大きなエゴノイズを生じるため，音源探査のプラットホームとしては問題がある．一方，固定翼機は動力を使わずに滑空飛行を行えば動力によるエゴノイズを生じないため，音源探査を実現できる可能性がある．本論文で用いるカイトプレーンは，このような固定翼機の一種で，凧型の主翼を有する無人航空機 [4, 5] である．この機体は機体サイズに比べ主翼が大きく，大きなペイロードを有するとともに，低速での飛行が可能という特徴があり，地上音源探査にも向いている．当然ながら，滑空だけでは飛行を継続できず，飛行高度を維持するためにはプロペラを回転させる動力飛行も必要で，音源探査と飛行の継続の間にはトレードオフがある．そこで，本論文では，プロペラの回転と停止を周期的に繰り返すことで飛行しながらエゴノイズの干渉を受けない音源定位を行う方法を提案する．また，音源位置をより正確に検出するために，音源に近づく飛行経路を生成する方法もあわせて考察する.

本論文の構成は以下のとおりである．次節でカイトプレーンについて簡単に説明し，その後，カイトプレーンからの音源探査方法の基本についてまとめている（第 3 節）．第 4 節ではプロペラを停止させるアプローチと，音源方向へと誘導する方法を提案する．これらの方法は第 5 節で数値シミュレーションで検証する．最後に第 6 節でまとめる.

2 カイトプレーン

2.1 ダイナミクス

本節では対象とするカイトプレーンの飛行特性を簡単に説明する．詳細は既報 [5, 6] を参照されたい.

カイトプレーンはデルタ形状の凧型の主翼を有する無

(a) Engine model (b) Electric motor model

図 1: Kiteplane

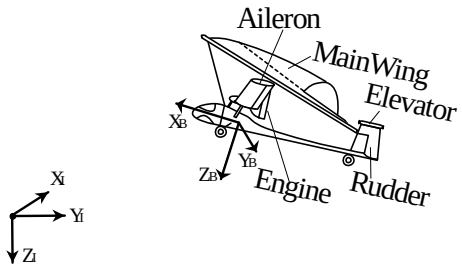


図 2: Kiteplane configuration (conventional type)

人航空機で（図 1），主翼は軽量な布製で，一般の固定翼機に比べ翼面積が広く大きなペイロードがあり，柔軟な翼構造から万一の墜落の際でも安全性が高い．

カイトプレーンの操作量にはエレベータ，ラダーとエルロンの 3 つの舵面と推力のためのプロペラの回転数がある．エルロンは主翼の左右の面積比を変化することで実現する構造になっている [7]．プロペラはエンジン (図 1 (a)) あるいはモータ (図 1 (b)) で駆動し，回転数によって推力を変化させて飛行速度を制御する．飛行速度に応じて揚力が変化するため，プロペラの回転数は主に飛行高度の制御に用いられる．また，飛行制御のために，GPS ならびに 3 軸加速度，3 軸角速度，3 軸磁気計が搭載されており，機体の姿勢情報を得ることが出来る．

以下，主翼に働く空気力を翼の左右それぞれについて $\boldsymbol{f}_{m,l}$, $\boldsymbol{f}_{m,r}$ と表し，エレベータ，ラダーに働く空気力，プロペラの推力をそれぞれ $\boldsymbol{f}_e$, $\boldsymbol{f}_r$, $\boldsymbol{T}$ と表すこととする．これらを用いて，機体を剛体と考え運動方程式をまとめると，

$$m\frac{d^2}{dt^2}\begin{bmatrix} x_I & y_I & z_I \end{bmatrix}^{\mathrm{T}} = \boldsymbol{f_I}, \tag{1}$$

$$\boldsymbol{I_B}\frac{d}{dt}\boldsymbol{\omega_B} + \boldsymbol{\omega_B} \times \boldsymbol{I_B}\boldsymbol{\omega_B} = \boldsymbol{n_B}, \tag{2}$$

のように表される．ここで，

$$\begin{aligned} \tilde{\boldsymbol{f}}_{\boldsymbol{I}} &= \boldsymbol{q} \odot \tilde{\boldsymbol{f}}_{\boldsymbol{B}} \odot \boldsymbol{q}^* \\ &= \boldsymbol{q} \odot \left(\tilde{\boldsymbol{f}_{m,l}} + \tilde{\boldsymbol{f}_{m,r}} + \tilde{\boldsymbol{f}}_{\boldsymbol{e}} + \tilde{\boldsymbol{f}}_{\boldsymbol{r}} + \tilde{\boldsymbol{T}} \right) \odot \boldsymbol{q}^* \\ \boldsymbol{n_B} &= \boldsymbol{l}_{m,l} \times \boldsymbol{f}_{m,l} + \boldsymbol{l}_{m,r} \times \boldsymbol{f}_{m,r} + \boldsymbol{l}_{\boldsymbol{e}} \times \boldsymbol{f}_{\boldsymbol{e}} \\ &\quad + \boldsymbol{l}_{\boldsymbol{r}} \times \boldsymbol{f}_{\boldsymbol{r}} + \boldsymbol{l}_{\boldsymbol{T}} \times \boldsymbol{T}, \end{aligned}$$

であり，$\tilde{\boldsymbol{x}}$ のように表される量は 3 次元ベクトル $\boldsymbol{x}$ の四元数での表現を与えるもので，$\tilde{\boldsymbol{x}} = \left[0, \boldsymbol{x}^T\right]^T$) と定義する．また，$x_I, y_I, z_I$ は機体の世界座標での位置を表し，$\boldsymbol{\omega_B}$ は機体座標での角速度を表す．機体姿勢は四元数 $\boldsymbol{q}$ で表現することとし，演算 $\odot$ は四元数同士の積とする．$\boldsymbol{f_I}$ ならびに $\boldsymbol{f_B}$ は質量中心に作用する合力を慣性座標と機体座標で表したもので，$\boldsymbol{n_B}$ は質量中心に作用するトルクを表す．$\boldsymbol{f_I}$ と $\boldsymbol{f_B}$ の間の変換は四元数を用いて表され，* は四元数の共役演算子を示すものとする．m および $\boldsymbol{I_B}$ は機体の質量と慣性行列をそれぞれ与え，$\boldsymbol{l}_{m,l}$, $\boldsymbol{l}_{m,r}$, $\boldsymbol{l}_e$, $\boldsymbol{l}_r$, $\boldsymbol{l}_T$ は翼での空気力の作用点を与える機体座標でのベクトルである．四元数の演算については [8] などを参照されたい．

姿勢のダイナミクスは四元数の変化として以下のように与えられる．

$$\frac{d}{dt}\boldsymbol{q} = \frac{1}{2}\tilde{\boldsymbol{\omega}}_{\boldsymbol{I}} \odot \boldsymbol{q} = \frac{1}{2}\boldsymbol{q} \odot \tilde{\boldsymbol{\omega}}_{\boldsymbol{B}} \tag{3}$$

2.2 制御器

[4] に示すようにカイトプレーンの姿勢動特性は安定しており，水平面と鉛直面の運動に分解してそれぞれ独立に制御することで現実的な経路追従が実現可能である．

所望の水平面内の飛行経路が与えられた時，飛行経路の単位接線ベクトルを $\boldsymbol{t}_p$ と表し，経路からの最短距離を与えるベクトルを経路誤差ベクトルと定義し，これを $\boldsymbol{e}$ と表すものとする．今，実現すべき飛行方向を $\boldsymbol{v}_d$ と表し，

$$\boldsymbol{v}_d = \exp^{-k_1|\boldsymbol{e}|^2}\boldsymbol{t}_p + k_2\frac{\boldsymbol{e}}{|\boldsymbol{e}|}, \tag{4}$$

と与えるものとする．ここで k_1 と k_2 は制御パラメータを表す．

$\boldsymbol{v}$ と θ_d を機体の水平面内での飛行速度と所望のバンク角と表すこととし，$\boldsymbol{v}$ と $\boldsymbol{v}_d$ のなす角に線形な形式で目標経路に追従するような所望のバンク角 θ_d を与えるものとする．つまり，θ_d は

$$\theta_d = k_3\mathrm{atan2}(\boldsymbol{v} \times \boldsymbol{v}_d, \boldsymbol{v} \cdot \boldsymbol{v}_d), \tag{5}$$

のように与えられる．ここで，k_3 は制御ゲインを表すものとし，(5) 中の $\boldsymbol{v}$ と $\boldsymbol{v}_d$ は計算上適宜 3 次元に拡張されるものとする．適当な姿勢制御器によって実際のバンク角を所望のバンク角 θ_d に追従することになるが，本論文では著者らの提案する非線形制御器 [6, 9] を用いることとした．

鉛直方向の運動については，本論文では飛行高度を一定の目標高度に追従させるものを考える．機体の飛行特性から，推力 $|\boldsymbol{T}|$ が釣り合いの値より増加すれば機体速度が増加し，その結果機体は上昇することとなり，逆もまた同様の関係があるので，プロペラ回転数を制御して推力 $|\boldsymbol{T}|$ を操作することで高度制御は実現される．例えば，[6, 9] などに示す簡単な PD フィードバック制御によって機体高度を制御できる．

3　無人航空機からの音源位置推定

無人航空機に搭載したマイクロホンアレイでの音源方向の推定の研究には Okutani[3] らのクアドロータヘリコプタで収録した音信号を MUSIC 法 [10] を適用した例があり，音源と無人航空機が近いなどの条件下で音源方向を推定することが可能である．このことから，本論文では，無人航空機の機体から見た音源の方向がある程度推定可能との仮定の下で音源の位置を推定する方法を考える．以下，推定された音源方向は機体から音源に向けた単位ベクトル $\boldsymbol{u}_s$ で表されるとする．ただし，適当な座標変換によって $\boldsymbol{u}_s$ は慣性座標系で表現されるものとする．

今，地表面が平らな平面で，地上からの機体の高度が分かるとすると, 音源の位置 $\boldsymbol{p}_s$ は

$$\boldsymbol{p}_s = \frac{z_I}{\begin{bmatrix} 0 & 0 & -1 \end{bmatrix} \boldsymbol{u}_s} \boldsymbol{u}_s + \boldsymbol{p}, \tag{6}$$

のような関係がある．ここで $\boldsymbol{p}$ は無人航空機の位置を表し $\boldsymbol{p} = \begin{bmatrix} x_I & y_I & z_I \end{bmatrix}^T$ と定義した．

一般に方向推定や姿勢情報などに不確かさがあるため，音源位置の推定情報を与える (6) の計算はこれらの不確かさを考慮する必要がある．そこで，(6) の与える点 $\boldsymbol{p}_s$ に替えて，点 $\boldsymbol{p}_s$ を含む小領域を音源位置として考える．対象とする探索空間を格子状のグリッドに分割し，$\boldsymbol{x}$ をあるグリッドの代表点の座標とすれば，当該のグリッドを $S(\boldsymbol{p}_s)$ と書くこととすれば，推定された小領域と共通部分を有するグリッドに音源が存在する可能性があるとして音源位置を表現する．具体的には以下のようにして計算する．

k 回目の観測を $\boldsymbol{p}_s(k)$ と表し，それまでにグリッド $g_{\boldsymbol{x}}$ が音源を含むと想定された回数を $N(\boldsymbol{x})$, つまり

$$N(\boldsymbol{x}) = \sum_k 1(\boldsymbol{x}, \boldsymbol{p}_s(k)), \tag{7}$$

とする．ここで $1(\boldsymbol{x}, \boldsymbol{y})$ は $S(\boldsymbol{y}) \cap g_{\boldsymbol{x}} \neq \phi$ であれば 1 を与えそうでなければ 0 となる関数である．これを用いれば, 音源位置は次に示す頻度の分布 $\hat{\boldsymbol{p}}_s$ で与えられる．

$$\hat{\boldsymbol{p}}_s(\boldsymbol{x}) = \frac{N(\boldsymbol{x})}{\sum_{\boldsymbol{y}} N(\boldsymbol{y})}. \tag{8}$$

