



NTT Communication Science Laboratories
OPEN HOUSE 2019

May

30Thr

12:00 ~ 17:30

31Fri

9:30 ~ 16:00

Venue: NTT Keihanna Building



Innovative R&D by NTT



corevo

Science of Machine Learning

- 01 Learning and finding congestion-free routes
~Online shortest path algorithm with binary decision diagrams~
- 02 Efficient and comfortable AC control by AI
~Environment reproduction and control optimization system~
- 03 Recover urban people flow from population data
~People flow estimation from spatiotemporal population data~
- 04 Improving the accuracy of deep learning
~Larger capacity output function for deep learning~
- 05 Which is cause? Which is effect? Learn from data!
~Causal inference in time series via supervised learning~
- 06 Forecasting future data for unobserved locations
~Tensor factorization for spatio-temporal data analysis~
- 07 Search suitable for various viewpoints
~“Pitarie”: Picture book search with graph index based search~

Science of Communication and Computation

- 08 We can transmit messages to the efficiency limit
~Error correcting code achieving the Shannon limit~
- 09 New secrets threaten past secrets
~Vulnerability assessment of quantum secret sharing~
- 10 Analyzing the discourse structure behind the text
~Hierarchical top-down RST parsing based on neural networks~
- 11 When children begin to understand hiragana
~Emergent literacy development in Japanese~
- 12 Measuring emotional response and emotion sharing
~Quantitative assessment of empathic communication~
- 13 Touch, enhance, and measure the empathy in crowd
~Towards tactile enhanced crowd empathetic communication~
- 14 Robot understands events in your story
~Chat-oriented dialogue system based on event understanding~

Science of Media Information

- 15 Voice command and speech communication in car
~World's best voice capture and recognition technologies~
- 16 Learning speech recognition from small paired data
~Semi-supervised end-to-end training with text-to-speech~
- 17 Who spoke when & what? How many people were there?
~All-neural source separation, counting and diarization model~
- 18 Changing your voice and speaking style
~Voice and prosody conversion with sequence-to-sequence model ~
- 19 Face-to-voice conversion and voice-to-face conversion
~Crossmodal voice conversion with deep generative models~
- 20 Learning unknown objects from speech and vision
~Crossmodal audio-visual concept discovery~
- 21 Neural audio captioning
~Generating text describing non-speech audio~
- 22 Recognizing types and shapes of objects from sound
~Crossmodal audio-visual analysis for scene understanding~

Science of Human

- 23 Speech of chirping birds, music of bubbling water
~Sound texture conversion with an auditory model~
- 24 Danswing papers
~An illusion to give motion impressions to papers~
- 25 Measuring visual abilities in a delightful manner
~Self eye-check system using video games and tablet PCs~
- 26 How do winners control their mental states?
~Physiological states and sports performance in real games~
- 27 Split-second brain function at baseball hitting
~Instantaneous cooperation between vision and action~
- 28 Designing technologies for mindful inclusion
~How sharing caregiving data affects family communication~
- 29 Real-world motion that the body sees
~Distinct visuomotor control revealed by natural statistics~
- 30 Creating a walking sensation for the seated
~A sensation of pseudo-walking expands peripersonal space~

01

Learning and finding congestion-free routes

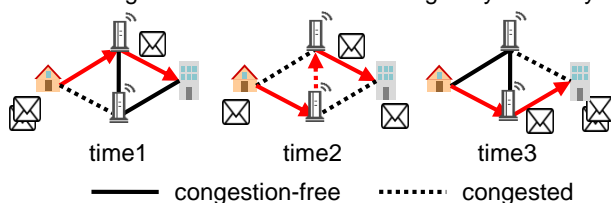
- Online shortest path algorithm with binary decision diagrams -

Abstract

We consider **adaptively finding congestion-free routes** connecting specified two locations on a network. In many practical scenarios, congestion on a network, or transmission time taken to send messages, changes dynamically. Therefore, we need to effectively learn congestion using past congestion data and efficiently find a congestion-free route each time we send a message. While there exist learning algorithms that can be used for predicting congestion, they incur **too much computation cost due to the presence of a huge number of possible routes**. We overcome this difficulty by using the **zero-suppressed binary decision diagram (ZDD)**, which is a compact representation of all possible routes. We develop a learning algorithm that can work on ZDDs without examining all possible routes explicitly, which enables us to **find congestion-free routes far more efficiently than the existing algorithms**.

Problem Setting

We choose a route every time we send a message, where congestion on the network changes dynamically.



We aim to find a congestion-free route at each time.

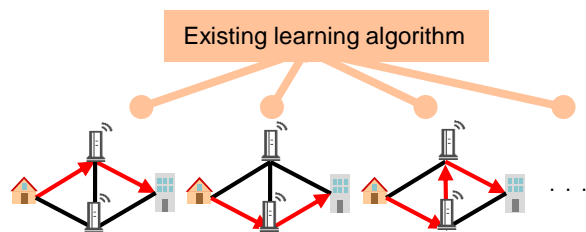
Difficulties of the Problem

First, we cannot see how congested each edge is when sending a message.

For example, cyberattacks may cause sudden congestion, which is sometimes hard to observe without sending a message and getting a feedback.

Second, since there are a huge number of possible routes, predicting congestion for each route is too costly.

Existing methods (e.g., [2]) learn and predict congestion by examining all possible routes, which takes too long time.

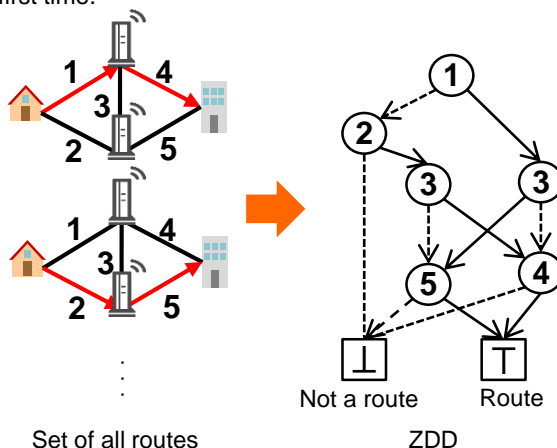


Efficient Algorithm with ZDDs

Our algorithm first compactly represents the set of all possible routes using the **zero-suppressed binary decision diagram (ZDD)**, and then performs learning algorithm [2] on the ZDD without examining all routes.



We have achieved to find congestion-free routes adaptively on a network with dozens of nodes for the first time.



Point 1. Can learn congestion-free routes efficiently.

All operations are performed on compact ZDDs, and thus our algorithm can run faster than existing algorithms.

Point 2. Need not reconstruct ZDDs at each time.

Once a ZDD is constructed, we can reuse it at each time. This makes our algorithm so efficient as to deal with sudden congestion.

References

- [1] S. Sakaue, M. Ishihata, S. Minato, "Efficient bandit combinatorial optimization algorithm with zero-suppressed binary decision diagrams," in *Proc. 21st International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.
- [2] N. Cesa-Bianchi, G. Lugosi, "Combinatorial bandits," *Journal of Computer and System Sciences*, Vol. 78, No. 5, pp. 1404–1422, 2012.

Contact

Shinsaku Sakaue Email: cs-liaison-ml at hco.ntt.co.jp
Linguistic Intelligence Research Group, Innovative Communication Laboratory



Innovative R&D by NTT
Open House 2019

Copyright © 2019 NTT. All Rights Reserved.

02

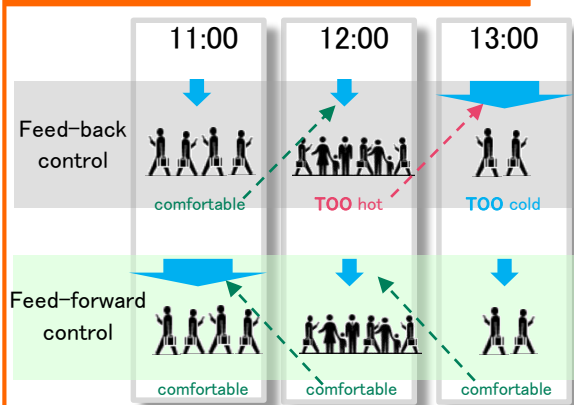
Efficient and comfortable AC control by AI

- Environment reproduction and control optimization system -

Abstract

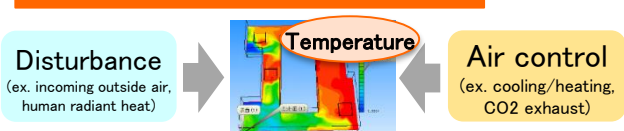
We propose an air-conditioning control system by AI to save more energy and to be more comfortable. In a large-scale facility, it takes several time to stabilize temperature. Traditional and typical way of control system, commonly known as **feed-back control**, makes sometimes uncomfortable and consumes extra energy by the time-delay. On the other hand, **feed-forward control** determines suitable control with predicting environment status of the facility. For example, if congestion is predicted, the air-flow could be increased or decreased in advance, which would make the facility's temperature suitable. We developed AI consisting of **environment reproduction system** and **control optimization system** to calculate the optimal operation schedule for multiple air-conditioning flows, and **demonstrated the importance of feed-forward control through field trial** at "COREDO Muromachi", which is one of the largest-scale commercial facilities, with NTT-Facilities and MITSUI FUDOSAN.

Background: feed-back or feed-forward



Prediction is needed to keep temperature comfortable.

Problem1: complex environment

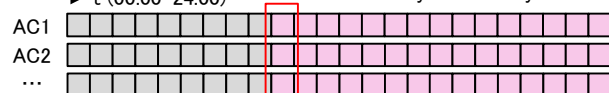


It is hard to build prediction model.

Problem2: insufficient data variation

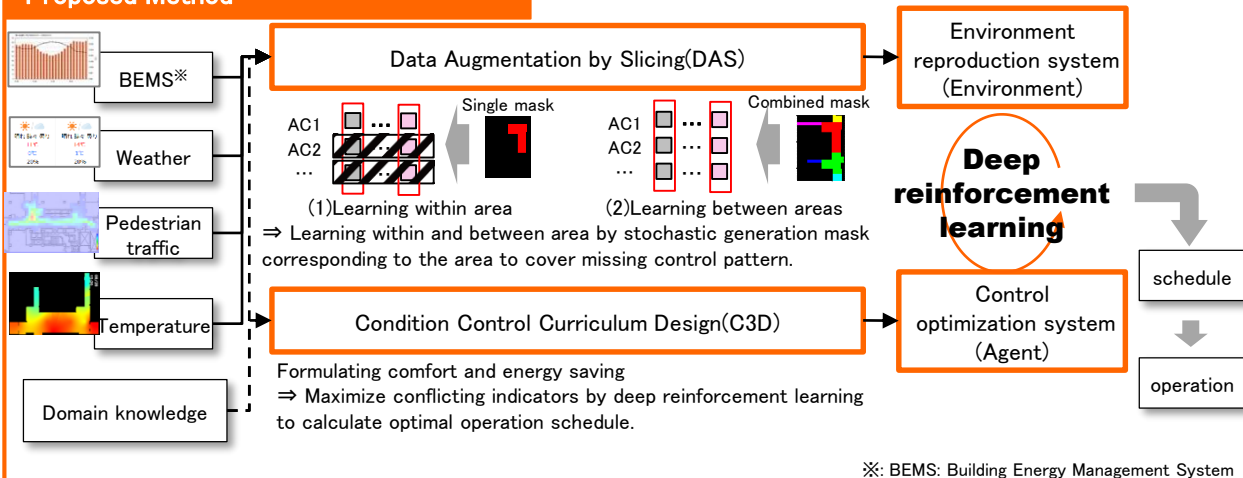
☐ : ON ☐ : OFF

→ t (00:00-24:00) All air flows are synchronously controlled



There are several missing teacher data like when air-conditioning flows are partially or periodically turned off, because normally all air flows are synchronously controlled.

Proposed Method



References

- [1] I. Shake, K. Kawase, Y. Suzuki, "NSRI × NTT × MITSUI FUDOSAN Collaboration results: Commenced joint experiments to utilize urban big data and AI in Nihonbashi Muromachi area," *NTT technical journal*, Vol. 29, No. 11, pp. 63-65, 2017.

Contact

Nobuhiko Matsuura Email: cs-liaison-ml at hco.ntt.co.jp
Network Innovation Laboratories



Innovative R&D by NTT
Open House 2019

Copyright © 2019 NTT. All Rights Reserved.

03

Recover urban people flow from population data

- People flow estimation from spatiotemporal population data -

Abstract

Real-time spatiotemporal population data is attracting a great deal of attention for understanding crowd movements in cities. The data is the aggregation of personal location information and consists of just areas and the number of people in each area at certain time instants. Accordingly, it does not explicitly represent crowd movement. We propose a **probabilistic collective graphical models** that can estimate crowd movement from spatiotemporal population data. There are two technical challenges: (i) poor estimation accuracy as the traditional approach means the model would have too many degrees of freedom, (ii) excessive computation cost. Our key idea is to model the transition probability between areas by using **three factors: departure probability of areas, gathering score of areas, and geographical distance between areas**. These advances enable us to **reduce the degrees of freedom of the model appropriately** and derive an **efficient estimation algorithm**.

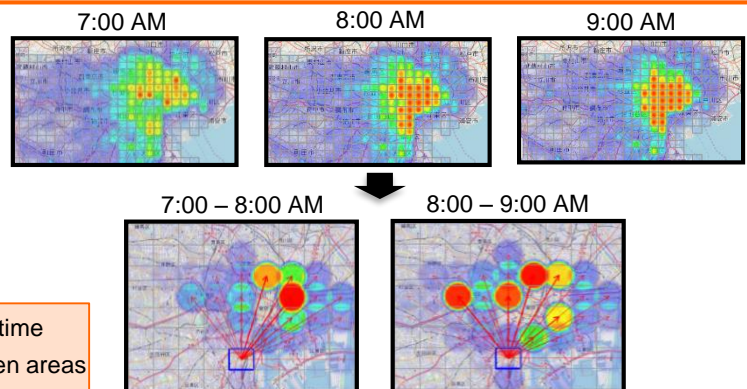
What we are working on?

- ✓ It is difficult to utilize people movement data across various services and enterprises because of privacy issues.
- ✓ In many practical situations, only aggregate information is available.

Our technology

Input: population of each area at each time

Output: # of people who moved between areas



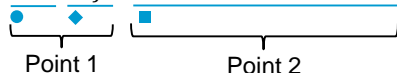
Technical points

Estimation reflects nature of human movements

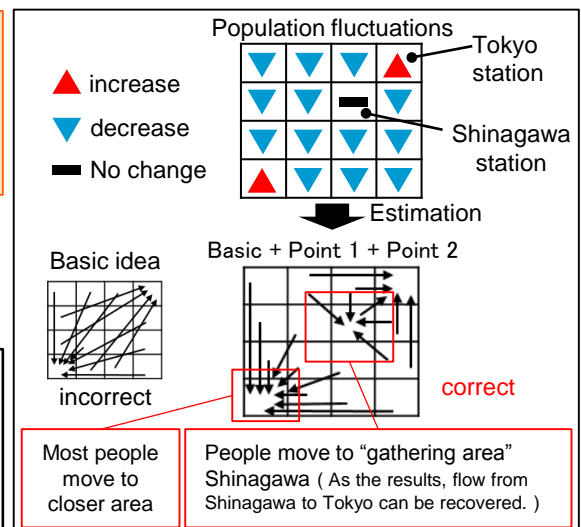
- **Basic idea:** Allocate people flow from decreasing area to increasing area
- **Point 1:** Consider **area characteristics**
 - **Gathering** area (areas where people are likely to gather)
 - **Emissive** area (areas where people are likely to leave)
- **Point 2:** Consider **distance** between areas

Transition probability from area i to j

$$\theta_{ij} \propto \pi_i \times s_j \times \exp(-\beta \cdot \text{dist}(i, j))$$



- Departure probability π_i determines whether the person leaves area i or stay
- If the person is deemed to leave i , (probability of next area j to be chosen) \propto (gathering factor of j) \times (distance between i and j)



References

- [1] Y. Akagi, T. Nishimura, T. Kurashima, H. Toda, "A fast and accurate method for estimating people flow from spatiotemporal population data," in Proc. the 27th International Joint Conference on Artificial Intelligence and the 23rd European Conference on Artificial Intelligence (IJCAI-ECAI-18), pp. 3293-3300, 2018.

Contact

Yasunori Akagi Email: cs-liaison-ml at hco.ntt.co.jp
Proactive Navigation Project, NTT Service Evolution Laboratories



Innovative R&D by NTT
Open House 2019

Copyright © 2019 NTT. All Rights Reserved.

