

多チャネル時系列信号に対する ブラインド信号源推定

分離精度と計算速度を両立した最先端の汎用手法

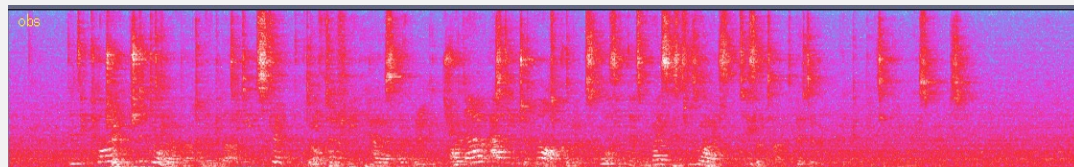
吉井 和佳

理研 革新知能統合研究センター (AIP)

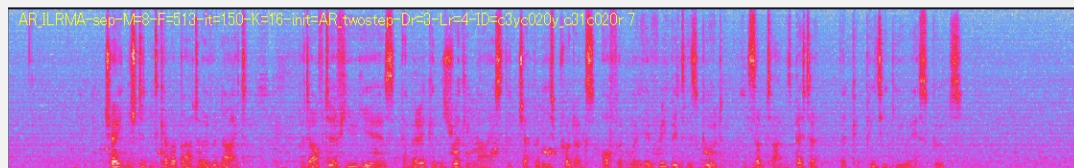
kazuyoshi.yoshii@riken.jp

音声強調・雑音除去の同時実行

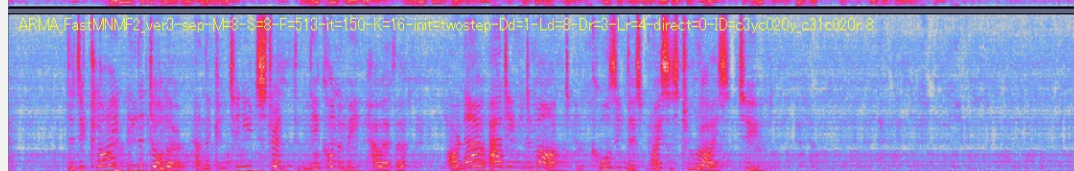
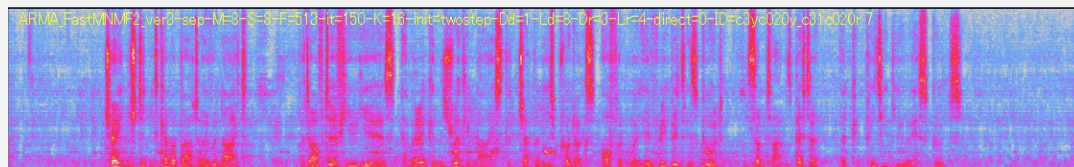
雑音・残響を含む
二話者の混合音



AR-ILRMA
(従来法)

