音響情景理解チーム　吉井 和佳
Sound Scene Understanding Team
Kazuyoshi Yoshii

チームの目標：実世界における音環境理解

課題：実環境の音源・残響・雑音は未知かつ多様性に富む → 遠隔では限定された種類のコマンド音声認識が限界

目標：**あらゆる環境**に適応し**あらゆる音響信号**（音声・音楽・環境音・雑音）を対象とした音環境理解技術の開発

音源分離・音源定位・認識・雑音抑圧・残響除去・マイクキャリブレーションなどを自動的に実行・同時最適化

アプローチ：**深層ベイズ学習の新しい枠組みと音環境理解への応用展開**

言語モデル：音声スペクトルの深層生成モデル（事前学習）

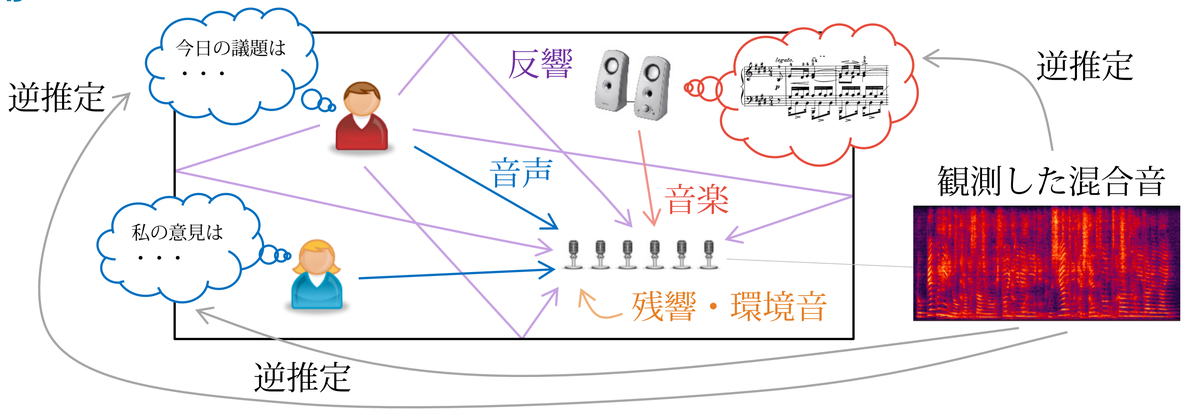
音響モデル：音響信号の重畳・伝播モデル（on-the-fly 学習）

現在の成果：**オンライン環境認識に必要となるコア技術を開発**

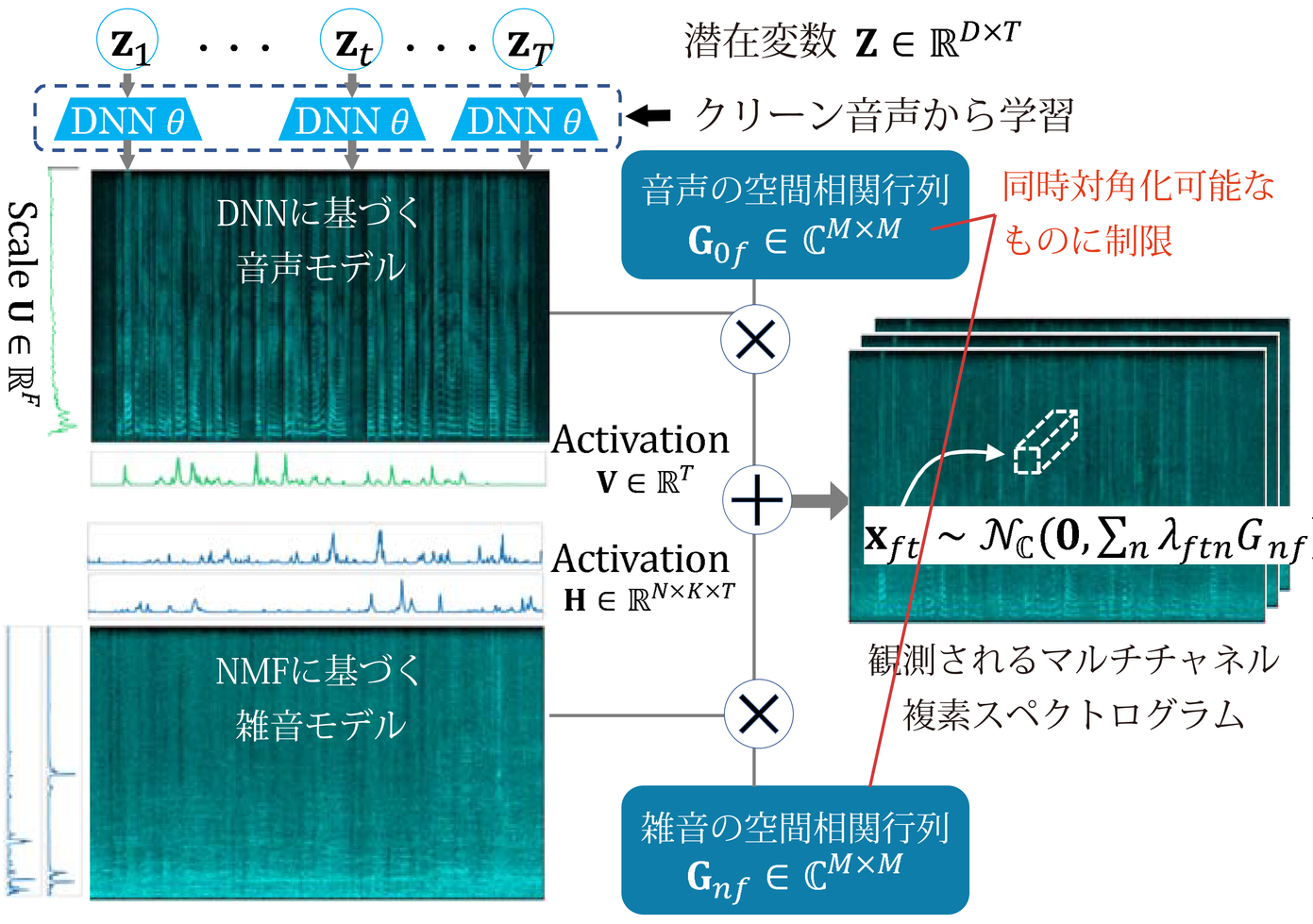