04

Improving the accuracy of deep learning

- Larger capacity output function for deep learning -

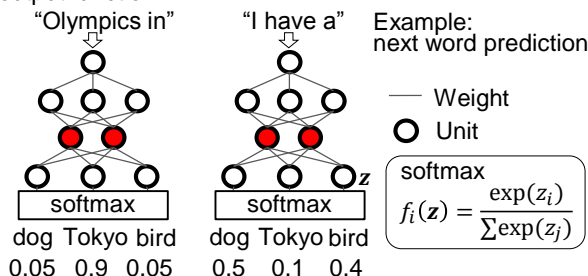
Abstract

Deep learning is used in a lot of applications, e.g., image recognition, speech recognition, and machine translation. In many applications of deep learning, softmax is used as an output activation function for modeling categorical probability distributions. To represent various probabilities, models should output various patterns, i.e., **models should have sufficient representation capacity**. However, **softmax can be a bottleneck of representational capacity** (the softmax bottleneck) under a certain condition. In order to break the softmax bottleneck, we propose a novel output activation function: **sigsoftmax**. To break the softmax bottleneck, sigsoftmax is composed of sigmoid and exponential functions. Sigsoftmax can **output more various patterns** than softmax **without additional parameters and additional computation costs**. As a result, the model with sigsoftmax can be more accurate than that with softmax.

Deep Learning

Deep learning is used in a lot of applications. (e.g., image recognition or machine translation)

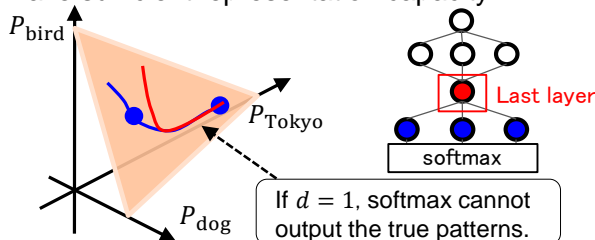
To represent the probability, softmax is used as an output function.



For accurate prediction, models should output a lot of patterns of probabilities, i.e., models require sufficient representation capacity.

Bottleneck of Representation

If the **number of units in the last layer d** < the **number of outputs M** - 1, softmax does not have sufficient representation capacity.



In natural language processing, **M = the number of vocabulary** is very large, and a lot of parameters are required if we set $d = M - 1$.

Problem

We assume the number of outputs M is 3.

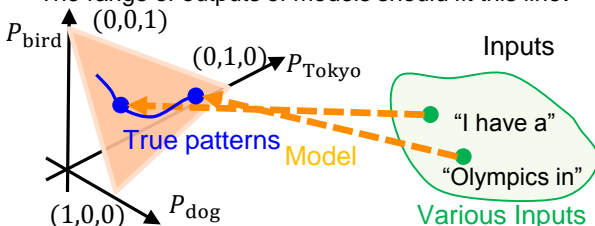
Probability of outputs: $(P_{\text{dog}}, P_{\text{Tokyo}}, P_{\text{bird}})$

→ It can be represented as a point on the triangle.

Models connect these points and inputs.

We assume the **blue line** represents the true patterns of probabilities (output set of various inputs).

→ The range of outputs of models should fit this line.



Novel output function: sigsoftmax

Multiplying sigmoid to have larger representation capacity

sigsoftmax

$$f_i(z) = \frac{\sigma(z_i) \exp(z_i)}{\sum \sigma(z_j) \exp(z_j)}$$

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

: sigmoid monotonically increasing to 1 from 0

Sigsoftmax has a larger representation capacity than softmax^[1] without additional parameters and computation costs.

References

- [1] S. Kanai, Y. Fujiwara, Y. Yamanaka, S. Adachi, "Sigsoftmax: Reanalysis of the softmax bottleneck," in Proc. 32nd Conference on Neural Information Processing Systems (NeurIPS), 2018.

Contact

Sekitoshi Kanai Email: cs-liaison-ml at hco.ntt.co.jp
NTT Software Innovation Center



Innovative R&D by NTT
Open House 2019

Copyright © 2019 NTT. All Rights Reserved.

05

Which is cause? Which is effect? Learn from data!

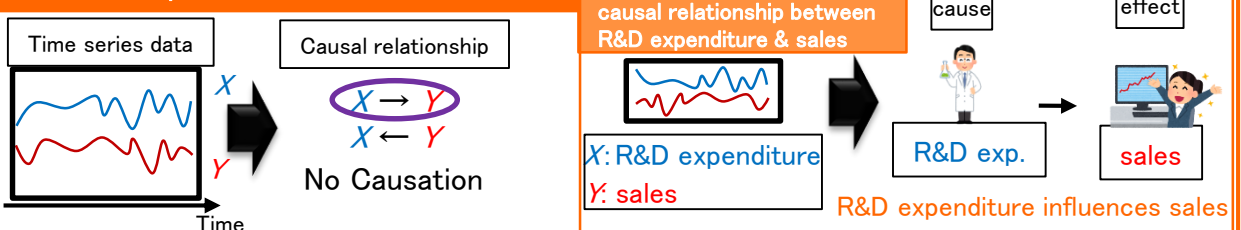
Causal inference in time series via supervised learning

Abstract

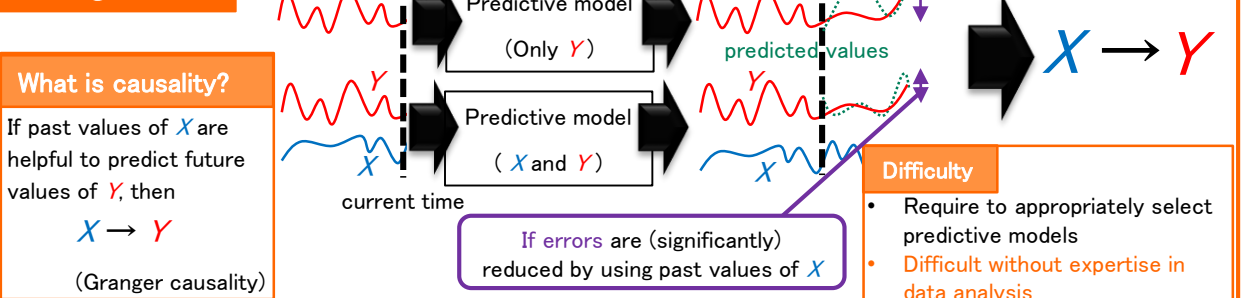
Our goal is to automatically discover “causal relationships” from time series data, i.e., a sequence of data measured over time. Discovering causal relationships has key applications in various fields: e.g., finding that “R&D expenditure influences sales” is useful for decision making in companies; discovering gene regulatory relationships provides a key insight for drug discovery researches.

To infer causal relationships, existing methods require us to select an appropriate mathematical expression (i.e., auto-regressive model) for each time series data, which is difficult without expertise in data analysis. For this problem, we build a novel approach that trains a machine learning model by using various data. Our method does not require a deep understanding of data analysis and therefore will help us to effectively make an important decision making in several situations.

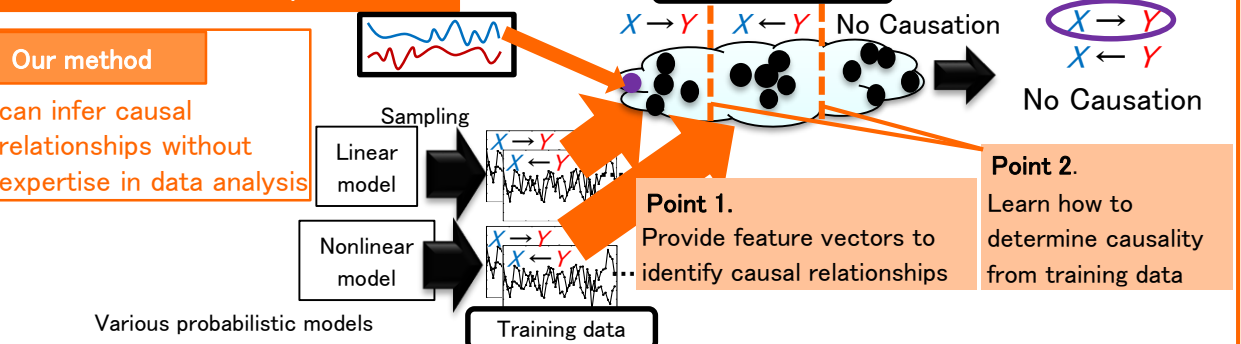
Problem setup: causal inference in time series



Existing methods



Learn causal relationships from data



References

- [1] Y. Chikahara, A. Fujino, “Causal Inference in Time Series via Supervised Learning,” in *Proc. 27th International Joint Conference on Artificial Intelligence (IJCAI)*, 2018.

Contact

Yoichi Chikahara Email: cs-liaison-ml at hco.ntt.co.jp
Learning and Intelligent Systems Research Group, Innovative Communication Laboratory



Innovative R&D by NTT
Open House 2019

Copyright © 2019 NTT. All Rights Reserved.

06

Forecasting future data for unobserved locations

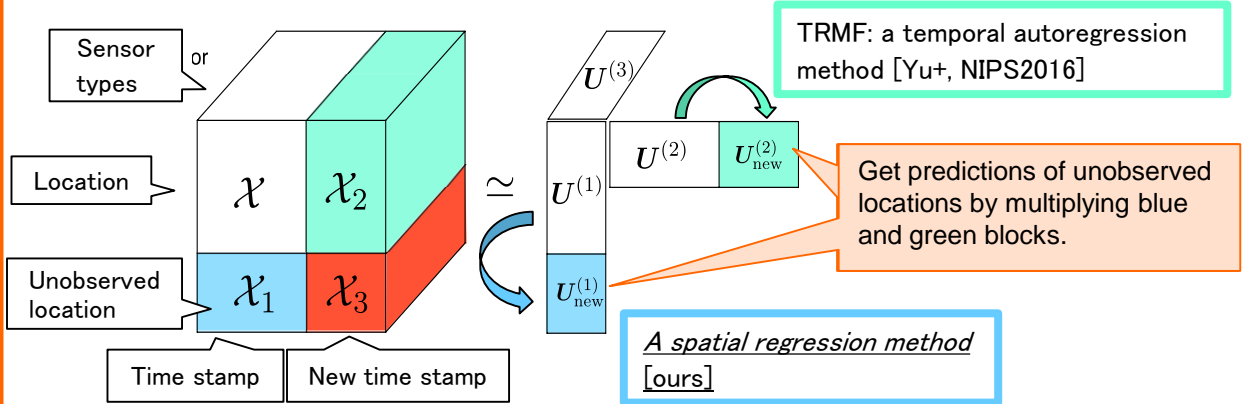
- Tensor factorization for spatio-temporal data analysis -

Abstract

Analysis of spatio-temporal data is a common research topic that **requires the interpolations of unobserved locations and the predictions of feature observations** by utilizing information about where and when the data were observed. One of the most difficult problems is to make **future predictions of unobserved locations**. Tensor factorization methods are popular in this field because of their capability of handling multiple types of spatio-temporal data, dealing with missing values, and providing computationally efficient parameter estimation procedures. We propose a new tensor factorization method that estimates low-rank latent factors by **simultaneously learning the spatial and temporal correlations**. We introduce **new spatial autoregressive regularizers** based on existing spatial autoregressive models and provide an efficient estimation procedure.

Spatio-Temporal Regression Problem

Our tensor factorization method estimates **factors of unobserved locations (blue)** with a spatial regression and employ it as a spatial regularizer. By combining it with **future actors (green)** obtained from an autoregression model, we enable to get **predictions of unobserved locations (red)**.



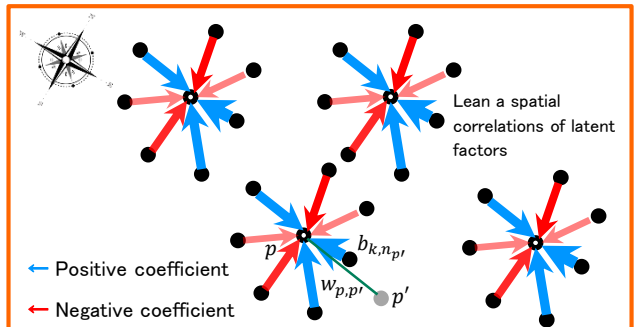
Our spatial regression method can deal with both grid and non-grid sensor locations by assigning the same coefficients based on the angle between a source and a target sensor locations.

Our angle dependent coefficient learning enables to get factors of unobserved locations $u_{p,k}^{(1)}$.

Spatial regression regularizer

$$\sum_{k=1}^K \sum_{p=1}^P \left(u_{p,k}^{(1)} - \sum_{p' \in E_p} b_{k,n_{p'}} w_{p,p'} u_{p',k}^{(1)} \right)^2 + \frac{\eta}{2} \|u_k^{(1)}\|_2^2,$$

A regression coefficient $b_{k,n_{p'}}$ is assigned by the angle between p and p' (red and blue arrows)



References

- [1] K. Takeuchi, H. Kashima and N. Ueda, "Autoregressive Tensor Factorization for Spatio-Temporal Predictions," in *Proc. of 2017 IEEE International Conference on Data Mining (ICDM)*, 2017.
- [2] 竹内孝, 鹿島久嗣, 上田修功, "自己回帰テンソル分解による時空間データ予測," *2018年度人工知能学会全国大会(第32回)*, 2018.

Contact

Koh Takeuchi Email: cs-liaison-ml at hco.ntt.co.jp
Ueda Research Laboratory



Innovative R&D by NTT
Open House 2019

Copyright © 2019 NTT. All Rights Reserved.

07

Search suitable for various viewpoints

- “Pitarie”: Picture book search with graph index based search -

Abstract

We propose a similarity search method for **finding similar objects in a large-scale database**. The search method is based on a graph index, where each vertex corresponds to an object and two vertices are connected by an edge when they satisfy a certain similarity condition. The graph index shows **small-world behavior**, that is, vertices can be reached from every other vertex by a small number of steps. Hence, searching the graph results in **quick termination of the search process**. Furthermore, since the graph index is constructed based on similarity between two objects, the search method is **versatile and can be applied to wide variety of media** such as text, images and audio. When applied to complex objects that are more than two media combined, such as picture books which consists of text and illustration, **users can search from various viewpoints**; users can find picture books that are not only similar in content but also similar in style of illustration.

Picture book search system “Pitarie”



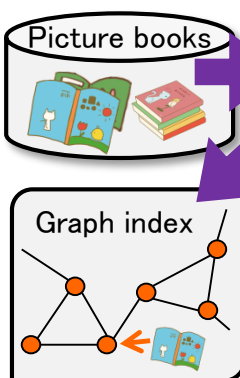
Search picture books from various viewpoints.

Case 1: Input a summary and search

Case 2: Search for books with similar contents

Case 3: Search for books with similar style of illustration (Example below)

Preparation of graph index in advance of search (Off line)



Graph index construction

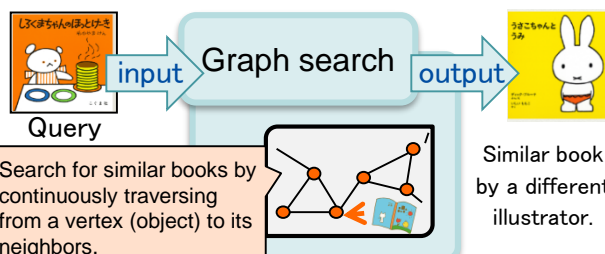
Construct a graph by connecting similar objects, and use it as an index for search.

- ✓ “Small-world behavior” : Any two objects are within a small number of edge hops.
- ✓ Since the graph construction is independent of media’s characteristic, the proposed method is applicable to various media.

Fast search by utilizing graph index

Example

You can find books with similar style of picture by various illustrators.



Search for similar books by continuously traversing from a vertex (object) to its neighbors.

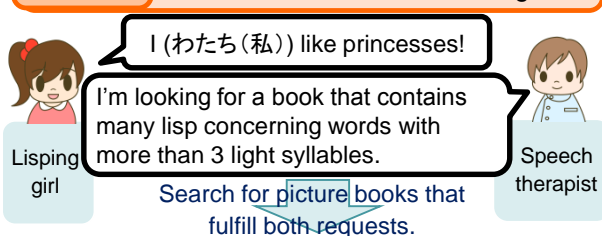
Cited picture books:

しろくまちゃんのほっとけーき、わかやまけん作、こぐま社、1972
うさこちゃんとうみ、ディックブルーナ作、福音館書店、1964
(The cover-front-like illustration of “The snow queen” was in-house illustrated.)