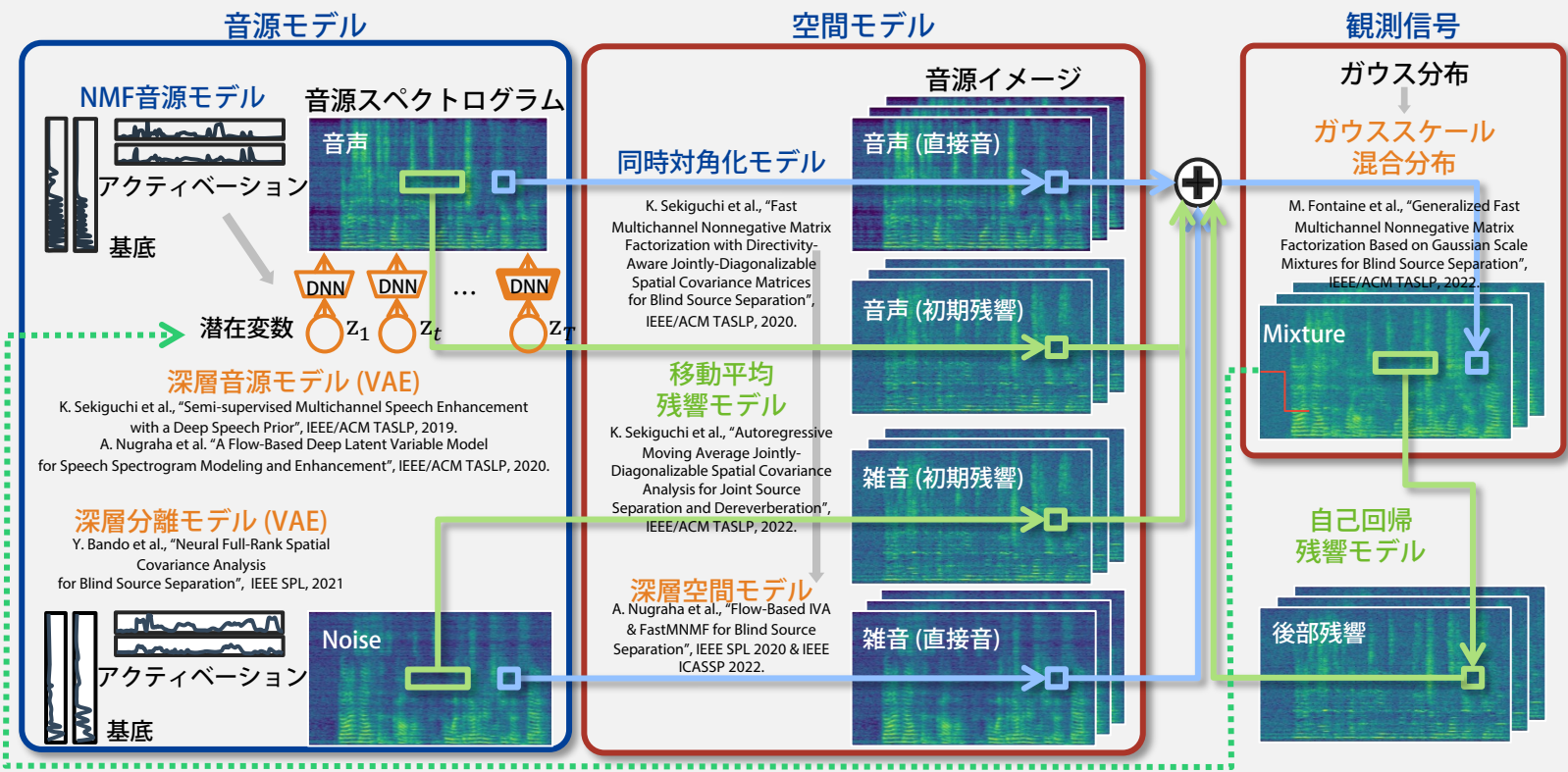


ARMA-FastMNMF
(提案法)



保有技術

- 世界最先端のブラインド音源分離・残響/雑音除去技術
- 統一的な確率的定式化・深層学習との融合 (多数のIEEE TASLP発表)



目標：視聴覚環境理解支援

- ・ スマートグラスを用いた人の知覚能力 (視覚/聴覚) の拡張
 - ・ 人の認知能力 (記憶/理解) の拡張も視野
 - ・ 注目話者を強調・認識・翻訳
 - ・ 接近する音源や物体を検知・注意喚起

システムの存在を意識せず
あたかも自身の能力で
知覚・認知している感覚



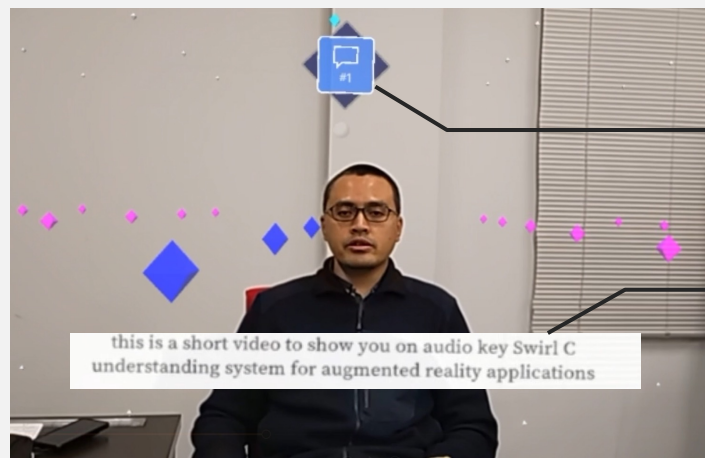
日常生活における透明なパートナー

- スマートグラスを用いた音声コミュニケーション支援
 - 現実の雑音環境下で目的話者のみの音声を強調・認識
 - キーコンセプト：使えば使うほど賢くなる
 - 「正解」が得られない実環境で破綻せずに動作することが重要
 - ユーザ・環境に対する教師なしオンライン適応