音源が空間に固定されると仮定すれば，$\hat{\boldsymbol{p}}_s$ の最大値を音源の推定位置とするのは自然な解釈である．

4　音源探査のための無人航空機の制御

本節ではカイトプレーンで音源を探査するため，これまでに述べてきたシステムに加えて，エゴノイズを抑制するプロペラの回転数制御と音源に向けた飛行経路計画について考える．

4.1　プロペラの回転制御

MUSIC 法はノイズに対して一定のロバスト性があるものの，騒音源のプロペラはマイクロホンの近くにあり，非常

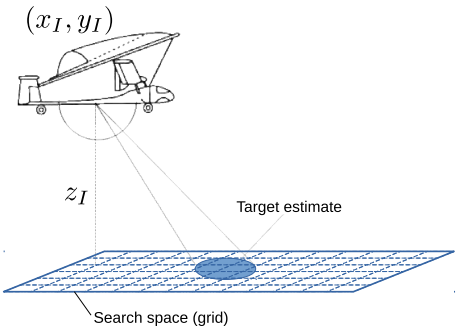


図 3: Grid space sound source localization from UAV

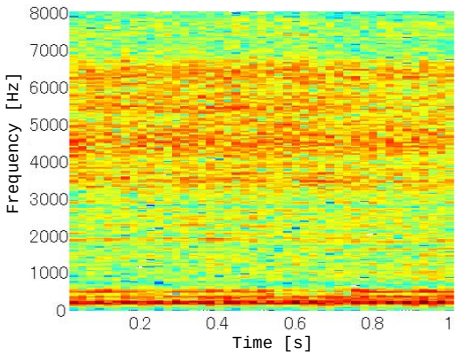


図 4: Frequency characteristics of rotor-noise

に大きな駆動音を生じるため MUSIC 法であっても定位性能を損なうあるいは定位が出来ない可能性がある．実際，図 4 は飛行中に測定した音信号のスペクトログラムを示すが，ノイズが広い帯域にわたって干渉していることが分かり，このノイズが対象音を覆い隠す可能性がある．逆に，プロペラの回転を停止し，駆動騒音のない状態を作り出せれば，音源定位能の向上に大きな効果があると考えられる．勿論，カイトプレーンではプロペラによる推力は高度の制御に関係しているので，プロペラを長時間にわたって停止したままにすることは出来ない．そこで，プロペラの回転と停止を周期的に繰り返すことで，高度を制御しつつ，音源の探査を実現する方法を考える．

無人航空機が安定した飛行状態にあるとし，簡単のため x_I 軸に沿って飛行しているものとする．ここでは飛行高度が問題となるため，高度に関するサブシステムを元のダイナミクス (1) から近似的に取り出せば

$$\begin{aligned} m\frac{d^2}{dt^2}x_I &= -k_x\frac{d}{dt}x_I + |\boldsymbol{T}|, \\ m\frac{d^2}{dt^2}z_I &= -mg + k_z v_x \frac{d}{dt}x_I, \end{aligned} \tag{9}$$

のように書ける．ここで，k_x, k_z ならびに v_x は線形化に伴う係数とノミナルな飛行速度を表すものとする．(9) に示されるように, 制御入力 $|\boldsymbol{T}|$ は速度 $\frac{d}{dt}x_I$ を介して高度を制御するのでこのダイナミクスはローパス特性があり，高い周波数で $\boldsymbol{T}$ を切り替えてもすぐには飛行高度 z_I が大きく変動することはなく，小さな脈動に止まることに

なる．

本論文では，プロペラの回転と停止を一定のデューティ比 $d \in [0,1]$ を持つ周期 P の繰り返しと定義する．つまり

$$T(t) = \begin{cases} u & t \in \left[nP, nP + \frac{d}{P}\right) \\ 0 & \text{otherwise} \end{cases}, \qquad (10)$$

のように与えるものとし，n と u はそれぞれ 0 以上の整数と元々の制御入力を表す．エゴノイズの無い時間区間は長いほど望ましいので，より大きな d が望ましいが，大きな高度の脈動が生まれることになるので適当な P, d を設定する必要がある．このシステムは非線形で解析的には解けないため，以下ではパラメータを経験的に調整し定めることとした．なお，提案法では音源定位のデータはプロペラの停止している時間区間, つまり $T(t) = 0$, の収録音を用いて行うこととする．

4.2 飛行経路計画

音源位置がある程度推定されれば，この情報に基づいて機体を音源に近づけることでより鮮明に対象音を測定することが出来，定位性能も改善されると期待される．また，一般に音源の十分に近くに無人航空機が飛行している場合は，音源に漸近する代わりにその音源の周囲を飛行した方が位置推定性能が良い．そこで，推定された音源位置を中心とする適当な半径の円軌道を所望の経路とし，これに追従させることを提案する．ただし，推定の初期段階では，音源位置の事前情報がないため，一定時間，適当に与えられた経路にそって飛行をするものとする．

提案する円軌道は以下のように与えられる．

$$\boldsymbol{x}_d = r \begin{bmatrix} \cos\psi & \sin\psi & 0 \end{bmatrix}^T + \begin{bmatrix} s_x & s_y & z_d \end{bmatrix}^T, \qquad (11)$$

ここで，r, s_x, s_y, z_d, ψ は経路の半径，推定された音源位置の X，Y 座標所望の高度と $[0, 2\pi)$ の区間内の適当なパラメータを表している．

この場合，第 2.2 節の制御器を適用する上で，経路誤差 $\boldsymbol{e}$ は以下のように定められる．

$$\boldsymbol{e} = \begin{bmatrix} x_I & y_I \end{bmatrix}^T - \frac{r}{|\boldsymbol{\Delta}|}\boldsymbol{\Delta} - \begin{bmatrix} s_x & s_y \end{bmatrix}^T \qquad (12)$$

ここで，

$$\boldsymbol{\Delta} = \begin{bmatrix} x_I & y_I \end{bmatrix}^T - \begin{bmatrix} s_x & s_y \end{bmatrix}^T,$$

であり

$$\pm|\boldsymbol{e}| = |\boldsymbol{\Delta}| - r$$

である．

音源位置情報は観測ごとに更新されているので，一定の観測数ごとに上の目標経路も周期的に更新することとする．

5 数値シミュレーション

提案法の有効性を数値シミュレーションを通じて検証した．

5.1 シミュレータ

非線形の飛行ダイナミクス (1) と (2)，(3) を数値積分によって実行した．(5) で与えられる目標バンク角 θ_d を [6] で提案される制御器への規範値とした．

音響信号のシミュレーションでは，音源が十分に遠くにあり，アレイ付近では平面波で近似出来ることから，物理的に正しいものではないが伝達特性が方向と距離に分解出来ると仮定した．また無人航空機の飛行に伴うマイクロホンと音源の相対位置 $\boldsymbol{e}$ の時間変化は音信号処理の観点からは比較的ゆっくりとしていることから，近似的に線形応答が成立するものと考えた．これらの仮定から，伝達関数行列 $\boldsymbol{H}$ を

$$\boldsymbol{H}(\boldsymbol{e}) = \boldsymbol{H}_d(\phi)\boldsymbol{H}_r(|\boldsymbol{e}|) \qquad (13)$$

のように方向性伝達関数 $\boldsymbol{H}_d$ と 距離依存性の伝達関数 $\boldsymbol{H}_r$ の積でモデル化する．ここで ϕ はマイクロホンアレイから見た音源方向を示している．

音源の位置と収録した音信号を s_s と $\boldsymbol{s}_m$ とすれば，

$$\boldsymbol{s}_m = \boldsymbol{H}(\boldsymbol{e})s_s + a_T|\boldsymbol{T}|\boldsymbol{n} \qquad (14)$$

の関係を用いて信号をシミュレートする．ここで $\boldsymbol{n}$ および $a_T|\boldsymbol{T}|$ はノイズデータとプロペラによる騒音信号を表しており，エゴノイズが推力 $|\boldsymbol{T}|$ に比例するものとモデル化している（a_T は比例係数）．また，機体の姿勢変化はジンバル等で補正されていると考え，マイクロホンアレイの姿勢については考えない．

5.2 対象環境

400m×400m の平面を探査空間とし，音源はこの中央に位置するものとした．無人航空機の初期位置は $(-200\text{m}, -200\text{m})$ にあり，初期の目標経路は 図 5 に示すような'S' 字の曲線を与えている．

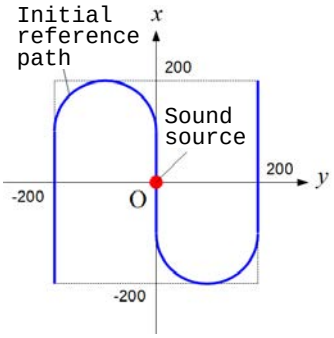
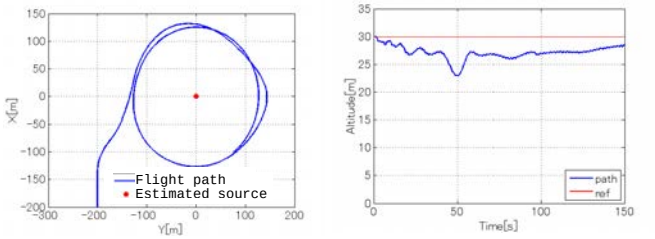


図 5: Initial flight path and search space

プロペラの駆動・停止のデューティー比は $d = 0.5$ とし，周期 P は 1.0s とした．また，全探査飛行時間は 150s とした．

5.3 結果

図 6 は提案法で実現された飛行結果を示す．図 6(a) はカイトプレーンが音源に近づき，音源の周囲を飛行した様子



(a) Flight path (b) Altitude

図 6: Flight path and altitude

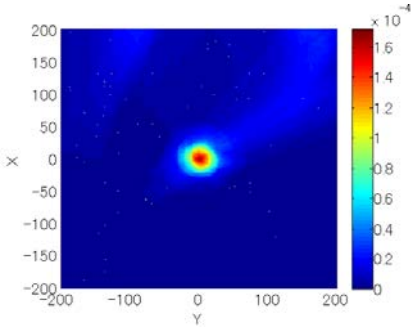


図 7: Sound source localization result

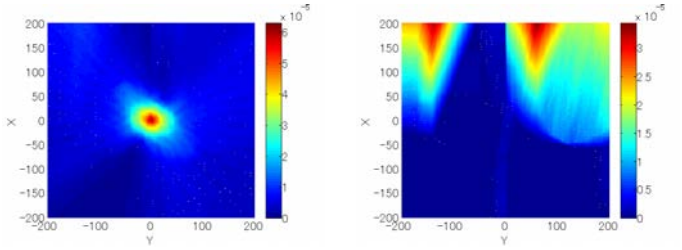
を示している．図 6(b) では，所望の飛行高度 30m （赤破線）に対し，実際の飛行高度 (青曲線）を示しており，目標高度近傍での飛行が達成されていることが分かる．

図 7 は提案法で推定された音源の位置を表す．音源位置分布を色で示しており，最大値が音源位置に対応しているため正しく推定出来たと言える．

提案法の効果を明らかにするため，プロペラの周期的な回転・停止，および経路生成を行わず初期経路のまま飛行を続けた場合のそれぞれでシミュレーションを行い，図 7 に対応する推定結果を図 8 に示す．図 8(a) によれば，経路制御をせずともある程度音源位置を推定は可能であったが，図 8(b) によればプロペラの停止は音源定位に不可欠であることが分かる．

