システム情報 猿渡・齋藤研究室
（創造情報 猿渡研究室）の紹介

東京大学大学院・情報理工学系研究科
システム情報学・創造情報学専攻　猿渡・齋藤研

（2025年5月）

猿渡・齋藤研（システム情報第一研究室）

教授
猿渡洋

専門分野

・教師無し最適化
・統計・機械学習
　論的信号処理

講師
齋藤佑樹

専門分野

・統計的機械学習
・音声知覚モデリング
・Human Computation

助教
山岡洸瑛

専門分野

・多チャネル信号処理
・方位時間差推定
・補助関数最適化

特任助教
岡本悠希

専門分野

・環境音合成
・環境音検出認識
・環境音DB構築

特任准教授
高道慎之介

専門分野

・音声コミュニケーション拡張
・音声言語情報処理

協力教員

・北村大地先生（香川高専）
・小山翔一先生（NII）
・中村友彦先生（産総研）
・伊藤信貴先生（産総研）

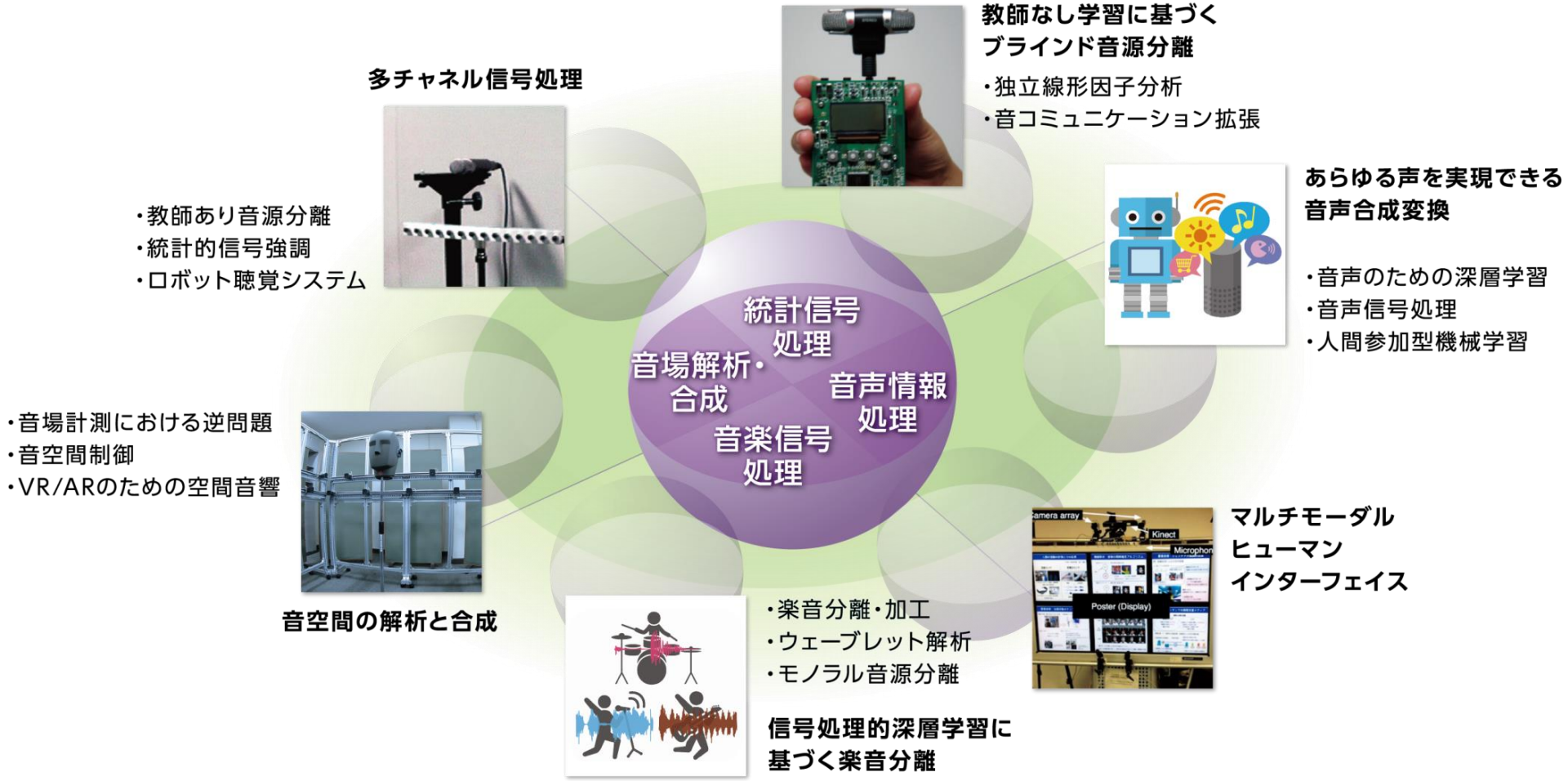
学術専門職員　高宗さん
秘書　丹治さん

学生

・博士課程学生12名
・修士課程学生9＋7名

研究俯瞰図

- 音声・音響・音楽メディアに関する信号処理・情報処理
- ヒューマンインターフェイス・コミュニケーションシステムの構築
- 統計的・機械学習論的信号処理、数理最適化問題等を研究



研究紹介1．ブラインド音源分離

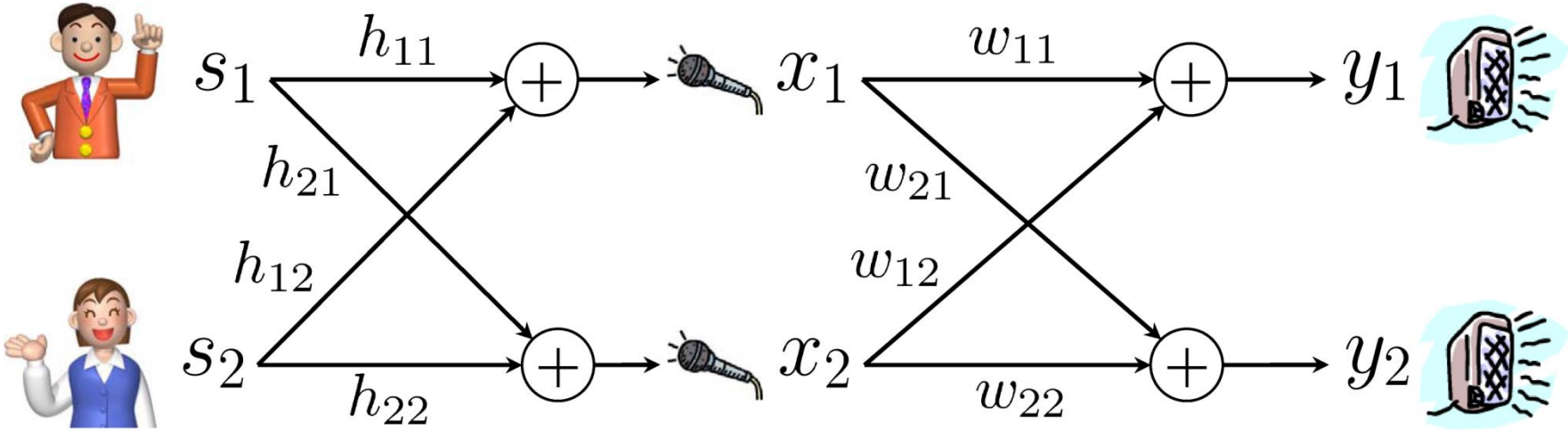
- 音の方向・声質・音量など、事前に何も分かっていなくても、瞬時に音を「聞き分ける」ことの出来るシステムを目指す。

- 独自に開発した高速独立成分分析（ICA）、独立低ランク行列分析（ILRMA）という教師無し数理最適化アルゴリズムに基づいて、音を統計的に独立な成分に分解することにより、別々の音声信号を見つける。

スモールデータ
教師無し最適化
低ランクモデル

ブラインド音源分離（BlindSourceSeparation）：聞き分けるAI

- 混ざり合った信号 x_1，x_2 から元の信号を取り出す
- どのように混ざったかに関する情報 $\boldsymbol{H}$ は利用できない
- 事前トレーニング出来ない⇒ビッグデータではなくスモールデータ



実は上記は**2つのことを同時に推定**している

- [空間] 統計的に独立な音源の分類問題（分離行列Wの推定）
- [音源] 各音源が属する確率分布$p(y)$や構造の推定問題

上記を閉形式で解く方法は存在せず凸問題でもない⇒**大変困難！**

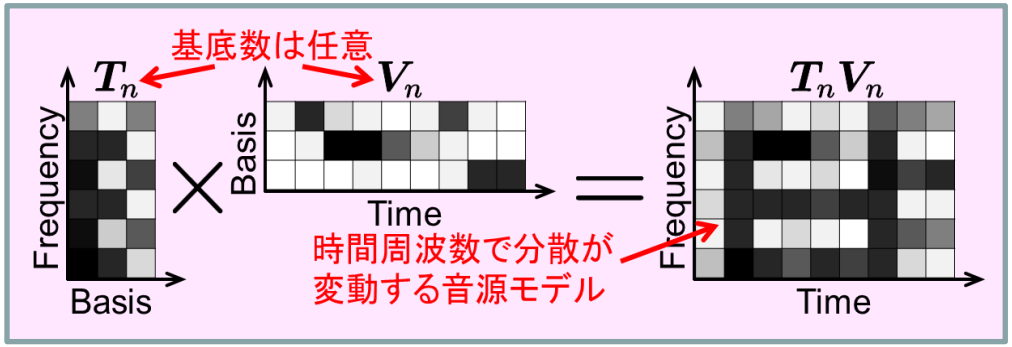
ILRMA：音源の独立性と低ランク性に着目したBSS

[IEEE Trans. ASLP 2016、IEEE SPS論文賞・ASJ粟屋賞・JSPS育志賞]