RGBカメラ・マイク5個を搭載



Microsoft HoloLens 2



目的話者を
示すマーカ

認識結果

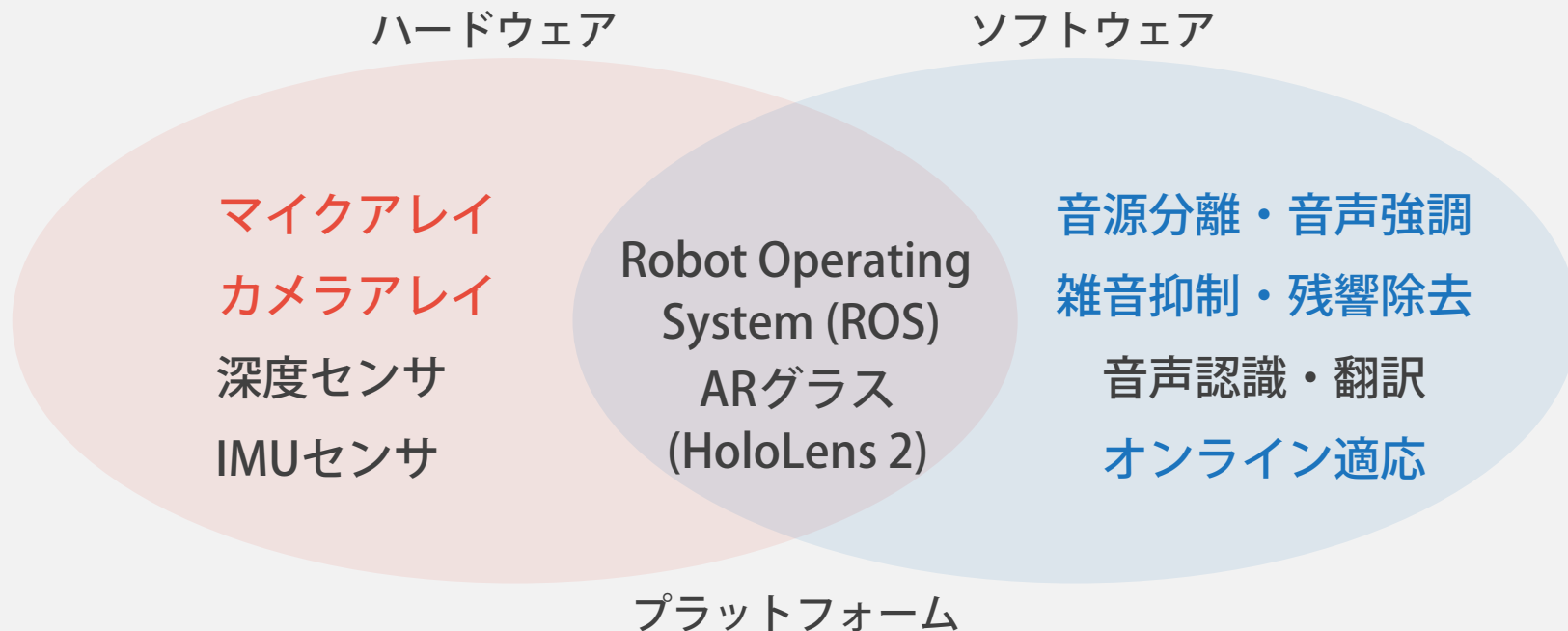
ARディスプレイの表示

リアルタイム音声強調・認識



ウェアラブルデバイスの未来

- ハードとソフトの融合によるスマート化・透明化
 - 環境「理解」に向けてはLLM技術の活用も重要
 - 人と計算機は以心伝心になれるか？ LLMが介入？

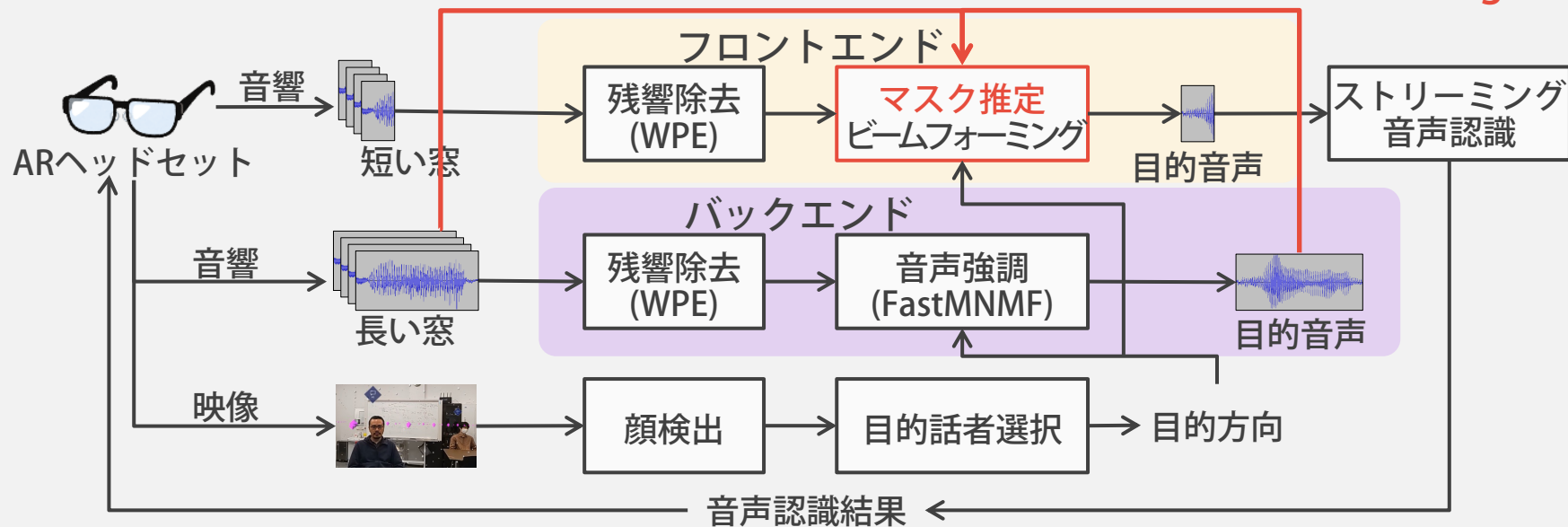


FastMNMF + Neural Beamforming

[Sekiguchi+
IROS2022]

