様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科研費
KAKENHI

科学研究費助成事業　　研究成果報告書

令和　５ 年　６ 月 １４ 日現在

機関番号：８２４０１

研究種目：若手研究

研究期間：2020～2022

課題番号：２０Ｋ１９８３３

研究課題名（和文）A Unified Computational Model for Audio-Visual Recognition of Human Social Interaction

研究課題名（英文）A Unified Computational Model for Audio-Visual Recognition of Human Social Interaction

研究代表者

Ｎｕｇｒａｈａ　Ａｄｉｔｙａ（NUGRAHA, Aditya Arie）

国立研究開発法人理化学研究所・革新知能統合研究センター・研究員

研究者番号：６０８５８０２５

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：人間の言語コミュニケーションを理解するために、視聴覚情報処理に関する確率的な計算モデルを策定することを目的とした。音声強調に利用するために、話者・音素特徴から音声信号を生成するモデルを提案した。また、古典的なIVAや最先端のFastMNMFなどの時不変のブラインド音源分離（BSS）手法に対して、Normalizing Flowに基づく原理的な時不変の拡張を紹介した。最後に、スマートグラスを用いた適応的なオーディオビジュアル音声強調を開発した。カメラ映像の指向性情報によって制御され、高速な環境依存型ビームフォーミングと低速な環境依存型BSSによって、ロバストで低遅延な音声強調を実現した。

研究成果の学術的意義や社会的意義
One key achievement is the prototype of adaptive speech enhancement for real-time speech transcription with head-worn smart glasses. It involves challenging egocentric information processing with non-stationary sensors. This technology may benefit older adults and people with hearing impairment.

研究成果の概要（英文）：We aimed for a probabilistic computational model of audio-visual information processing for understanding human verbal communication. We proposed a model for generating speech signals from speaker labels controlling the voice characteristics and phone labels controlling the speech content. For speech enhancement, it potentially improves not only the signal quality but also the speech intelligibility. We also introduced principled time-varying extensions, based on a novel deep generative model called normalizing flow, of time-invariant blind source separation (BSS) methods, including the classical independent vector analysis and the state-of-the-art FastMNMF. Finally, we developed adaptive audio-visual speech enhancement with augmented reality smart glasses. Camera images allow speakers of interest to be identified to control direction-aware enhancement. We achieve robust low-latency enhancement via a fast environment-sensitive beamforming governed by a slow environment-agnostic BSS.

研究分野：Audio-visual speech enhancement for smart glasses

キーワード：Audio-visual processing　Smart glasses　Adaptive system　Blind source separation　Speech enhancement　Speech recognition　Neural spatial model　Generative model

1. 研究開始当初の背景

Human communication relies on information obtained by the different senses in a complementary manner. Although human senses include sight, hearing, taste, smell, and touch, visual and auditory systems are typically predominant. A human can interact with a group of people, even in an unfamiliar crowd (environment), by visually localizing and identifying the sound sources, especially the active speakers, to improve the listening focus. **Could we formulate a computational model of audio-visual information processing so that a computer can robustly do the same in different environments?**

While conventional studies on audio-visual information typically address different tasks separately, such as speaker identification, speech separation, and lip reading, **jointly considering multiple tasks based on audio-visual information in a complementary manner should be beneficial**. For example, visual-based speaker localization is more reliable than audio-based one when a speaker is seen. In contrast, audio-based localization can still work when a person is not seen but actively speaking. Associating speech characteristics obtained from speaker identification to facial features from face detection could improve the robustness of audio-visual speaker localization in unideal cases that trouble localization based on audio or visual data only, such as when speakers are occluded or silent at times. Unified processing of both modalities promises a superior performance that benefits applications, e.g., adaptive speech enhancement.