- ILRMAのコスト（対数尤度）関数→これを最小化

$$\mathcal{J} = \sum_{i,j} \left[\sum_m \log \sum_k z_{mk} t_{ik} v_{kj} + \sum_m \frac{|y_{ij,m}|^2}{\sum_k z_{mk} t_{ik} v_{kj}} - 2 \log |\det \boldsymbol{W}_i| \right]$$

音源の低ランク性コスト関数
（音源NMFモデルの推定に寄与）



音源の独立性コスト関数
（空間モデルWの推定に寄与）

$$p(\boldsymbol{y}) = p(y_1)\, p(y_2) \cdots p(y_m)$$

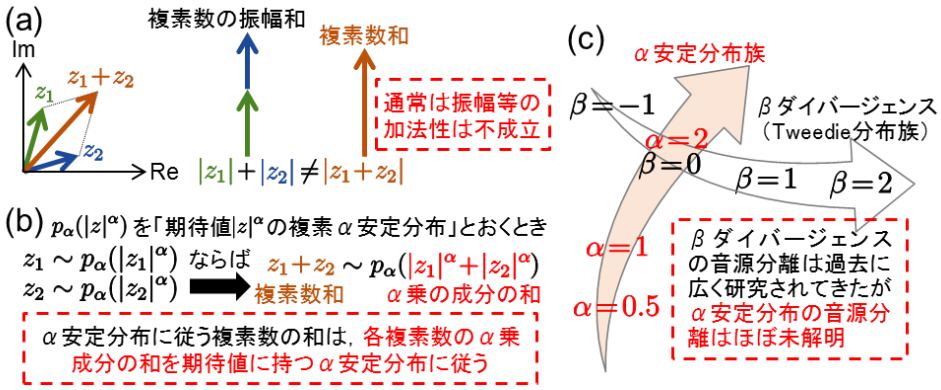
となるWを推定

両者を交互にMajorization-Minimization（補助関数法）アルゴリズムで反復最小化

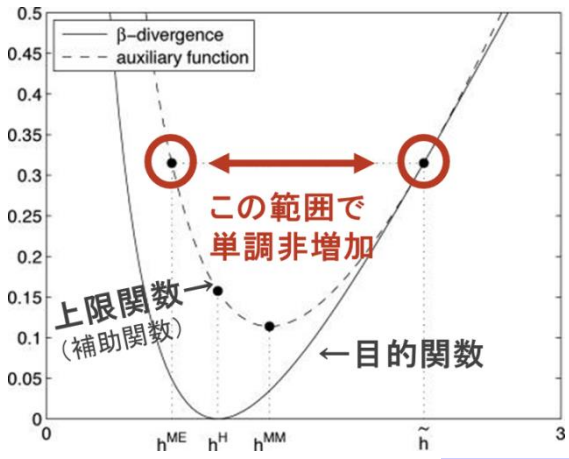
- ✓ コスト値の単調減少性を保証（勾配法には無い特徴）
- ✓ 高速かつ安定な求解法を実現（従来の多入力NMFと比較して2ケタ速い）

モデルの多様化・数理解法の開拓

- 音源生成モデルの多様化　IEEE SPS Tokyo Joint Chapter学生賞!
  - 複素波形重ね合わせと整合するα安定分布の導入
    ⇒t-ILRMA [EURASIP-JASP2018]
  - 複素球状ポアソン分布の導入
    ⇒β=1-divergence最小化ILRMA
    [IEICE Trans. 2018]

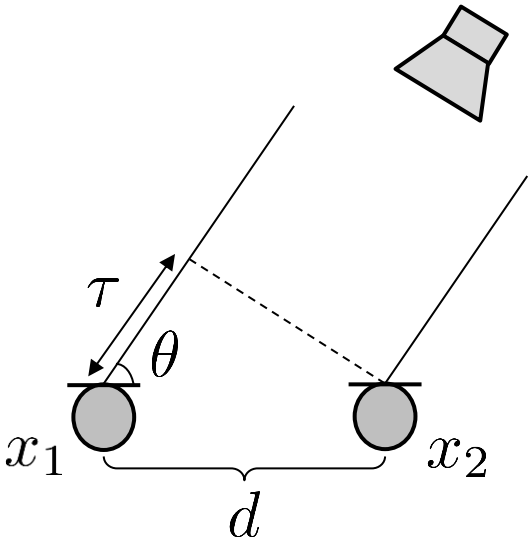


- 座標降下法におけるバリエーション
  - パラメトリックMajorization-Equalizationアルゴリズムによる音源・空間最適化の「バランス」化
    [CAMSAP2017], [IEEE Trans. ASLP2019]
- 深層学習（DNN）との融合
  - 独立深層学習行列分析（IDLMA）の提案
    [IEEE Trans. ASLP2019]



補助関数型時間差推定 [Yamaoka+, 2019]

- 音の到来時間差
  - 音の到来方向に対応 ($\tau = d\cos(\theta)/c$)
  - ブラインドに推定する必要あり
  - 音源分離とは相補的な関係

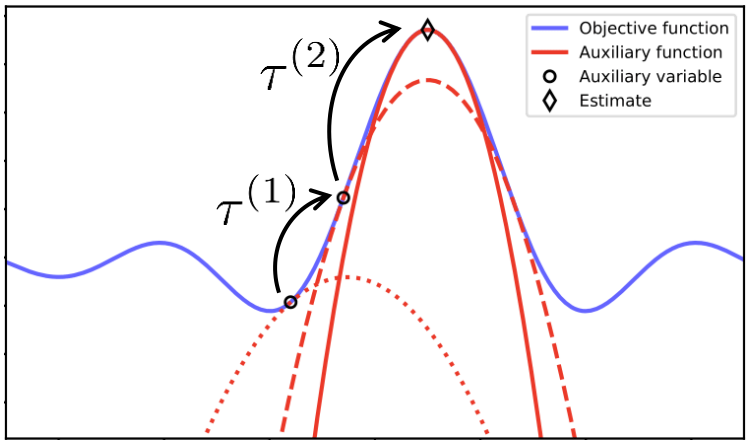


- 補助関数法に基づく推定
  - 目的関数は非凸で閉形式の解はない

$$\arg\max_{\tau} \sum_k x_{1k} x_{2k}^* e^{-j2\pi \frac{k}{K}\tau}$$

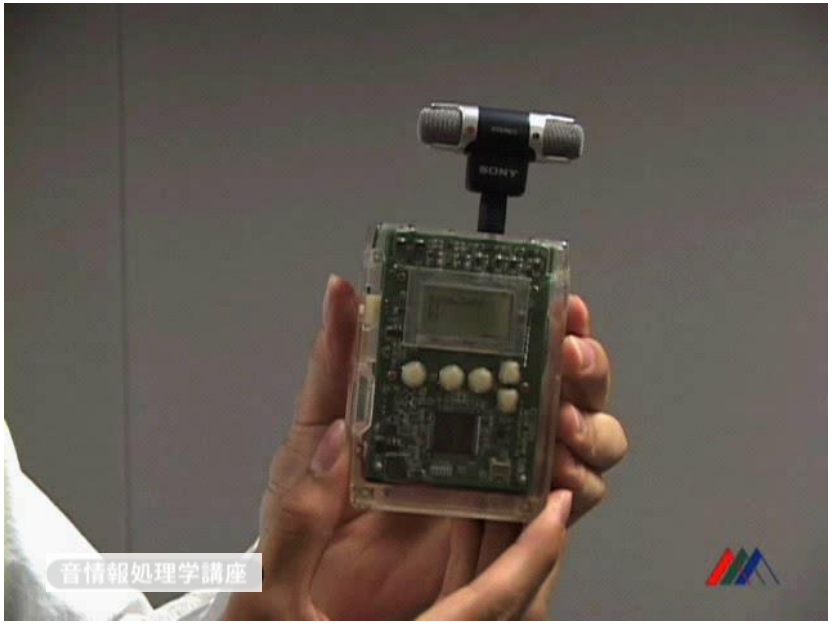
  - 二次の補助関数を設計，
    閉形式の解による更新則を提案

$$\tau^{(\ell+1)} \leftarrow \tau^{(\ell)} - \frac{\sum_k A_k \mathrm{sinc}\,\theta_k^{(\ell)} \frac{\theta_k^{(\ell)}}{\omega_k}}{\sum_k A_k \mathrm{sinc}\,\theta_k^{(\ell)}}$$