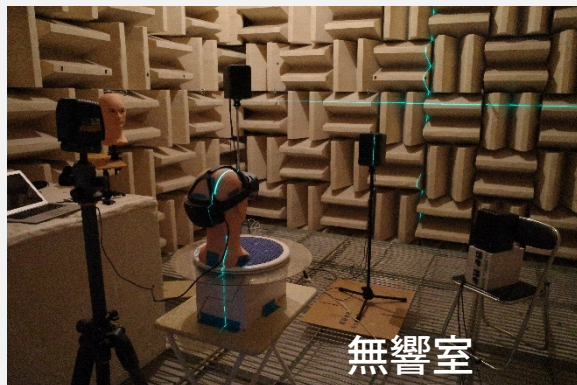
- 二重プロセスに基づくオンライン音声強調
 - フロントエンド：DNNマスク推定 + MVDRビームフォーミング
 - Teacher-Student学習に基づくオンライン適応
 - バックエンド：長いセグメントに対するブラインド音声強調
 - 環境に依存せず安定した性能

音声強調部のFine-tuning



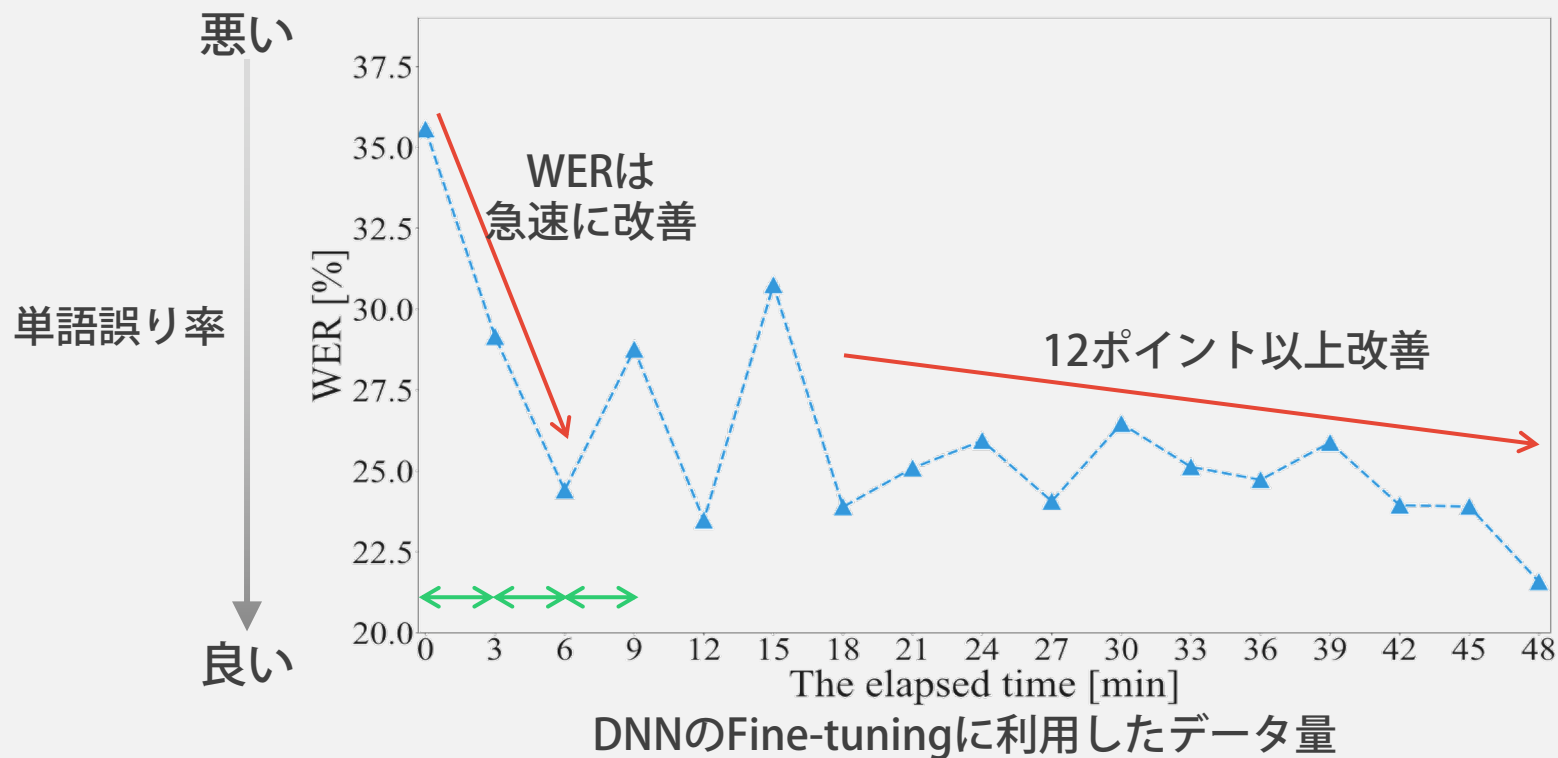
実験環境

- マスク推定DNNのオンライン適応効果を評価
- 学習データ
 - 音声 x 2 + 雑音