Many studies on audio-visual information processing employ deep neural networks (DNNs) trained in a fully supervised manner, so the robustness of the DNNs to unseen inputs would be limited, depending on the parallel training data diversity. We, therefore, want to formulate a probabilistic model to cope with the variability of audio-visual data and its interrelationship. In the DNN-based generative modeling framework called the variational autoencoder (VAE), **a generation DNN and a recognition DNN are jointly trained on the observations in an unsupervised manner to discover the latent variables**. This unsupervised training does not need parallel data, so it is practical for real-world applications.

2. 研究の目的

This research aimed **to formulate and realize a probabilistic computational model of audio-visual information processing**. We wanted to learn the underlying representation (latent variables), such as *message (idea)* $\mathbf{z}_\mathrm{m}$, *voice feature* $\mathbf{z}_\mathrm{s}$, and *facial feature* $\mathbf{z}_\mathrm{f}$, from the audio-visual inputs, such as *speech signal* $\mathbf{s}$ and *facial expression* $\mathbf{f}$, in an unsupervised manner. It translates to modeling the joint probability distribution given by:

$$p(\mathbf{s},\mathbf{f},\mathbf{z}_s,\mathbf{z}_f,\mathbf{z}_m) \;=\; p(\mathbf{s}|\mathbf{z}_s,\mathbf{z}_m)\,p(\mathbf{f}|\mathbf{z}_f,\mathbf{z}_m)\,p(\mathbf{z}_s)\,p(\mathbf{z}_f)\,p(\mathbf{z}_m).$$

This unified model may simultaneously take into account different tasks that are often separately considered in existing studies. The tasks may include *recognition* ones, such as speaker identification (speech signal $\mathbf{s}$ → speaker identity $\mathbf{z}_\mathrm{s}$), face identification (face image $\mathbf{f}$ → human identity $\mathbf{z}_\mathrm{f}$), and speech recognition (speech signal $\mathbf{s}$ → text $\mathbf{z}_\mathrm{m}$), and also *generation* ones, such as speech synthesis (text $\mathbf{z}_\mathrm{m}$ + speaker identity $\mathbf{z}_\mathrm{s}$ → speech $\mathbf{s}$) and voice conversion (text $\mathbf{z}_\mathrm{m}$ + another speaker identity $\mathbf{z}_\mathrm{s}$ → speech $\mathbf{s}$). **We then aim to use the model for downstream tasks, such as audio-visual speech enhancement,** and expect the model competes well with models trained in a supervised manner.

3. 研究の方法

This research wanted **to elucidate whether our idea of human-inspired computational audio-visual information processing is achievable**. The simple model above may not be able to capture the complex relationship between the latent and observed variables because it is merely based on intuition about human capability in exploiting audio-visual inputs.

By having audio-visual speech enhancement and speaker diarization as the target downstream tasks, we planned to do research in the following order. We first wanted to develop *speech generative models*. Building upon the existing model, in which time-varying speech is generated by time-varying latent variables, we wanted to have time-invariant voice representation. We then wanted to develop *face- or lip-movement generative models*. The main challenge for all of these models is the disentanglement of the time-varying latent variables from the time-invariant ones. Finally, we wanted to have *joint speech and face- or lip-movement generative models*.

In FY2020, we developed a generative model of time-varying speech from disentangled representations of time-invariant speaker labels and time-varying phone labels. In addition to this model based on the variational autoencoder (VAE), we also developed another one based on a combination of VAE and normalizing flow (NF). See **Section 4(1)** for further description.

In FY2021, we improved blind source separation (BSS) techniques, which are useful for speech enhancement. We integrated normalizing flow (NF) into the state-of-the-art joint diagonalization techniques for spatial covariance matrices. See **Section 4(2)** for further description.

In FY2021 and FY2022, we worked on the visual aspect of audio-visual information processing. Learning from recently published works on lip-movement-informed speech enhancement, we decided to shift our focus to tackling issues in real-world scenarios where lip movement detection is often not reliable, e.g., because the target speaker is too far away from the camera or the image resolution is too low due to the camera hardware limitations. In this case, instead of lip movement, human faces or bodies can still be detected from the camera images to inform speech enhancement about the locations of the target speakers. See **Section 4(3)** for further description.