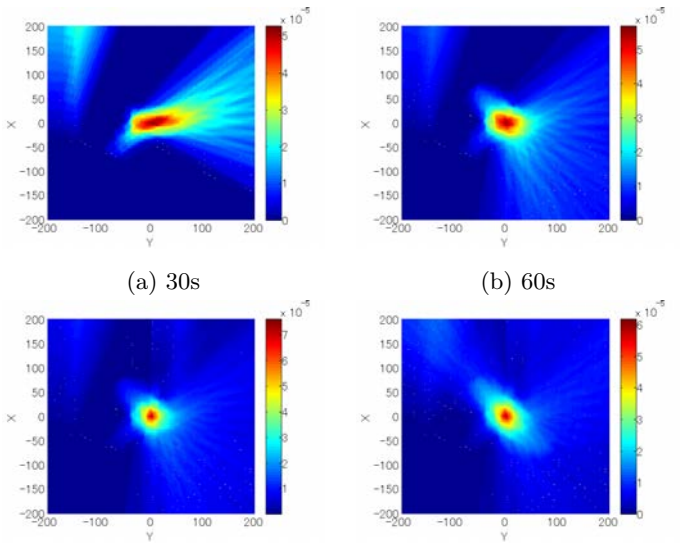
音源位置推定の推移の様子を図 9 に示している．この場合は，推定の時間発展を明らかにするために，初期時刻での目標経路をたどることとしている（図 8(a) に対応). この図より，中央にある音源推定結果は対称ではないことが示されており，正規分布のような対称な分布を仮定する例えばカルマンフィルタのような手法では不適当であることが示唆される．

図 7 と 図 8(a) はともに正しい音源位置の推定しており，これらの間に明確な差を見ることは容易ではないが，音源位置分布のピークの値を比べた　表 1　によれば，音源位置を想定して経路生成したものの方が鋭いピークを形勢していることが分かる．なお，これは探査空間全体で正規化しているので，値の大きさそのものは重要ではないが，二つの方法の間での比較には意味があることに注意されたい．



(a) without path generation (b) conventional flight

図 8: Sound source localization result



(a) 30s (b) 60s

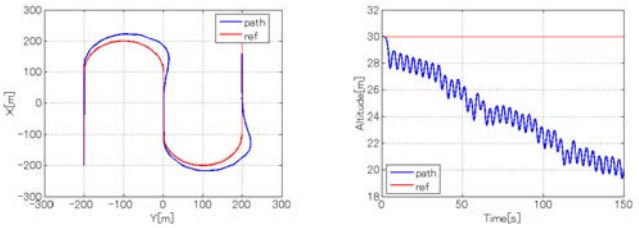
(c) 90s (d) 120s

図 9: Evolution of sound source localization

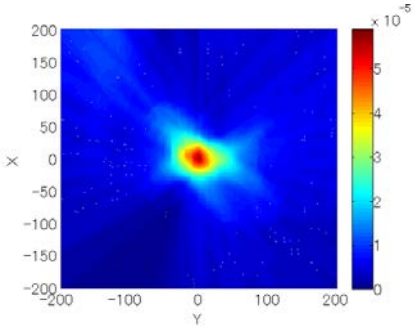
また，プロペラの回転・停止 (10) を行う飛行でも高度の変動は安定的であったが，上の結果で示されるように，無人航空機はバンクしながら旋回するよう制御されており，モデル化の際考慮しなかった機体ダイナミクスの影響によって，水平面内での運動と高度方向に干渉が生じる可能性がある．特に，長時間，プロペラを停止しながら旋回すると，この干渉は顕著になると考えられるので，プロペラの回転・停止の周期 P は十分に注意して設計する必要がある．このことを示すため，$P = 2.0$s の場合のシミュレーション結果を Fig. 10 に示す．ここでは水平面内の目標経路は初期に与えた'S' 字のものである．この場合でも音源位置の推定は可能であったが，飛行高度を保つことが出来ず徐々に下降してしまっており，不適切な結

表 1: Sound source localization clarity

	Proposed	without path generation
Maximum value of SSL	1.7035×10^{-4}	8.0865×10^{-5}



(a) Flight path (b) Altitude



(c) Sound source localization

図 10: Flight result with longer rotor stall period

果となった.

6 おわりに

本論文では，凧型の主翼を有する無人航空機にマイクロホンアレイを搭載し，地上の音源を探査する方法として，プロペラを周期的に停止しながら音源に向かって誘導する手法を提案した．数値シミュレーションを通じて，プロペラを停止することが広い範囲の音源定位に重要であること，また音源周辺を旋回する円軌道を設計することで定位性能が改善されることが示された．また，本論文ではレスキューなどのタスクを考え音源が固定されている場合を考えたが，このため頻度に基づいて音源位置を推定する方法が適用可能であった.

今後はより一般的な場合として，移動音源を対象とすることが考えられる．この場合は，音源の運動を推定することになるが，この運動に伴う不確かさが生じるため繰り返しベイズ推定などの運動モデルを用いた推定法を採り入れる必要があろう．また，複数の音源が存在する場合は，単に音源方向だけでなく，その種類などを判じ，音源同士を混同する必要となるが，これも今後の課題の一つである.

謝辞

本研究の一部は科研費基盤研究（S）24220006　ならびに内閣府 ImPACT プログラム「タフ・ロボティクス・チャレンジ」の助成を受けました.

参考文献

[1] T. Ishiki and M. Kumon, “A microphone array configuration for an auditory quadrotor helicopter system,” in *Safety, Security, and Rescue Robotics (SSRR), 2014 IEEE International Symposium on*, Oct 2014.

[2] M. Basiri, F. Schill, P. U. Lima, and D. Floreano, “Robust acoustic source localization of emergency signals from micro air vehicles,” in *IROS*, 2012, pp. 4737–4742.

[3] K. Okutani, T. Yoshida, K. Nakamura, and K. Nakadai, “Outdoor auditory scene analysis using a moving microphone array embedded in a quadrocopter.” in *IROS*. IEEE, 2012, pp. 3288–3293.

[4] M. Kumon, M. Nagata, R. Kohzawa, I. Mizumoto, and Z. Iwai, “Flight path control of small unmanned air vehicle,” *Journal of Filed Robotics*, vol. 23, no. 3-4, pp. 223–244, 2006.

[5] M. Kumon, Y. Udo, H. Michihira, M. Nagata, I. Mizumoto, and Z. Iwai, “Autopilot system for kiteplane,” *IEEE/ASME Transactions on Mechatronics*, vol. 11, no. 5, pp. 615–624, oct 2006. [Online]. Available: http://ci.nii.ac.jp/naid/120002464294/

[6] S. Tajima, T. Akasaka, M. Kumon, and K. Okabe, “Guidance control of a small unmanned aerial vehicle with a delta wing,” in *Proceedings of Australasian Conference on Robotics and Automation*, 2013.

[7] Y. O. S. T. M. K. K. Nakashima, K. Okabe, “Small Unmanned Aerial Vehicle with Variable Geometry Delta Wing,” 2014.

[8] R. M. Murray, Z. Li, and S. S. Sastry, *A Mathematical Introduction to Robotic Manipluation*. CRC Press, 1994.

[9] T. Akasaka and M. Kumon, “Robust attitude control system for kite plane,” in *Proceedings of System Integration 2012*, 2012, pp. 1623–1626, (in Japanese).

[10] R. Roy and T. Kailath, “Esprit - estimation of signal parameters via rotational invariance techniques,” *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 37, no. 7, pp. 984–995, 1989.

複数のマイクロフォンアレイとロボット聴覚ソフトウェア HARK を用いた野鳥の位置観測精度の検討

Assessing the accuracy of bird localization derived from multiple microphone arrays and robot audition HARK

松林志保 鈴木麗璽
名古屋大学大学院情報科学研究科
小島諒介
東京工業大学大学院情報理工学科
中臺一博
（株）ホンダ・ リサーチ・インスティチュート・ジャパン，東京工業大学大学院情報理工学科

要旨

本研究は、３つのマイクロフォンアレイとロボット聴覚ソフトウェア HARK を用いた野鳥の観測精度に関する予備的調査の結果を報告する。1 つ目の調査では、愛知県豊田市の森林内でスピーカーからの鳥の歌の再生音を用いて HARK による定位精度の検討を行った。2 つ目の調査では、人間による野鳥の観測結果と HARK により定位された野鳥の位置を比較し、その定位精度を調べた。

1 はじめに

近年、複数のマイクロフォンで構成されるデバイスであるマイクロフォンアレイを用いて音源の方向や位置を定位したり、定位した音源を分離する技術が急速に発展している。この技術の野鳥研究への応用は、従来の単一のマイクロフォンによる録音と比べてより豊富な生態情報の記録を可能にするため、生態理解へ大きな貢献を果たすことが期待される。

しかし、独自開発のシステムに基づく研究[1, 2]等はなされているものの、この技術の野鳥研究への活用は未だ限定的な状況にあるといえる。その要因として，録音のためのデバイスの入手と分析のためのソフトウェアの利用が容易でない点や、野鳥の鳴き声の定位・分離性能の評価が十分でない点等が挙げられる。

我々はこれらの課題を克服すべく、ロボット聴覚オープンソースソフトウェアである HARK（Honda Research Institute Japan Audition for Robots with Kyoto University）[3]と、市販の会議用マイクロフォンアレイを用いたシステムを構築し、野鳥の歌に基づく方角の定位等を試みている[4]。HARK は、音響工学等に関する詳細な知識を必要とせずに市販のマイクロフォンアレイ等を用いて PC 上に音源定位や分離等を含むシステムを柔軟に構築可能なソフトウェアであり、これまでロボットの聴覚のために開発されてきた。現在様々な文脈における応用が展開中であるが、野鳥の歌などの野外での音源定位の精度は十分検証されていない。

本研究は、複数のマイクロフォンアレイと HARK を用いた野鳥の二次元位置推定精度を検討することを目的とする。その手段として２種類の実験を行った。まず、事前に録音された野鳥１１種の歌を野外でスピーカーを用いて流し、マイクロフォンアレイからの距離や鳥の種類が HARK による定位精度におよぼす影響を測った。次に、HARK が実際のさえずりに基づいて定位した野鳥の位置を人間による野鳥観測結果と比較し、その定位精度を調べた。

2 手法

2.1. 複数のマイクロフォンアレイの設置

野鳥の録音実験は２０１５年６月、スピーカーテストは同年１０月に名古屋大学フィールド科学教育研究センター稲武フィールド（愛知県豊田市稲武町）内の標高約１０００m、樹齢６０~７０年のカラマツ植林と広葉樹の混合林内において、晴天ほぼ無風の気象条件下で行われた。録音には３つのマイクロフォンアレイ（Dev-audio 社 Microcone）を用いた。

各マイクロフォンアレイは、林内に定めた一辺が１０m の正三角形の各頂点の位置に設置した三脚上に配置した。録音に用いた Microcone はそれぞれ７つ（水平方向に６つ、天井部分に一つ）のマイクロフォンから成る。

2.2 再生音源とスピーカーテスト

事前に録音された音源として、野鳥大鑑[5]付属の CD から調査地で営巣する野鳥１１種の代表的なさえずりとその他の声（地鳴き等）を用いた（Table 1）。これらの音を、iPod と地上約１mに設置した外付けスピーカー（Sanwa supply bluetooth wireless speaker MM-SPBTBK）から一辺が１０mの正三角形の中心、中心から西、北東、南東方向に２５、５０m離れた地点から流した（Figure 1）。２５m地点でのマイクロフォンアレイの位置、スピーカーの距離および角度が正確に設置された場合に、鳥の歌の再生音が各

