

[研究助成 (S)]

スモールデータ機械学習に基づく リアルタイム音コミュニケーション能力拡張システム

2238001



研究代表者
(助成金受領者)

共同研究者

東京大学
大学院情報理工学系研究科

慶應義塾大学 理工学部情報工学科

東京大学
大学院情報理工学系研究科

教授

准教授

講師

猿 渡 洋

高 道 慎之介

齋 藤 佑 樹

[研究の目的]

本研究では、スモールデータ機械学習に基づく音情報処理及びその音バーチャルリアリティ(VR)・拡張現実感(AR)システムへの展開を通じて、人間の音コミュニケーション能力拡張の実現を主たる目的とする。具体的には、「なるべく少ない事前情報から複雑な音情景を独立成分に分解し、加工・拡張再現する」という自律的音入出力システムの開発を基盤とし、このシステムの実証的アプリケーションとして「音メディア VR・AR」を開発する。両耳補聴器やスマホ等の不特定多数の音波動センサが一致団結してユーザの受聴を助ける音コミュニケーション能力拡張システムの実現を通じて、人間と機械の融和によるライフイノベーションへ貢献する。

本研究の基盤技術は、近年のビッグデータ機械学習では不可能であった「スモールデータ(事前データが全く用意できない・非常に少ない)」における新しい数理工学・情報処理の発展に関するものであり、データ利活用で遅れをとった本邦における学術的な貢献が大きい。またライフイノベーション技術の側面から眺めた場合、視力に関しては眼鏡という補助機器が普及しているのに対し、聴覚に関してはその支援技術の開発は大変遅れているという現実がある(例えば 2025 年度において補聴器の利用率は約

16%に留まる)。本研究で提唱するシステムは、従来の専門家を介したハイコストなオーダーメイド装置(従来補聴器等)を大きく改善する技術であると言える。本研究では、新しい形態「不特定多数の音波動センサが一致団結してユーザの受聴能力拡張へ貢献する自律カスタムメイドかつ協創型システム」を最新の数理工学及びIoT 情報処理を駆使して実現する点に大きな社会貢献がある。また健聴者においても、音 VR・AR により対話対象固有の声質を加工して、明瞭性補償のみならず好みの人格への変更(安心声の実現)・感情の平滑化(受聴者ストレス軽減)等により、カスタムメイドかつ安心な新しい対話コミュニケーション能力拡張が人類にもたらされる。

[研究の内容、成果]

1. 開発システムの概要

提案する聴覚能力拡張システムは大きく分けて2つの要素(1)音情景分解処理、(2)音情景拡張・合成 VR・AR 処理から成る(システム概念図は図1 参照)。(1)では、申請者らが長年研究してきた音源分離理論である「独立低ランク行列分析(ILRMA)」等を活用し、両耳補聴器やスマホマイク等の位置不定センサ群で観測された多次元信号の中から統計的に独立な音源をリアルタイムで抽出する。次に(2)で

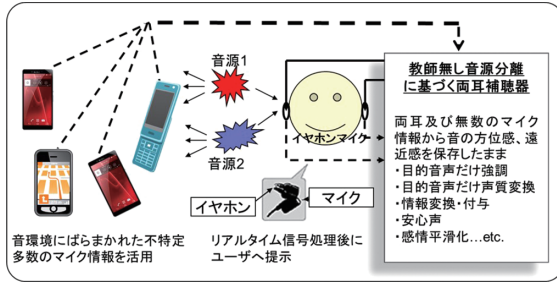


図1 音コミュニケーション能力拡張システムの概観。

はリアルタイム DNN 声質変換処理を基盤とし、(1) で推定された目的音声の話者性を自由に変換する。この統合によりリアルタイム音コミュニケーションに適用可能な聴覚能力拡張システムが実現できる。次章以降において、開発した具体的な処理系の説明を行う。

2. 音情景分解処理 1 (教師無し音源分離)