4. 研究成果

Due to space limitations, we can only briefly describe a few representative research results. We grouped the results into speech generative models, time-varying blind source separation, and adaptive audio-visual speech enhancement with smart glasses.

(1) Speech Generative Models

Du et al. (2020) proposed **a phone- and speaker-aware speech generative model based on variational autoencoder (VAE)** given three latent variables: (1) time-invariant speaker label, encoding the voice characteristic, (2) time-varying phone label, conveying the message, and (3) additional time-varying latent variables, holding other unspecified aspects in speech. The generated speech sounded natural with a good set of latent variables, and voice conversion could be done by modifying the speaker label. The model also worked well for speech separation outperforming the classical model without phone and speaker information. Additionally, *Nugraha et al.* (2020b) presented **another speech generative model based on a combination of VAE and the normalizing flow (NF)**. We showed that this novel model represents and produces better speech harmonics and improves a speech enhancement system utilizing it.

(2) Time-Varying Blind Source Separation Based on Normalizing Flow

Nugraha et al. (2020a) introduced **NF-IVA, a time-varying extension of independent vector analysis (IVA) based on the normalizing flow (NF) for determined BSS of multichannel audio signals**. It estimates demixing matrices that transform mixture spectra to source spectra in the complex-valued spatial domain such that the likelihood of those matrices for the mixture spectra is maximized under some non-Gaussian source model by gradient descent in an unsupervised manner. *Nugraha et al.* (2022) then proposed **NF-FastMNMF that integrates NF into the multichannel nonnegative matrix factorization with jointly-diagonalizable spatial covariance matrices, a.k.a. FastMNMF, for determined and non-determined BSS**. NF-FastMNMF successfully performed separations of multiple speech utterances by stationary or non-stationary speakers in noisy-reverberant environments.

(3) Adaptive Audio-Visual Speech Enhancement with Smart Glasses

The *dual process theory* in human cognition study postulates that human thought can arise from two different processes: a fast, intuitive process and a slow, deliberate process. **We adopt the concept to develop dual-process systems for audio-visual speech enhancement with augmented reality (AR) head-mounted display, a.k.a. smart glasses.** Using smart glasses equipped with cameras and microphones, sources (speakers) of interest can be identified from the camera images. We use this directional information to drive speech enhancement. The downstream task is a streaming automatic speech recognition (ASR) for cocktail-party conversation assistance that would benefit older adults and people with hearing impairment.

Nugraha et al. (2022) proposed **a practical dual-process visually-informed speech enhancement system that adapts environment-sensitive frame-online beamforming (front end) with help from environment-agnostic block-online BSS (back end)**. To use minimum variance distortionless response (MVDR) beamforming, one may train a deep neural network (DNN) that estimates time-frequency masks used for computing the covariance matrices of sources (speech and noise). Instead, one may try directly estimating the source covariance matrices with a BSS method, such as the state-of-the-art FastMNMF. In practice, however, neither the DNN nor the FastMNMF can be updated in a frame-online manner due to its computationally-expensive nature. Our DNN-free system, as shown in **Figure 1**, leverages the posteriors of the latest source spectrograms given by block-online FastMNMF to derive the current source covariance matrices for frame-online beamforming. This method is applied with a blind dereverberation method called weighted prediction error (WPE). The evaluation shows that our system can respond to scene changes due to interfering speaker movements and outperformed a system with DNN-based beamforming in terms of an ASR performance metric called word error rate (WER).