マイクロフォンアレイに届く理想的な角度を Table 2 にまとめる。

再生音は正三角形の中心に向けて流した。野鳥の歌の大きさは鳥の種類や競争相手の存在、外部音の有無などによる状況で異なるが、本実験では再生音の大きさは約３５～４０ｄＢ の周辺音に対して約０～２０ｄＢ 大きい音（周辺音とほぼ同じかわずかに大きい音）で流した。音の大きさは無料の騒音測定器アプリ[6]を用いて測定した。同時に正三角形の頂点に設置した３つのマイクロフォンアレイを用いてスピーカーから流れる再生音を録音した。

Table 1 スピーカーテストに用いた野鳥リストと歌の種類。

鳥の種類	英名	鳥の名前コード	歌の種類
ウグイス	Japanese bush warbler	JBWA	さえずり、間奏
オオルリ	Blue-and-white flycatcher	BAWF	さえずり、間奏
ソウシチョウ	Red-billed leiothrix	RBLE	さえずり
キビタキ	Narcissus flycatcher	NAFL	さえずり、間奏
ヤマガラ	Varied tit	VATI	さえずり、威嚇音
ヒガラ	Coal tit	COTI	さえずり、地鳴き
センダイムシクイ	Eastern crowned willow warbler	ECWW	さえずり
ヒヨドリ	Brown-eard bulbul	BEBU	さえずり、時鳴き
カッコウ	Common cucko	COCU	さえずり
ツツドリ	Oriental cucko	ORCU	さえずり
ホトトギス	Lesser cucko	LECU	さえずり

Figure 1 ３つのマイクロフォンアレイと鳥の歌の再生音を流す地点の位置関係。一辺が１０ｍの正三角形の北の頂点にマイクロフォンアレイ１、南の頂点にマイクロフォンアレイ２、東の頂点にマイクロフォンアレイ３を配置した。

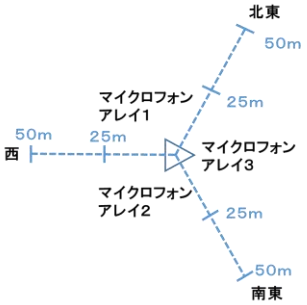


Table 2 三角形の中心、および各方向２５ｍ地点から鳥の歌の再生音を流した場合、音源が定位される理想的な角度。マイクロフォンアレイを設置した三角形の中心から北方向は０°、南方向は１８０°、東方向は－９０°、西方向は９０°とする。

	中心	西方向	北東方向	南東方向
マイクロフォンアレイ1	-150°	103°	-47°	-150°
マイクロフォンアレイ2	-30°	77°	-30°	-137°
マイクロフォンアレイ3	90°	90°	-17°	-163°

2.3 野鳥の位置観測と歌の分析

人間の観測者がマイクロフォンアレイを設置した正三角形の中心に立ち、録音開始と同時に周辺で観測された鳥の種類、中心からの大まかな位置等を約５分毎に記録した。野鳥の位置や種類は歌から推定し、正三角形の中心から２５、５０、１００ｍの同心円を用いてフィールドノートに記録した[7]。野鳥観測は目視と聞き取りに基づく（Figure 2）。

Figure 2 野鳥の観測と３つのマイクロフォンアレイを用いた録音風景。図中黄色い線は一辺が１０ｍの正三角形を示す。

録音した野鳥の歌から、歌の再生と音声分析ソフトウェア Praat[8]を用いてのスペクトログラム（声紋）分析により手動で鳥種を分類し、個々の歌の始まりと終わりの時間を抽出した。これらの手動分析の結果を後述の HARK による音源定位結果と比較することで自動音源定位精度の検討を行った。

2.4 HARK による音源定位・分離・位置の推定

３つの各マイクロフォンアレイで収録した音声信号から方向・分離音を抽出するために、HARK を用いて音源定位・分離を行った。まずそれぞれのマイクロフォンアレイについて７チャネルの音声信号を読み込み、短時間フーリエ変換によって得た各チャネルのスペクトログラムから MUSIC 法[9] を用いて音源定位を行った。次にその定位結果をもとに Geometric High order Decorrelation based Source Separation(GHDSS)法[10]を用いて各音源方向に対応した音源を分離する音源分離を行った。

最後に、音源定位によって得た各マイクロフォンアレイを起点とした３つの方向（半直線）の交点を計算することで、音源の二次元空間内での位置を求めた。この時、音源定位の方向の誤差を許容するために、３つの半直線のすべての中心を音源とした３つの半直線のうちひとつでも交点を持たない半直線の組み合わせがある場合は誤検出として棄却した。

3 実験結果

3.1 スピーカーによる録音再生テスト

3.1.1. 定位音源の位置分布の確認

HARK により鳥の歌を自動定位した結果を参考にして、個々の分離音を人間の耳で確認することで再生音源との比較作業を行った。実際に再生音が定位された方角と、マイクロフォンアレイや音源の設置位置が理想的な場合の音源とマイクロフォンアレイの角度をTable 3に示す。各方向毎に HARK が再生音を定位した角度と理想的な角度との差異に着目すると、西方向からの音源は北方向、北東方向からの音

源は南方向、南東方向からの音源は北方向と一定方向にずれが生じていた。このずれは HARK により定位されたた音源の２次元空間位置分布（Figures 3~5）にも反映された。しかしながら Figures 3~5 に示されるように、システマティックな位置のずれはあるものの再生音はおおむね各スピーカーの位置付近で定位された。

再生音以外にマイクロフォンアレイ付近で定常的に定位された音源は、定位の際に３つのマイクロフォンアレイが異なる音源を同一の音源として定位したことにより生じたものと推測される。この現象は特に西、北東方向から再生音を流した場合に南東方向で頻繁に発生した（Figures 3 & 4）。これらの音源は南東方向に密生する笹群の葉音と推定される。逆に南東方向から再生音を流した場合には、北東方向にも再生音以外の音源が定位された（Figure 5）。北東方向の分離音を調べるとその多くは再生音の反響音であった。この反響音は付近のプレハブ小屋に起因すると考えらえる。また、南東方向から再生音を流した地点は笹群内に位置する。笹群内では鳥の歌は定位されたものの分離しきれず、前後のさえずりや周囲の音と結合する現象も確認された。

Table 3 各マイクロフォンアレイで定位された実際の音源の方角と理想の方角の比較。音源とマイクロフォンアレイの位置関係は Figure 1 を参照のこと。マイクロフォンアレイを設置した三角形の中心から北方向は０°、南方向は１８０°、東方向は－９０°、西方向は９０°とする。

西方向(録音#137)	A:HARKが定位した方向	B:理想の方向	AB間のずれ
マイクロフォンアレイ1	95°	103°	北方向に8°
マイクロフォンアレイ2	55°	77°	北方向に22°
マイクロフォンアレイ3	75°	90°	北方向に15°
北東方向(録音#138)	A:HARKが定位した方向	B:理想の方向	AB間のずれ
マイクロフォンアレイ1	-60°	-47°	南方向に13°
マイクロフォンアレイ2	-45°	-30°	南方向に15°
マイクロフォンアレイ3	-50°	-17°	南方向に33°
南東方向(録音#139)	A:HARKが定位した方向	B:理想の方向	AB間のずれ
マイクロフォンアレイ1	-145°	-150°	北方向に5°
マイクロフォンアレイ2	-120°	-137°	北方向に17°
マイクロフォンアレイ3	-150°	-163°	北方向に13°

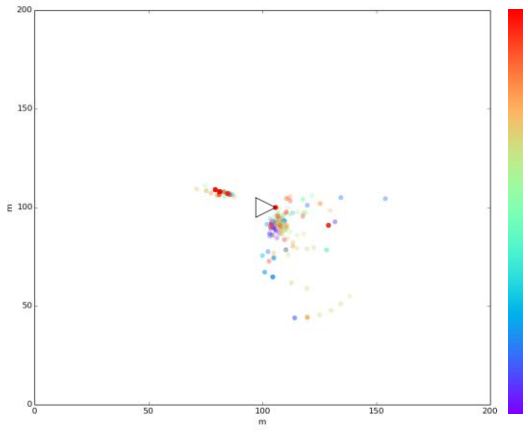


Figure 3 西方向２５m 地点からの音源を定位した二次元位置分布。各色点は定位された音源を時系列で示す。これらの音源は再生音以外の音源も含む。

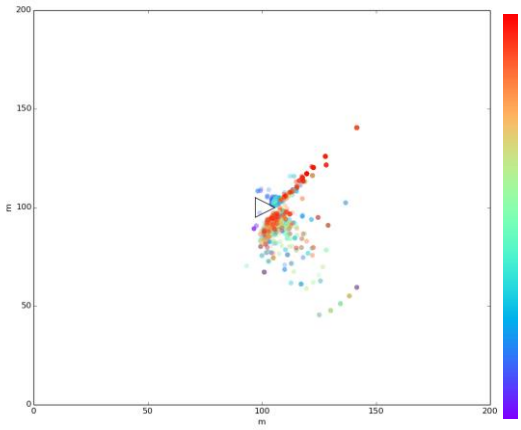


Figure 4 北東方向２５m 地点からの音源を定位した二次元位置分布。各色点は定位された音源を時系列で示す。これらの音源は再生音以外の音源も含む。

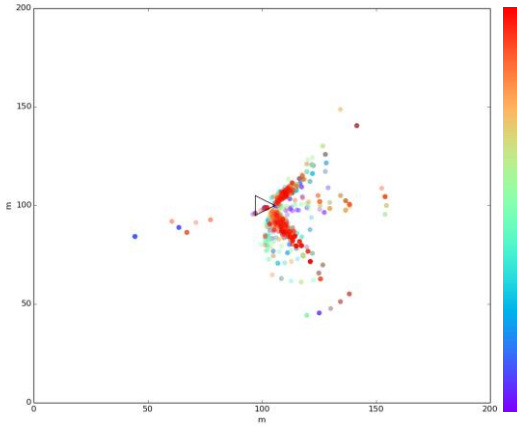


Figure 5 南東方向２５m 地点からの音源を定位した二次元位置分布。各色点は定位された音源を時系列で示す。これらの音源は再生音以外の音源も含む。

3.1.2. 各方向における音源定位性能の検討

HARK による音源自動定位の精度は、音源からの距離と鳥の種類の影響を受けたようである。西、北東、南東方向における、２５、５０m離れた地点から鳥の歌の再生音流した場合の各方向の３つのマイクロフォンアレイの平均定位指数を Table 4 に示す。音源との距離が２５mの場合は、３方位とも全鳥種が定位され、全１１鳥種の平均定位指数は各方向ともに９０を超えた。一方、音源との距離が５０mに伸びると、全鳥種の平均定位指数は西方向で３７．９、北東方向で６８．２、南東方向で５９．１に低下した。

鳥種による違いに着目すると、５０m地点における三方向の平均定位指数に見られるように、ウグイス、キビタキ、ヒガラなど比較的高音でさえずる種は７０を超えたが、低周波の音域の歌を持つカッコウやツツドリは２０以下となった（Table 4）。

Table 4 ３つのマイクロフォンアレイによる平均定位指数。指数は、HARK により分離された音を人間が耳で確認した際に、各マイクロフォンアレイが音源とほぼ同質の質を保ちつつ各鳥の歌を定位した場合にはそのマイクロフォンアレイに２、歌を部分的に定位した場合や分離精度が不十分な場合は１、全く定位しなかった場合は０のスコアを与えた後、３つのマイクロフォンアレイのスコアを合計した最大可能スコア（６）との比率を計算しその平均値を０から１００までの値で正規化した。例えば、西方向５０m地点での場合、ウグイスのスコアはマイクロフォンアレイ１では２、マイクロフォンアレイ２と３では各１ずつとなり、３つのマイクロフォンアレイの平均定位・分離指数は６６．７となる。音源と３つのマイクロフォンアレイの位置関係はFigure 1を参照のこと。