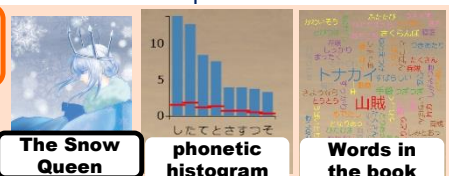
Various applications

Example

Search books suitable for articulation disorder training



Joint Research with a hospital.



References

- [1] T. Hattori, T. Kobayashi, S. Fujita, Y. Okumura, K. Aoyama, “Pitarie: Picture Book Search with Interdisciplinary Approach,” *NTT Technical Review*, vol.14, no.7, pp. 1-8, 2016.
- [2] K. Aoyama, K. Saito, H. Sawada, N. Ueda, “Fast approximate similarity search based on degree-reduced neighborhood graphs,” in *Proc. The 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1055-1063, 2011.

Contact

Takashi Hattori Email: cs-liaison-ml at hco.ntt.co.jp

Learning and Intelligent Systems Research Group, Innovative Communication Laboratory



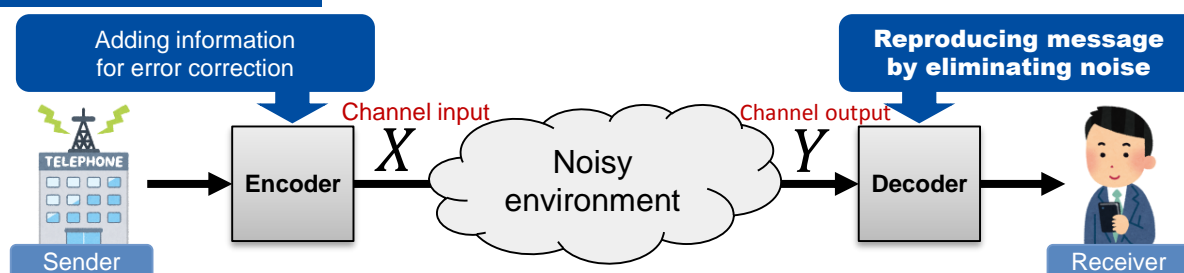
Innovative R&D by NTT
Open House 2019

Copyright © 2019 NTT. All Rights Reserved.

Abstract

For the realization of high-speed digital communication, it is necessary **transmitting messages reliably with high efficiency under noisy environment**. The limit of efficiency is derived by a computer scientist C. E. Shannon and it is called the Shannon limit. It is known that we can achieve the limit for a particular class of channels with LDPC (Low Density Parity Check) codes or the Polar codes, which are used in the 5G mobile communication technology. However, it is impossible to achieve the limit for a general class of channels with these codes. We propose a novel technology called CoCoNuTS (Code based on Constrained Numbers Theoretically-achieving the Shannon limit). With this technology, we can construct a code **achieving the Shannon limit for a general class of channels**. Our goal is **realizing future high-speed digital communication** by establishing related peripheral technologies.

Error Correcting Code

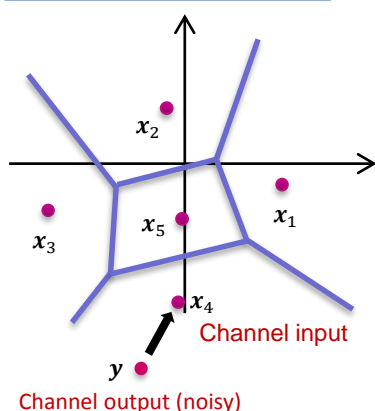


$$\text{Efficiency} = (\text{no. of message symbols}) / (\text{no. of transmitted signals})$$

We want to maximize efficiency to the Shannon limit.

Essence of proposed technology

Geometric illustration



Difficulty in encoding and decoding

- To achieve the Shannon limit, it is necessary to place channel inputs $\{x_i\}$ efficiently (as far as possible in the left figure).
- When the optimum distribution of a channel input X is uniform, we can achieve the limit with LDPC codes or the Polar codes. However, when the optimal input distribution is not uniform, it is impossible to achieve the limit by using these codes.
- In the naïve decoding method, we have to guess a channel input from a channel output (x_4 from y in the left figure) by using the brute-force search, which is impractical.

Proposed technology

- We can realize an ideal layout of channel inputs by using the constrained-random-number generator.
- We can avoid the brute-force search by using the constrained-random-number-generator, which provides a practical decoding method.

References

- [1] J. Muramatsu, "Channel coding and lossy source coding using a generator of constrained random numbers," *IEEE Transactions on Information Theory*, Vol. IT-60, No. 5, pp. 2667-2686, May 2014.
- [2] J. Muramatsu, S. Miyake, "Channel code using constrained-random-number generator revisited," *IEEE Transactions on Information Theory*, Vol. IT-65, No. 1, pp. 500-508, Jan. 2019.

Contact

Jun Muramatsu Email: cs-liaison-ml at hco.ntt.co.jp
Learning and Intelligent Systems Research Group, Innovative Communication Laboratory



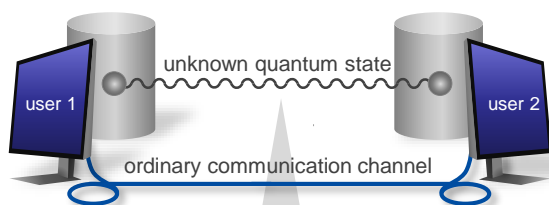
Innovative R&D by NTT
Open House 2019

Abstract

We investigate a **counter-intuitive phenomenon** of quantum state discrimination that the success probability of identifying all the unknown quantum states **increases** even when the number of unknown states **increases**. The phenomenon is known for vulnerability of **quantum secret sharing (QSS)**, which enables one to distribute a secret amongst untrusted participants securely, however, the necessary and sufficient condition for the phenomenon was unknown. We **show the condition** for a specific discrimination task and **construct a practical method** to realize the phenomenon. These results **advance the analysis** of the phenomenon and **reveal the vulnerability** of QSS. Since quantum state discrimination lies at the heart of many quantum information processing tasks, our research **widely contributes to the future information society based on quantum technologies**, where people would obtain the benefits from genuine quantum information processing.

Quantum state discrimination

1. Distribute a randomly chosen quantum state to two users
2. The users try to identify the state by using an ordinary communication channel



Suppose the quantum state is randomly chosen from four **Bell states** (standard and useful quantum states)

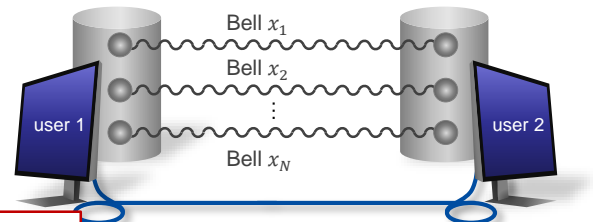
Bell 1
Bell 2
Bell 3
Bell 4

→ Bell x

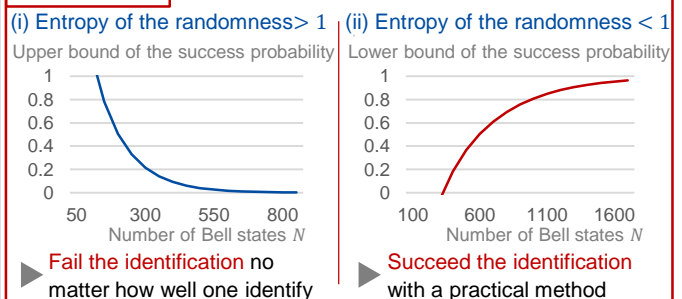
Success probability of the identification < 1

Discrimination of independent Bell states

1. Distribute randomly chosen Bell states to two users
2. The users try to identify **all the states** x_1, x_2, \dots, x_N by using an ordinary communication channel

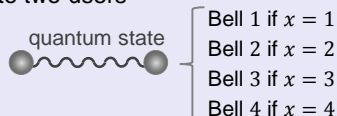


Main Results



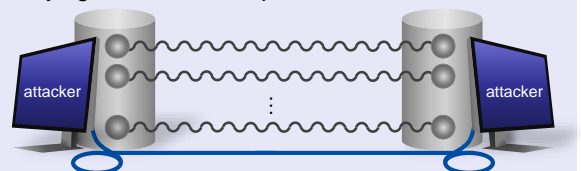
Vulnerability assessment of QSS

1. Distribute quantum states each of which encodes **secret x** to two users



Main results → **The more secrets one distribute to attacking users, the more vulnerable the secrets become**

2. Attacking users try to read secret x_1, x_2, \dots, x_N by identifying the distributed quantum states



References

- [1] S. Akiyue, G. Kato, "Bipartite discrimination of independently prepared quantum states as a counterexample to a parallel repetition conjecture," *Physical Review A*, Vol. 97, No. 10, 042309, 2018.

Contact

Seiseki Akiyue Email: cs-liaison-ml at hco.ntt.co.jp
Computing Theory Research Group, Media Information Laboratory



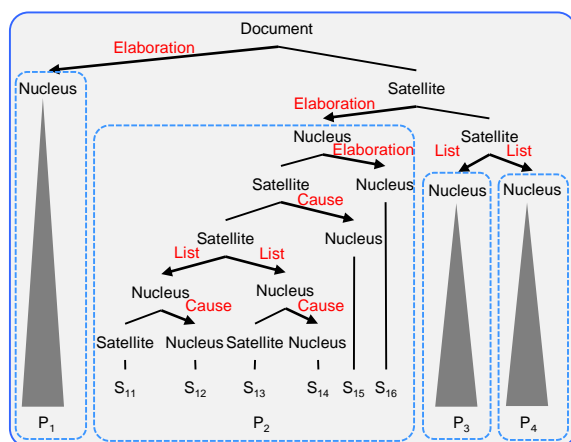
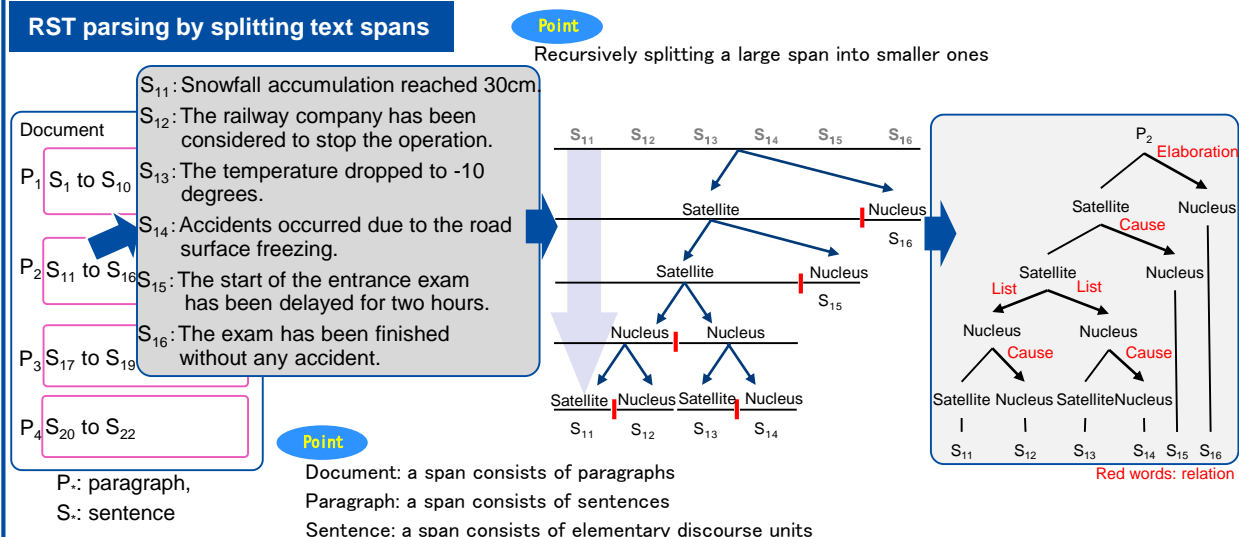
Innovative R&D by NTT
Open House 2019

Copyright © 2019 NTT. All Rights Reserved.

Abstract

Analyzing a discourse structure behind the document is crucial for context aware Natural Language Processing (NLP) tasks including machine translation and automatic summarization. We propose a neural discourse parsing method based on Rhetorical Structure Theory (RST) that regards a document as a constituent tree. Our parser builds RST trees at different levels of granularity in a document and then replace leaves of upper-level RST trees with lower-level RST trees that were already constructed. The parsing is performed in a top-down manner for each granularity level by recursively splitting a larger text span into two smaller ones while predicting nuclearity labels and rhetorical relations. Unlike previous discourse parsers, our parser can be fully parallelized at each granularity in a document and does not require any handcrafted features such as syntactic features obtained from full parse trees of sentences.

RST parsing by splitting text spans



Evaluation results on RST-DT corpus

	Unlabeled	Nuclearity	Relation
Proposed	72.0	58.6	46.7
w/o granularity	66.5	53.4	43.3
Previous	68.6	55.9	45.8
Human	78.3	66.8	57.1

Bracket F-score

This is a collaborative research project between Okumura Lab at Tokyo Institute of Technology and NTT CS labs.

References

- [1] N. Kobayashi, T. Hirao, M. Okumura, M. Nagata, "Top-down RST Parsing Utilizing Granularity Levels in Documents," in *Proc. of 25th Annual Meeting of Natural Language Processing*, pp. 1002-1005, 2019.

Contact

Tsutomu Hirao Email: cs-liaison-ml at hco.ntt.co.jp
Linguistic Intelligence Research Group, Innovative Communication Laboratory



Innovative R&D by NTT
Open House 2019

Copyright © 2019 NTT. All Rights Reserved.

11

When children begin to understand hiragana

- Emergent literacy development in Japanese -

Abstract

Although many studies have reported child literacy development, **it remains unclear when and how toddlers acquire letters well before starting formal education**. We focus on Japanese hiragana letters to investigate (a) when toddlers begin to understand hiragana, and (b) what kind of letters is easily acquired. This work's eye-tracking experiment shows that **toddlers at 32–39 months begin to understand hiragana letter-sound mapping**. Moreover, our large-scale corpus analysis found that various factors, such as letter frequency in picture books and visual complexity, contribute to the acquisition of hiragana reading and writing. We aim to extend our findings to develop an early detection method and letter-learning method for children with reading difficulties.

1 Hiragana understanding measured by eye-gaze



- **Participants:** 80 toddlers at 24–48 months of age
- **Method:** Measuring eye-gaze patterns during gazes at the screen
- **Analysis:** Calculating viewing ratio of target letters

Although toddlers at 32–39 months could barely read hiragana letters (Fig. 1), they start to understand letter-sound correspondence (Fig. 2).

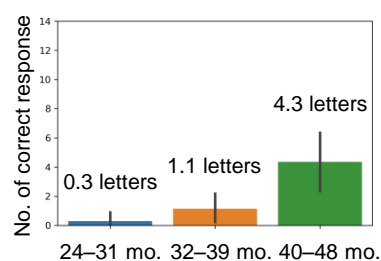


Fig. 1 Hiragana reading task

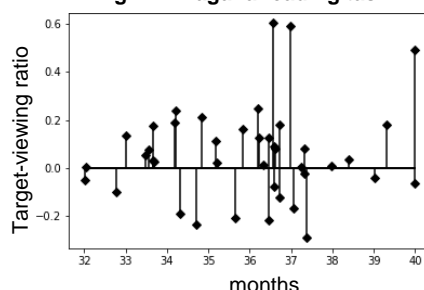
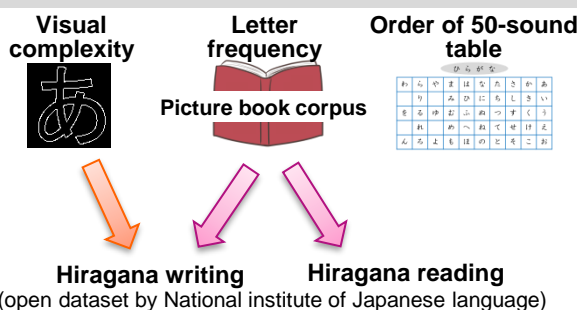


Fig. 2 Understanding of letter-sound correspondence (32–39 mo.)

2 Factors contributing to hiragana reading and writing acquisition



While hiragana reading acquisition relates to letter frequency in picture books, hiragana writing acquisition relates to visual complexity as well as letter frequency (Fig. 3).