 - 音声：クリーン音声に無響室のインパルス応答を畳み込み
 - 雑音：36個の音源を配置してシミュレーション
- DNNの学習：15時間
- テストデータ
 - 音声 + 雑音
 - 音声：ダミーヘッドで録音
 - 雑音：同様にシミュレーション
- DNNの適応：0~48 分
- 評価用：18分



実験結果

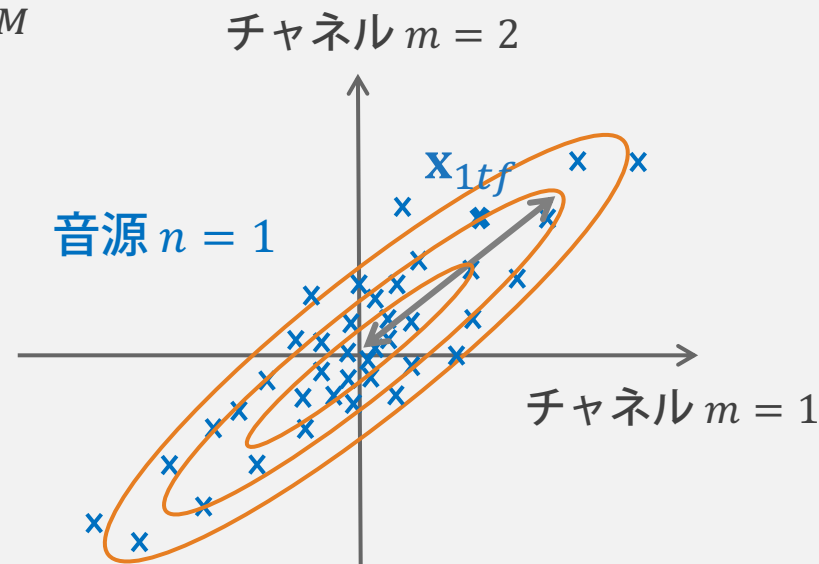
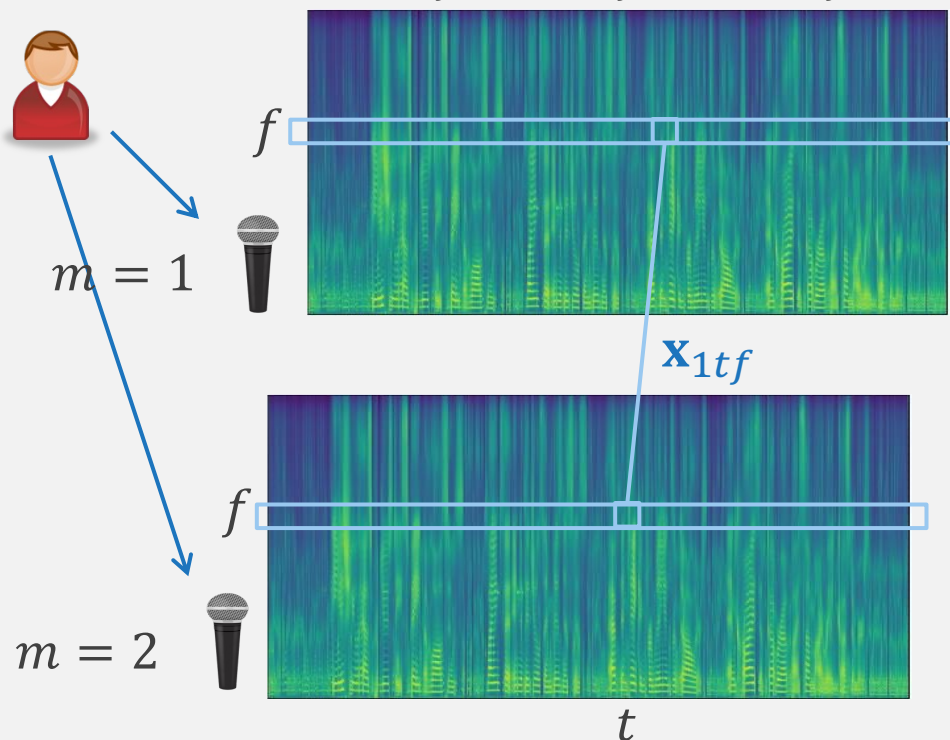
- 適応が進むにつれて単語誤り率 (WER) が低下 (性能改善)
- マスク推定DNNは3分ごとに更新 (計算量が多いため)