ILRMA を含む一般的なブラインド音源分離アルゴリズムでは、全ての音源を個別の点音源生成モデルで表現し、それらの独立性に着目して最適な分離系を推定する。しかしより一般的な環境を考慮する場合、特に音声以外の環境音(騒音・BGM・生活音等)は点音源では表現できず、かつそれらを個別に扱うのは効率的ではない。よってこれを「音声+それ以外」の2音源群モデルへ拡張する。本研究では、前記の機能を実現するため、著者らが提案していたランク制約付き空間共分散行列推定法(RCSCME) [Kubo+, IEEE Trans. ASLP 2020] を用いる。図2に基本となる音源分離手法の概観を示す。

この音源分離処理を音コミュニケーション能力拡張システムに導入する場合、人間のコミュニケーション処理の即時性より、リアルタイム信号処理系(入力から数十～百 ms 程度の遅延を許容した上ですぐに分離出力が出ること)として実装することが必須である。ここで図2の信号処理概観図を見ると、以下の特徴がある事が分かる。

- (a) 前段の ILRMA 部は音源の統計的独立性を測るためある程度長いデータバッファが必

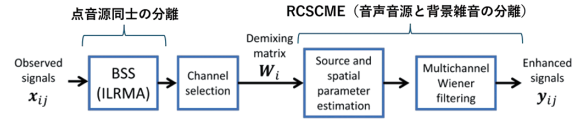


図2 音源分離手法概観。前段では ILRMA を用いてランクの欠落した雑音推定及び目的音声の情報を推定し、後段の高速パラメータ推定部分(RCSCME)において背景雑音を除去する。

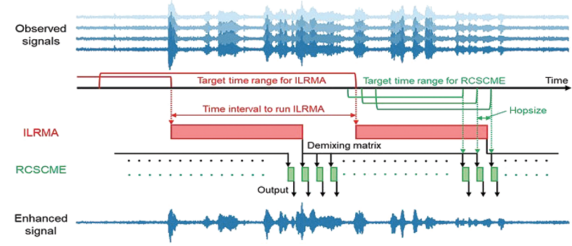


図3 リアルタイム音源分離処理フロー。前段の ILRMA は最適化に時間がかかるので長い周期で更新し、後段の RCSCME は短時間フレームで処理することによりリアルタイム音源分離を実現する。

要となり、その最適化に時間がかかる。つまり、リアルタイム処理化におけるボトルネックであると言える。

- (b) 後段の RCSCME 部は、前段の ILRMA で推定された特徴量を基に欠落している情報を再推定するものであり、高速に稼働することが出来る。

これらの特徴に鑑み、前段の ILRMA 部はある程度長い周期でゆっくり更新し、それを補うため後段の RCSCME は短時間(即時)時間フレームで処理するという2層構造の信号処理フローを提案した(図3参照) [Ishikawa+, IEEE Access 2025]。この構造を導入することにより、リアルタイム音源分離を実現する。またボトルネックとなる前段 ILRMA 部において、その数値演算を高速もしくは数値誤差安定に実行するアルゴリズムも独自に開発した。このリアルタイム教師無し音源分離アルゴリズムを補聴器模擬システム(図4参照)へ導入し、より現実的な状況における音声強調性能評価を行った。