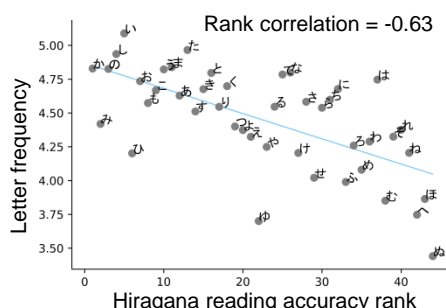


Fig. 3 Letter frequency and reading accuracy

References

- [1] H. Higuchi, Y. Okumura, T. Kobayashi, "Acquisition of letter-sound correspondence in Japanese-speaking 2-year-olds: An eye-tracking study" in Biennial Meeting of Society for Research in Child Development (SRCD), 2019.
- [2] H. Higuchi, Y. Okumura, T. Kobayashi "Influence of letter properties for Japanese hiragana reading and writing acquisition" *The Japan Journal of Logopedics and Phoniatrics*, Vol.60 No.2, 113-120, 2019.
- [3] H. Higuchi, Y. Okumura, T. Kobayashi "Influence of letter properties for Japanese katakana reading and writing acquisition" *The Japan Journal of Logopedics and Phoniatrics*(in press).

Contact

Hiroki Higuchi Email: cs-liaison-ml at hco.ntt.co.jp
Interaction Research Group, Innovative Communication Laboratory



Innovative R&D by NTT
Open House 2019

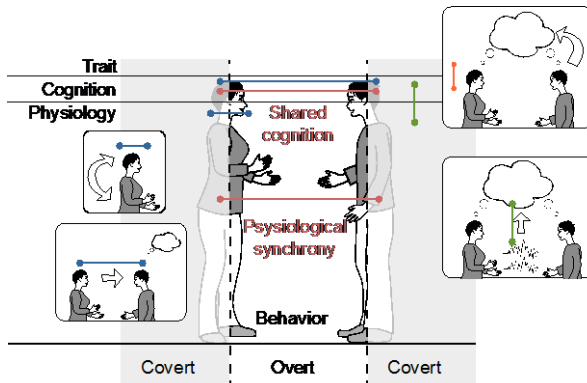
Copyright © 2019 NTT. All Rights Reserved.

Abstract

Empathy is the basis of people's social lives. However, the mechanism has not yet been fully elucidated because it is a complex phenomenon consisting of subjectivity, physiology, and behavior. The purpose of this study is to quantify empathy from a multifaceted point of view considering individual differences. We examined how physiology, behavior and cognition are related in an individual, and how they are shared with other individuals. In order to deal with the large individual difference in the subjective judgment, we built a computational model that explains the individual difference from their personality traits, exploiting the wisdom of the group approach which aggregates the judgment of multiple individuals. This kind of framework for quantitative measurement of empathy including individual differences will make it possible to assess and predict the effects of interventions that promote empathy tailored to individuals. We believe this is an essential step toward improving human well-being.



Empathy involves a variety of phenomena, from the low-level phenomenon (e.g. physiological synchronization) to the higher-level phenomenon (e.g. cognitive sharing). We try to understand these from various aspects.

Individual's physiology and behavior ^[1]

* joint research with Tsukuba University

In the task of distinguishing between posed and spontaneous faces, we directly compared humans with machines using video or EMG signal in terms of accuracy.

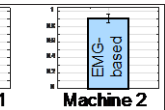
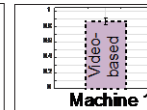
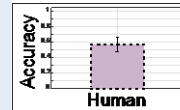
Is facial expression (FE) of this person spontaneous or posed?

Target :



FE was measured using
Video & Electromyography

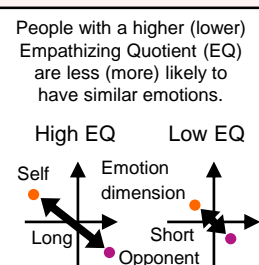
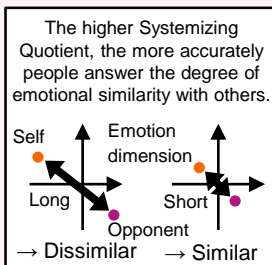
Respondents :



Accuracy: human < Video \approx EMG

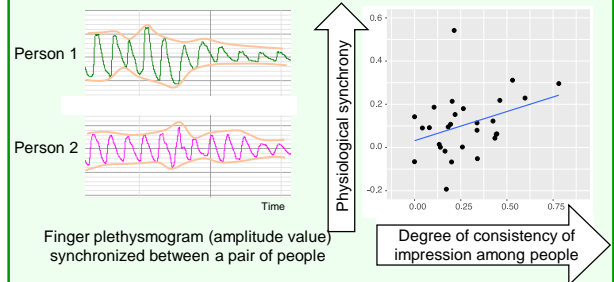
Cognition and personality trait on emotion sharing ^[2]

Based on the similarity judgment theory, we clarified which types of people tend to have emotion similar to others and to recognize their emotional similarities accurately.



Physiological synchrony and cognitive sharing

We found that the physiological response is synchronized during cooperative work, and that the higher the physiological synchrony, the more consistent the impression of the interaction.



References

- [1] M. Perusquia-Hernandez, S. Ayabe-Kanamura, K. Suzuki, S. Kumano, "The Invisible Potential of Facial Electromyography: a Comparison of EMG and Computer Vision when Distinguishing Posed from Spontaneous Smiles," in *Proc. Conf. Human Factors in Computing Systems (CHI)*, 2019.
- [2] L. Antakiet, M. Matsuda, K. Otsuka, S. Kumano, "Analyzing Generation and Cognition of Emotional Congruence using Empathizing Systemizing Quotient," *International Journal of Affective Engineering*, Vol. 17, No. 3, pp. 183-192, 2018.

Contact

Shiro Kumano Email: cs-liaison-ml at hco.ntt.co.jp
Sensory Resonance Research Group, Human Information Science Laboratory



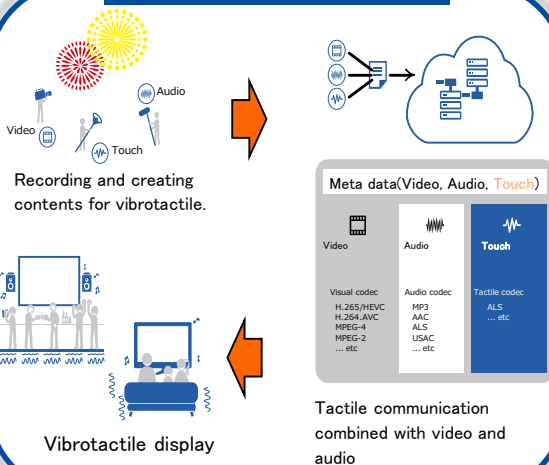
Innovative R&D by NTT
Open House 2019

Copyright © 2019 NTT. All Rights Reserved.

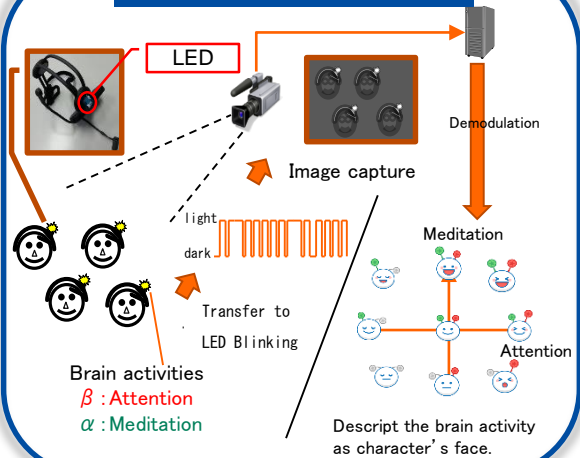
Abstract

This study targets the **empathetic communication** that occur in sharing the same field. To provide quantitative evaluations, physiological changes are observed using **optical camera communication**. **Vibrotactile stimulation** was presented simultaneously to enhance the viewing experience. Thanks to the vibrotactile communication technologies, we can record, distribute, and display tactile information in accordance with audiovisual contents and optical camera communication enabled us to simultaneously observe physiological responses from crowd of people. By combining these interdisciplinary technologies we can run cyclical research processes of sense intervention, measurement, evaluation, and factor analysis to progress the research on empathetic communication. Based on these research results, we will make a design theory for making field that can enhance the wellbeing of the people gathered in the field.

Vibrotactile communication



Mass measurement by Optical Camera communication

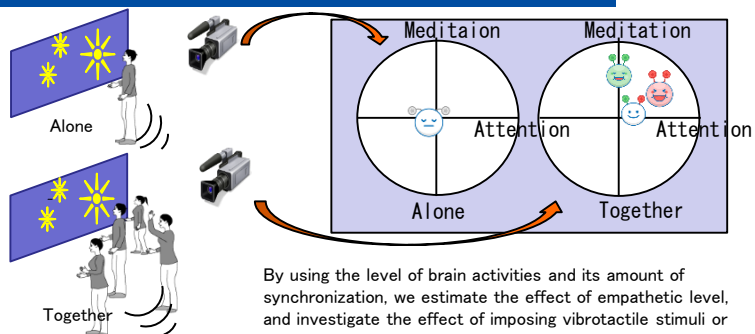


Observation of empathetic communication using OCC and vibrotactile



We measure the brain activities of the users who are **experiencing vibrotactile intervention** by using **optical camera communication**.

We compare the condition between alone and together to elucidate the effect of **empathetic communication caused by sharing a field**.



By using the level of brain activities and its amount of synchronization, we estimate the effect of empathetic level, and investigate the effect of imposing vibrotactile stimuli or the effect of the number sharing the field.

References

- [1] Y. Shiraki, T. G. Sato, T. Moriya, "Flexible synchronization in optical camera communication with on-off keying," in *Proc. IEEE GLOBECOM Workshops*, pp. 1-6, 2017.
- [2] T. G. Sato, Y. Shiraki, T. Moriya, "Audience Excitement Reflected in Respiratory Phase Synchronization," *IEEE Int. Conf. on SMC*, pp. 2856-2860, 2017.

Contact

Takashi G Sato Email: cs-liaison-ml at hco.ntt.co.jp
Moriya Research Laboratory



Innovative R&D by NTT
Open House 2019

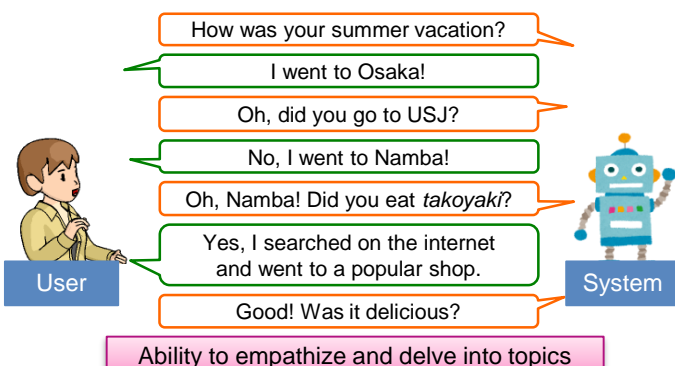
Copyright © 2019 NTT. All Rights Reserved.

Abstract

The proposed chat-oriented dialogue system can make users think **the system understands the user's utterances**. By understanding the user's utterances as an event structure (a group of time, place, person, etc.), we achieve a chat-oriented dialogue system that can sympathize and delve into topics during a chat. To understand a user's events from the user's utterances, a system must understand various words/phrases in user utterances. To tackle this problem, we focus on **general words and phrases that are familiar in a chatting situation** but difficult to extract by conventional methods. Using this technology, systems can extract a user's utterances by organizing the extracted information. In the future, we aim to **foster a world where humans can converse with systems like humans with mutual understanding** by grounding the extracted information to the system and external knowledge.

System utterance generation based on event understanding

The system generates its next utterances corresponding to the user's event, extracted from the user's utterances as structured event information, by comparing the event with system knowledge.



Example of an extracted event structure

Time	Summer vacation
Place	Osaka/Namba
Person	
Action	ate <i>takoyaki</i>
Feeling	

Example of similar events (system knowledge)

Time	in vacation	Time	in September
Place	Osaka/USJ	Place	Osaka/Namba
Person	with family	Person	with friends
Action	saw turtles	Action	ate <i>takoyaki</i>
Feeling	was cute	Feeling	was delicious

This system was developed based on the dialogue system that won *first prize* in a live competition in Japan (2019).

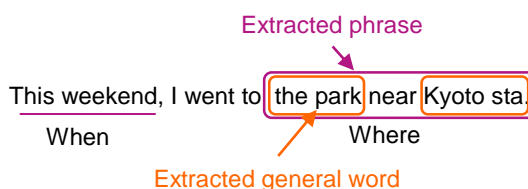
Phrase detection in user utterance

To understand user's events from his/her utterances, various types of words/phrases must be extracted. By analyzing such words and phrases in chats, we achieve this extraction.

Example of location words/phrases in chats

User utterance (Red is location phrase)	Conventional NE extraction	Proposed extraction
I went to Italy .	Italy	Italy
We went to the park near Kyoto station .	Kyoto station	park near Kyoto Station
I often go to electricity shop .	(none)	electricity shop

70% chance that phrases that are not Named Entity (NE) appear in chats (in case of location words/phrases)



Ability to extract general words and phrases

References

- [1] H. Narimatsu, H. Sugiyama, M. Mizukami, "Detecting Location-Indicating Phrases in User Utterances for Chat-Oriented Dialogue Systems," in *Proc. The Fourth Linguistic and Cognitive Approaches to Dialog Agents Workshop (LACATODA)*, 2018.
- [2] H. Sugiyama, H. Narimatsu, M. Mizukami, T. Arimoto, "Empirical study on domain-specific conversational dialogue system based on context-aware utterance understanding and generation," *JSAI SIG-SLUD*, 2018. (in Japanese)
- [3] M. Mizukami, H. Sugiyama, H. Narimatsu, "Event Data Collection for Recent Personal Questions," in *Proc. LACATODA*, 2018.

Contact

Hiromi Narimatsu Email: cs-liaison-ml at hco.ntt.co.jp
Interaction Research Group, Innovative Communication Laboratory



Innovative R&D by NTT
Open House 2019

Copyright © 2019 NTT. All Rights Reserved.

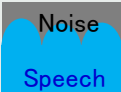



Abstract

Our technology can **support a speech command and hand-free communication** even in noisy environment such as road noise without any stresses. Clear speech can be picked up from the noise-mixed sound in order to realize speech command with high accuracy. A lot of computational complexity and memory was required to keep speech quality and reduce only noise so far. This problem can be solved by **using our acoustical knowhow**, moreover, low latency was able to be also achieved. In addition, **a sign of howling was able to be detected rapidly** by combining multiple microphone array. Our goal is **to improve an in-car acoustical environment** by reducing noises which are road noise, engine noise, and any sound from other cars. We will also try to establish an event detection technology in order to **help a driving assistant or an early maintenance** by detecting emergency car or anomalous in sound.

Differences from conventional

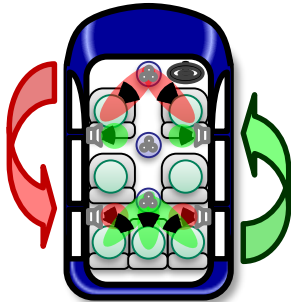
IM : Intelligent Microphone

ASTER : Anti-distortion Suppression of noise with mask-based TransER function estimation

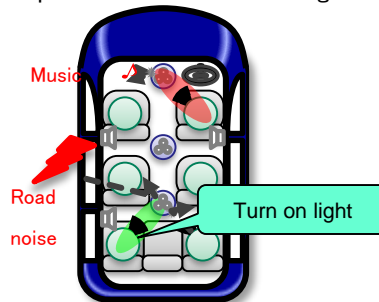
	Conventional NTT technology 1 (IM*)	Conventional NTT technology 2 (ASTER*)	This technology (IM-ASTER)
Abstract	Extracting target sound from noise-mixed sound by combining a linear and non-linear process.	Reducing noise from noise mixed speech signal while minimizing speech distortion.	Achieving low computational complexity and small amount of memory, and having advantages of IM and ASTER.
Image of process  (Input signal)	 Speech is degraded by too much noise reduction	 Noise is removed while minimizing speech distortion	 Noise is removed while minimizing speech distortion
Minimize distortion	△	○	○
High level noise	○	△	○
Computational complexity / memory	○	×	○

Demonstration

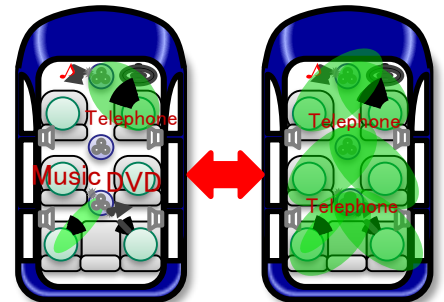
In car communication



Speech command from target seat



Hand-free communication



References

- [1] Y. Hioka, K. Furuya, K. Kobayashi, K. Niwa, Y. Haneda, "Underdetermined sound source separation using power spectrum density estimated by combination of directivity gain," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 6, pp. 1240-1250, 2013.
- [2] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. Yu, W. H. Fabian, M. Espi, T. Higuchi, S. Araki, T. Nakatani, "The NTT CHIME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *Proc. IEEE ASRU*, pp. 436-443, Dec. 2015.