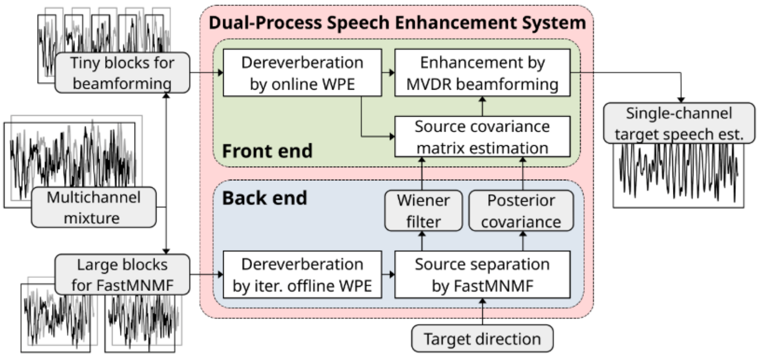

Dual-Process Speech Enhancement System
Tiny blocks for beamforming
Dereverberation by online WPE
Enhancement by MVDR beamforming
Source covariance matrix estimation
Front end
Single-channel target speech est.
Multichannel mixture
Back end
Wiener filter
Posterior covariance
Large blocks for FastMNMF
Dereverberation by iter. offline WPE
Source separation by FastMNMF
Target direction

Figure 1: *Dual-process, low-latency speech enhancement system with a frame-online front end (beamforming) informed by a block-online back end (FastMNMF).*

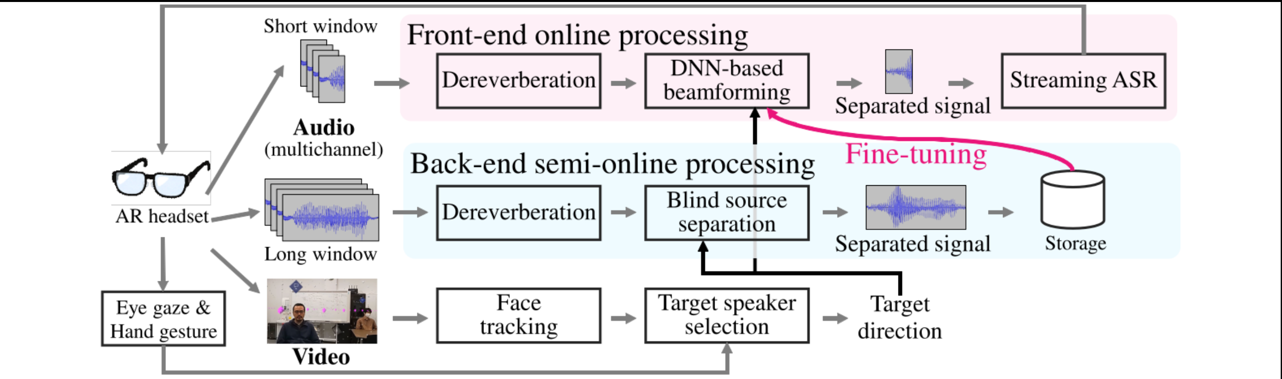
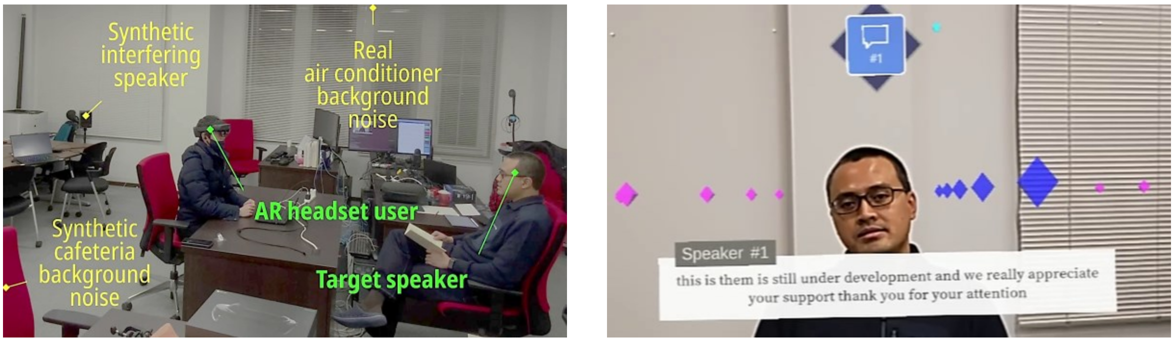


