人間と情報環境の共生インタラクション基盤技術の創出と展開　　2022 年度
2019 年度採択研究代表者　　年次報告書

戸田 智基

名古屋大学 情報基盤センター
教授

# 音メディアコミュニケーションにおける共創型機能拡張技術の創出

主たる共同研究者:
小野 順貴（東京都立大学 システムデザイン学部 教授）
亀岡 弘和（日本電信電話(株) NTTコミュニケーション科学基礎研究所 特別研究員）

## 研究成果の概要

ユーザとシステム間の質の高い即時インタラクションの実現に向け、低遅延かつ高精度なリアルタイム処理を可能とする基盤技術の研究を継続しつつ、発声・聴覚機能拡張プロトタイプの構築に取り組んだ。

発声機能拡張グループ（名大　戸田）では、音声変換処理の高精度化・高速化に継続して取り組むとともに、発声支援・機能拡張プロトタイプ開発に取り組んだ。系列ベース深層音声変換による発声支援技術、低遅延リアルタイム深層音声変換による発声機能拡張技術、物理的制約を加味した深層音声波形生成による歌唱表情制御技術、実環境応用に向けた各種基盤技術、国際チャレンジによる研究分野の活性化など、多くの技術的進展が得られた。

聴覚機能拡張グループ（都立大　小野）では、選択的聴取を可能とする低遅延リアルタイム音響信号処理基盤技術を中心に研究に取り組んだ。その結果、リアルタイム音源分離の新たな効率的アルゴリズムの開発、深層学習を用いた音源分離の高精度化、頭部回転への頑健性の向上、低演算量化のための次元削減技術、end-to-end での音声認識応用における最適化、「ブリンキー」による音光変換を用いたスパーススペクトルセンシングなど、多様な成果が得られた。

機械学習基盤グループ（NTT　亀岡）では、物理パラメータや画像などによる直感的な音声制御を可能にする音声変換の実現に向け、声質／韻律特徴の個別変換を可能にする深層音声変換モデルの解きほぐし学習法、音声と映像を併用したマルチモーダル音声スタイル変換法を開発した。また、高速かつ高品質な音声変換の実現に向け、低遅延リアルタイム系列ベース深層音声変換法、FFT 演算の効率性を活かした深層音声波形生成法を開発した。加えて、深層生成モデルを用いた多チャンネル音源分離手法も開発した。

2022 年 12 月に中間シンポジウムを開催し、発声・聴覚機能拡張の各種プロトタイプの体験型デモ展示を行った。

【代表的な原著論文情報】


1) W.-C. Huang, S.-W. Yang, T. Hayashi, T. Toda, "A comparative study of self-supervised speech representation based voice conversion," IEEE Journal of Selected Topics in Signal Processing, Vol. 16, No. 6, pp. 1308-1318, 2022.
2) W.-C. Huang, E. Cooper, Y. Tsao, H.-M. Wang, T. Toda, J. Yamagishi, "The VoiceMOS Challenge 2022," Proc. INTERSPEECH, pp. 4536-4540, 2022.
3) Y. Masuyama, X. Chang, S. Cornell, S. Watanabe, N. Ono, "End-to-end integration of speech recognition, dereverberation, beamforming, and self-supervised learning representation," Proc. IEEE SLT, pp. 260-265, 2023.【IEEE SLT 2023 Best Student Paper Award】
4) L. Li, H. Kameoka, S. Makino, "FastMVAE2: on improving and accelerating the fast variational autoencoder-based source separation algorithm for determined mixtures," IEEE/ACM Transactions on Audio, Speech and Language Processing, Vol. 31, pp. 96-110, 2022.
5) H. Kameoka, T. Kaneko, S. Seki, K. Tanaka, "CAUSE: Crossmodal action unit sequence estimation from speech with application to facial animation synthesis," Proc. INTERSPEECH, pp. 506-510, 2022.