Contact

Noboru Harada Email: cs-liaison-ml at hco.ntt.co.jp
Media intelligence laboratory



Innovative R&D by NTT

Open House 2019

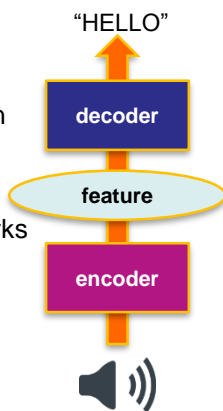
Copyright © 2019 NTT. All Rights Reserved.

Abstract

We propose a semi-supervised end-to-end method for learning speech recognition from small paired data and large unpaired data. This is because preparing the paired data of a speech and its transcription text requires a large amount of human effort. In our method, we introduce speech and text autoencoders that share encoders and decoders with an automatic speech recognition (ASR) model to improve ASR performance using speech-only and text-only training datasets. To build the speech and text autoencoders, we leverage state-of-the-art ASR and text-to-speech (TTS) encoder-decoder architectures. These autoencoders learn features from speech-only and text-only datasets by switching the encoders and decoders used in the ASR and TTS models. Simultaneously, they aim to encode features to be compatible with ASR and TTS models using a multi-task loss.

Speech Recognition

- Prepare paired data of speech and transcription text for training
- Train speech encoder and text decoder networks
- Perform speech recognition on given speech input



Problems

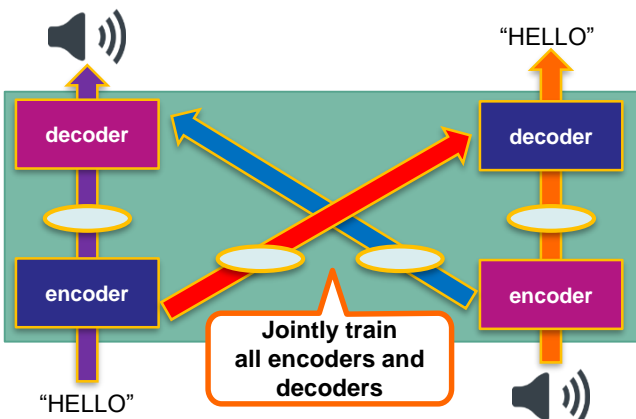
- The encoder-decoder model requires a large amount of the paired speech and transcription text dataset.
- Preparation of such dataset needs huge amounts of time and money
- If the networks can learn speech-only and text-only datasets, the data preparation becomes much easier



Semi-supervised training method

Point 1: Combine speech-to-text task with text-to-speech

- Train with paired data: speech-to-text, text-to-speech
- Train without paired data: speech-to-speech, text-to-text



Training task				
speech to text	text to text	speech to speech	text to speech	char error rate
✓				15.0 %
✓	✓			9.0%
✓		✓		8.7%
✓	✓	✓	✓	8.4%

Point2: Training to encode or decode

features  that look similar to each other

References

- [1] S. Karita, S. Watanabe, T. Iwata, A. Ogawa, M. Delcroix, "Semi-supervised end-to-end speech recognition," in Proc. of 2018 Interspeech, pp. 2-6, 2018.
- [2] S. Karita, S. Watanabe, T. Iwata, M. Delcroix, A. Ogawa, T. Nakatani, "Semi-supervised end-to-end speech recognition using text-to-speech and autoencoders," in Proc. of 2019 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2019.

Contact

Shigeki Karita Email: cs-liaison-ml at hco.ntt.co.jp
Signal Processing Research Group, Media Information Laboratory



Innovative R&D by NTT
Open House 2019

Copyright © 2019 NTT. All Rights Reserved.

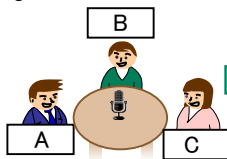
Abstract

We propose a method to accurately estimate "who spoke when" based on speaker's voice characteristics. It works even in a situation where multiple speaker's speech signals overlap, and accurately counts the number of speakers in such cases. Conventional methods with the similar functionality works only when the observed signal satisfies certain a priori (unrealistic) assumptions (e.g. the number of speaker known in advance, speakers never change their locations). However, these assumptions cannot be often satisfied in realistic scenarios, which leads to performance degradation. On the other hand, the proposed method, which is based purely on deep learning, can theoretically learn and deal with any realistic conversation situations. It is expected to serve as a fundamental technology for automatic conversation analysis systems, and will contribute to realization of automatic meeting minutes generation systems and communication robots.

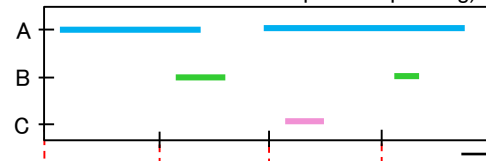
Difficulty of automatically analyzing 'who spoke when'

Conversation data is dynamic & diverse

(Target environment example)



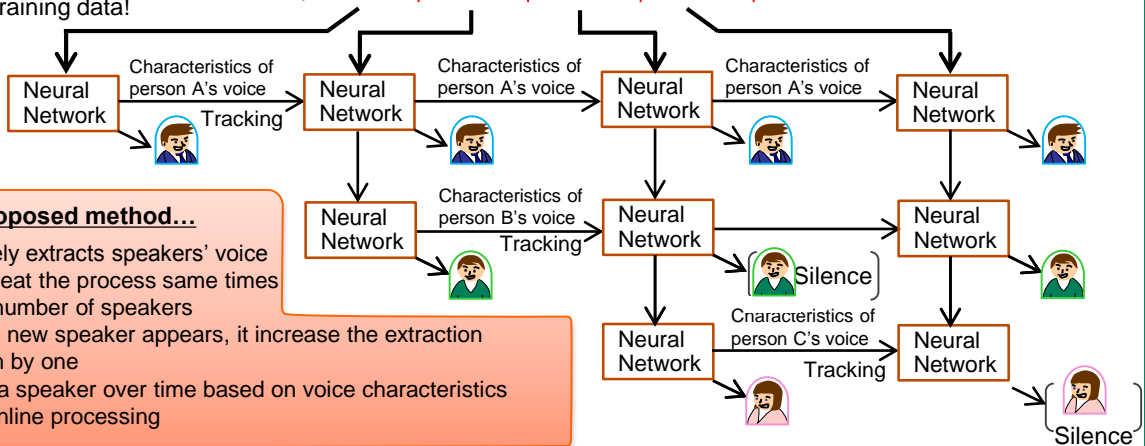
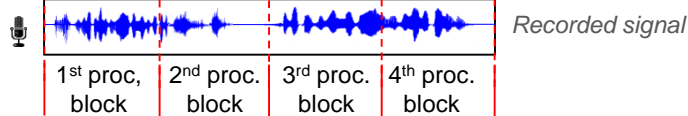
(Periods where each speaker speaking)



- The number of speakers is arbitrary and changes over time
 - People speak intermittently
 - Voice often overlaps each other
 - Speaker location changes randomly
- Difficult for conventional method to handle

Proposed method

- Estimating 'who spoke when' based on deep learning
- Entire process optimized with training data!



Proposed method...

- Iteratively extracts speakers' voice and repeat the process same times as the number of speakers
- When a new speaker appears, it increase the extraction iteration by one
- Tracks a speaker over time based on voice characteristics
- Block online processing

Advantage of the proposed method in comparison with conventional method

- The proposed method achieves source separation and source number counting simultaneously.
- The proposed method can track speaker's voice over time based on voice characteristics. It can keep tracking the speaker even if the speaker changes his/her location.

References

- [1] K. Kinoshita, L. Drude, M. Delcroix, T. Nakatani, "Listening to each speaker one by one with recurrent selective hearing networks," in *Proc. IEEE International Conference on Acoustics, Speech & Signal Processing (ICASSP)*, pp. 5064-5068, 2018.
- [2] T. von Neuman, K. Kinoshita, M. Delcroix, S. Araki, T. Nakatani, R. Haeb-Umbach, "All-neural online source separation, counting, and diarization for meeting analysis," in *Proc. IEEE International Conference on Acoustics, Speech & Signal Processing (ICASSP)*, 2019.

Contact

Keisuke Kinoshita Email: cs-liaison-ml at hco.ntt.co.jp
Signal Processing Research group, Media Information Laboratory



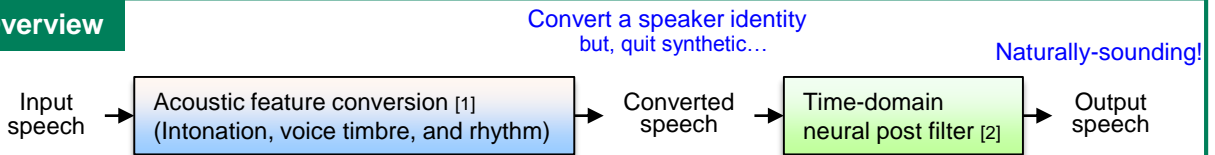
Innovative R&D by NTT
Open House 2019

Copyright © 2019 NTT. All Rights Reserved.

Abstract

We propose an voice and prosody conversion method for **impersonating a desired speaker's identity** and **hiding a speaker's identity**. The conversion method consists of acoustic feature conversion and time-domain neural postfilter. The acoustic feature conversion is based on a sequence-to-sequence learning with attention mechanism, which makes it possible to **capture the long-range temporal dependencies** between source and target sequences. The later post filter employs a cyclic model based on adversarial networks, which **requires no assumption for the speech waveform modeling**. In contrast to current voice conversion techniques, the proposed method makes it possible to **convert not only voice timbre but also prosody and rhythm** while achieving high-quality speech waveform generation due to the proposed time-domain neural post filter. The remaining challenge is the real-time voice conversion which is our ongoing work.

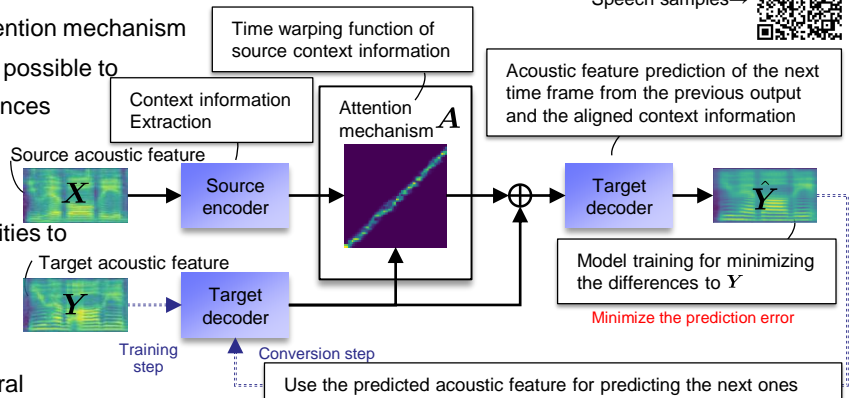
Overview



Acoustic Feature Conversion [1]

(e.g., Impersonating a speaker's identity and modifying pronunciation)

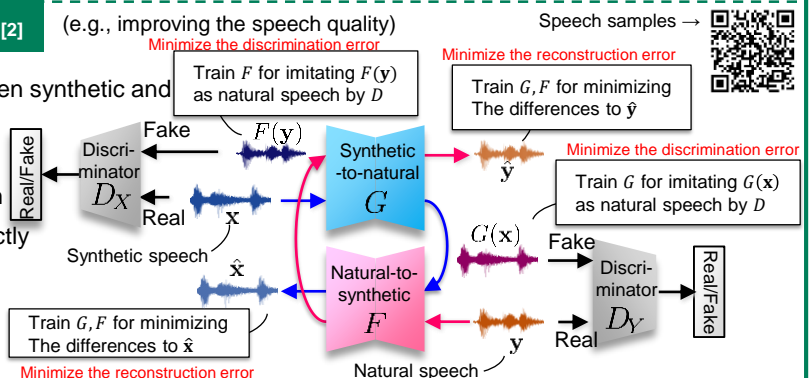
- Train encoders, decoder, and attention mechanism
- Encoder-decoder model makes it possible to
 1. Handle input and output sequences of different lengths
 2. Convert not only voice timbre but also rhythm
- Attention mechanism has the abilities to
 1. Select critical information from the encoded representation in accordance with the output sequence representation
 2. Consider the long-range temporal dependencies for converting intonation



Time-domain Neural Post Filter [2]

(e.g., improving the speech quality)

- Train conversion functions G, F between synthetic and natural speech
- Cyclic model makes it possible to
 1. Train the models with non-parallel data of synthetic and natural speech
 2. Handle the phase information correctly due to need for the reconstruction of speech waveform
- Generative adversarial learning helps to generate clear speech



References

- [1] K. Tanaka, H. Kameoka, T. Kaneko, N. Hojo, "AttS2S-VC: Sequence-to-Sequence Voice Conversion with Attention and Context Preservation Mechanisms," in *Proc. 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2019)*, May 2019.
- [2] K. Tanaka, H. Kameoka, T. Kaneko, N. Hojo, "WaveCycleGAN2: Time-domain Neural Post-filter for Speech Waveform Generation," *arXiv:1904.02892*, Apr. 2019, (submitted to *INTERSPEECH2019*).

Contact

Kou Tanaka Email: cs-liaison-ml at hco.ntt.co.jp

Learning and Intelligent Systems Research Group, Innovative Communication Laboratory



Innovative R&D by NTT
Open House 2019

Copyright © 2019 NTT. All Rights Reserved.

Abstract

Humans are able to imagine a person's voice from the person's appearance and imagine the person's appearance from his/her voice. In this work, we take an information-theoretic approach using deep generative models to develop a method that can convert speech into a voice that matches an input face image and generate a face image that matches the voice of the input speech by leveraging the correlation between faces and voices. We propose a model, consisting of a speech encoder/decoder, a face encoder/decoder and a voice encoder. We use the latent code of an input face image encoded by the face encoder as the auxiliary input into the speech decoder and train the speech encoder/decoder so that the original latent code can be recovered from the generated speech by the voice encoder. We also train the face decoder along with the face encoder to ensure that the latent code will contain sufficient information to reconstruct the input face image.

Crossmodal Voice Conversion/Face Image Generation

Leverage underlying correlation between voices and appearances to

- • convert speech into a voice that matches an input face image, and
- generate face image that matches input speech.