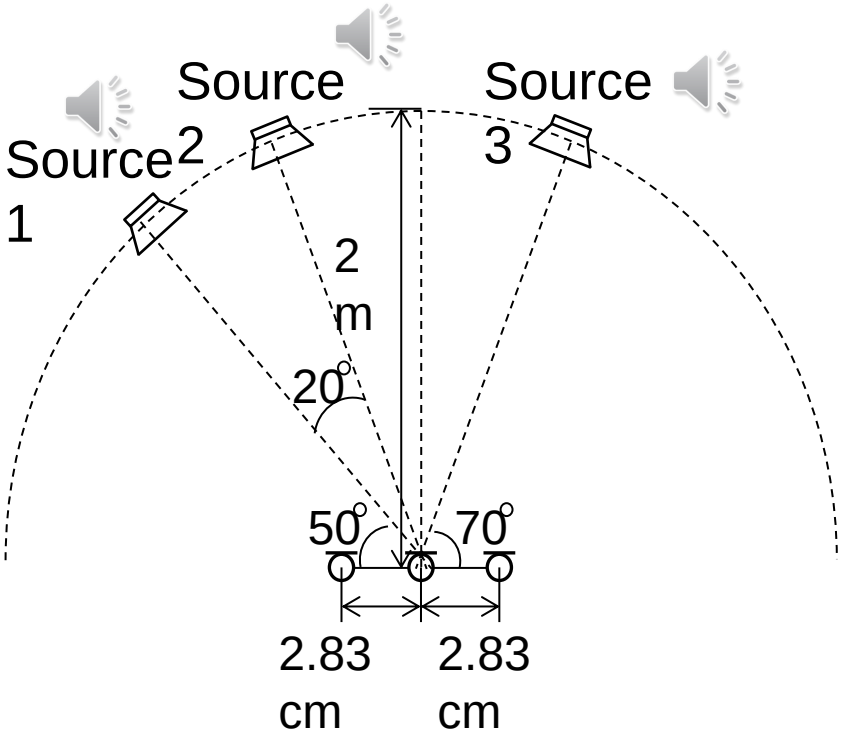


高速ICA、独立低ランク行列分析によるデモ

- リアルタイム音声聞き分け（警察備品に採用）



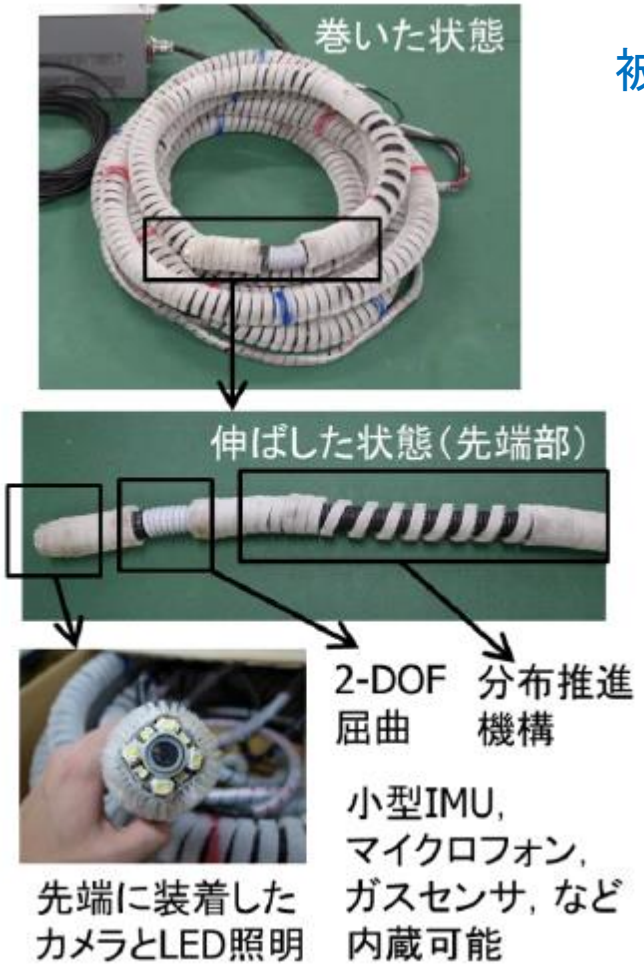
- ドラム、弦楽器、音声からなる複合音の分離

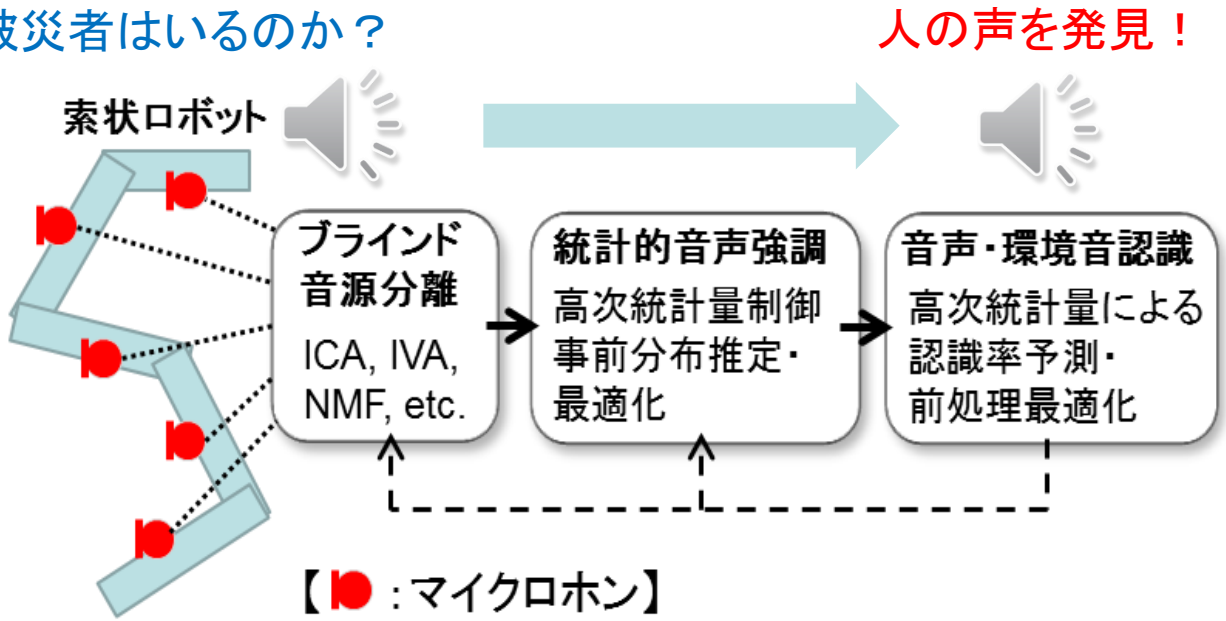


内閣府ImPACT災害対応タフロボット [2016年6月プレスリリース]

■災害時の倒壊家屋に入り込んで被災者発見

■環境音認識による状況把握・救助支援





いかなる曲がりくねった形状においてもマイク同士が協調して騒音の中から被災者の声を見つけ出す

独立深層学習行列分析

Independent Deeply Learned Matrix Analysis
(IDLMA: 発音はアイドルエムエー）
[Makishima, Saruwatari+, IEEE-Trans. 2019]

ILRMAにおける問題点：音源の低ランク性？

$$\mathcal{J}_{\mathrm{ILRMA}} = \frac{1}{J}\sum_{i,j,n}\left[\log\sum_{l} t_{il,n}v_{lj,n} + \frac{|\boldsymbol{w}_{i,n}^{\mathsf{H}}\boldsymbol{x}_{ij}|^2}{\sum_{l} t_{il,n}v_{lj,n}}\right] - \sum_{i}\log|\det \boldsymbol{W}_i|^2$$

音源モデル (低ランク性)

音源によっては低ランク性が
成り立たない場合がある

ならば！

事前に学習データを用いて音源モデルの分散を推定する写像を作る

空間モデル (音源間が独立)

音源・マイク位置，部屋の形状，
残響時間などの膨大な物理要因に依存

学習データの用意は非現実的
ブラインドに推定

ILRMAにおける問題点：音源の低ランク性？

$$\mathcal{J}_{\mathrm{ILRMA}} = \frac{1}{J}\sum_{i,j,n}\left[\log\sum_{l} t_{il,n}v_{lj,n} + \frac{|\boldsymbol{w}_{i,n}^{\mathsf{H}}\boldsymbol{x}_{ij}|^2}{\sum_{l} t_{il,n}v_{lj,n}}\right] - \sum_{i}\log|\det \boldsymbol{W}_i|^2$$

音源モデル (低ランク性)

空間モデル (音源間が独立)