実装した補聴器模擬リアルタイム音源分離処理は、合計8個のマイクロホンから構成される(両耳に各3マイク+不特定位置スマホ内蔵2

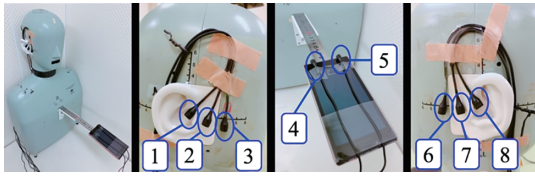


図4 補聴器を模擬した音源分離システム概観。頭部・胴体とマイクロホン配置を示す。ここで4・5番マイクロホンは手元の補助デバイス（スマホ等）を模擬している。

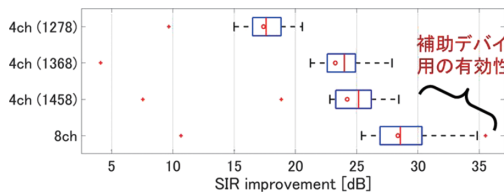


図5 リアルタイム目的音声抽出性能結果。特に4・5番マイク（手元の補助デバイスを模擬）の併用が効果的である事が分かる。

マイク)。なお、マイクロホン位置座標は使用しないので位置不定マイクロホンアレーとみなすことが出来る。そこで得られた最大8チャネルの入力信号において、前述のリアルタイム音源分離処理を実行した。音響環境としては、補聴器ユーザの正面から1～1.5 m離れたところに居る1人の話者が発話しており、その周辺で不特定多数の会話騒音が発生していると仮定した。観測音における入力SN比は0 dBである。

リアルタイム音源分離処理は汎用PC上のGPUもしくはCPUで計算されている。その遅延時間を計測したところ、GPUを使用した場合は4マイク使用時に40 ms、8マイク使用時に80 ms程度であった。これは人間の知覚の観点から見ても、十分に速いリアルタイム処理系であると言える。一方CPUを使用した場合、4マイク使用時にはリアルタイムで動作するものの、8マイク使用時はリアルタイム動作が不可能であることが分かった。目的音声抽出性能結果を図5に示す。ここではSignal-to-Interference (SIR) 比を示している。4マイク使用時と8マイク使用時を比較すると、明らかに8マイク使用時の性能が高い。また、同じ4マイク使用時においても、両耳付近のマイク群

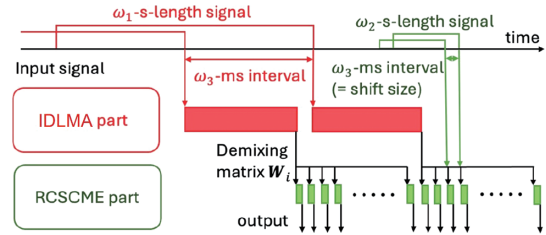


図6 リアルタイム半教師有り音源分離の概観。

を使用するだけでなく、補助デバイス（手元のスマホ）に付設されたマイクロホンも併用した場合が最も音声抽出性能が高く、位置不定マイクロホン群を活用する提案システムの有効性が確認された。

3. 音情景分解処理2（半教師有り音源分離）

前述の教師無し音源分離では、大量の音響データベースを使って事前にモデルを学習する必要がないという利点がある。一方、近年においては、典型的な雑音下において音声信号を推定する学習済みDNNは比較的容易に入手可能である。そこで教師無しILRMAの音源モデル部を音声推定用DNN（Denoiser）に置き換えた半教師有り音源分離「独立深層学習行列分析（IDLMA）[Makishima+, IEEE TASLP 2019]」をリアルタイム拡張して実装した。このリアルタイムIDLMAにRCSCMEを接続したシステムの評価を行った。

実装したリアルタイム半教師有り音源分離の概観を図6に示す。ここでは前述の教師無しシステムと同様、前段のIDLMAは最適化に時間がかかるので長い周期で更新し、後段のRCSCMEは短時間フレームで処理することによりリアルタイム音源分離を実現している。また、IDLMAにおいて使用される音声推定用DNNは、音声のスペクトログラムを推定するだけでなく、音声の無音区間も提供する。これにより、雑音空間相関行列を自己推定し、その情報を雑音事前分布として後段のRCSCMEに用いることとする。