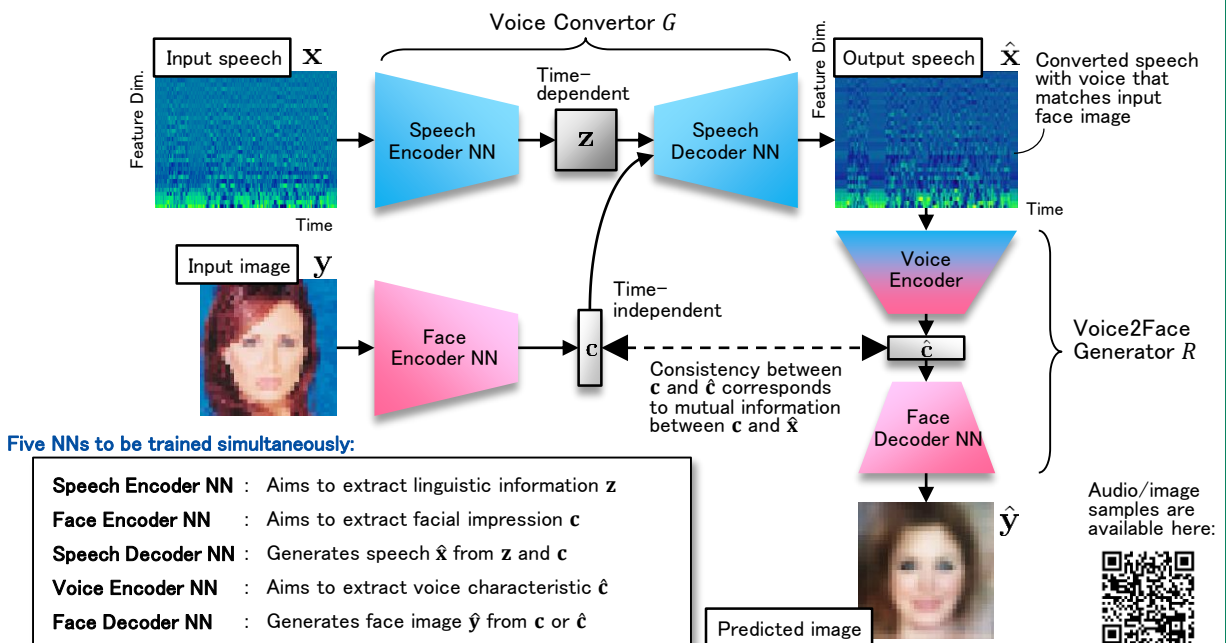
Information-theoretic approach using deep generative models

Voice Converter G : Neural network (NN) that converts input speech \mathbf{x} into $\hat{\mathbf{x}} = G(\mathbf{x}, \mathbf{y})$ by using face image \mathbf{y} as auxiliary input

Training objective: Train G so that mutual information between $\hat{\mathbf{x}} = G(\mathbf{x}, \mathbf{y})$ and \mathbf{y} is maximized

$$I[G(\mathbf{x}, \mathbf{y})|\mathbf{y}] \geq \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim p(\mathbf{x}, \mathbf{y})} [\log R(\mathbf{y}|G(\mathbf{x}, \mathbf{y}))] \rightarrow \text{Maximize lower bound w.r.t. } G \text{ and } R$$

Speech and face image pair \rightarrow NN that approximates posterior $p(\mathbf{y}|\mathbf{x})$



References

- [1] H. Kameoka, T. Kaneko, K. Tanaka, N. Hojo, "StarGAN-VC: Non-parallel many-to-many voice conversion using star generative adversarial networks," in *Proc. 2018 IEEE Workshop on Spoken Language Technology (SLT 2018)*, pp. 266–273, Dec. 2018.
- [2] H. Kameoka, T. Kaneko, K. Tanaka, N. Hojo, "ACVAE-VC: Non-parallel voice conversion with auxiliary classifier variational autoencoder," *arXiv:1808.05092 [stat.ML]*, 2018.
- [3] H. Kameoka, K. Tanaka, A. Valero Puche, Y. Ohishi, T. Kaneko, "Crossmodal Voice Conversion," *arXiv:1904.04540 [cs.SD]*, 2019.

Contact

Hirokazu Kameoka Email: cs-liaison-ml at hco.ntt.co.jp
Media Recognition Research Group, Media Information Laboratory

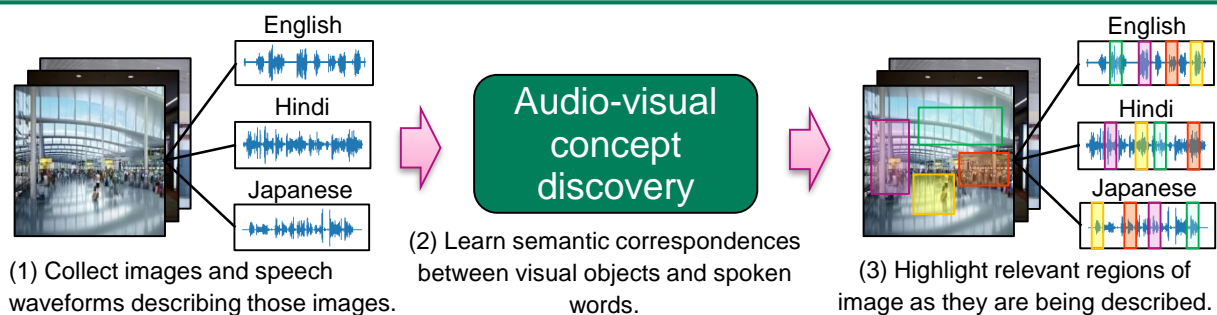


Innovative R&D by NTT
Open House 2019

Copyright © 2019 NTT. All Rights Reserved.

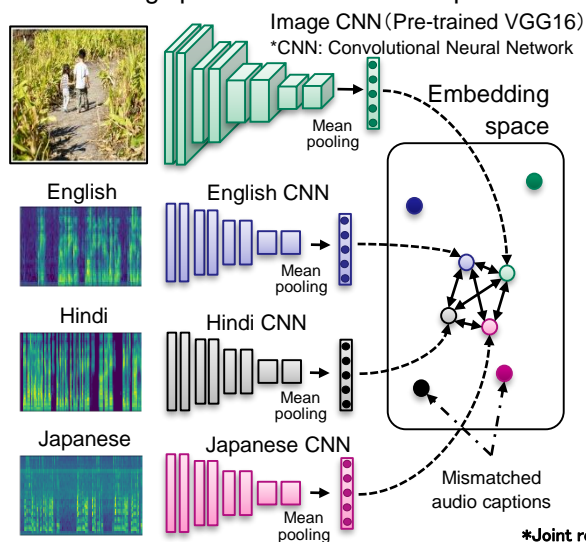
Abstract

In order for AI to visually perceive the world around it and to use language to communicate, it needs a dictionary that associates the visual objects in the world with the spoken words that refers to them. We explore **a neural network models that learn semantic correspondences between the objects and the words** given images and multilingual speech audio captions describing those images. We show that training a trilingual model simultaneously on English, Hindi, and newly recorded Japanese audio caption data offers improved retrieval performance over the monolingual models. Further, we demonstrate **the trilingual model implicitly learns meaningful word-level translations based on images**. We aim for a future in which AI discovers concepts autonomously while finding the audio-visual co-occurrences by simply providing media data that exists in the world such as TV broadcasting. We also consider **the application to large-scale archive retrieval and automatic annotation** that involves interactions between different sensory modalities such as vision, audio, and language.



Learning neural network embeddings for images and spoken audio captions

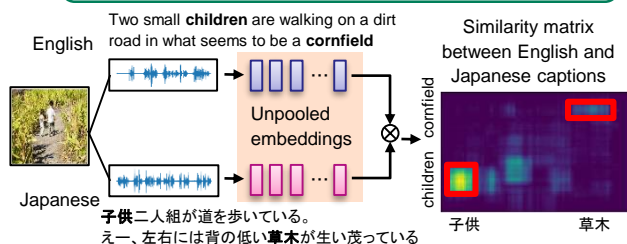
Paired image and audio captions are more similar in embedding space than mismatched pairs



Evaluation of embedding space learned from image and audio captions

- (1) Audio-visual retrieval performance**
Recall scores for the top 10 hits (1,000 image/caption pairs)
Monolingual model: 0.45 → Multilingual model: 0.50
- (2) Cross-lingual audio2audio retrieval performance**
Recall scores for the top 10 hits (1,000 cross-lingual caption pairs)
w/o using images: 0.01 → w/ using images: 0.50

Exploring visually grounded speech-to-speech translation



*Joint research results with MIT Computer Science and Artificial Intelligence Laboratory

References

- [1] Y. Ohishi, A. Kimura, T. Kawanishi, K. Kashino, D. Harwath, and J. Glass, "Crossmodal Search using Visually Grounded Multilingual Speech Signal," *IEICE Technical report on Pattern Recognition and Media Understanding (to appear)*
- [2] D. Harwath, G. Chuang, and J. Glass, "Vision as an Interlingua: Learning Multilingual Semantic Embeddings of Untranscribed Speech," In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2018)*, April 2018.

Contact

Yasunori Ohishi Email: cs-liaison-ml at hco.ntt.co.jp
Media Recognition Group, Media Information Laboratory



Innovative R&D by NTT
Open House 2019

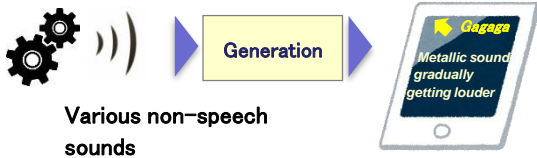
Copyright © 2019 NTT. All Rights Reserved.

Abstract

Recently, detection and classification of various sounds has attracted many researchers attention. We propose an **audio captioning system that can describe various non-speech audio signals in the form of natural language**. Most existing audio captioning systems have mainly focused on “what the individual sound is,” or classifying sounds to find object labels or types. In contrast, **the proposed system generates (1) an onomatopoeia, i.e. a verbal simulation of non-speech sounds, and (2) a sentence describing sounds**, given an audio signal as an input. This allows the description to include more information, such as **how the sound sounds and how the tone or volume changes over time**. Our approach also enables directly measuring the distance between a sentence and an audio sample. The potential applications include sound effect search systems that can accept detailed sentence queries, audio captioning systems for videos, and AI systems that can hear and represent sounds as humans do.

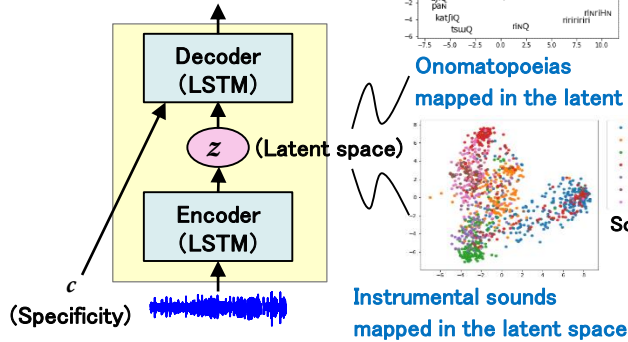
Detailed audio captioning

- Given an audio input, the system describes not only what the sound source is, but also how it is sounding and how it is changing over time, as a natural language sentence. **NEW**



Method

A high-pitched fricative noise is ...



1. Onomatopoeic mode: output a word simulating the sound
2. Sentence mode: output a sentence describing the sound

【Point】 No unique correct answer for captioning
→ 【Proposal】 Can control the degree of detail by conditioning the decoder by “Specificity” input
 Specificity: Sum of the amount of information contained in the output text.

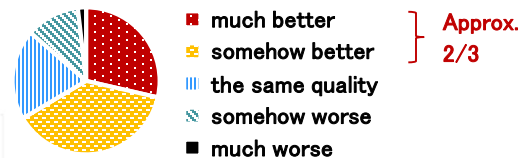
Experimental results

The description matches the sound:



Appropriateness of the output sentence (without specificity conditioning)

With the specificity control, the sentence is:



Effectiveness of the specificity conditioning

Examples of description for a base drum sound (English translation from Japanese)

c	Generated sentence
—	A low sound rings for a moment
20	A low sound sounds for a moment
50	A low, striking sound sounds as if something is dashed on a mat
80	A very low-pitched drum is played only once
110	A faint, low-pitched sound sounds as if something is hit dully, and it soon disappears

References

- [1] Shota Ikawa, Kunio Kashino, “Generating sound words from audio signals of acoustic events with sequence-to-sequence model,” In Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018), April 2018.
- [2] Shota Ikawa, Kunio Kashino, “Acoustic event search with an onomatopoeic query: measuring distance between onomatopoeic words and sounds,” In Proc. Detection and Classification of Acoustic Scenes and Events (DCASE 2018), November 2018.

Contact

Kunio Kashino Email: cs-liaison-ml at hco.ntt.co.jp
 Media Information Laboratory



Innovative R&D by NTT
 Open House 2019

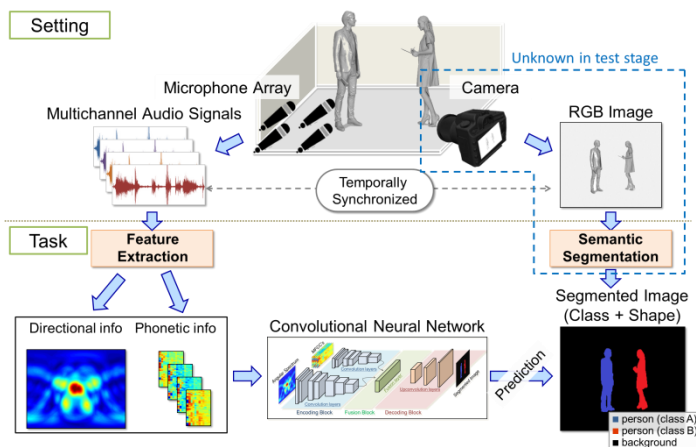
Copyright © 2019 NTT. All Rights Reserved.

Abstract

Sounds provide us with vast amounts of information about surrounding objects and scenes and can even remind us visual images of them. Is it possible to implement this noteworthy ability on machines? We addressed this task and developed **a crossmodal scene analysis method that can predict the structures and semantic classes of objects/scenes from auditory information alone**, i.e., without actually looking at the scene. Our approach uses a convolutional neural network that is designed to directly output semantic and structural information of objects and scenes by taking low-level audio features as its inputs. An efficient feature fusion scheme is incorporated to model underlying higher-order interactions between audio and visual sources. Our method allows users to visually check the state of the scene even in a case where they cannot or do not want to use a camera. Our method will contribute to expanding the availability of monitoring applications in various environments.

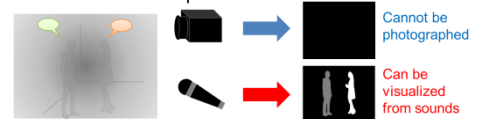
Crossmodal Scene Understanding

Predicting structures and semantic classes of objects from multi-channel audio signals



Visualizing scenes where photographing impossible or prohibited

Dark rooms or spaces with high privacy levels can be visualized with microphones.



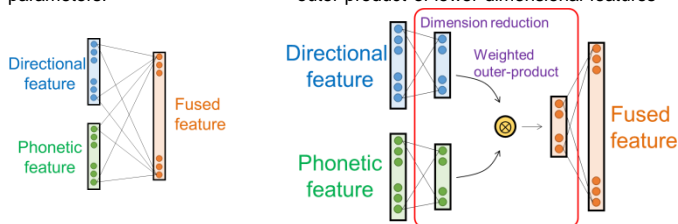
Demonstrated recognition of limited number of object classes possible so far



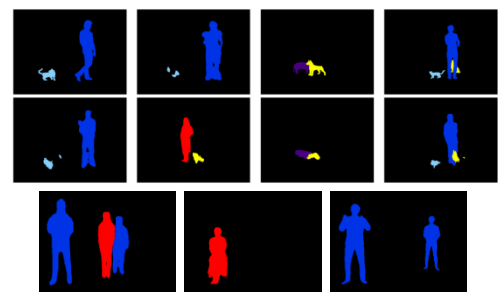
Features

Feature fusion layer for efficiently modeling higher-order interactions between audio and visual sources

Typical fusion scheme has prohibitive number of parameters! Our approach reduces the number of parameters by considering weighted outer-product of lower-dimensional features



Recognition of various classes of objects from real sound sources possible



References

- [1] G. Irie, M. Ostrek, H. Wang, H. Kameoka, A. Kimura, T. Kawanishi, K. Kashino, "Seeing through sounds: predicting visual semantic segmentation results from multichannel audio signals," in Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2019.
- [2] H. Wang, G. Irie, H. Kameoka, A. Kimura, K. Hiramatsu, K. Kashino, "Audio-based Semantic Segmentation based on Bilinear Feature Fusion," Meeting on Image Recognition and Understanding (MIRU), 2018.

Contact

Go Irie Email: cs-liason-ml at hco.ntt.co.jp
Recognition Research Group, Media Information Laboratory



Innovative R&D by NTT
Open House 2019

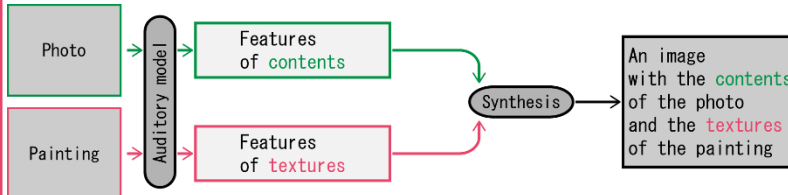
Copyright © 2019 NTT. All Rights Reserved.

Abstract

Natural scenes such as bubbling water and rustling trees give us specific perception of **sound textures**. We developed a method to **artificially give such textures** to speech and music. Inspired by research on manipulating image textures, we improved the method so that we can apply it to sound. A computational model that takes into account our hearing mechanism enabled **effective control of sound textures** in terms of hearing sensation. The method is realized in the same framework as the image texture manipulation. This indicates that, in the brain, seen and heard textures are processed by the similar mechanisms. From a scientific viewpoint, this study leads to **understanding of the mechanisms of sound texture perception** by comparing the model's internal states with the brain activities induced by hearing sounds. From an application perspective, the proposed method enables us to speak in a voice that does not actually exist or to play music with an instrument that does not exist.