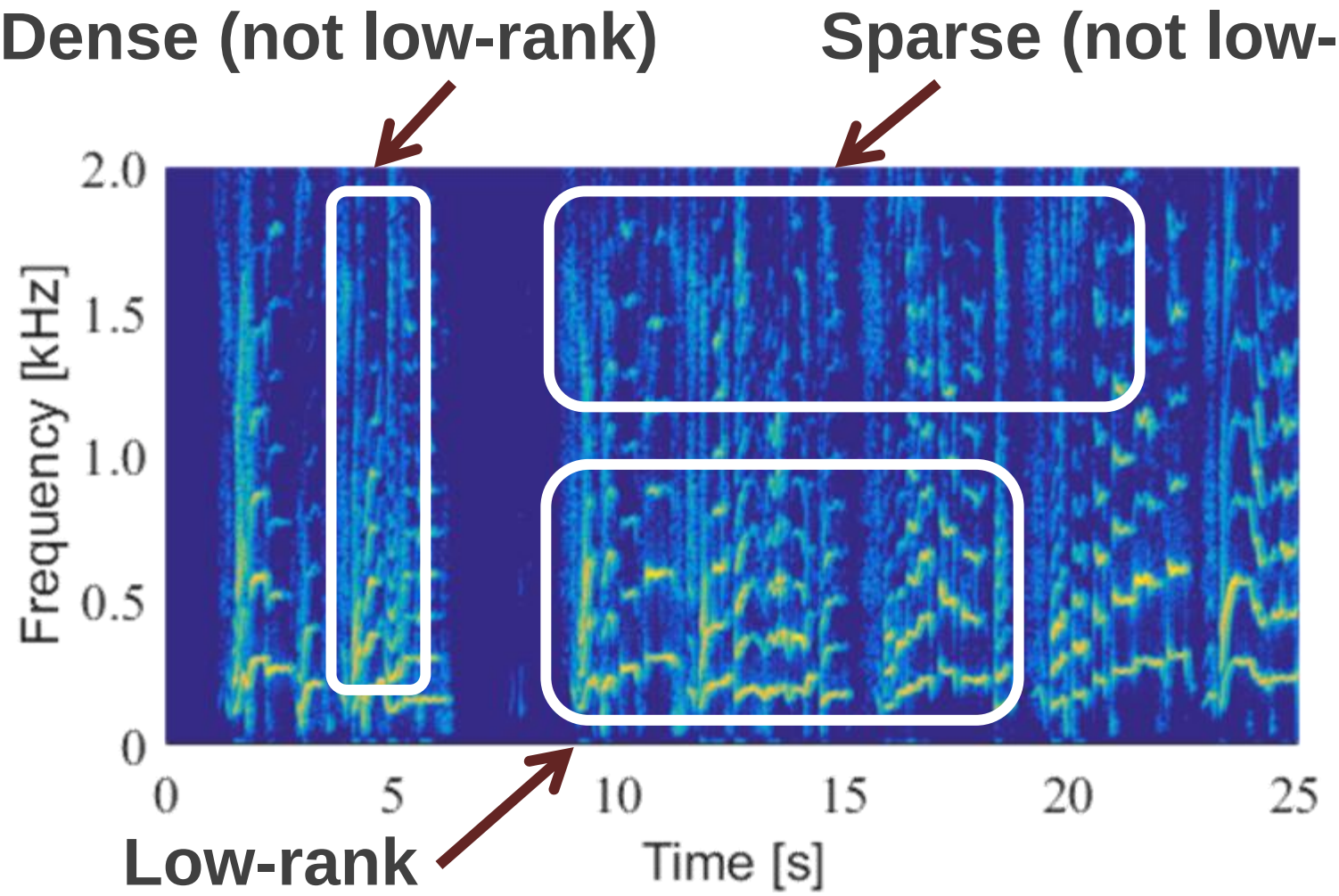
- 深層学習（DNN）による強力なモデリング能力を活用する！
- 今まで培ってきた「教師あり音源分離（例：教師ありNMF）」の技術を昇華させる形で研究を発展できる。
- 急速に発展するDNN研究を我々ならではの視点で拡張する。

事前に学習データを用いて音源モデルの分散を推定する写像を作る

学習データの用意は非現実的

ブラインドに推定

音源の低ランク性？ (例：音声信号)



このような複雑な構造を持つ信号はDNNでモデリングする

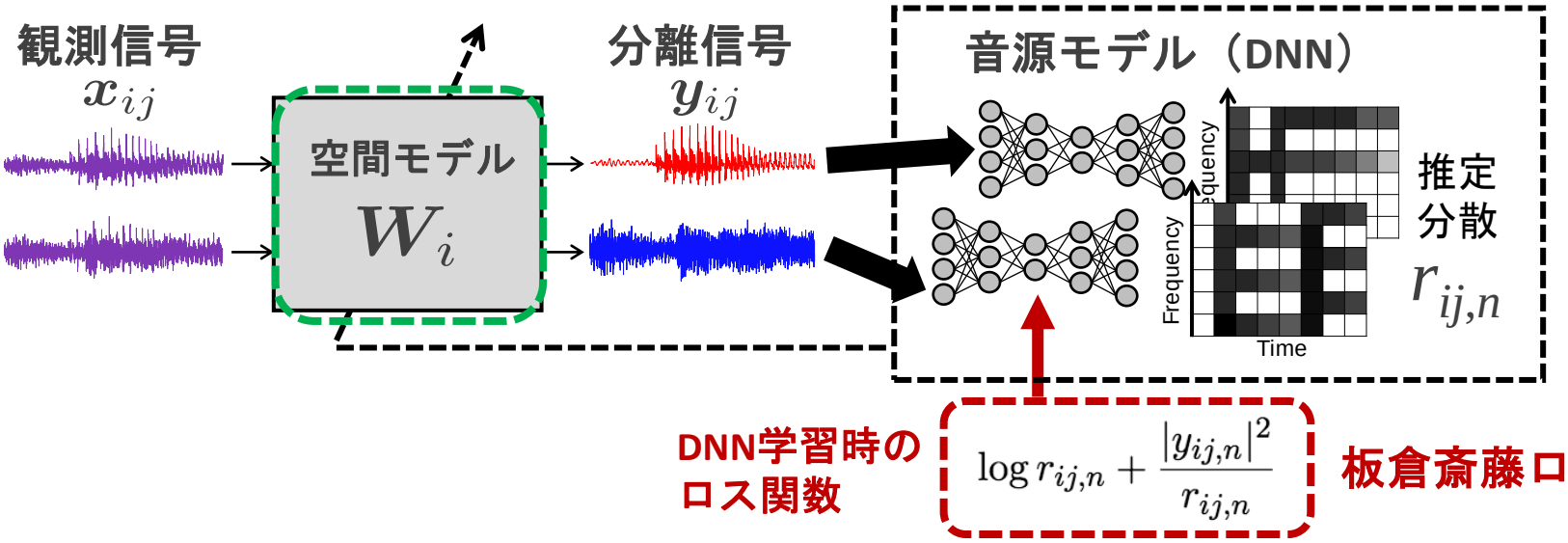
提案手法：DNN音源モデルによる最尤推定

■ 独立深層学習行列分析（IDLMA）

$$y_{ij,n} = \boldsymbol{w}_{i,n}^{\mathrm{H}} \boldsymbol{x}_{ij}$$

$$\mathcal{J} = \frac{1}{J} \sum_{i,j,n} \left(\log r_{ij,n} + \frac{|\boldsymbol{w}_{i,n}^{\mathrm{H}} \boldsymbol{x}_{ij}|^2}{r_{ij,n}} \right) - \sum_{i} \log |\det \boldsymbol{W}_i|^2$$

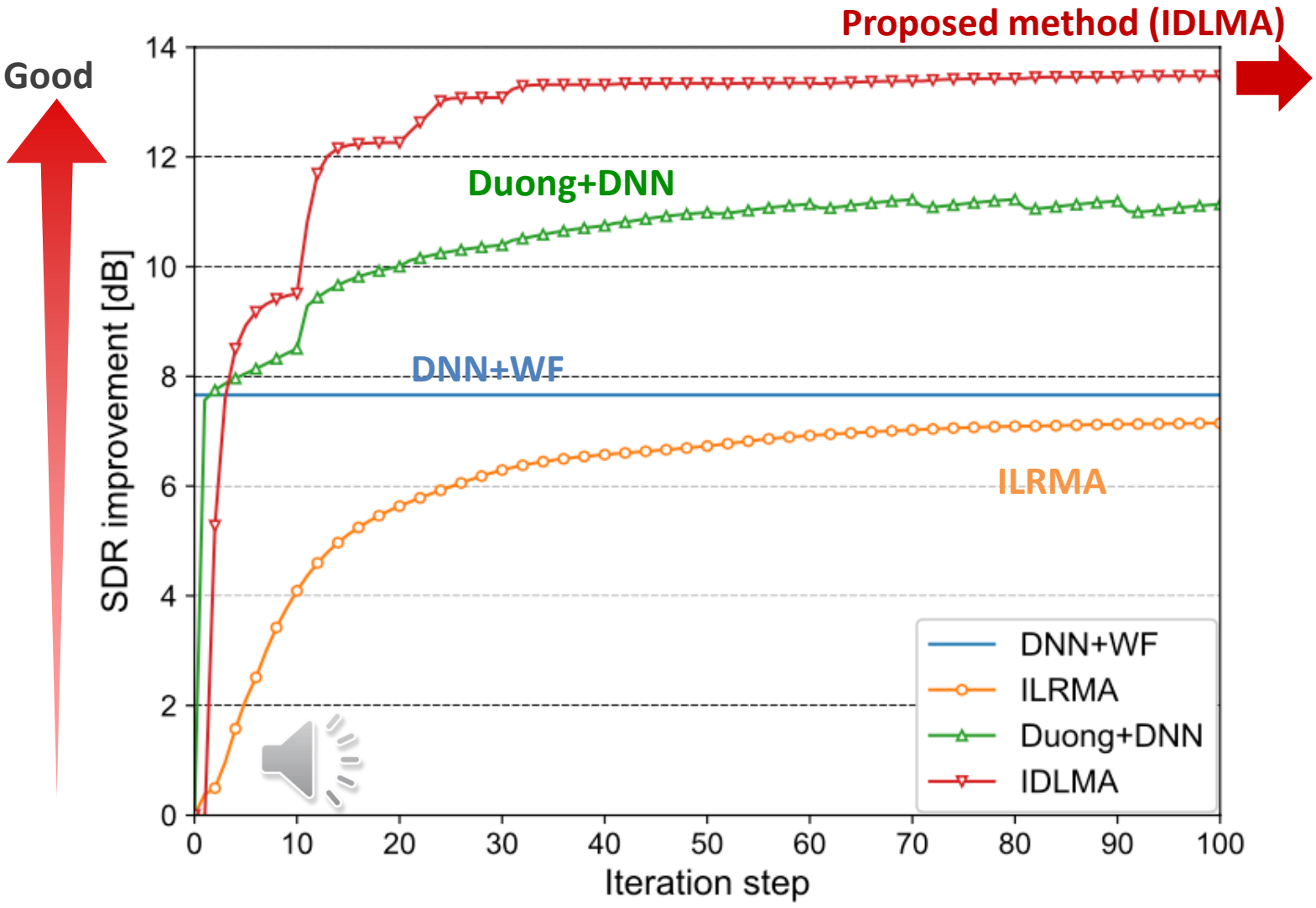
音源モデル (DNN)　⟷　交互に最適化　空間モデル (音源間が独立)