図4に示す補聴器模擬システムを用いて半教

師有り音源分離処理の評価を行った。まず遅延時間を計測したところ、GPUを用いた場合においては、前章で述べた教師無し音源分離処理系とほぼ同等となり十分なリアルタイム性が確保された。一方で、CPUを用いた場合において、DNNが高速に推論できなくなるため、リアルタイム動作は確認できなかった。目的音声抽出性能を測定したところ、通常のリアルタイムILRMAと比べて約2 dBほどのSIR改善がみられた。ただし、前章で述べたリアルタイムILRMAは音源方位に関する事前分布を導入しており、さらに性能が高かった。これらの結果をまとめると、以下のようになる。

(1) 事前学習済みDNNを用いる半教師有り音源分離は、目的音声のスペクトログラム情報を確率モデルとして保持するものの、空間的な情報は持っていない。一方、雑音に関しては、無音区間推定によってその空間相関行列（空間情報）を自己推定してRCSCME部の事前分布を構成する。(2) 一方で先に開発した教師無し音源分離は、目的音声の空間的事前分布を活用しており、その方が分離性能に与える影響が大きい。つまり、この「目的音の空間情報」の存在が前記の事前学習されたスペクトログラムモデルよりも強力に働くと考えられる。

4. 音情景拡張・合成VR・AR処理および統合システム評価

本研究で目指す補聴器等の音コミュニケーション能力拡張システムにおいては、前段の音情景分解処理のみならず、対話コミュニケーションの即時性よりリアルタイム声質変換処理の開発・高精度化は必須となる。本研究ではこれを音メディアVR・ARレイヤーととらえ、前章までに述べた音情景分解処理（センシング・解析レイヤー）と統合し、最終的な統合システム評価を行った（図7参照）。

本研究では、声質変換を行うDNNモデルとして、著者らの技術をベースに開発されたParavo model v2を基礎にし、JVS音声コーパ

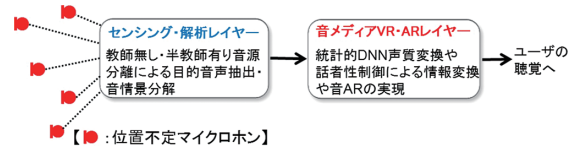


図7 音コミュニケーション能力拡張システムにおける2レイヤー構造。

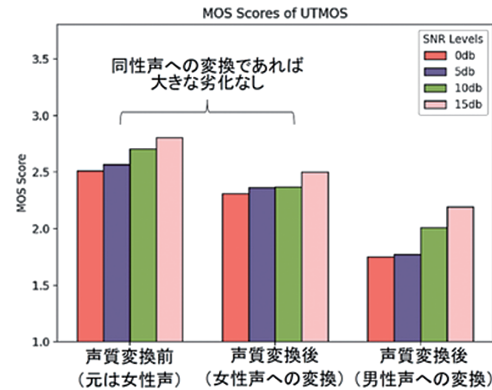


図8 UTMOSによる統合システムの品質評価結果。

ス（51名話者）を用いて話者の多様性を拡大したものをを用いた。このParavo声質変換はCPU上で動作し、遅延は60 ms以下である。よって、前段の音源分離処理系（遅延は40 ms）と合わせても100 ms以下の遅延でユーザーに音響拡張された音を提示することができる。この遅延量は人間の聴覚においてはほぼ許容できる量であり、リアルタイム性は十分に確保されている。

統合システム評価として、まず初めに、音質評価を行った。ここでは、前段の音情景分解として2章で述べた教師無し処理系を用いた。ここで抽出された目的音声に対してDNN声質変換を施し、その変換音声の品質をUTMOSによって数値化した。このUTMOSはMean Opinion Score（MOS：5段階主観評価）値を自動で算出するものであり、人間の聴覚印象に近い特性を持つ。元音声は若年女性声であるが、それを様々な女性声もしくは男性声に変換するタスクを採用した。入力SN比別の実験結果を図8に示す。