	音源との距離25m				音源との距離50m			
	西方向	北東方向	南東方向	3方位平均	西方向	北東方向	南東方向	3方位平均
ウグイス	100	100	100	100	66.7	83.3	83.3	77.8
オオルリ	100	100	100	100	50.0	66.7	66.7	61.1
ソウシチョウ	100	100	100	100	33.3	83.3	83.3	66.7
キビタキ	100	100	100	100	66.7	83.3	83.3	77.8
ヤマガラ	100	100	100	100	33.3	83.3	50.0	55.6
ヒガラ	100	100	100	100	66.7	83.3	66.7	72.2
センダイムシクイ	100	100	100	100	33.3	66.7	50.0	50
ヒヨドリ	100	100	100	100	33.3	83.3	33.3	50
ツツドリ	33.3	66.7	100	66.7	0	0	33.3	11.1
カッコウ	100	100	100	100	0.0	33.3	16.7	16.7
ホトトギス	100	100	100	100	33.3	83.3	83.3	66.7
全11種平均	93.9	97.0	100	97.0	37.9	68.2	59.1	55.1

3.2. 野鳥の自動音源定位結果と人間による鳥観測結果の比較

Figure 6 は HARK により定位された音源の２次元位置分布を、Figure 7 は人間の野鳥観測に基づく個々の野鳥の種類とその推定位置を示す。HARK による音源の定位結果と人間の観測者による野鳥の推定位置を比較すると、その空間的分布パターンには類似性が見られた。

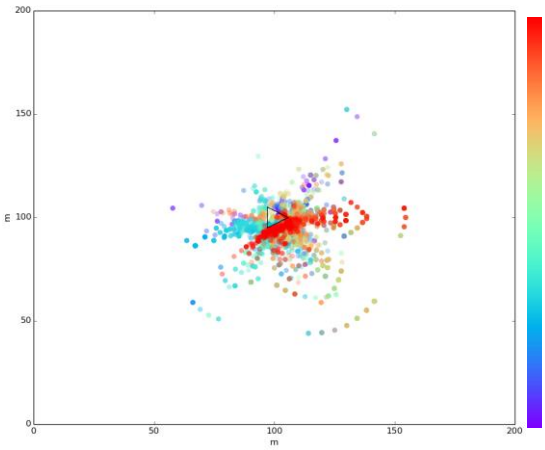


Figure 6 HARK により定位された音源。各色点は定位された音源を時系列で示す。これらの音源は観測者の足音や周辺音等も含む。図中の正三角形は、マイクロフォンアレイを設置した正三角形に対応する。

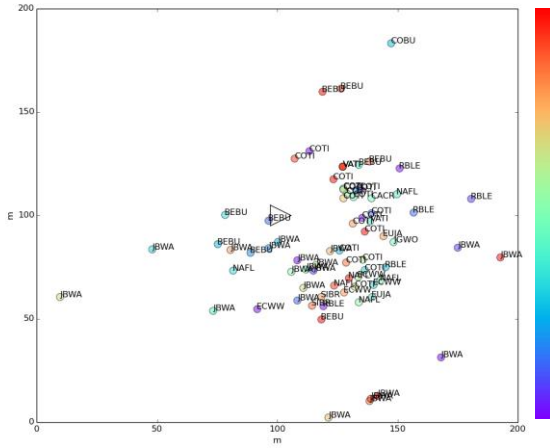


Figure 7 人間による野鳥観測結果。各色点は観測された鳥の種類と大まかな位置を時系列で示す。鳥の位置は１６方位で表示した。鳥の名前コードはTable 1参照のこと。図中の正三角形は、マイクロフォンアレイを設置した正三角形に対応する。

HARK による自動音源定位精度の検討のため、録音全体のスペクトログラムとその再生に基づく手動分析（歌の開始・終了時間の抽出と種の分類）と比較した。その一例を Figure 8 に示す。この比較により、数個体の歌が一定方向で定位・分離されることを確認したが、その精度にはばらつきが見られた。

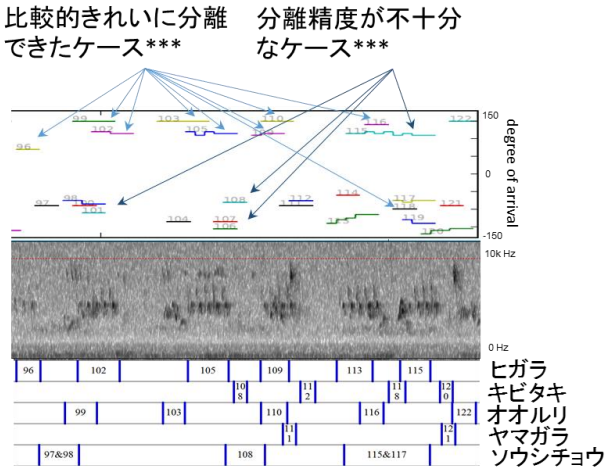


Figure 8 HARK による鳥の自動音源定位・分離 **vs.** 人間による手動分析の一例。上段：**HARK** を使った自動音源定位。図中の各色線に対応する数字は分離されたファイル名を示す。中段：録音全体のスペクトログラム。数字は上記のファイル名に対応する。下段：スペクトログラムとその再生に基づく手動分析。人間の耳による分類が正しいという仮定に基づく。

精度にはばらつきが見られたものの、HARK による音源定位結果は、人間による観測を補完する可能

性を示した。例えば、連続的なヤマガラの威嚇音にかき消され聞き落としていた他の種のさえずりがスペクトログラムにより明らかになった例や、さえずりの音質・バリエーションがよく似たキビタキとオオルリの判別に迷う際に、定位された鳥の位置を前後の時間帯の位置と比較することで区別が容易になった例などがある。

4 考察

スピーカーテストの結果、周辺音よりわずかに大きい音で流した野鳥１１種の歌の定位距離の限界はおよそ５０~７５mと推定された。この結果をもとに、HARK が定位した音源の二次元位置分布と人間による観測結果を比較すると、その分布パターンには類似性が見られた。この距離限界を超えると、人間の耳では容易に識別できる種、例えば比較的大きく特徴的な歌をもつウグイス（JBWA）やソウシチョウ（RBLE）も定位されなかった（Figures 6 & 7）。

定位限界距離の推定に加え、スピーカーテストは HARK による野鳥の定位精度（accuracy）の検討には音源の位置だけではなく音源の分類作業が不可欠であることを明示した。その顕著な例として、３方向の中で最も高い、位置の正確さ（precision）を示した西方向からの再生音実験（Figure 3~5）が、最も低い平均定位指数(Table 4)を示したケースが挙げられる。つまり西方向では定位された再生音自体が少なく、逆に北東および南東方向では比較的多くの再生音が定位されたが、定位された音源の中には再生音以外の多くの音源も含まれていた。

常時定位される野鳥の歌以外の音源はスピーカーテストだけではなく、HARK による定位結果と人間の観測者による野鳥観測の比較実験でも確認された。この現象は、各マイクロフォンアレイが異なる音源を定位しているにもかかわらず、同一の音源として処理することに起因する。これらの音源を除去するためには、ひとつひとつの分離音を人間が聞き分け各マイクロフォンアレイが同じ音源を指しているかを確認する作業が必要であるが、耳作業による音源の聞き分けは多大な労力と時間を要する。この事例は HARK による音源の自動分類性能の必要性を強く示唆する。

HARK による音源の定位性能は、音源からの距離だけではなく環境要因に大きく左右された。音が空気を振動して伝わる際には、空気そのものに加え、伝達途中にぶつかる障害物による減衰、吸収、拡散の影響を受けてひずみが生じる。音の伝達、ひいては HARK の定位性能に影響を及ぼした最も影響力の大きい障害物としては、録音現場付近のプレハブ小屋、地形、植生の３つが挙げられる。Figure 9 は、調査地周辺の航空写真と地形を示す。プレハブ小屋は北方向のマイクロフォンアレイ（Figure 1）にほぼ隣接し、北東方向からの音の伝達の障壁になっただけではなく、その他の方向からの音にも影響を及ぼした。小屋に起因する音の拡散や反響効果は、同じさえずり音が複数の方向で細切れに定位される事例や再生音以外の音源が一定方向で定位される事例（Figure 5）から確認できた。

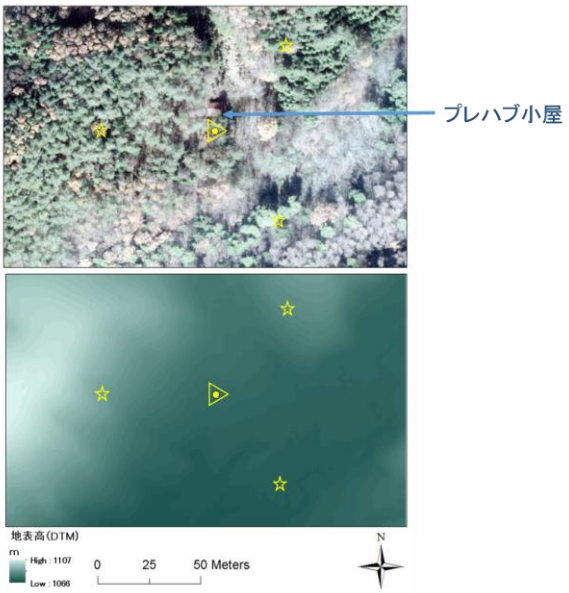


Figure 9 調査地周辺図。上 : 航空写真（撮影年不明、落葉期）。下 : 航空機 LiDAR による数値地形モデル（２０１４年）。地形図に関する観測と作成は中日本航空株式会社による。図中の黄色い三角形は、３つのマイクロフォンアレイを設置した一辺が１０mの正三角形に対応し、星印は各方向５０mから再生音を流した地点を示す。

西、北東方向に広がる急勾配な地形(Figure 9)も HARK による音源定位性能に影響を及ぼしたと考えられる。斜面に加え、音源とマイクロフォンアレイの間に位置する植生の影響も無視できない。特に比較的単純な森林構造を持つ針葉樹内に位置する西、北東方向から鳥の歌の再生音を流した場合、再生音以外の周辺音は主に南東方向で顕著に見られた。これは南東方向に位置する広葉樹林の林床に密生する笹群の葉音と推察される。笹の葉音は南東方向からの再生音を流した際には、HARK による分離性能を低下させ複数の音源を結合する現象を起こした。

音源からの距離に加えて、鳥の鳴き声の種類も定位精度に影響を与えた。一般的に低周波の歌は高周波の音に比べて減衰やそのほかの干渉を受けにくいため効率よく遠くに届く。また、森林への依存度が特に高い鳥は残響の影響や音のひずみを避けるため、比較的単純な構造の歌を歌うことが知られている[9]。本来であれは、スピーカーテストに使われた、ツツドリ、カッコウ、ホトトギス（いずれも Cuckoo 科）は順に約５００、８００、１５００Hz [8]と比較的低周波かつ単純な構造の歌を歌うため、定位される可能性は高いことが予測された。しかしながら実際の平均定位指数は低い結果となった。これは、HARK による音限定の際にノイズカットのため２０００Hz

以上に注目して定位を行ったため、特に低周波の音域で鳴くツツドリとカッコウの歌がカットされる結果となったためである。近距離で定位された音源は、これらの種の歌の一部のうち比較的高音部分がノイズカットをすり抜けた、もしくは偶然同方向の周辺音を拾った可能性がある。一方、高周波でさえずるヒガラや、比較的複雑な歌構造を持つキビタキは高い定位指数を示した。この一因としては、ヒガラやキビタキの歌の周波数が、風などの周辺の低周波の音とはっきり異なることがあげられる。今後のスピーカーテストで考慮すべき点として、音源の音質を鳥の周波数に絞ること、そして鳥種毎に歌の大きさを調整する必要がある。今回の実験では全鳥種を一定の地上高、大きさで流したが、実際には音の伝達効率は鳥の鳴く位置や環境の影響を受け、同時に鳥の歌の大きさは鳥の体の大きさなどに比例するからである。