■ **空間モデル**：各音源が統計的に**独立**となる分離行列を推定

■ **音源モデル**：J を最小化するような分散 $r_{ij,n}$ を推定するDNNを各音源ごとに構成

実験結果例（反復最適化）


Proposed method (IDLMA)
Good
SDR improvement [dB]
Iteration step
Duong+DNN
DNN+WF
ILRMA
DNN+WF
ILRMA
Duong+DNN
IDLMA
0
2
4
6
8
10
12
14
0
10
20
30
40
50
60
70
80
90
100


Vo.
Ba.

研究紹介2. 統計的音声合成

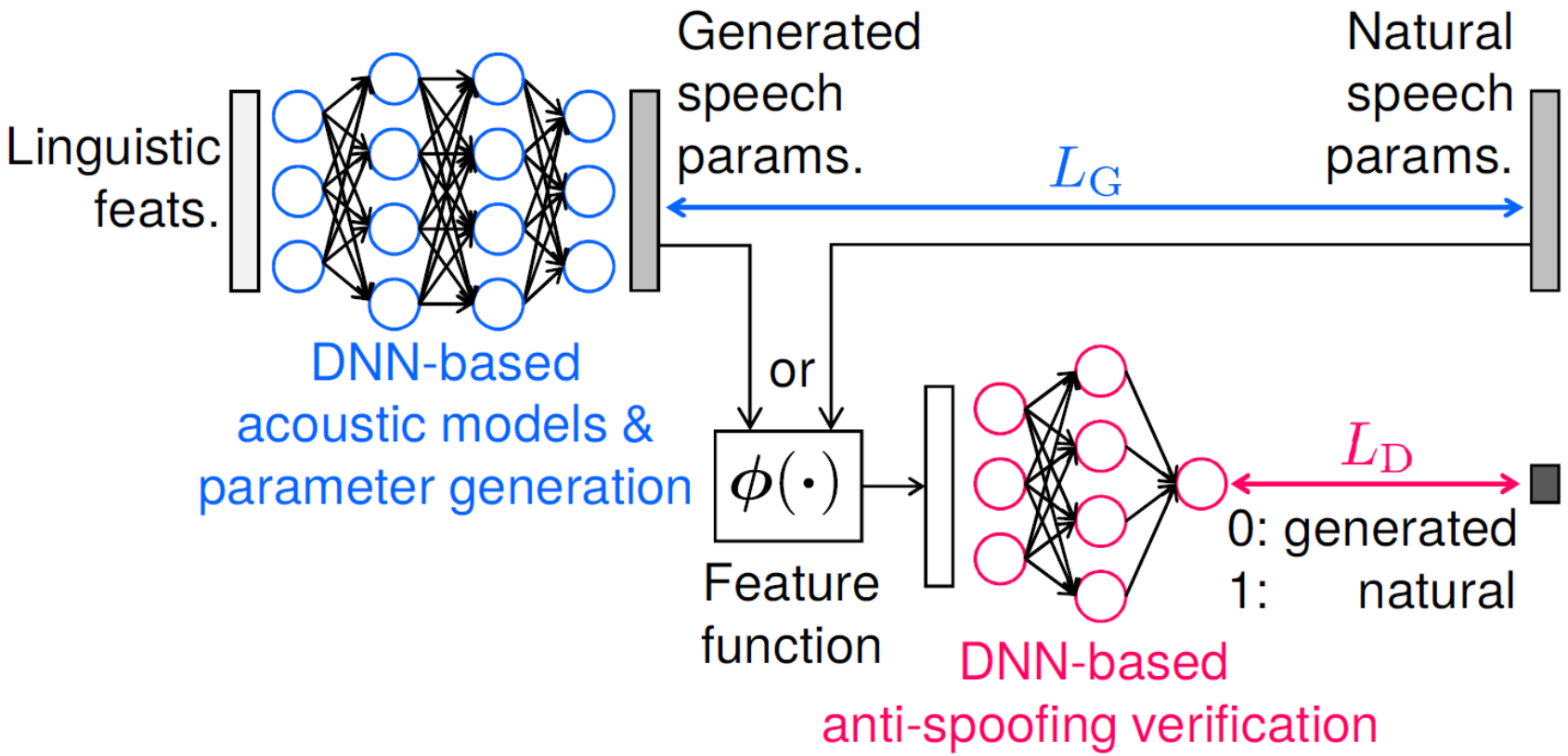
- テキストもしくは自分の声を入れるだけで、誰の声でも、どんな訛りでも、何語でも、喋ることが出来るような統計的音声合成システムを実現する。

- 深層学習（Deep Neural Net）の枠組みを活かし、「AIオレオレ詐欺師」と「AI防犯課刑事」を対決させて、お互いに精度を高めるAnti-Spoofing 敵対学習理論を独自に提唱

ビッグデータ
深層学習
敵対学習

Anti-Spoofingと敵対する音響モデル学習理論

[IEEE SPS Young Author Best Paper賞・IEEE SPSJ 学生論文賞他]



人間の声に似せようと努力

ウソ（合成音）に騙されまいと攻防

リアルタイムDNN声質変換

2019年3月 日経xTECHプレスリリース

さらに柔軟な音声合成へ

- 松任谷由実＋AI荒井由実「Call me back」人工声の提供（2022年の紅白にて紹介）

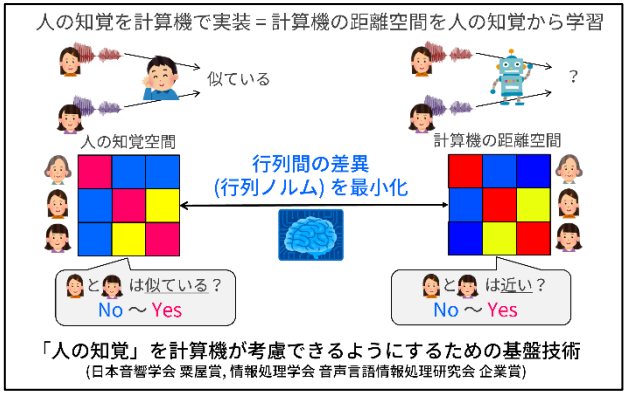
出典：YouTube 松任谷由実 – Call me back/松任谷由実with 荒井由実

- フィラー挿入付き自発音声合成（言いよどむAI）

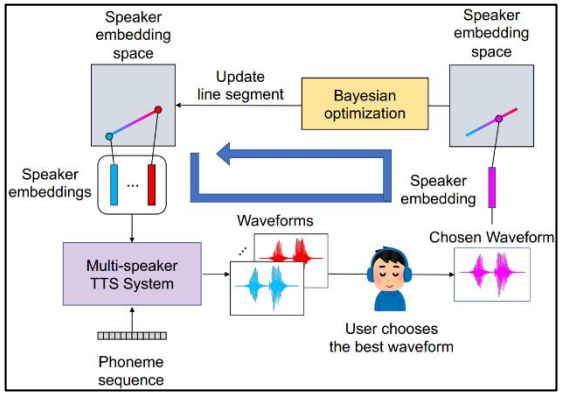
ざっくりいうと，先ほど少しお話ししましたけども，戦後のそういうサブカルチャーのイメージという…

ざっくりいうと，(アノー)先ほど(アノー)少し(アノー)お話ししましたけども，戦後のそういうサブカルチャーのイメージという…

人と機械の相互作用を通じた音声合成・変換の最適化

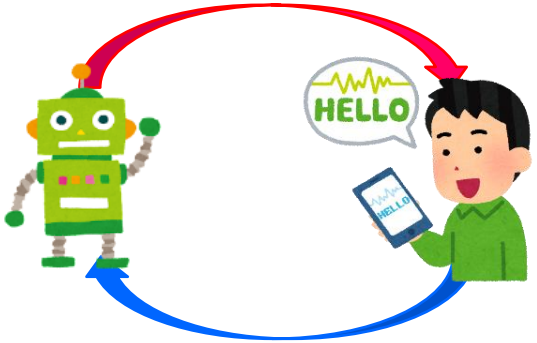


話者知覚に基づく
音声表現学習 [Saito+TASLP21]