音源信号の観測モデル

- 音源イメージ $\mathbf{x}_{1tf}$ が多変量ガウス分布に従うことを仮定

音源イメージ $\mathbf{x}_{1tf} = [x_{1tf1}, \dots, x_{1tfM}] \in \mathbb{C}^M$



$$\mathbf{x}_{1tf} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \lambda_{1tf} \mathbf{G}_{1f})$$

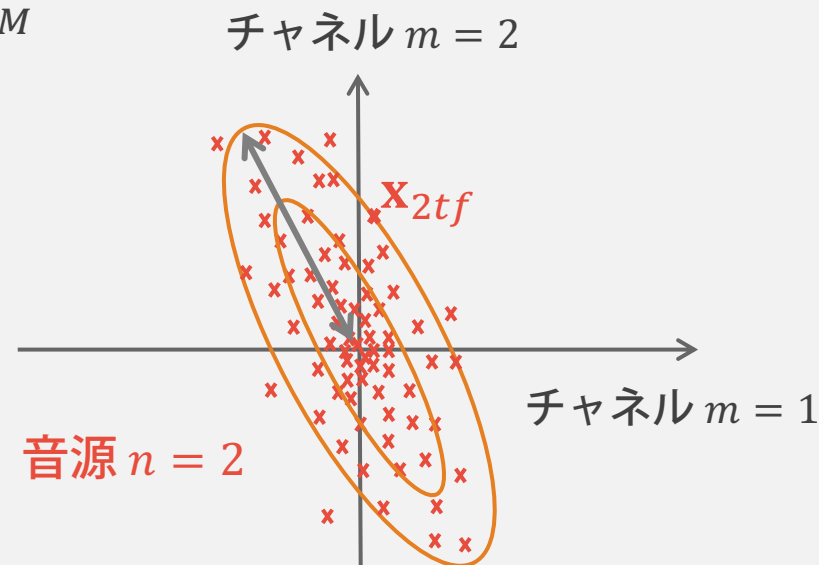
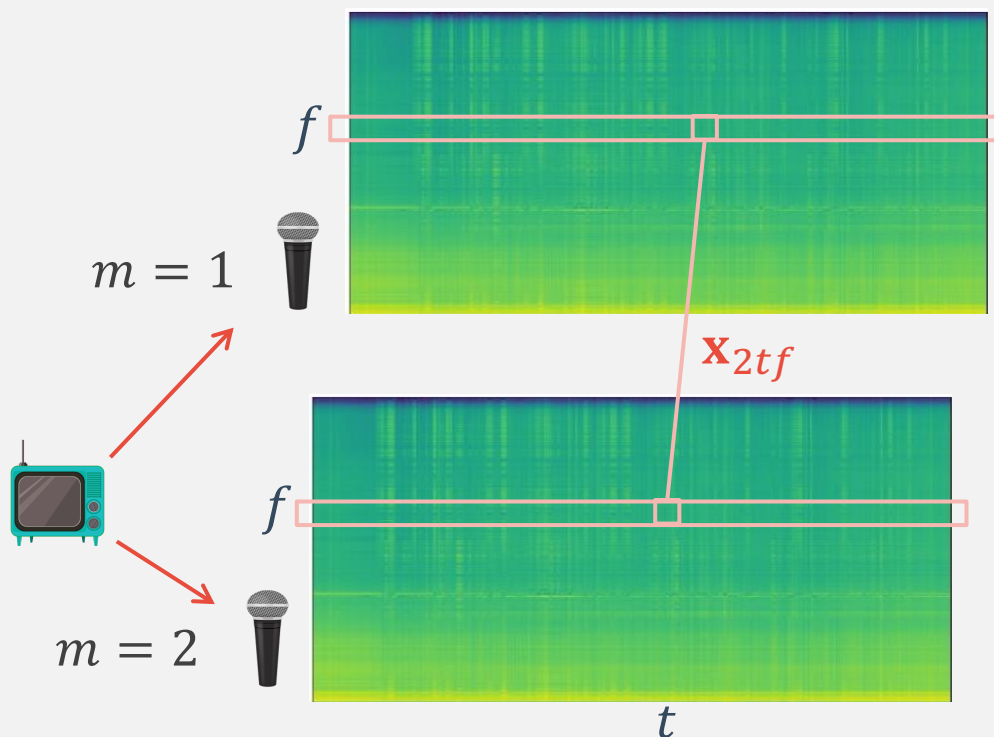
パワースペクトル
密度(PSD)

空間共分散行列
(SCM)

音源信号の観測モデル

- 音源イメージ $\mathbf{x}_{2tf}$ が多変量ガウス分布に従うことを仮定

音源イメージ $\mathbf{x}_{2tf} = [x_{2tf1}, \dots, x_{2tfM}] \in \mathbb{C}^M$



$$\mathbf{x}_{2tf} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \lambda_{2tf} \mathbf{G}_{2f})$$

パワースペクトル
密度(PSD)

空間共分散行列
(SCM)