5 おわりに

本稿は、複数のマイクロフォンアレイを用いた野外に置けるスピーカーテストおよび野鳥の音源定位精度の予備的調査の結果を報告した。スピーカーテストでは、晴天の無風状態下で周辺音よりわずかに大きな音で鳥の歌の再生音を流した場合、その種類によりマイクロフォンアレイから約５０～７５ｍの距離まで定位できることが明らかになった。また、音源定位性能は周辺の人工物、地形、植生に加え、鳥の歌の周波数に影響を受けることが示された。これらの点を考慮した上で、HARK による音源定位結果と人間による野鳥観測結果を比較すると相互間には類似した二次元位置分布が示された。さらに、分離された音源のスペクトログラムとその再生に基づく手動分析による種類の識別や歌の始まりと終わりの切り出しは、鳥がいつどこで鳴いたかという情報をより明確にした。このことは、HARK が人間による野鳥観測を補完する可能性を示唆している。

いずれの実験も初期的な段階にあるが、位置情報をもつ音声データを解析することは、野鳥の生態理解へ向け重要な意義を持つ。野鳥観測においては、瞬時に識別が難しい場合が頻繁に起こりうる。例えばオオルリやキビタキなど声の音質や歌のフレーズが似た個体が交互に鳴きその識別に迷う場合、さえずりの位置情報を前後の情報と比較することで２種の聞き分けが容易になった。このような例では特にデータの再現性が大きな意義を持つ。

今後の課題として、HARK による自動分類機能の充実があげられる。野鳥のさえずりは、同種でも様々なレパートリーがあり、人間による手動分析は多大な時間と労力を要する。また、人的エラーの可能性も否めない。自動分類の機能が加われば分析の効率は格段に向上し、野鳥の位置的空間及び時間的空間利用の解明に向けた応用の可能性が高まると考えられる。

謝辞

高部直紀氏、近藤崇氏（名古屋大）のフィールドワークへのご協力に謝意を表する。また航空写真とLiDAR 地表モデルをご提供いただいた山本一清先生（名古屋大学）に感謝申し上げる。本研究の一部はJSPS 科研費 15K00335，24220006 の助成を受けたものである。

参考文献

[1]Collier,T.C., Kirschel, A.N.G., and Taylor, C.E. (2010). Acoustic localization of antbirds in a Mexican rainforest using a wireless sensor network. Journal of Acoustical Society of America. 128(1). 182-189.

[2]Blumstein D.T., Mennill, D.J., Clemins, P., Girod,L., Yao, K., Patricelli, G., Deppe, J.L., Krakauer, A.H., Clark, C., Cortopassi, K.A., Hanser, S.F., McCowan, B., Ali, A.M., and Kirschel, A.N.G. (2011). Acoustic monitoring in terrestrial environmental using microphone arrays: applications, technological considerations and prospectus. Journal of Applied Ecology, 48. 758-767.

[3]Nakadai, K., Takahashi, T., Okuno, H.G., Nakajima, H., Hasegawa, Y., and Tsujino, H. (2010). Design and implementation of robot audition system “HARK”-open source software for listening to three simultaneous speakers. Advanced Robotics, 24. 739-761.

[4]Suzuki, R. Hedley, R., and Cody, M.L. (2015). Exploring temporal sound-space partitioning in bird communities emerging from inter- and intra-specific variations in behavioral plasticity using a microphone array. Abstract book of Joint Meeting of the American Ornithologists' Union & Cooper Ornithological Society. 86.

[5]蒲谷鶴彦・松田道生著 日本野鳥大鑑鳴き声 420. (2011).小学館.

[6]Noise level meter. Retrieved October 1, 2015, from https://itunes.apple.com/jp/app/noiselevelmeter/id694670057?ign-mpt=uo%3D5

[7]Ralph, C.J., Droege, S., and Sauer, J.R. Managing and monitoring point counts: standards and applications. (1995). USDA Forest Service general technical report. PSW-GTR 149.

[8]Boersma, P and Weenink, D. (2015). Praat: doing phonetics by computer (Version 5.4.20) [Computer program]. Retrieved July 26, 2015, from http://www.praat.org/

[9]Schmidt, R.O. (1986). Multiple emitter location and signal parameter estimation. Antennas and Propagation, IEEE Transactions, 34.3. 276-280.

[10]Nakajima, H., Nakadai,K., Hasegawa,Y. and Tsujino,H. (2008). Adaptive Step-size Parameter Control for real World Blind Source Separation, In Proc. ICASSP 149-152.

[11]Gill, F.B. (2007). *Ornithology*. NY: W.H. Freeman and Company.

HARK SaaS: ロボット聴覚ソフトウェア HARK のクラウドサービスの設計と開発

HRAK SaaS: Design and Implementation of Robot Audition Software HARK as a Service

水本武志, 中臺一博

Takeshi MIZUMOTO, Kazuhiro NAKADAI

株式会社 ホンダ・リサーチ・インスティチュート・ジャパン

Honda Research Institute Japan, Co., Ltd.

t.mizumoto@jp.honda-ri.com, nakadai@jp.honda-ri.com

Abstract

本稿では、2008 年より公開を開始したロボット聴覚ソフトウェア HARK[1] をクラウドサービスとして実装した HARK SaaS (Software as a Service) について報告する。HARK SaaS は多チャネル音ファイルを受信して HARK が提供する音源定位や音源分離などの結果を返すクラウドサービスである。従来の HARK で必須であったローカル計算機へのインストール作業や、高負荷の信号処理が実行できる高い性能要求が不要となるため、より簡単に HARK を利用できる。評価実験では、Amazon Web Services (AWS) を用いてサーバ 6 台構成で応答時間と処理時間を計測した。その結果、応答時間は 100 並列アクセスまでは 100msec 程度であること、処理時間はオーバーヘッドが無視できるほど入力データが長い場合は実時間処理が可能であることを確認した。

1 はじめに

ロボット聴覚分野で研究開発されてきた音源定位・音源分離などのマイクロホンアレイ処理技術が実装されたロボット聴覚ソフトウェア HARK が 2008 年から公開されている [Nakadai 09]。公開以来 HARK は様々なシステム、例えばクイズの司会 [Nishimuta 15] やテレプレゼンスロボット [Mizumoto 12] に応用されている。また、ユーザビリティの面でもインストーラやドキュメントの整備を行うなど、利便性向上の継続的な努力が続けられている。しかし、既存システムへの組み込みには ROS (Robot OS) 等の別ソフトウェアやソケット通信の実装が必要となるなど、依然ハードルは高い。さらに、現在の HARK は信号処理を単一の計算機上で実行する必要があるので、計算機への要求スペックも高い。そのため、低スペックの計算機、例えば組み込みデバイスでの利用には専用の実装が必要であった [中臺 15]。

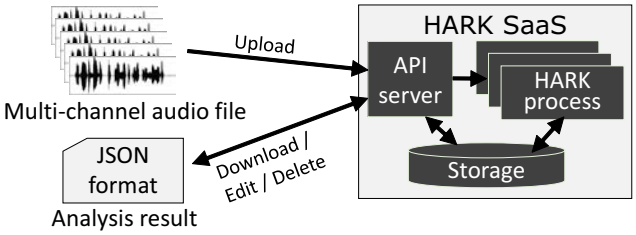


図 1: HARK SaaS の概要

一方、近年の無線ネットワーク環境の普及や Internet of Things の流行などにより、ネットワーク接続できるセンサデバイスや小型計算機が多く出回っている。例えば、Intel® Edison (Intel 社) や Raspberry Pi® (Raspberry Pi Foundation)、BeagleBoard ™ (テキサス・インスツルメンツ社、Digi-Key) に代表されるようなネットワーク接続可能な小型計算機は容易に入手できる一般的なものになっている。

これらの状況にもとづいて、本稿では、HARK をクラウドサービスとして設計・開発した **HARK SaaS** について報告する。図 1 に示す本サービスの概要のとおり、ユーザは HARK SaaS への多チャネル音ファイルアップロードと、処理結果の取得・更新・削除ができる。本サービスはサーバ側で全処理を行うため、ローカル計算機にネットワーク接続が必須となるものの、サービス利用をするだけなので従来のインストール作業を回避でき、信号処理をクラウド上のサーバへ移譲するので要求スペックが低くなる。このため、従来のローカル型 HARK の課題の解決が期待できる。

HARK SaaS 設計上の要求条件は次の 3 点である。

[1] Honda Research Institute Japan Audition for Robots with Kyoto University

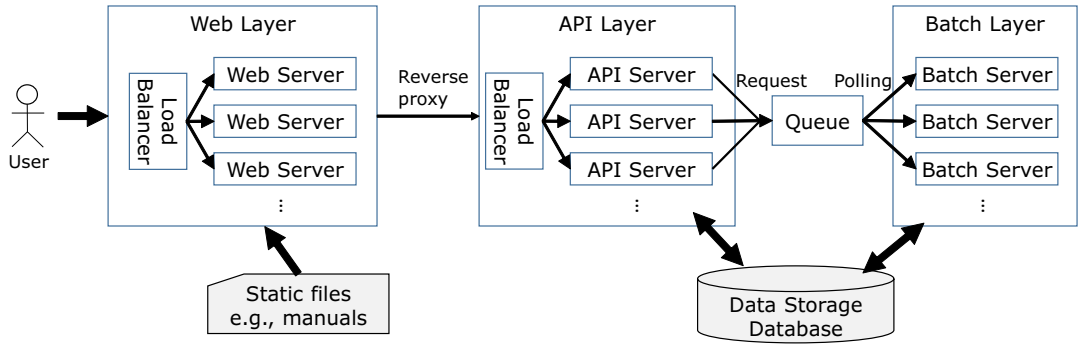


図 2: HARK SaaS のアーキテクチャ

インタフェースの汎用性

他のソフトウェアやクラウドサービスとの組み合わせが容易であれば、HARK SaaS の応用を行いやすくなる。そこで、標準規格に準拠した汎用的なインタフェースを持たせる必要がある。

ユーザビリティ

プログラムを作成せずに HARK の機能を利用したいユーザと、本サービスと組み合わせたソフトウェアを開発したいユーザの両方に利用しやすいユーザインタフェースを設計する必要がある。

信頼性

サービスとしてのセキュリティや安定性を高める設計が必要である。また、処理結果についても、従来のローカル型 HARK との互換性を持たせる必要がある。

本稿の構成は次のとおりである。まず、2 章で関連するクラウドサービスについて議論する。次に、3 章で本サービスのアーキテクチャやデータ構造を設計する。4 章で本サービスの基本性能に関する実験結果について議論し、5 章で本稿をまとめる。

2 音のクラウドサービス

本章では、音声や音楽などの音データを利用するクラウドサービス (以下、音のクラウドサービスと呼ぶ) について議論し、本サービスの位置付けを明らかにする。音のクラウドサービスには、(1) アップロードされた音データをそのまま用いるサービスと、(2) 音データの処理を伴うサービスがある。本サービスはマイクロホンアレイ処理を行うので後者に分類される。以下で述べるように様々な音のクラウドサービスが公開されているが、後者のサービスは単一チャネル処理のみであり、本サービスのように多チャネル音データを処理するサービスは筆者らの知る限り存在しない。

アップロードされた音データをそのままに用いるサービスは、ソーシャルネットワーキングや音の共有を目的としたサービスが一般的である。例えば、 SoundCloud® [2] や YouTube® [3] は、ユーザがアップロードした音データを、他のユーザと共有することができる。これらのサービスでは、音データ以外にもユーザ自身が追加したタグやコメント、5 段階評価などの音データに対する付加情報が合わせて提供される。