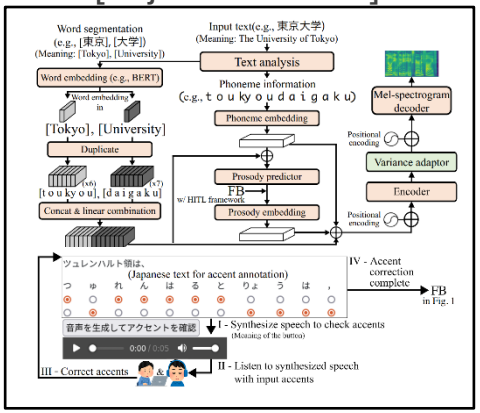
Human-in-the-loop
話者適応 [Udagawa+IS22]

機械による
音声情報処理能力の拡張

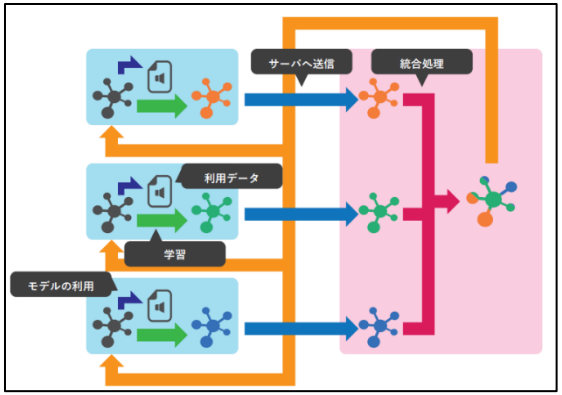


人の知覚を導入した最適化

アクセント訂正FB音声合成
[Fujii+APSIPA22]

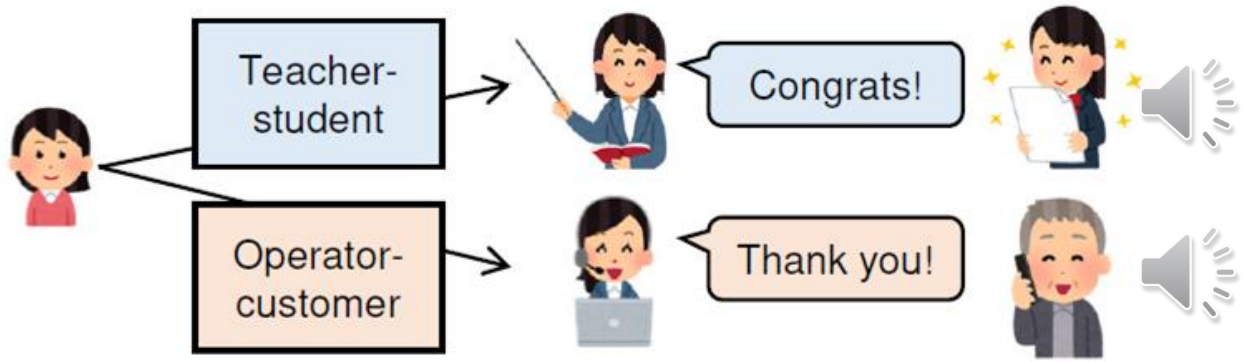


Federated 声質変換
[平井+SLP23]

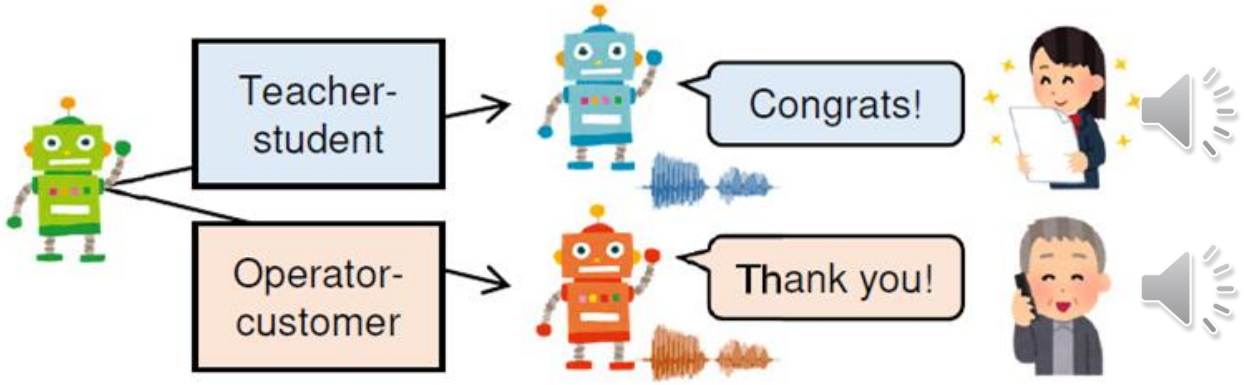


人と機械が協調して情報伝達を行うための基盤技術を構築

研究例: 多ドメイン共感的対話音声合成



人間同士のコミュニケーション:
様々な対話ドメインで相手と共感的に会話可能



人間・ロボット間のコミュニケーションでもこれを再現

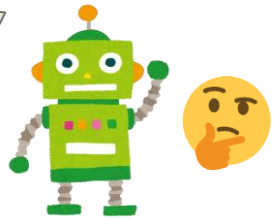
ChatGPT-EDSS: 大規模言語モデル (LLM) と対話して発話スタイルを制御する音声合成

対話相手
(人間)

聞き手
(AI)

Hi, teacher!

Oh, did you get a good score?



Bingo!!

LLM に
「どう応答すべきか?」
を質問

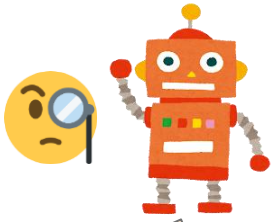
Speaker: Hi, teacher!

Listener: Oh, did you get a good score?

Speaker: Bingo!!

Listener: Congrats!!

対話
アドバイザー
(LLM)



喜んで,
祝うように

Congrats!!

対話履歴を考慮し,
相手にどう応答すべきかを回答

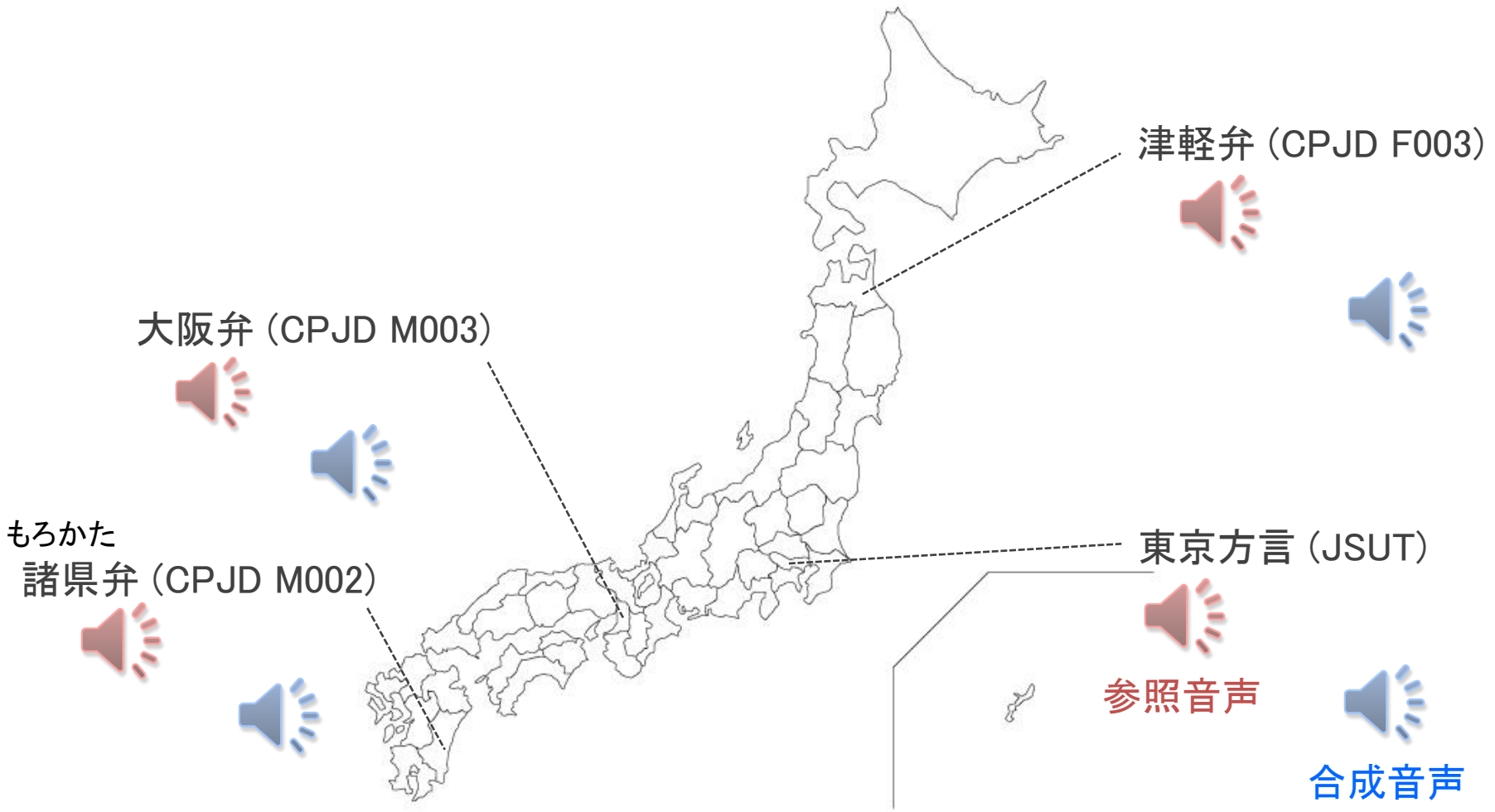
合成 w/ 感情ラベル

合成 w/ 対話履歴

合成 w/ 対話履歴 + LLM

[Saito+23INTERSPEECH]