Figure 2: *Dual-process, low-latency speech enhancement system with a frame-online front end based on beamforming with a DNN-based mask estimator. The DNN is occasionally updated using the signals enhanced (separated) by a block-online back end based on FastMNMF.*

Sekiguchi et al. (2022) introduced another **dual-process visually-informed speech enhancement system that utilizes block-online FastMNMF (back end) for run-time adaptation of frame-online beamforming with DNN-based mask estimator (front end)**. FastMNMF works well in various environments thanks to its unsupervised nature. In contrast, beamforming that uses a pre-trained DNN for estimating spatial information of sources (speech and noise) suffers from drastic performance degradation in mismatched conditions. We thus propose speech enhancement based on DNN-based beamforming with FastMNMF-guided adaptation, as shown in **Figure 2**. Pairs of a noisy speech and its enhanced speech obtained by FastMNMF (back end) are used together with the original parallel training data for updating the DNN (front end) at a computationally-allowable interval. Our experiments showed that a run-time adaptation using only 12 minutes of observation could improve the WER by more than 10 points. By taking the idea of model adaptation further, *Du et al.* (2022) jointly fine-tuned the DNN-based mask estimator and ASR models to maximize the ASR objective function.

We developed **a prototype system using Microsoft HoloLens 2**, which was the only developer-friendly holographic smart glasses available in the market. Our system uses visual information from the camera images to provide possible target speakers from which the user can select using eye gaze or hand gestures. The speech enhancement front end extracts possibly multiple speech signals coming from the target directions corresponding to the user selection and input the extracted speech signals into an ASR system. The generated transcriptions are then displayed as virtual content on the AR headset, as shown in **Figure 3**.


Synthetic interfering speaker
Real air conditioner background noise
AR headset user
Synthetic cafeteria background noise
Target speaker
#1
Speaker #1

this is them is still under development and we really appreciate your support thank you for your attention

Figure 3: *We developed a prototype system with augmented reality smart glasses (Microsoft HoloLens 2). The left figure illustrates one of the evaluation setups and the right figure depicts the smart glasses user point of view showing some virtual markers and transcription of the enhanced speech.*

５．主な発表論文等

〔雑誌論文〕　計6件（うち査読付論文　6件／うち国際共著　6件／うちオープンアクセス　4件）

１．著者名 Fontaine Mathieu、Sekiguchi Kouhei、Nugraha Aditya Arie、Bando Yoshiaki、Yoshii Kazuyoshi	４．巻 30
２．論文標題 Generalized Fast Multichannel Nonnegative Matrix Factorization Based on Gaussian Scale Mixtures for Blind Source Separation	５．発行年 2022年
３．雑誌名 IEEE/ACM Transactions on Audio, Speech, and Language Processing	６．最初と最後の頁 1734～1748
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/TASLP.2022.3172631	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

１．著者名 Sekiguchi Kouhei、Bando Yoshiaki、Nugraha Aditya Arie、Fontaine Mathieu、Yoshii Kazuyoshi、Kawahara Tatsuya	４．巻 30
２．論文標題 Autoregressive Moving Average Jointly-Diagonalizable Spatial Covariance Analysis for Joint Source Separation and Dereverberation	５．発行年 2022年
３．雑誌名 IEEE/ACM Transactions on Audio, Speech, and Language Processing	６．最初と最後の頁 2368～2382
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/TASLP.2022.3190734	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Bando Yoshiaki、Sekiguchi Kouhei、Masuyama Yoshiki、Nugraha Aditya Arie、Fontaine Mathieu、Yoshii Kazuyoshi	４．巻 28
２．論文標題 Neural Full-Rank Spatial Covariance Analysis for Blind Source Separation	５．発行年 2021年
３．雑誌名 IEEE Signal Processing Letters	６．最初と最後の頁 1670～1674
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/LSP.2021.3101699	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Nugraha Aditya Arie、Sekiguchi Kouhei、Yoshii Kazuyoshi	４．巻 28
２．論文標題 A Flow-Based Deep Latent Variable Model for Speech Spectrogram Modeling and Enhancement	５．発行年 2020年
３．雑誌名 IEEE/ACM Transactions on Audio, Speech, and Language Processing	６．最初と最後の頁 1104～1117
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/TASLP.2020.2979603	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