音データの処理を伴うサービスを、対象とするデータが音声のものと音楽のものに分類して議論する。音声を対象としたサービスには、Google 社の音声検索サービス、Apple 社のスマートフォン上で動作する音声対話サービス Siri®、ロボットインタラクションを目的とした音声認識と音声合成サービス Rospeex [杉浦 13] などがある。これらは入力された音声を認識し、認識結果そのものや、認識結果に基づく検索結果、音声応答などを返すサービスである。また、インターネット上のポッドキャストや動画を音声認識によってテキスト化し検索できるサービス PodCastle [Goto 13] は、ユーザによるアノテーションを利用した性能向上を組み合わせたサービスである。音楽を対象としたサービスには、その基礎技術となる歌声や音楽の分析・検索に関する研究が数多くされており [後藤 08]、公開されているサービスにも、SoundHound 社の ハミング検索サービス midomi®[4] や、音楽からサビ区間やメロディを推定し、表示することで能動的な音楽鑑賞を可能とする Songle [5] [Goto 11] などがある。

3 HARK SaaS の設計と実装

本章では、 HARK SaaS の詳細について述べる。まず 3.1 節で、 1 章で述べた 3 点の要求条件を検討しながらアーキテクチャを設計する。次に 3.2 節でデータ構造を設計し、3.3 節でサービスの実装について述べる。

3.1 アーキテクチャ設計

インタフェースの汎用性について検討する。まず、本サービスの全機能は HTTPS リクエストを用いること、送受

[2] https://soundcloud.com
[3] https://www.youtube.com
[4] http://www.midomi.co.jp
[5] http://songle.jp

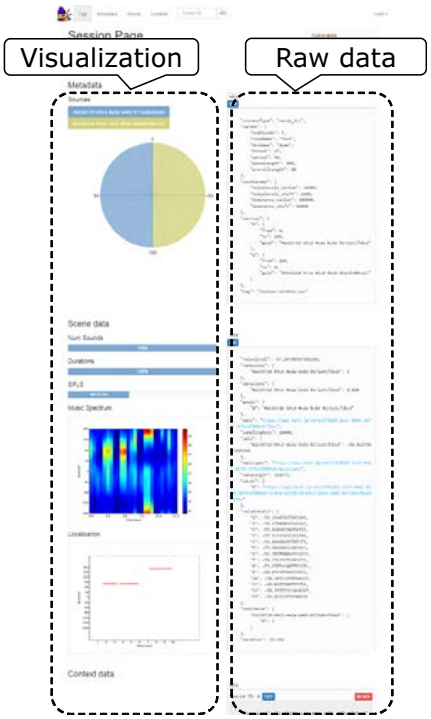


図 3: 可視化機能を備えた Web UI

```
import pyhark.saas
h = pyhark.saas.PyHarkSaaS("API_KEY", "API_SECRET")
h.login()                                # 認証
h.createSession(metadata)                # パラメータ設定
h.uploadFile(open(filename, 'rb'))       # ファイル送信
h.wait()                                 # 処理終了待ち
result = h.getResults()                  # 処理結果受信
```

図 4: HARK SaaS SDK を用いたサンプルコード

信データは JSON フォーマットを用いることとする。これらはいずれも標準規格なので、ほぼあらゆるプログラミング言語からこれらのインタフェースを介して本サービスを利用することが可能となる。次に、インタフェースの複雑さ制限するため、ローカル型 HARK の自由に信号処理のデータフローを構成できる機能に制限を加える。代わりに標準的な音源定位と音源分離を行う構成を提供し、多くのパラメータ、例えば音源定位・音源分離用伝達関数、音源定位閾値、定位長などを提供することで、インタフェースの汎用性とサービスの利便性の両立を図る。

ユーザビリティについて検討する。プログラミングを行わずに HARK の機能を利用したいユーザに対しては、Web インタフェースに解析結果の可視化機能を提供する (図 3)。本インタフェースを用いれば、ブラウザ操作のみで音ファイルの送信と結果の確認ができる。一方、プログラミングを行ってソフトウェアに組み込みたいユーザについては、Software Development Kit (SDK) を提供する。SDK は Python モジュールとして提供し、認証や HARK 処理リクエスト、結果の取得ができる。SDK を用いたサンプルコードを図 4 に示す。

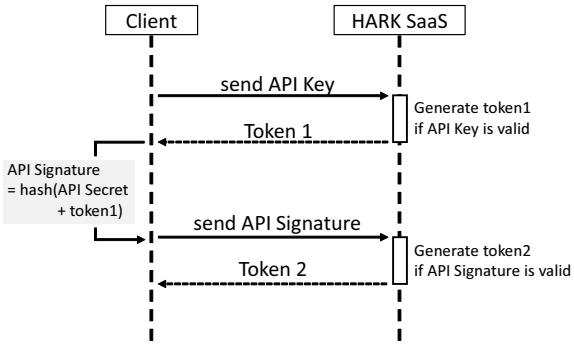


図 5: 認証シーケンス

信頼性について検討する。まず、サービスの安定性を実現するためには、処理の特性ごとにレイヤを分割しそれらを疎結合させる構造と、負荷変動に追従できるスケーラビリティを持たせる必要がある。これらを満たすアーキテクチャを図 2 に示す。前者については、(1) ドキュメントなどの静的ファイルを配信とリクエストの後段への転送を行う Web レイヤ、(2) データベースアクセスが必要なリクエストの処理と HARK 実行リクエストの後段への転送を行う API レイヤ、(3) HARK の実行や後処理などの時間のかかる処理を行う Batch レイヤに分割する。各レイヤの疎結合構造は次のように実現する。まず、Web レイヤから API レイヤへの転送には API レイヤのロードバランサを介することとする。この設計によって、Web レイヤの API レイヤサーバ台数への依存性を排除できる。次に、API レイヤから Batch レイヤへのリクエスト転送をキューを介することとする。この設計によって API レイヤのサーバと Batch レイヤのサーバはキューのみにアクセスすればよく、互いのサーバ台数への依存性がなくなる。後者のスケーラビリティについては、 Web レイヤと API レイヤの前面にロードバランサを配置する。この設計によって、負荷が高ければロードバランサに接続するサーバ台数を増やし、負荷が低ければサーバ台数を減らすことでレイヤ全体の処理性能を制御できる。

次に、サービスのセキュリティを実現するために次の設計を行う。(1) ユーザ認証。ユーザごとに 2 つの情報 (API Key, API Secret) を提供し、全てのサービスへのアクセスについて図 5 に示す手順で得られた一時認証トークン (Token2) の提供を要求する。また、一時認証トークンは短時間で無効化することで、流出時の影響を制限する。(2) 暗号化。通信を暗号化するために、本サービスへのリクエストを全て HTTPS アクセスのみに制限する。

最後に、ローカル版 HARK との互換性については、Batch レイヤのサーバでローカル版 HARK 自体を実行し、得られる結果を全て次節で設計するデータ構造で表現することとする。これによって、ローカル版と同じ実装を用いるので結果の互換性を確保できる。

3.2 データ構造の設計

本節では HARK SaaS で利用するデータ構造を設計する。まず、本サービスにアップロードされたひとつの音ファイルをデータ単位と定義し、**セッション** と呼ぶ。本設計では全ての処理結果や処理パラメータは全てセッション単位で表現する。

ひとつのセッションに関するデータを 3 種類に分類する。

メタデータ

ユーザが与えるデータ。例えば、HARK に与えるパラメータや音源方向ラベルが含まれる。音源方向ラベルとは、方向範囲ごとに定めるラベルのことで、これを適切に設定すれば、マイクロホンアレイと音源の位置関係が変化しない場合 (会議など) に、音源定位された音イベントへラベルを自動付与できる。

コンテキスト情報

音イベント毎のデータ。HARK によって音源定位された音イベント毎に定義される。例えば、音イベントの開始時間と終了時間、仰角と方位角、音量、分離音などが含まれる。

シーン情報

シーン全体のデータ。コンテキスト情報を集計した結果など、ひとつのセッション全体に対して定義される。例えば、処理される音ファイルの長さやサンプリングレート、音量時系列等の音ファイルそのものに関する情報や、音源方向ラベルごとの音イベント数、その合計時間などの音源方向ラベルごとの情報が含まれる。

3.3 HARK SaaS 実装

本サービスを AWS 上に実装した。各コンポーネントに用いた AWS のサービスは次のとおりである。Web レイヤと API レイヤの負荷分散には Elastic Load Balancer (Amazon ELB) を、Batch レイヤが監視するキューには Simple Queue Service (Amazon SQS) を、アップロードされる音声ファイルや処理結果の保存には Simple Storage Service (Amazon S3) を、処理結果などのその他のデータの保存は Amazon RDS を利用した。

4 評価実験

本章では、HARK SaaS の評価実験について述べる。実験では、応答時間と処理時間の計測を通して本サービスの基本性能を明らかにし、アプリケーションのサンプルによって本サービスの応用例を示す。

4.1 実験設定

実験に使用した HARK SaaS サービスの構成は 3 章で述べたとおりである。実験では、 Web, API, Batch 各レイヤに、サーバを 2 台ずつ割り当てた。また、全ての実験でローカル計算機は 1 台のみを使用し、大規模な負荷試験で標準的に行われる複数台の計算機を用いたアクセスは行っていない。実験に使用した音ファイルには、標準的な室内で 8 チャネルのマイクロホンアレイで収録した 6 名の自由会話を用いた。処理時間計測で用いる音ファイルは 10 秒, 10 分, 30 分のデータとし、応答時間計測で用いる音ファイルはすべて 10 秒のデータとした。

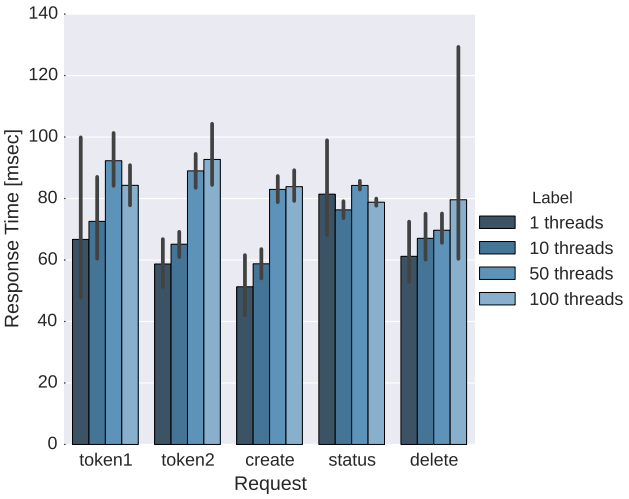


図 6: 実験 1: リクエストの応答速度

4.2 実験 1: 応答時間

応答時間計測には、 オープンソース・ソフトウェア JMeter [6] を利用した。試験シナリオは、認証から処理リクエスト、結果の取得と削除までの一連の処理とした。本シナリオは次の 6 種類のリクエストで構成される。

1. Token 1 (認証)
2. Token 2 (認証)
3. Create (セッションの作成)
4. Upload (データアップロード)
5. Status (処理状態確認)
6. Delete (セッション削除)

上記シナリオを、同時並列実行数を 1, 10, 50, 100 と変化させながら各 10 回ずつ実行し、応答時間を計測した。

応答時間を図 6 と 図 7 に示す。横軸のラベルは試験シナリオの各リクエストを表し、ラベル中の棒グラフはそれぞれ同時アクセス数に対応した応答時間を表し、エラーバーは当該リクエストの応答時間の標準偏差を表す。

図 6 に示す通り、 100 並列アクセスの場合でも応答時間は 100msec 程度を維持しているので、この負荷であれば安定した処理ができているといえる。一方で、ファイルアップロードについては 図 7 に示す通り、同一のファイルをアップロードしたにもかかわらず応答時間が伸びて