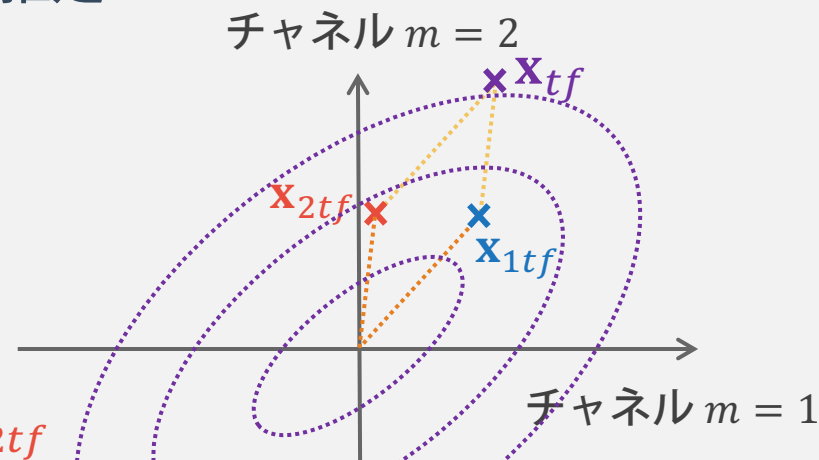
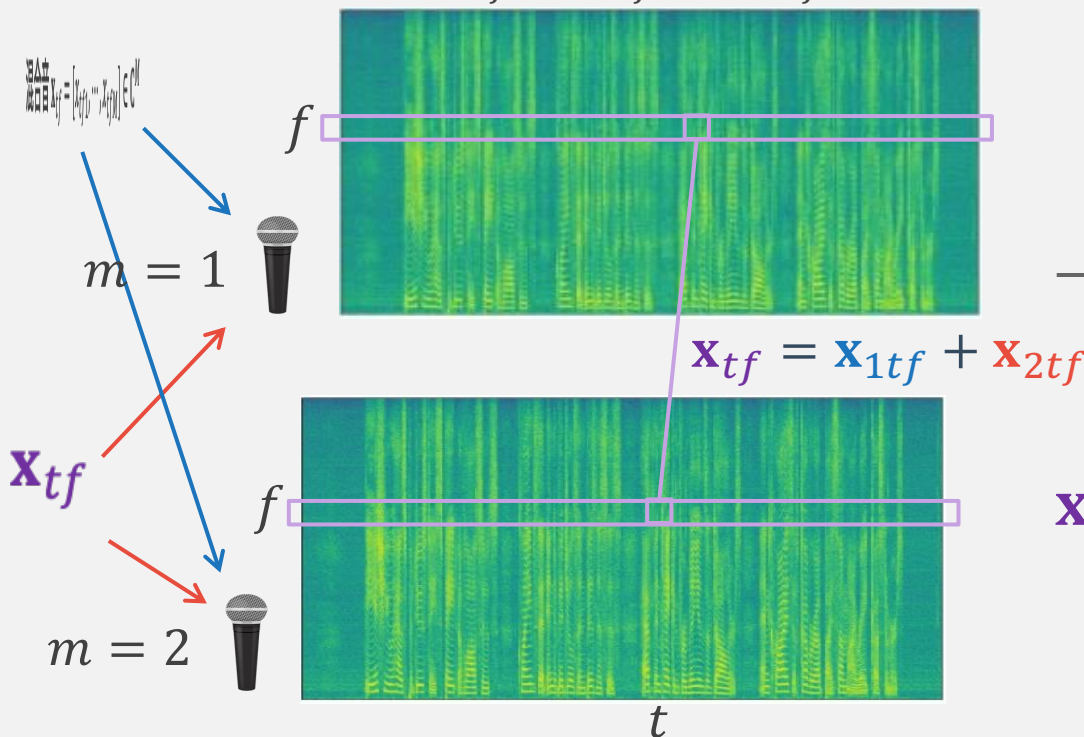
フルランク共分散分析 (FCA)

[Duong+ 2010]

- 混合音スペクトル $\mathbf{x}_{tf}$ も多変量ガウス分布に従う

- $\lambda = \{\lambda_{ntf}\}$ および $\mathbf{G} = \{\mathbf{G}_{nf}\}$ を最尤推定

混合音 $\mathbf{x}_{tf} = [x_{tf1}, \dots, x_{tfM}] \in \mathbb{C}^M$



$$\mathbf{x}_{tf} \sim \mathcal{N}_{\mathbb{C}} \left(\mathbf{0}, \sum_{n=1}^N \lambda_{ntf} \mathbf{G}_{nf} \right)$$