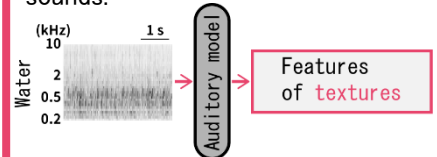
Prior study: image texture conversion

By combining **contents** in a photo and **textures** in a painting, an image that has both of them is synthesized. (Converting the **textures** of the photo to that of the painting while preserving its **contents**.)



Prior study: representation of sound textures

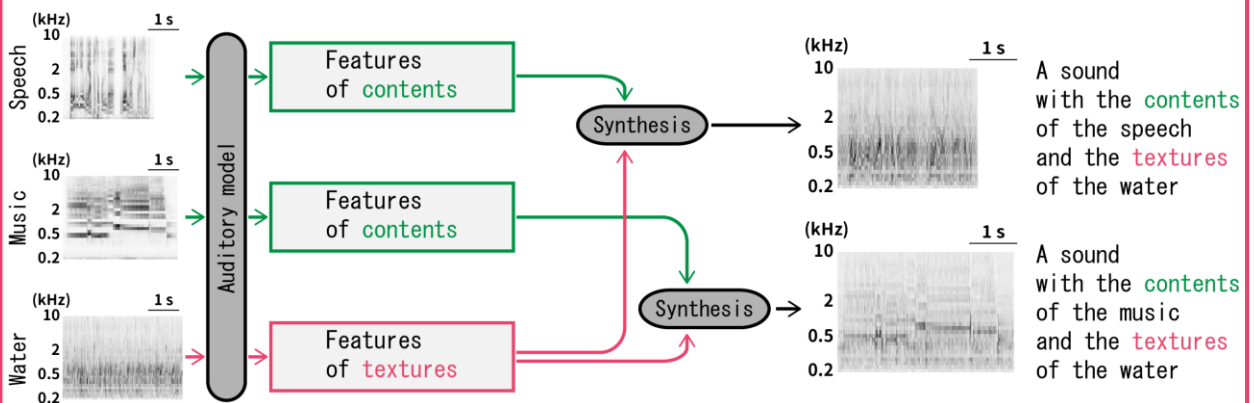
A model of the auditory system calculates the features that represent **textures** of environmental sounds.



This study: sound texture conversion

A model of the auditory system calculates the features that represent **textures** of environmental sounds and the features that represent **contents** of speech and music.

By combining **contents** in a speech or music and **textures** in an environmental sound, a sound that has both of them is synthesized. (Converting the **textures** of the speech or music to that of the environmental sound while preserving its **contents**.)



References

- [1] T. Koumura, H. Terashima, S. Furukawa, "Chimeric sounds with shuffled "texture" and "content" synthesized by a model of the auditory system," in Proc. International Symposium on Universal Acoustical Communication, 2018.
- [2] T. Koumura, H. Terashima, S. Furukawa, "Sound texture transfer using a model of the auditory system," in Proc. Annual Conference of the Japanese Society for Artificial Intelligence, 2017.

Contact

Takuya Koumura Email: cs-liaison-ml at hco.ntt.co.jp
Sensory Resonance Research Group, Human Information Science Laboratory



Innovative R&D by NTT
Open House 2019

Copyright © 2019 NTT. All Rights Reserved.

Abstract

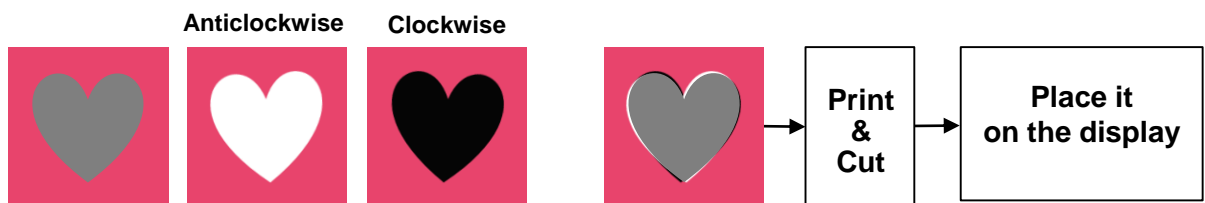
We propose a technique to give motion illusions to static paper objects. Previous studies have reported visual illusions wherein a static “virtual” object apparently moves on the basis of the luminance interaction between object’s contours and the object’s background. However, no studies have proposed a method to give motion illusion to a static “real” object. This study found a phenomenon in which a paper objects having bright and dark contours apparently moved against the background with dynamic luminance modulation. Manipulating the contour patterns could also produce not only a simple illusory movement such as translation but also relatively complex illusory movements such as expansion, contraction, and rotation. We call this technique Danswing (Dance + swing) papers. By utilizing the Dancing papers, it is possible to gather customer’s attention towards an actually static, but perceptually dynamic, objects.

How to create Danswing papers

● Apparently rotating heart object

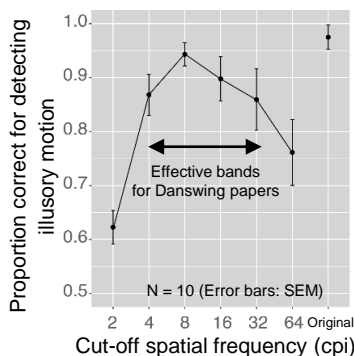
① We digitally create images of a gray heart-shape figure, a slightly anticlockwise rotated white heart-shape object, and a slightly clockwise rotated black heart-shape object.

② The three heart-shape objects are digitally synthesized so that the gray object is set at the most front layer. The synthesized image is printed out, cut, and placed on the display with dynamic luminance modulation.



Visual mechanism for Danswing papers

● Specific bands of spatial frequency are related to Danswing papers.



Using stimulus clips wherein a specific band of spatial frequency was extracted, we asked observers to detect illusory motion, and found that the observers’ performance was good when the clips contained the specific bands of spatial frequency.

Please scan the right QR code, and check our YouTube clip of Danswing papers!



References

- [1] T. Kawabe, “Danswing papers,” in *Proc. SIGGRAPH Asia 2018 (SA '18) Posters* Article No. 4.
 [2] T. Kawabe, “Danswing papers,” Top 10 finalist of Best illusion of the year contest. <http://illusionoftheyear.com/2018/10/danswing-papers/>

Contact

Takahiro Kawabe Email: cs-liaison-ml@hco.ntt.co.jp
 Sensory and Representation Research Group, Human Information Science Laboratory



Innovative R&D by NTT
 Open House 2019

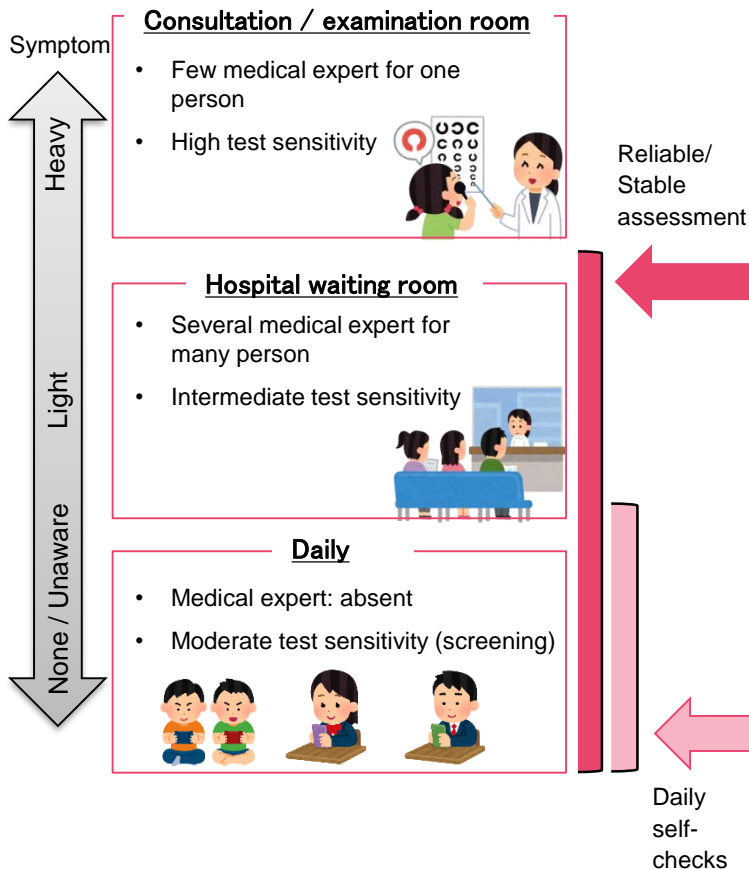
Copyright © 2019 NTT. All Rights Reserved.

Self eye-check system using video games and tablet PCs

Abstract

We explore an enjoyable and simple way to measure functions of the eye. We combined know-hows accumulated through vision-science-experiments with technologies for drawing precise computer graphics on web browsers, and **created a system for testing visual functions with a generic tablet device. Our system can be utilized for self-checking of eye functions in a delightful way like a video game.** Previous tests of visual function are often time-consuming and normally require the help of medical experts. **Our system allows users to measure each visual function in about 3 minutes.** This system can be utilized to self-check users' eye condition routinely. In addition, by accumulating knowledge through data of many people including patients with eye diseases in simple and short-time measurement, we can expect an early detection of eye diseases, rehabilitation application, and scientific findings about complex visual processes.

Usage of vision test in various places



Simple tablet tests

- Easy tests with short measuring time
- Run on tablet PCs.
- Modified conventional measurement protocols.



Gamified vision tests

- Gamification with good graphical and protocol designs
- Aiming at repeated use in daily situations



References

- [1] K. Hosokawa, K. Maruya, S. Nishida, "Testing a novel tool for vision experiments over the internet," *Journal of Vision*, Vol. 16, p. 967, 2016.

Contact

Kazushi Maruya Email: cs-liaison-ml at hco.ntt.co.jp
Sensory Representation Research Group, Human Information Science Laboratory



Innovative R&D by NTT
Open House 2019

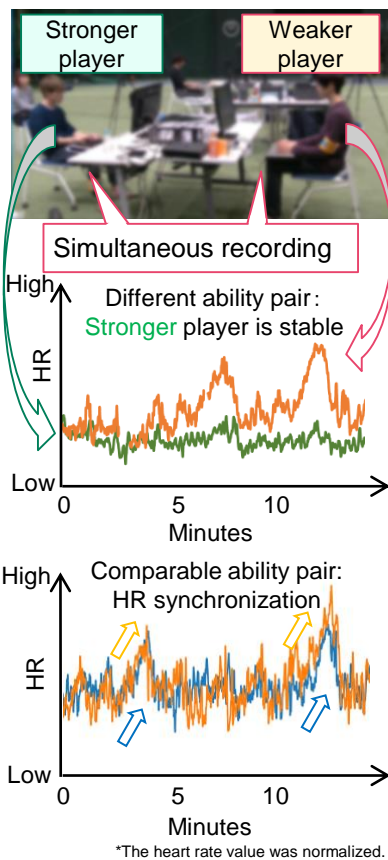
Copyright © 2019 NTT. All Rights Reserved.

Abstract

It is said that mental state is an important factor if one is to be a winner in sports. Although mental and physiological states are related, **the relationship between physiological state and sporting performance, especially in real games**, remains unclear. Here, we investigate this relationship for real competition in esports, baseball, and snowboarding by focusing on the heart rate (HR) as an indicator of mental state. The results show **a strong relationship between sporting performance and HR**, such as the huge variation in HR that occurs when the opponent is a higher-level player, the stable performance that accompanies a stable heart rate regardless of the situation, and a top player delivering a good performance with a high HR. Further investigation will reveal the component of the mental state related to performance and will enable us to develop ways of improving athletes' performance by adjusting their physiological state.

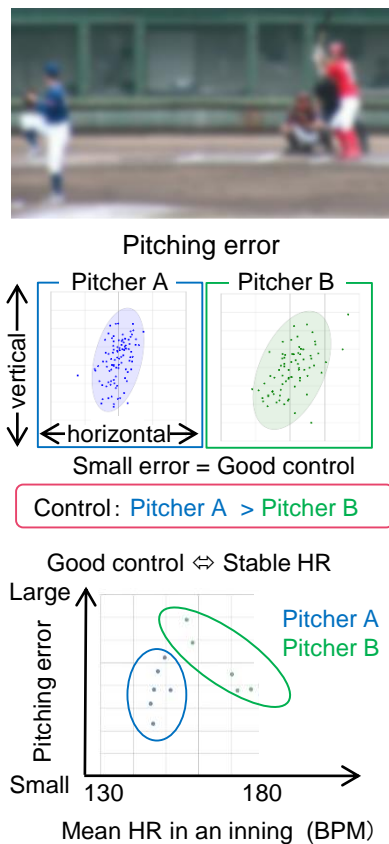
Esports (Fighting game)

Relative abilities of opponents define the HR variation



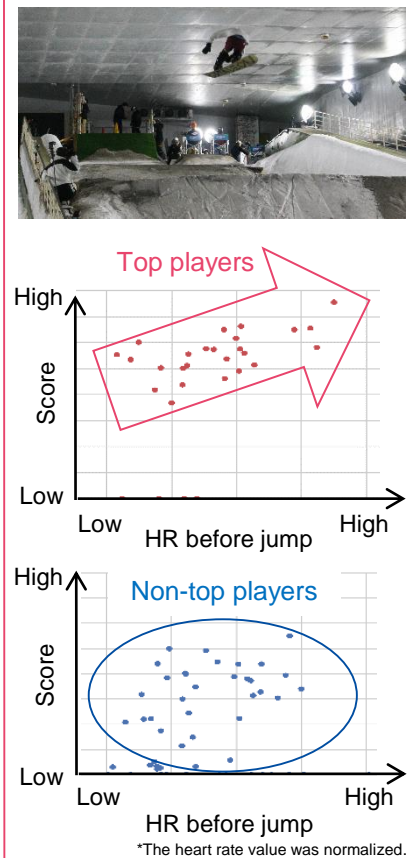
Baseball (Pitching)

Stable HR leads to stable high-performance pitching



Snowboard (Jumping)

Top players HRs increase to make a high-scoring jump



References

- [1] K. Watanabe, N. Saijo, M. Kashino, "The physiological change reflecting the fight-or-flight response of an esports player correlates strongly with that of the opponent," in Proc. NEURO2019, 2019.
- [2] T. Fukuda, T. Mochida, N. Saijo, M. Kashino, "Game situation affects pitching control - variation of pitching error distribution in real games," Japan Society of Baseball Science, the 5th Annual Meeting, 2017.
- [3] S. Matsumura, K. Watanabe, T. Kimura, M. Kashino, "Relationship between pre-competitive physiological states and performance in snowboard jumping competitions," Japan Society of Ski Science, the 29th Annual Meeting, 2019.

Contact

Ken Watanabe Email: cs-liaison-ml at hco.ntt.co.jp
Sports Brain Science Project



Innovative R&D by NTT
Open House 2019

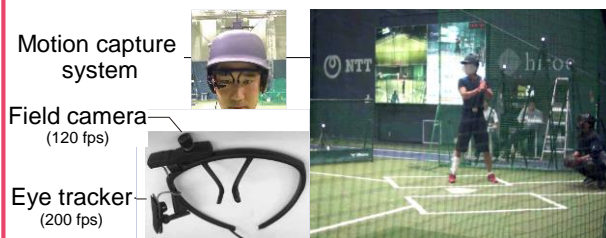
Copyright © 2019 NTT. All Rights Reserved.

Abstract

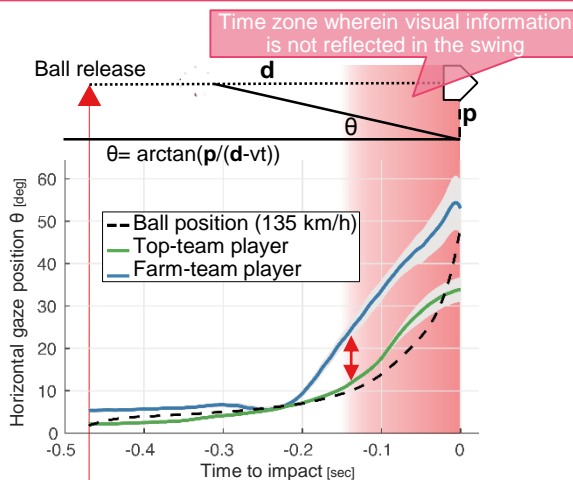
In ball games, it is necessary to move the body appropriately in reaction to a ball moving at high speed, but the mechanism of such movement is not known. In this study, we **examined the brain mechanism that captures a fast moving ball in a limited time** by 1) measuring the eye and body movements of professional baseball players while they were actually hitting and 2) performing basic experiments using an optical illusion. By measuring eye and body movements in a scenario close to the actual game, we succeeded in capturing the **sophisticated skills used by top athletes**. In basic experiments using the illusion, we clarified **how the brain uses visual information to control body movements**. Our goal is twofold: to uncover the implicit brain functions for vision and action and to establish a **new training method to train people in techniques for optimal body control according to the situation**. This will help improve the motor skills of a wide range of people, from children to the elderly, as well as top athletes.