[6] http://jmeter.apache.org/

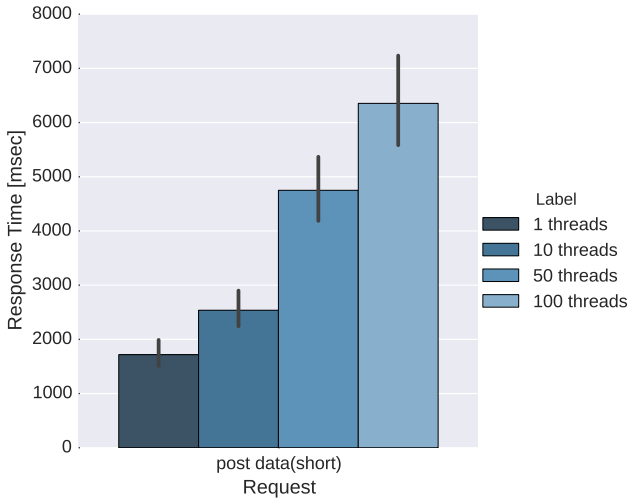


図 7: 実験 1: ファイルアップロードリクエストの応答速度

いる。つまり、本実験で用いた 2 台の構成では処理が間に合わず、待ちが発生している。

これより、本実験で用いた構成の性能では、結果取得や認証の処理には足るものの、同様の規模でファイルのアップロードの処理には足りない。したがって、アップロード量の状況に応じた台数の増減が必要である。

4.3 実験 2: 処理時間

ローカル版 HARK では実時間で音ファイルの処理を行えるが、HARK SaaS ではリクエストの処理やデータ転送等のオーバーヘッドが含まれるので、システム全体の処理時間はより長くなる。本実験では, 異なるデータ長の多チャネル音データの処理時間を計測し、HARK SaaS 全体としての処理速度を評価する。実験には、8 チャネルの音ファイルを使用し、その長さは 3 種類 (10 秒, 10 分, 30 分) とした。また、音源定位のされやすさに関する閾値を 26 - 30 まで 1 ずつ 5 段階に変化させ、音源定位数による処理時間の変化も評価した。

実験結果を図 8 に示す。縦軸が HARK SaaS へのリクエストが受理されてから処理結果が戻るまでの全体の処理時間を表し、横軸は入力データ長を表す。また、処理時間と入力データ長の比で表されるリアルタイムファクタ (RT) を表 1 に示す。RT とは実時間制を表す指標で、 1 より小さければ、入力データ長より処理時間が短いので、実時間性があると言える。

結果について議論する。まず、閾値を変化させてもほぼ処理時間に変化はなかった。これは、HARK 処理と音イベントの後処理を並行して行っているために後処理の時間の影響が小さいことが理由であると考えられる。次に、表 1 に示すとおり 10 分以上のデータでは実時間性を確保できているが、10 秒では確保できていない。これは、クラウドサービス化に伴うオーバーヘッドの影響の方が、HARK の処理時間の短さよりも大きいことが原因である。

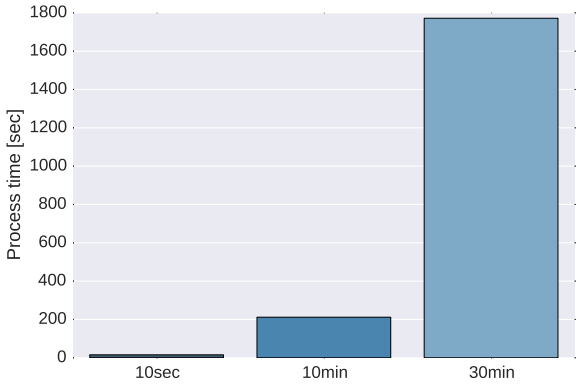


図 8: 実験 2: 入力データごとの処理時間計測結果

表 1: 実験 2: リアルタイムファクタ

データ長 [sec]	処理時間 [sec]	RT
10	15.6	1.56
300	211.7	0.35
1800	1771.7	0.98

4.4 実験 3: HARK SaaS サンプルアプリケーション

HARK SaaS の応用システムの例として、音環境可視化アプリケーションを構築した (図 9) 本アプリケーションは、録音された音ファイルを HARK SaaS へアップロードし、処理結果を受信し、処理結果とメタデータで設定された音源方向ラベルを利用して結果を可視化する。本アプリケーションは音源方向ラベルごとの音イベント集計結果を色分けして表示するので、方向ごとの音環境分析ができる。なお、本アプリケーションは HARK SaaS の Python SDK と、可視化ライブラリ Seaborn、 Matplotlib を使用している。

可視化画面は 4 部分から構成されている。まず、図左上は方向ごとの音イベント数を表し、音源方向ラベルごとに色分けがされている。この図から、音源方向ラベルの方向に関する傾向が分かる。例えば図 9 の場合、緑色の音源方向ラベルの音イベントは 0 度 方向から多く発生していることがわかる。次に、図右上部は音源方向ラベルごとの音イベントの継続時間のヒストグラムと平均値を表す。ヒストグラムから音源方向ラベルごとの継続時間の傾向を分析でき、右端の平均値から音源方向ラベル同士の継続時間の比較ができる。続いて、図右中部は音源方向ラベルごとの音イベント音量のヒストグラムと平均値を表す。ここでも継続時間と同様に音源方向ラベルごとの分析やそれぞれの比較ができる。最後に、図下部は時間・方向ごとの音イベントを表す。この図より、-120 度の方向からは 100 秒から 180 秒、280 から 380 秒、 430 秒から 500 秒の 3 回にわたって音イベントが連続的に発生していることがわかる (濃青で表示)。

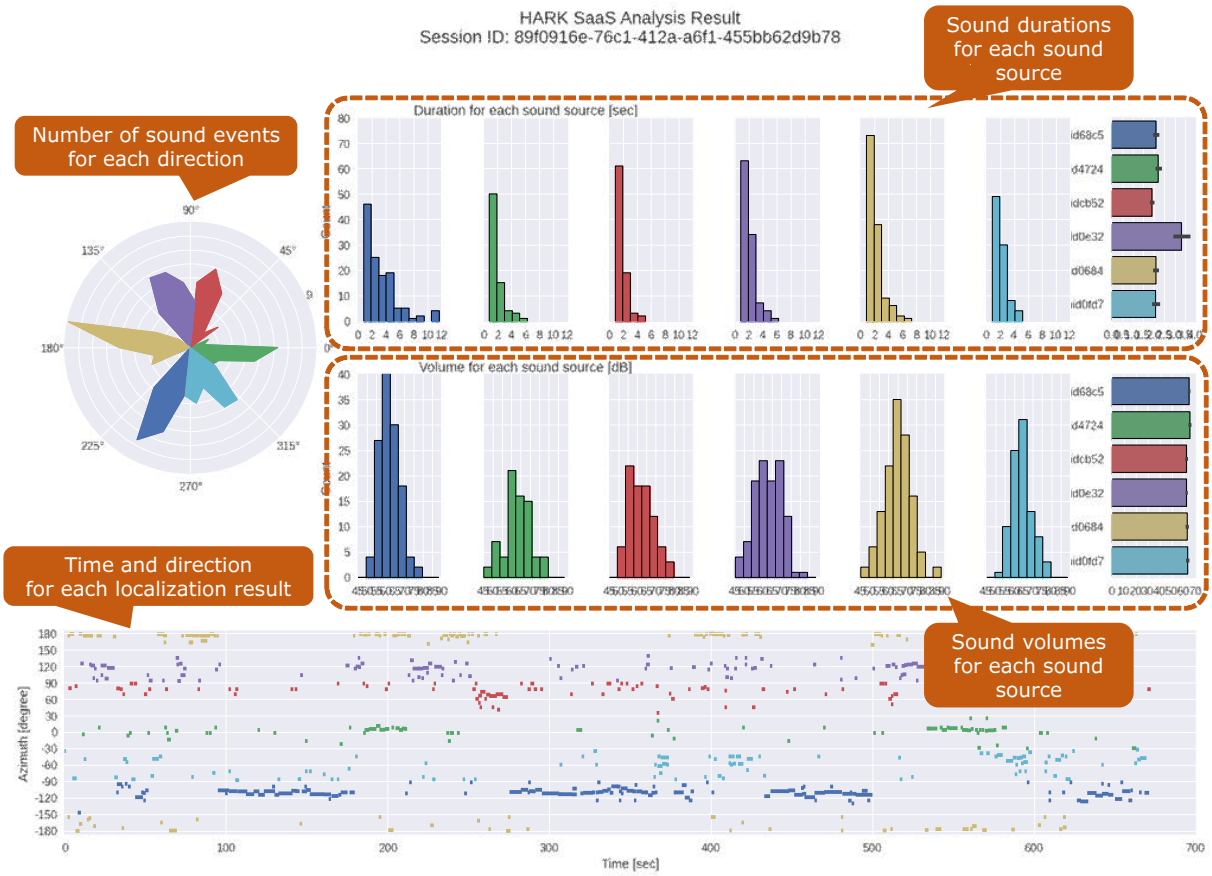


図 9: HARK SaaS サンプルアプリケーション： 音環境可視化

5 まとめ

本稿では、ロボット聴覚ソフトウェア HARK をクラウドサービスとして実装した HARK SaaS について報告した。実験の結果、100 並列アクセスまでは応答時間が 100msec 程度と安定していること、10 秒のデータの場合はオーバーヘッドのために実時間性が失われるものの、それより長いデータであれば実時間性が確保できることが明らかになった。今後は、応答速度の向上やより大規模なアクセスに耐えうる冗長設計、分離音の後処理の充実による機能拡大などを行う予定である。

参考文献

[Goto 11] Goto, M., Yoshii, K., Fujihara, H., Mauch, M., and Nakano, T.: Songle: A Web Service for Active Music Listening Improved by User Contributions, in *ISMIR*, pp. 311–316 (2011)

[Goto 13] Goto, M., Ogata, J., and Eto, K.: PodCastle: A web 2.0 Approach to Spech Recognition Research, in *Interspeech*, pp. 2397–2400 (2013)

[Mizumoto 12] Mizumoto, T., Nakadai, K., Yoshida, T., R. Takeda, T. T., T. Otsuka, and Okuno, H. G.: Design and Implementation of Selectable Sound Separation on the Texai Telepresence System using HARK, in *ICRA*, pp. 694–699 (2012)

[Nakadai 09] Nakadai, K., Okuno, H. G., Nakajima, H., Hasegawa, Y., and Tsujino, H.: Design and Implementation of Robot Audition System “HARK”, *Advanced Robotics*, Vol. 24, pp. 739–761 (2009), doi:10.1163/016918610X493561

[Nishimuta 15] Nishimuta, I., Yoshii, K., Itoyama, K., and Okuno, H. G.: Toward a Quizmaster Robot for Speech-based Multiparty Interaction, *Advanced Robotics*, Vol. 29, No. 18, pp. 1205–1219 (2015)

[後藤 08] 後藤 真孝, 齋藤 毅, 中野 倫靖, 藤原 弘将：歌声情報処理の最近の研究, 日本音響学会誌, Vol. 64, No. 10, pp. 616–623 (2008)

[杉浦 13] 杉浦 孔明, 堀 智織, 是津 耕司：rospeex:クラウド型音声コミュニケーションを実現する ROS 向けツールキット, 電子情報通信学会技術研究報告, クラウドネットワークロボット, 第 113 巻, pp. 7–10 (2013)

[中臺 15] 中臺 一博, 水本 武志, 中村 圭佑：モバイル端末用マクロホンアレイシステムの開発とコミュニケーション支援への適用, ロボット学会学術講演会 (2015)

〒162 東京都新宿区津久戸町 4-7 OS ビル 402 号室 03-5261-3401 Fax: 03-5261-3402

(本研究会についてのお問い合わせは下記にお願いします.)

SIG-AI-Challenges home page (WWW): http://winnie.kuis.kyoto-u.ac.jp/SIG-Challenge/