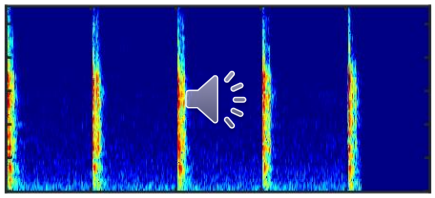
Cross-dialect TTS: 方言をまたいで喋らせる音声合成



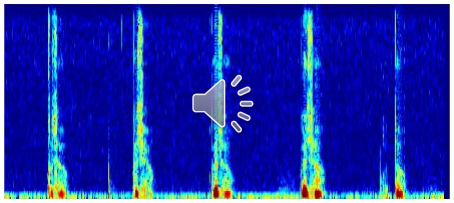
[山内+2403SP]

Voice-to-Foley: 声真似に基づく環境音合成

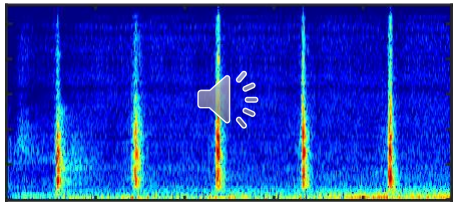
目標音

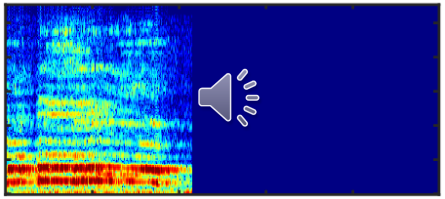


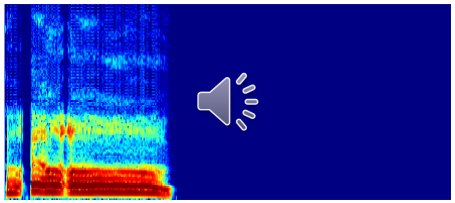
声真似

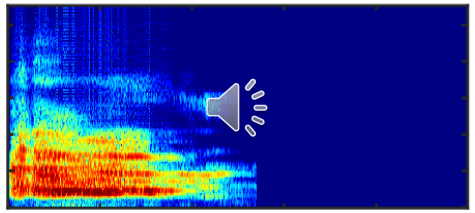


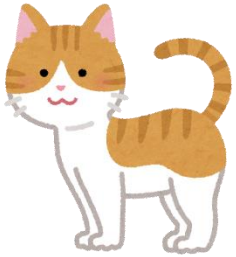
合成音

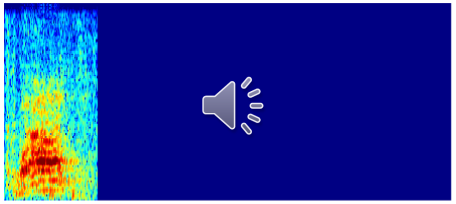


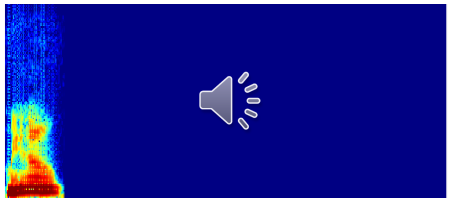


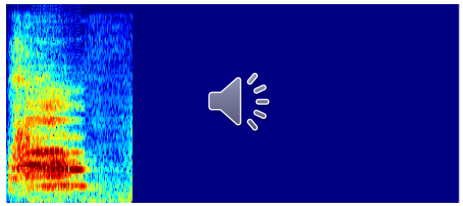








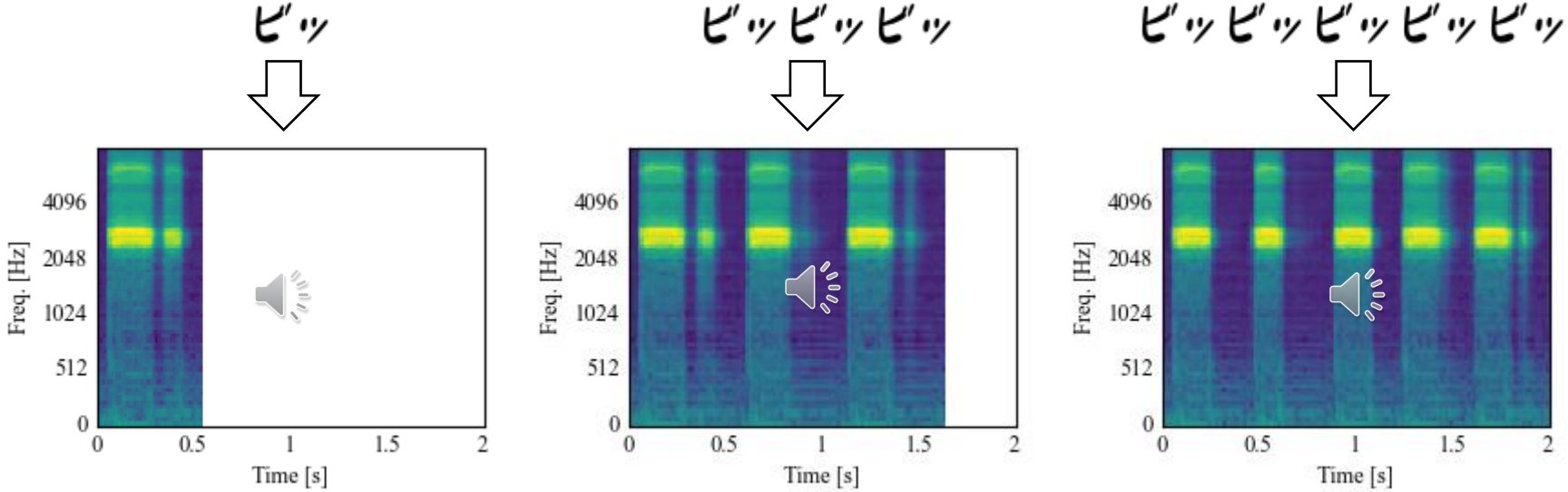




[Okamoto+24ICASSP]

Visual Onoma-to-Wave: 画像を考慮した環境音合成

オノマトペ (擬音) 文字画像からの合成



環境音イメージ画像で条件付けした合成

[Ohnaka+23ICASSP]

研究紹介3．音楽信号解析

- 様々な楽器がまじりあった音楽信号の中から、自分の好きな楽器を見つけ出し、自分の好みのリミックス版を製作する。

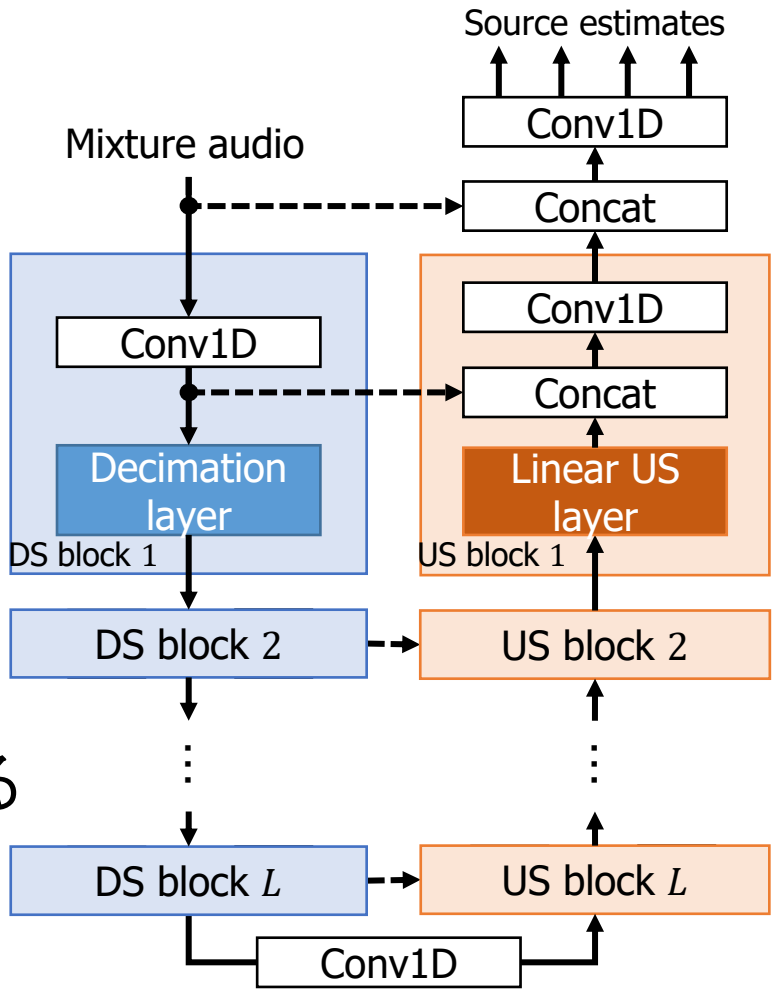
- 非負値行列因子分解（NMF）や深層学習（DNN）という教師有りアルゴリズムを用い、音を事前に学習したパターンに基づいて分解することにより、信号を解析する。