1．著者名 Sekiguchi Kouhei、Bando Yoshiaki、Nugraha Aditya Arie、Yoshii Kazuyoshi、Kawahara Tatsuya	4．巻 28
2．論文標題 Fast Multichannel Nonnegative Matrix Factorization With Directivity-Aware Jointly-Diagonalizable Spatial Covariance Matrices for Blind Source Separation	5．発行年 2020年
3．雑誌名 IEEE/ACM Transactions on Audio, Speech, and Language Processing	6．最初と最後の頁 2610～2625
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/TASLP.2020.3019181	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

1．著者名 Nugraha Aditya Arie、Sekiguchi Kouhei、Fontaine Mathieu、Bando Yoshiaki、Yoshii Kazuyoshi	4．巻 27
2．論文標題 Flow-Based Independent Vector Analysis for Blind Source Separation	5．発行年 2020年
3．雑誌名 IEEE Signal Processing Letters	6．最初と最後の頁 2173～2177
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/LSP.2020.3039944	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

〔学会発表〕　計14件（うち招待講演　0件／うち国際学会　11件）

1．発表者名 Nugraha Aditya Arie、Sekiguchi Kouhei、Fontaine Mathieu、Bando Yoshiaki、Yoshii Kazuyoshi
2．発表標題 Flow-Based Fast Multichannel Nonnegative Matrix Factorization for Blind Source Separation
3．学会等名 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)（国際学会）
4．発表年 2022年

1．発表者名 Fontaine Mathieu、Di Carlo Diego、Sekiguchi Kouhei、Nugraha Aditya Arie、Bando Yoshiaki、Yoshii Kazuyoshi
2．発表標題 Elliptically Contoured Alpha-Stable Representation for MUSIC-Based Sound Source Localization
3．学会等名 European Signal Processing Conference (EUSIPCO)（国際学会）
4．発表年 2022年

１．発表者名
Sumura Yoshiaki、Sekiguchi Kouhei、Bando Yoshiaki、Nugraha Aditya Arie、Yoshii Kazuyoshi

２．発表標題
Joint Localization and Synchronization of Distributed Camera-Attached Microphone Arrays for Indoor Scene Analysis

３．学会等名
International Workshop on Acoustic Signal Enhancement (IWAENC)（国際学会）

４．発表年
2022年

１．発表者名
Nugraha Aditya Arie、Sekiguchi Kouhei、Fontaine Mathieu、Bando Yoshiaki、Yoshii Kazuyoshi

２．発表標題
DNN-Free Low-Latency Adaptive Speech Enhancement Based on Frame-Online Beamforming Powered by Block-Online FastMNMF

３．学会等名
International Workshop on Acoustic Signal Enhancement (IWAENC)（国際学会）

４．発表年
2022年

１．発表者名
Du Yicheng、Nugraha Aditya Arie、Sekiguchi Kouhei、Bando Yoshiaki、Fontaine Mathieu、Yoshii Kazuyoshi

２．発表標題
Direction-Aware Joint Adaptation of Neural Speech Enhancement and Recognition in Real Multiparty Conversational Environments

３．学会等名
Annual Conference of the International Speech Communication Association (Interspeech)（国際学会）