高速多チャネル非負値行列分解 (FastMNMF) を考案

→ 汎用 BSS 法として非常に精度が高い・実環境でよく動く

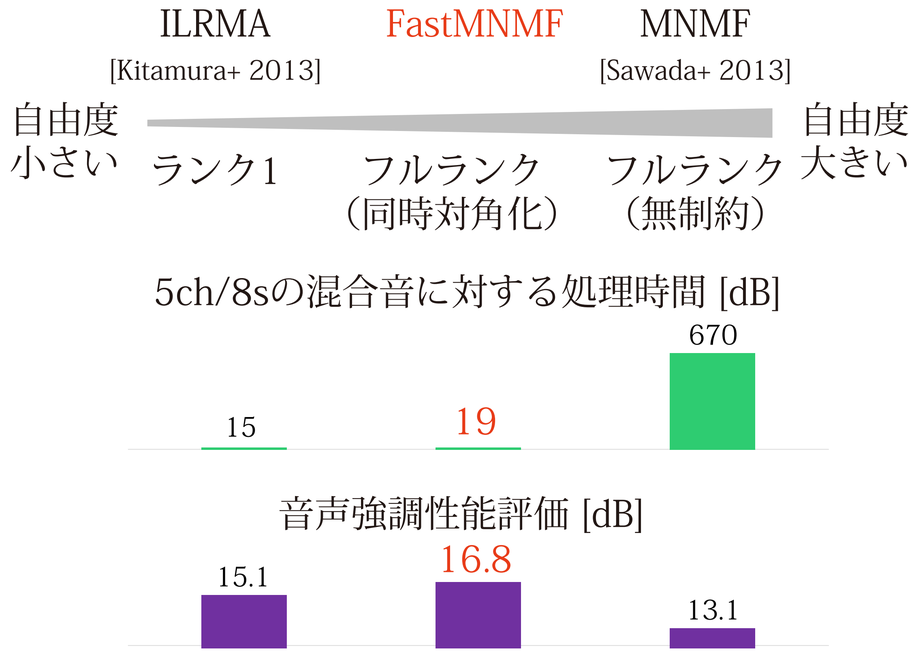
屋内ロボットのための視聴覚統合 SLAM 法を開発



深層生成モデルに基づく半教師ありマルチチャネル音声強調

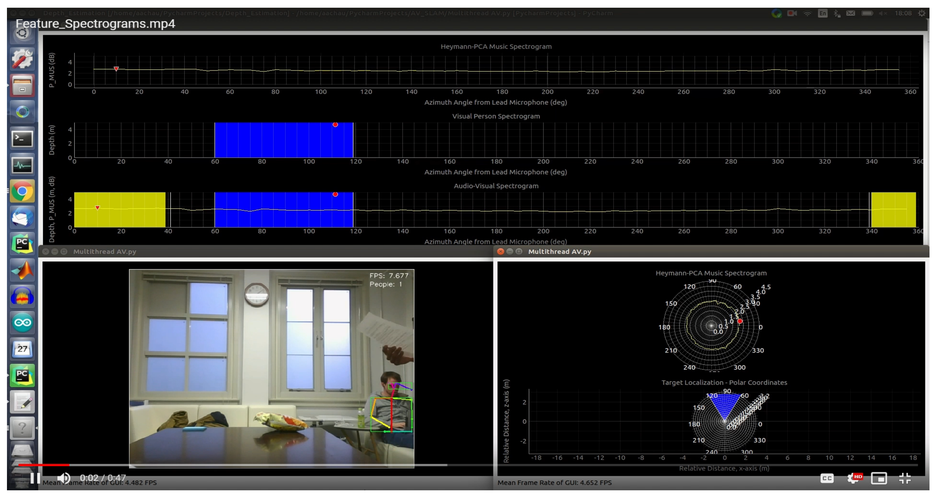


高速かつ精度の高い汎用 BSS 手法を提案
(実用上、現時点でベストな選択)



NMFに基づく音源モデル → DNNに基づく音源モデルに
置き換えれば 18.8 [dB] まで性能が向上

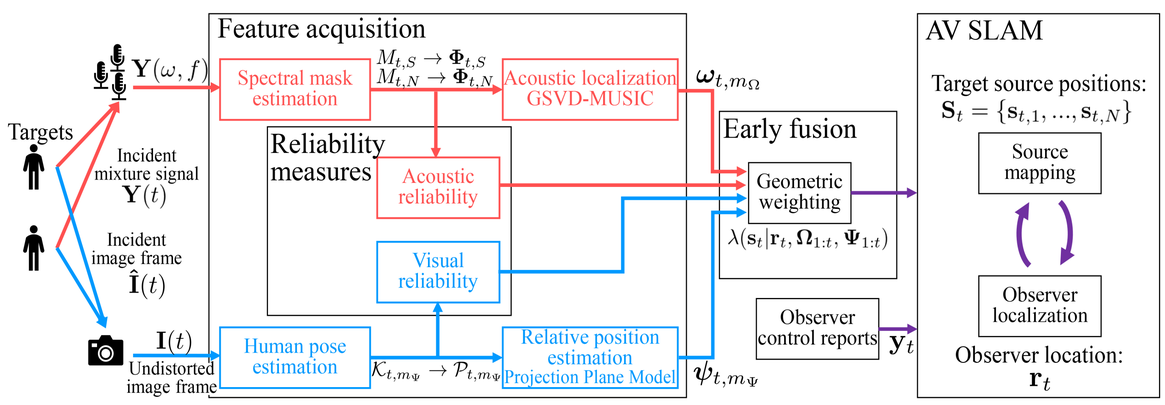
状態空間モデルに基づく視聴覚統合人物追跡



屋内ロボット（ドローン）を想定した SLAM 技術を開発

視聴覚状態空間モデルを用いた仮説統合とリアルタイム追跡

1. 視覚情報の解析：OpenPoseを用いて人物位置の仮説を生成
2. 聴覚情報の解析：Music法を用いて音源位置の仮説を生成
3. Early Fusion: 幾何的な制約を考慮して仮説を統合
4. Late Fusion: 自らの動作情報と時間的連続性を考慮しつつ、
   自己位置と人物位置を同時推定



A. Chau, K. Sekiguchi, A. A. Nugraha, K. Yoshii, K. Funakoshi, “Audio-Visual SLAM towards Human Tracking and Human-Robot Interaction in Indoor Environments,” IEEE International Conference on Robot and Human Interactive Communication (ROMAN), pp. 1–8, October 2019. **Best Conference Paper Award**

今後の展開：多人数会話分析への応用

ミーティングの自動書き起こし/可視化/要約システム

高齢者会話を対象とした認知機能の分析と改善支援

ヒューマノイドロボットのための視聴覚環境理解フロントエンド