音源モデル

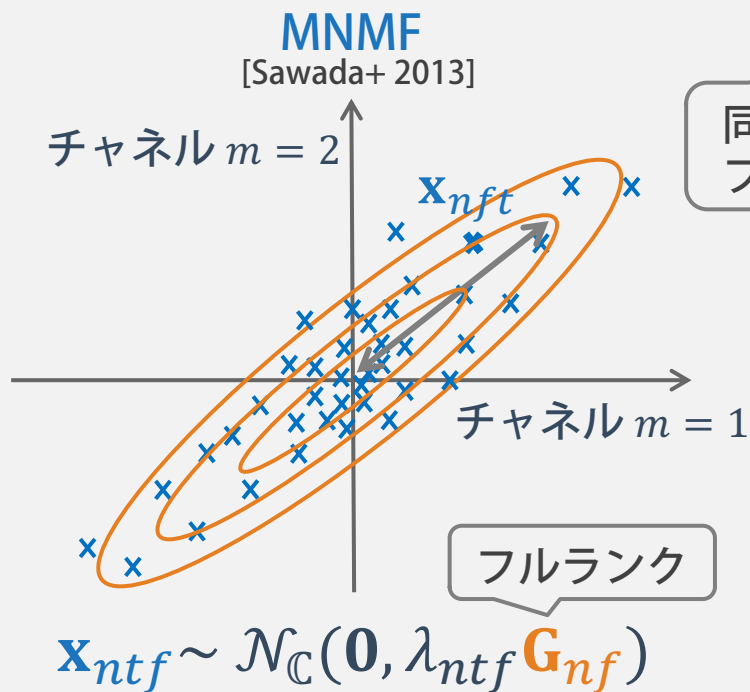
空間モデル

FastMNMFの位置づけ

- 計算速度と精度を両立する最先端の信号源分離手法

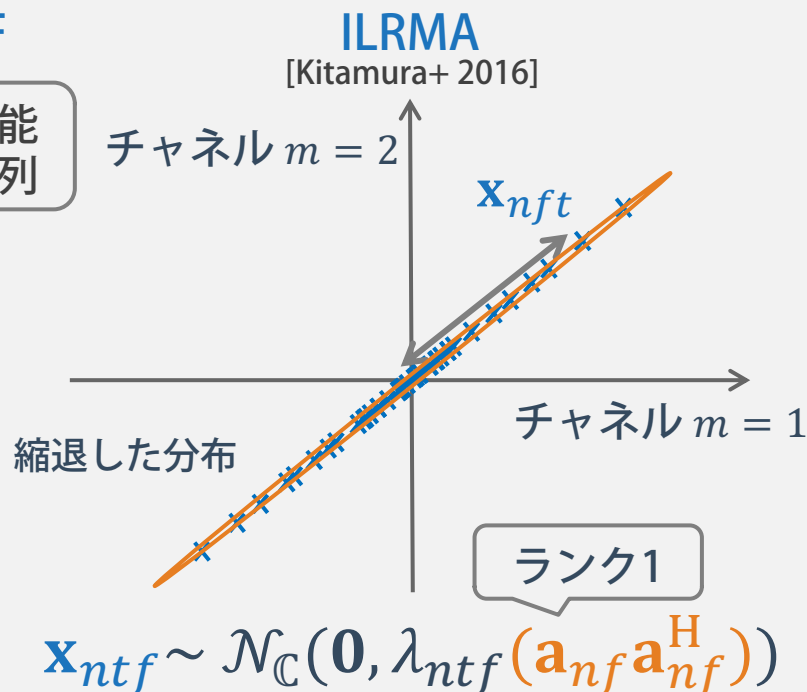
表現力

速度



FastMNMF

同時対角化可能
フルランク行列



新技術の特徴・従来技術との比較

- 多チャネル信号に対する汎用的な信号源分離技術

現在でも
BSSの最先端

- 高精度な分離

- 標準的なBSSである独立成分分析 (ICA) よりはるかに高精度
- 最新の類似のBSS手法 (ILRMA, MNMF) を上回る

- 高速な計算

- リアルタイム処理にも応用可能
- ビームフォーミングと組み合わせ可能

- 容易な実装

- <https://pyroomacoustics.readthedocs.io/en/pypi-release/pyroomacoustics.bss.html>

- 幅広い応用先

- 音響信号に限らず、多チャネル信号の分離・雑音除去等に使える

想定される用途

- センサアレイを用いた信号処理
 - 音響信号処理
 - 音声認識のフロントエンド：雑音混じりの混合音を処理
 - 発話区間同定・議事録自動生成
 - 生体信号処理
 - 脳波・脳磁図
 - 心電図・心磁図
 - 多チャネル計測信号処理
 - レーダーアレイシステム
 - 多入力多出力 (MIMO) 伝送システム
 - その他さまざまな複数センサからなる観測システム

本技術に関する知的財産権

- 出願番号：特願2021-025864
- 出願日：2021/2/22
- 公開番号：特開2022-127719
- 公開日：2022/9/1
- 発明の名称：ソース信号分離装置、ソース信号分離方法、プログラム、ならびに、情報記録媒体
- 特許出願人：国立研究開発法人理化学研究所

問い合わせ先

- 株式会社理研イノベーション 新技術説明会事務局
 - license-contact@innovation-riken.jp