図8より、(1) 単純に音源分離を施した音声は2.5以上のMOS値となっており十分な品質が確保されている、(2) 同性への変換でもおお

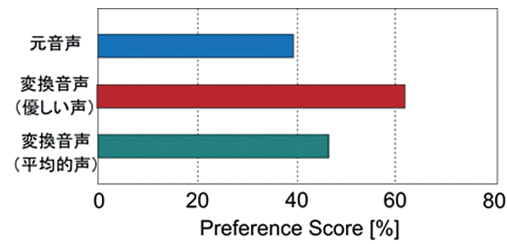


図9 「優しい声」への人格変更 AR に関する評価。

むね同等の値となっており劣化は少ない，(3) 異性への変換は劣化が見られるが特に入力 SN 比が高い場合にはそれほど問題ではない，という事が分かる。

次に VR・AR 的な拡張として，好みの人格への変更（安心声の実現・受聴者ストレス軽減）を想定し，51 変換話者のうち「優しさ」素性の一番高い話者への変換を評価した。ここでは「優しさ」に着目した二対比較による Preference Score を算出した。図9に評価結果を示す。本結果より，元音声と比べ20%以上好まれる音声に音響拡張できた事が分かる。よって，本統合システムは，目的音声を強調するという単純な補聴器機能だけでなく，ユーザの聴覚ストレス軽減機能も有する音メディア AR システムになっている事が示された。

[今後の研究の方向，課題]

本研究では，音コミュニケーション能力拡張の実現を目指し，リアルタイム性を有した音情景分解・声質変換処理を統合したシステムを提案した。今後の課題としては，(1) 更なる遅延時間の短縮とエッジデバイスで利用できる軽量 CPU 上での動作保証，(2) さらに大規模な位置不定センサ群の統合，(3) 感情まで含めた柔軟な声質制御の実現，等が挙げられる。

[成果の発表，論文など]

・原著論文

- [1] Y. Ishikawa, T. Nakamura, N. Takamune, D. Kitamura, H. Saruwatari, Y. Takahashi, K. Kondo, "Real-time speech extraction based on rank-constrained spatial covariance matrix estimation and spatially regularized independent low-rank matrix analysis with fast demixing matrix estimation," IEEE Access, vol. 13, pp. 88683-88706, 2025.
- [2] K. Seki, N. Ito, K. Yamauchi, Y. Okamoto, K. Yamaoka, Y. Saito, S. Takamichi, H. Saruwatari, "Language-queried target speech extraction using para-linguistic and non-linguistic prompts," Acoustical Science and Technology, vol. 46, no. 6, pp. 671-675, 2025.
- [3] K. Imamura, T. Nakamura, N. Takamune, K. Yatabe, H. Saruwatari, "Stride conversion algorithms for convolutional layers and its application to sampling-frequency-independent deep neural networks," Signal Processing, vol. 242, no. 110420, 12 pages, 2026.

・国際会議発表（全て査読有り）

- [1] K. Seki, S. Takamichi, N. Takamune, Y. Saito, K. Imamura, H. Saruwatari, "Spatial voice conversion: voice conversion preserving spatial information and non-target signals," Proc. Interspeech, 5 pages, 2024.
- [2] Y. Ishikawa, N. Takamune, K. Yamaoka, H. Saruwatari, "Theoretical formulation of online independent vector analysis using framewise probabilistic generative model," Proc. EUSIPCO, 5 pages, 2025.
- [3] Y. Haruyama, K. Yamaoka, N. Takamune, H. Saruwatari, "Decentralized independent low-rank matrix analysis for blind source separation on distributed microphone arrays," Proc. EUSIPCO, 5 pages, 2026.

他 23 件

・国内発表（査読無し）

合計 21 件

・受賞

- [1] 石川，日本音響学会学生優秀発表賞，2024 年
- [2] 小川，日本音響学会学生優秀発表賞，2024 年
- [3] 石川，IEEE Signal Processing Society Tokyo Joint Chapter Student Award，2024 年
- [4] 小川，応用音響研究会学生研究奨励賞，2026 年