４．発表年
2022年

１．発表者名
Sekiguchi Kouhei、Nugraha Aditya Arie、Du Yicheng、Bando Yoshiaki、Fontaine Mathieu、Yoshii Kazuyoshi

２．発表標題
Direction-Aware Adaptive Online Neural Speech Enhancement with an Augmented Reality Headset in Real Noisy Conversational Environments

３．学会等名
IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)（国際学会）

４．発表年
2022年

1．発表者名

Fontaine Mathieu、Sekiguchi Kouhei、Nugraha Aditya Arie、Bando Yoshiaki、Yoshii Kazuyoshi

2．発表標題

Alpha-Stable Autoregressive Fast Multichannel Nonnegative Matrix Factorization for Joint Speech Enhancement and Dereverberation

3．学会等名

Annual Conference of the International Speech Communication Association (Interspeech)（国際学会）

4．発表年

2021年

1．発表者名

Sekiguchi Kouhei、Bando Yoshiaki、Nugraha Aditya Arie、Fontaine Mathieu、Yoshii Kazuyoshi

2．発表標題

Autoregressive Fast Multichannel Nonnegative Matrix Factorization For Joint Blind Source Separation And Dereverberation

3．学会等名

IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)（国際学会）

4．発表年

2021年

1．発表者名

Nugraha Aditya Arie、Sekiguchi Kouhei、Fontaine Mathieu、Bando Yoshiaki、Yoshii Kazuyoshi

2．発表標題

Determined Blind Source Separation Based on NF-IVA with Time-Varying Linear Transformations

3．学会等名

The Spring Meeting of the Acoustical Society of Japan (ASJ)

4．発表年

2021年

1．発表者名

Sekiguchi Kouhei、Bando Yoshiaki、Nugraha Aditya Arie、Fontaine Mathieu、Yoshii Kazuyoshi

2．発表標題

Joint Blind Source Separation and Dereverberation Based on ARMA-FastMNMF

3．学会等名

The Spring Meeting of the Acoustical Society of Japan (ASJ)

4．発表年

2021年

１．発表者名
Nugraha Aditya Arie、Sekiguchi Kouhei、Fontaine Mathieu、Bando Yoshiaki、Yoshii Kazuyoshi

２．発表標題
Unsupervised Source Separation with Deep Spatial Models

３．学会等名
RIKEN-AIP Open Seminar

４．発表年
2021年

１．発表者名
Fontaine Mathieu、Sekiguchi Kouhei、Nugraha Aditya Arie、Yoshii Kazuyoshi

２．発表標題
Unsupervised Robust Speech Enhancement Based on Alpha-Stable Fast Multichannel Nonnegative Matrix Factorization

３．学会等名
Annual Conference of the International Speech Communication Association (Interspeech)（国際学会）

４．発表年
2020年

１．発表者名
Yoshii Kazuyoshi、Sekiguchi Kouhei、Bando Yoshiaki、Fontaine Mathieu、Nugraha Aditya Arie

２．発表標題
Fast Multichannel Correlated Tensor Factorization for Blind Source Separation

３．学会等名
European Signal Processing Conference (EUSIPCO)（国際学会）

４．発表年
2020年

１．発表者名
Du Yicheng、Sekiguchi Kouhei、Bando Yoshiaki、Nugraha Aditya Arie、Fontaine Mathieu、Yoshii Kazuyoshi、Kawahara Tatsuya

２．発表標題
Semi-supervised Multichannel Speech Separation Based on a Phone- and Speaker-Aware Deep Generative Model of Speech Spectrograms

３．学会等名
European Signal Processing Conference (EUSIPCO)（国際学会）

４．発表年
2020年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

Demo web page for NF-FastMNMF: https://aanugraha.github.io/demo/nffastmnmf/
Demo web page for Neural FCA: https://ybando.jp/projects/spl2021/
Demo web page for NF-IVA: https://aanugraha.github.io/demo/nfiva/
Demo web page for GF-VAE: https://aanugraha.github.io/demo/gfvae/

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関