スパース分解
低ランクモデル
教師有り学習

多重解像度深層分析(1/2):動機
[Nakamura+ IEEE Trans. ASLP2021]

- Wave-U-Net [Stoller+2018]
  - 時間信号を直接入力し分離音を得るEnd-to-end DNNモデル
  - 繰り返しダウン・アップサンプリングを行うU-Net構造をもつ

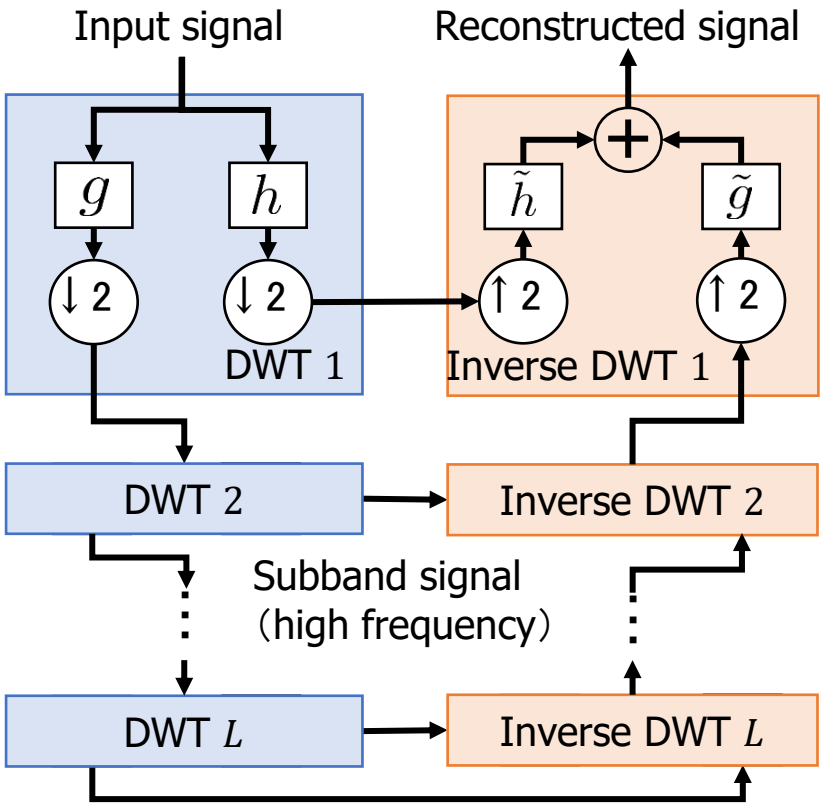
- しかし，信号処理の観点から見るとダウンサンプリングが問題…
  - 特徴量ドメインでエリアシングが発生
  - ダウンサンプリングで情報が欠落しうる
    ⇒ 分離性能の低下を招く 🤔



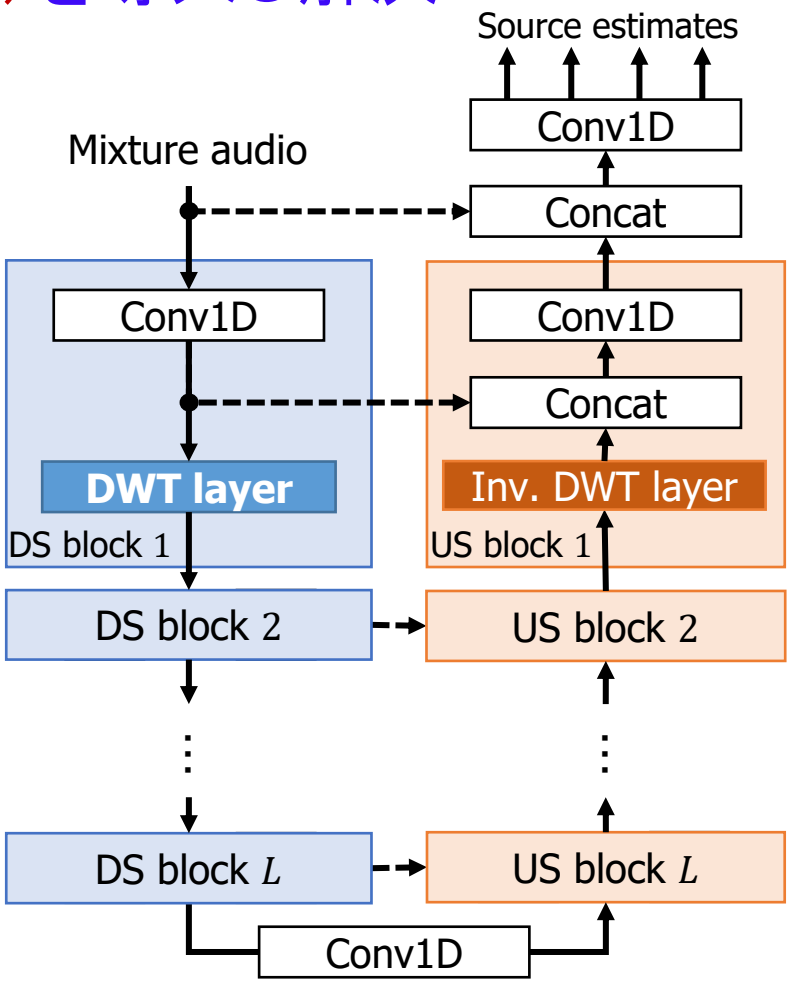
多重解像度深層分析(2/2):解決方法

[Nakamura+ IEEE Trans. ASLP2021]

- Wave-U-Netと多重解像度解析の構造の類似性に着目
  ⇒ 離散ウェーブレット変換(DWT)を導入し解決‼



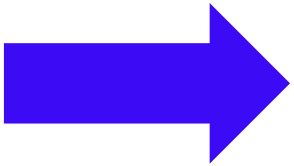
多重解像度分析



多重解像度深層分析(MRDLA)

多重解像度深層分析による楽音分離デモ

- Vocal，bass，drums，guitarの音に分離



	分離音	正解音
Vocal:		
Bass:		
Drums:		
Guitar:		

研究プロジェクト・国内外研究者コラボ

[2024年～]

- サイバネティックアバター（内閣府ムーンショット） with 京大河原・阪大石黒先生
- リアルタイム・スモールデータ音響AR（立石研究助成S）
- 楽音信号分解（ヤマハ研究開発センター） with 北村先生・高橋さん・近藤さん
- ランク制約付き空間相関行列推定（NTT-CS研） with 池下さん・中谷さん
- 極限音響ドローン災害救援システム（鹿島財団国際共同研究） with NZオークランド大
- 安心声変換（Beyond AIソフトバンク）
- 次世代音声翻訳（科研基盤S分担） with NAIST中村先生・名大戸田先生
- 深層学習を用いた音声デザイン（科研費基盤A）with 明治大 森勢先生
- サイレントスピーチ（産総研）with 産総研 持丸先生
- 音声言語情報処理の基盤モデル構築（産総研） with 産総研 深山さん・緒方さん
- 自己教師あり対話音声合成（Google）with Google 木下さん
- 音声スタイルキャプショニング（NTT人間情報研） with NTT 井島さん
- Talking head 合成（LINEヤフー） with LINEヤフー 橘さん
- ゲーム実況解説音声合成（科研費若手） with 産総研 石垣さん
- Human-in-the-loop 日本語方言音声合成(ACT-X)
- Human-in-the-loop音響情景分析（科研費基盤B） with 産総研 中村先生
- 人間に准ずる音声認識合成（JST創発）
- 歴史的音声アーカイブ（国語研異分野）
- ダークデータ音声処理（科研費基盤B）
- AIアバター音声合成(Imagica)
- 言語横断音声合成(AI Communis)

積極的な外部交流を

推奨しています！

とにかく音が好きな人は集まれ！

2014年以降の研究教育実績

- 原著論文：Top論文誌IEEE, ASA, EURASIP, ISCA 44編, 他25編
- 国際・国内会議発表：**数えきれないので略**
- 教員・指導学生が2014年以降に**131件の学術賞**を受賞

音の情報処理に興味がある人

統計的信号処理の数理に興味がある人

機械学習を使って生のデータを扱いたい人

一期一会なスモールデータの研究をしたい人

波動現象を数理的に考えたい人 etc.

そんなあなたにお勧めです。