How do professional baseball players hit a fastball?

Experiment Measured eye and body movements using wearable sensors in a situation close to a real game



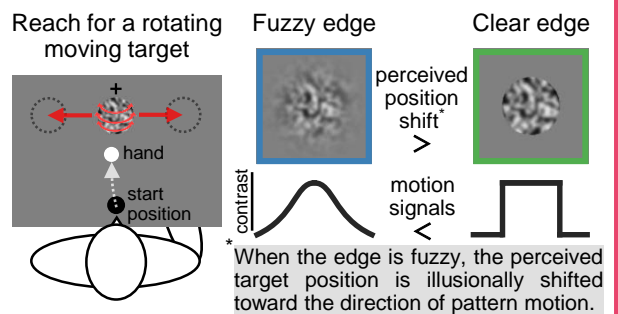
Results Good hitters capture a ball by linking eye and body movements until such a time as no visual information is reflected in the swing.



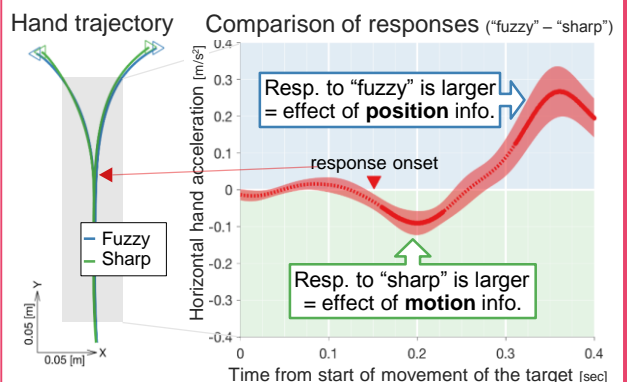
Not only visual information but also motor commands in the brain are used.

Why can a fastball be hit in a very limited time?

Experiment Tested whether motor responses to a moving target is driven by **position** or **motion** brain information



Results Arm movements are first driven by target motion information and then adjusted by the position information.



The bat's trajectory starts to change before the ball's position is located.

References

- 1) Y. Kishita, H. Ueda, M. Kashino, "Eye movements in real baseball batting by elite players," in Proc. *The 48th Annual Meeting of the Society for Neuroscience*, 2018.
- 2) H. Ueda, N. Abekawa, S. Ito, H. Gomi, "Temporal development of an interaction effect between internal motion and contour signals of drifting target on reaching adjustment," in Proc. *The 47th Annual Meeting of the Society for Neuroscience*, 2017.

Contact

Hiroshi Ueda Email: cs-liaison-ml at hco.ntt.co.jp
Sports Brain Science Project



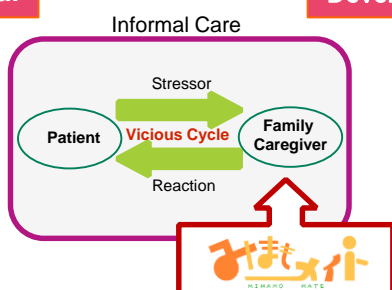
Innovative R&D by NTT
Open House 2019

Copyright © 2019 NTT. All Rights Reserved.

Abstract

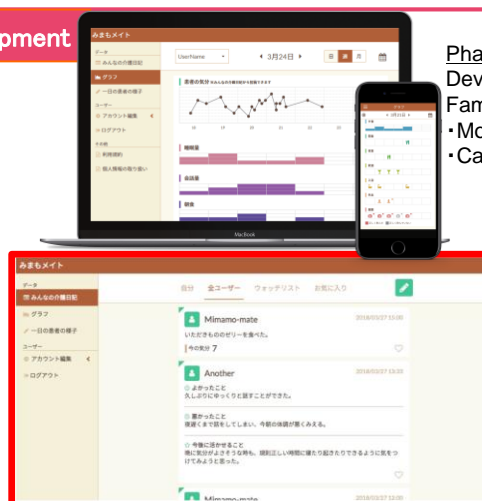
Previous research has shown that tracking technologies have the potential to help family caregivers optimize their coping strategies and improve their relationships with care recipients. In this research, we explore how **sharing the tracked data** (i.e., caregiving journals and patient's conditions) with other family caregivers affects home care and family communication. Although previous works suggested that family caregivers may benefit from reading the records of others, sharing patients' private information might fuel negative feelings of surveillance and violation of trust for care recipients. To address this research question, we added a sharing feature to the previously developed tracking tool and deployed it for six weeks in the homes of 15 family caregivers who were caring for a depressed family member. Our findings show how the sharing feature attracted the attention of care recipients and **helped the family caregivers discuss sensitive issues with care recipients**.

Goal



Our goal is to design a tool that helps family caregivers improve their care and communication with their care recipients at home.

Development



Phase 1

Development of a tracking tool.
Family caregivers record:
• Mood & activities of the patients
• Caregiving activities

Phase 2

• Add a **sharing** feature to the tracking tool
• Can **read others'** records.

Study

To investigate how information sharing about care recipients by family caregivers impacts family communication

Hypotheses:

1. Learn from other records
2. Gain emotional support
3. Conflict between caregiver-care recipient

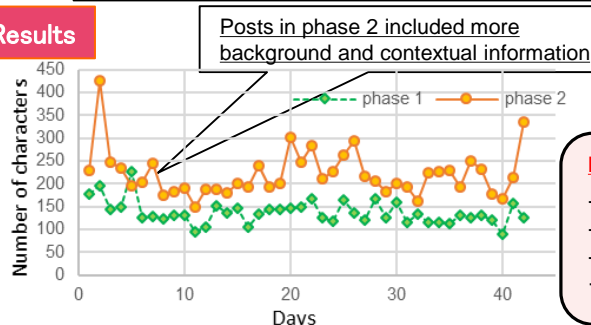
Participants (family caregivers):

- 14 adults that care for a depressed family member
- 11 women, 3 male, average age 43
- 6 housewives, 6 full-time employees, 2 self-employed

Patient condition:

- Onset of illness: On average, 8yrs ago, 8 experienced a relapse
- All take antidepressants, all given regular consultation

Results



Hypotheses 1 & 2: Supported.

Participants reported improvement of their coping strategies by learning from others. Hypotheses 3: Not supported.

Increased caregiver-care recipient communication

- Others' records triggered communication about depression
 - Care recipients suggested what to write on caregiving journal
 - Caregivers indirectly expressed feelings to care recipients
- Importance of supporting indirect communication

References

- [1] N. Yamashita, H. Kuzuoka, T. Kudo, K. Hirata, E. Aramaki, K. Hattori, "How Information Sharing about Care Recipients by Family Caregivers Impacts Family Communication," in *Proc. ACM Conference on Human Factors in Computing Systems (CHI)*, 2018.
- [2] N. Yamashita, H. Kuzuoka, K. Hirata, T. Kudo, E. Aramaki, K. Hattori, "Changing Moods: How Manual Tracking by Family Caregivers Improve Caring and Family Communication," in *Proc. ACM Conference on Human Factors in Computing Systems (CHI)*, 2017.

Contact

Naomi Yamashita Email: cs-liaison-ml at hco.ntt.co.jp
Interaction Research Group, Innovative Communication Laboratory



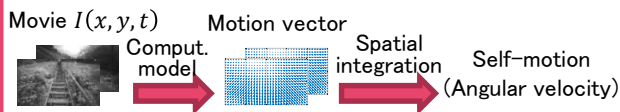
Innovative R&D by NTT
Open House 2019

Abstract

Visual motion has critical roles for quickly adjusting posture, eyes, and limbs in dynamic interactions with environments. By behavioral experiments and synthetic model simulations, we have tried to reveal fundamental mechanisms of implicit visuomotor processing. It is difficult to retrieve detailed information about the scene from highly blurred image. However, we found that **blurred image sequence can provide higher estimation accuracy of rapid self-motion than the original image sequence**. Interestingly, implicit motor responses of hands and eyes are highly sensitive for low-spatial frequency stimuli. These results suggest that the brain knows the importance of low-spatial frequency component to code the high-speed self-motion from the statistical relationship between visual motion and head/posture fluctuation. This type of visuomotor control would be helpful to realize a novel visual processing for moving robot.

Procedure of self-motion estimation

- We employed a computational model which reproduced human visual processing properties[1]



Natural statistics of self-motion & estimation accuracy of proposed model

- Blurred images provided highest estimation accuracy (center column of scatter plots)
- Motion signals in low-spatial frequency are important to estimate self-motion

Watching a poster



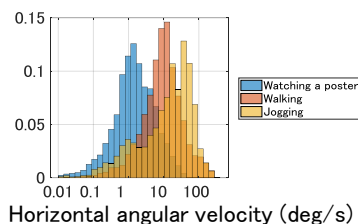
Walking



Jogging



Distributions of self-motion



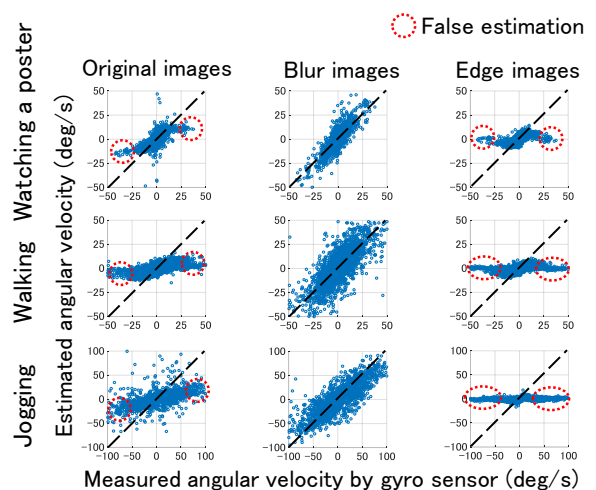
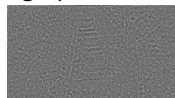
Original image



Blurred image (low-pass filtered)

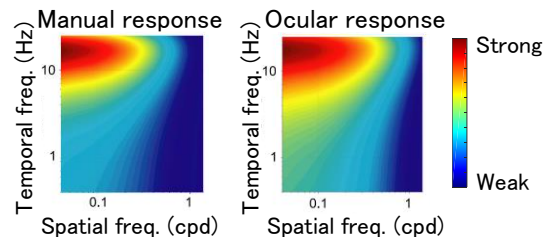


Edge image (high-pass filtered)



Human studies: implicit motor responses induced by visual motion[2]

- Spatiotemporal tunings of implicit motor responses are highly sensitive to low-spatial frequency stimulus



The body sees blurred images to estimate rapid self-motion information

References

- [1] 中村大樹, 佐藤俊治, “計算論的に最適な速度推定器によってMT野細胞の複雑な反応特性を説明する,” in 第27回日本神経回路学会全国大会, 2017.
- [2] H. Gomi, N. Abekawa, S. Nishida, “Spatiotemporal tuning of rapid interactions between visual-motion analysis and reaching movement,” *The Journal of Neuroscience*, vol. 26, No. 20, pp. 5301-5308, 2006.
- [3] D. Nakamura, H. Gomi, “Statistical analysis of optic flow induced by body motion characterizing OFR and MFR,” in *JNNS satellite meeting*, 2018.

Contact

Daiki Nakamura Email: cs-liaison-ml at hco.ntt.co.jp
Sensory and Motor Research Group, Human Information Science Laboratory



Innovative R&D by NTT
Open House 2019

Copyright © 2019 NTT. All Rights Reserved.

Abstract

Body action such as walking is known to extend the subjective boundaries of peripersonal space (PPS; the **space immediately surrounding our body**) and to facilitate the processing of audio-tactile multisensory stimuli presented within the PPS. However, it is unclear whether the boundaries change when a sensation of walking is induced with no physical body motion. Here, we presented several vibration patterns on the soles of the feet of seated participants to evoke a sensation of walking, together with a looming sound approaching the body. We measured reaction times for detecting a vibrotactile stimulus on the chest, which was taken as a behavioral proxy for the PPS boundary. Results revealed that a cyclic vibration consisting of lowpass-filtered walking sounds presented at the soles that clearly evoked a sensation of walking decreased the reaction times, indicating that the PPS boundary was expanded forward by inducing a sensation of walking.

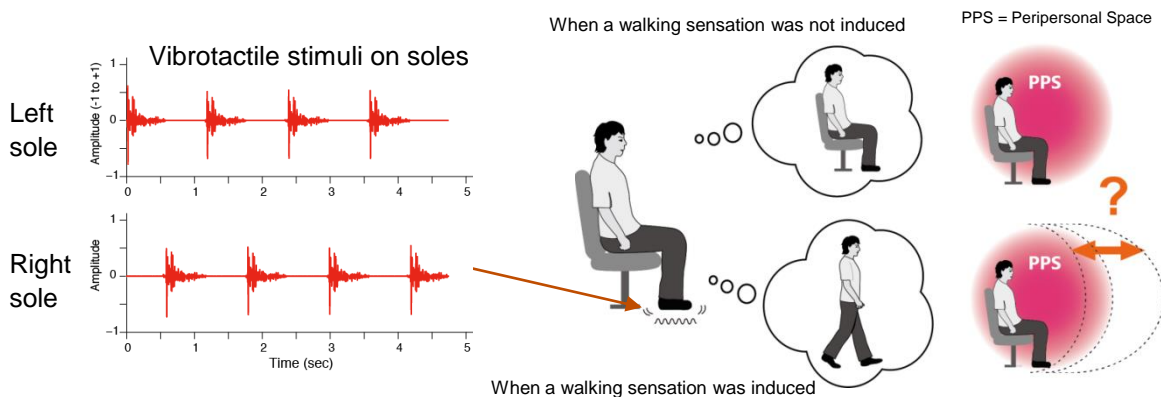
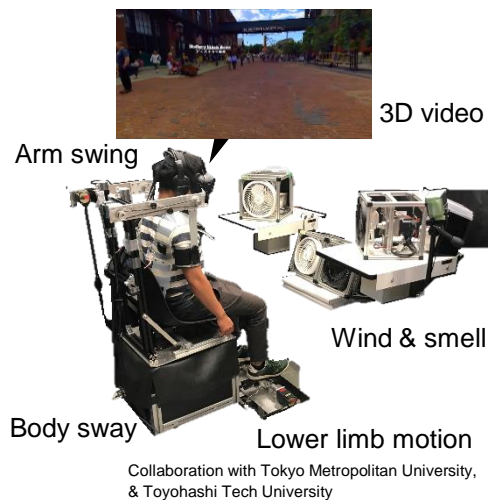
A sensation of pseudo-walking created by multisensory stimulation

We have developed a method for creating a sensation of walking by combining multisensory information, such as a mechanism for moving the upper and lower limbs, a motorized chair for body sway, and wind and smell presentation, with the aim of achieving “generating a sensation of walking while seated”.

We focused on the soles of foot and presented a vibration to create a sensation of walking.

PPS expanded by sensation of pseudo-walking

We found that reaction time to a stimulus approaching toward the body changed when a vibration stimulus was applied to the sole of the foot.



References

- [1] Tomohiro Amemiya, “Haptic Interface Technologies Using Perceptual Illusions,” in *Proc. of 20th International Conference on Human-Computer Interaction (HCI International 2018)*, pp.168-174, Las Vegas, NV, July 2018.
- [2] Koichi Shimizu, Gaku Sueta, Kentaro Yamaoka, Kazuki Sawamura, Yujin Suzuki, Keisuke Yoshida, Vibol Yem, Yasushi Ikei, Tomohiro Amemiya, Makoto Sato, Koichi Hirota, Michiteru Kitazaki, “FiveStar VR: shareable travel experience through multisensory stimulation to the whole body,” in *Proc. of SIGGRAPH Asia 2018 Virtual & Augmented Reality*, Article 2, Tokyo, Japan, Dec. 2018.

Contact

Tomohiro Amemiya Email: cs-liaison-ml at hco.ntt.co.jp
Sensory and Motor Research Group, Human Information Laboratory



Innovative R&D by NTT
Open House 2019

Copyright © 2019 NTT. All Rights Reserved.

