学術フロンティア講義 サイバネティクス入門
―物理・人・社会を繋げる情報科学の先端―
2024/07/02

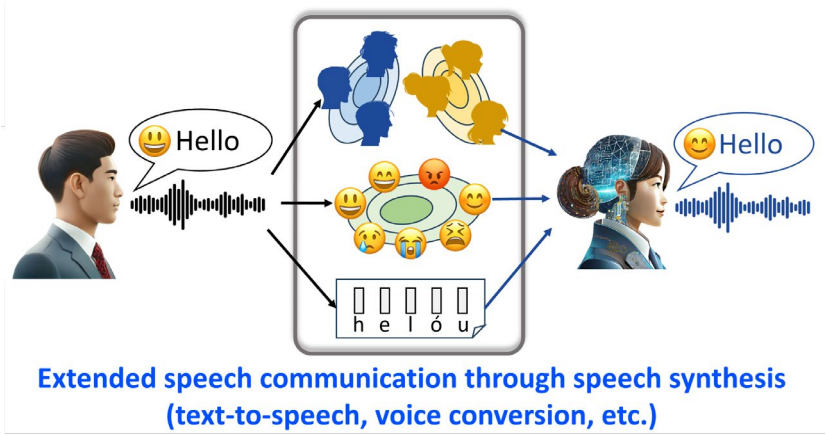
音を解析・合成する信号処理技術

計数工学科 講師
齋藤 佑樹

自己紹介 (齋藤佑樹: SAITO Yuki, PhD)

@ysaito_human

- 現職: 工学部 計数工学科 システム情報工学コース 講師
  - 特任助教 (2021 ~ 2022) → 助教 (2023) → 講師 (現在)
- 出身: 北海道釧路市 (1993 ~ 2016) → 東京 (2016 ~ 現在)
  - 釧路高専 (情報工学科・専攻科) → 東大院 情報理工学系研究科
- 専門: 音声合成・変換, 機械学習
  - コンピュータで人間の声を人工的に作る & 変える技術
  - (流行りの言葉を使うなら)人間の声に関する生成 AI の研究



Extended speech communication through speech synthesis
(text-to-speech, voice conversion, etc.)

計数工学科 システム第1研究室の紹介

- 音声・音響・音楽メディアに関する信号処理・情報処理
- 統計的・機械学習論的信号処理, 数理最適化問題等を研究

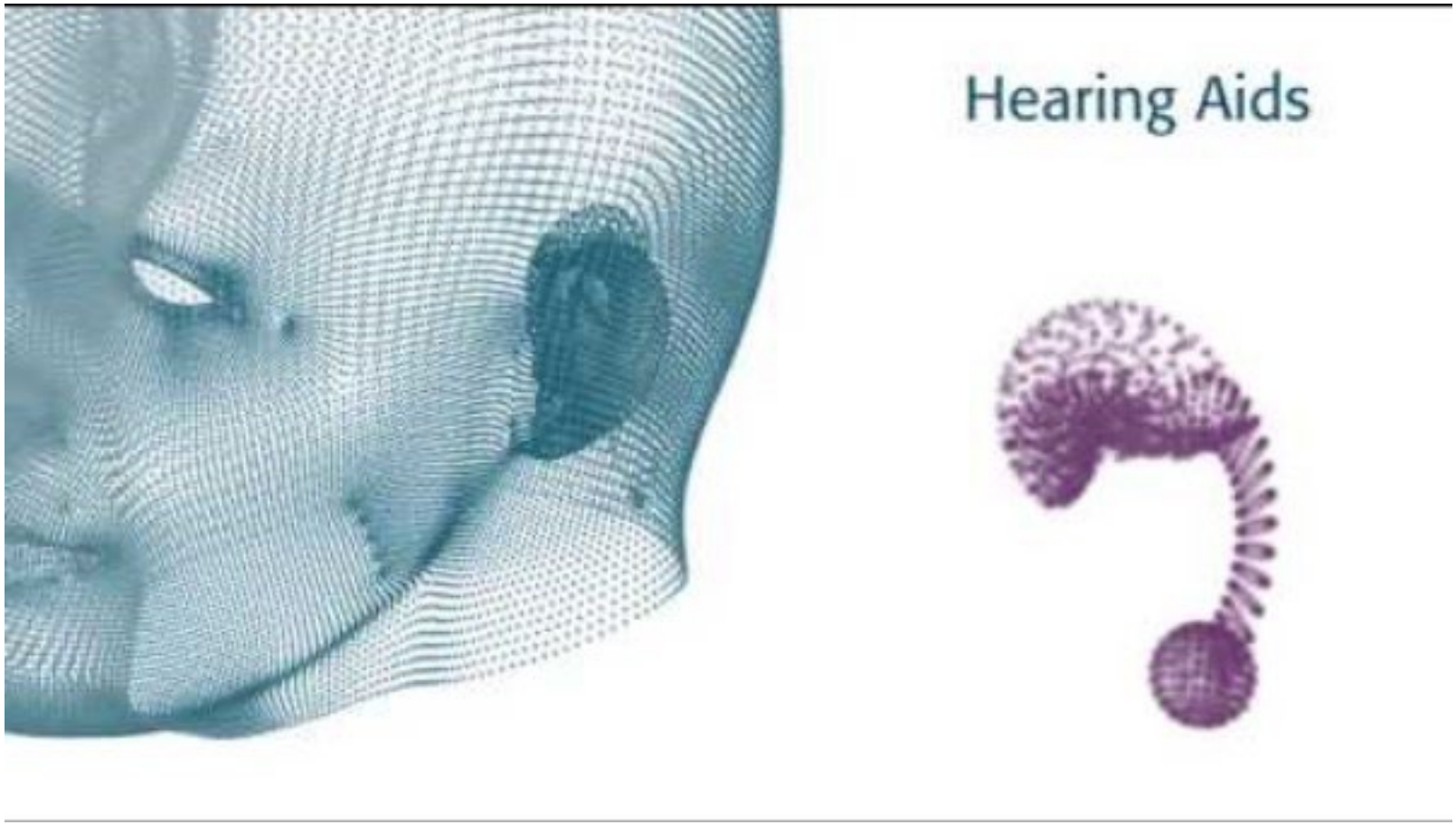
時を超えて蘇る50年前の歌声 ～スモールデータを用いたタスク混合深層学習による歌唱再現～ (2022)

“高道慎之介助教を中心とした研究チームは、(中略)、**歌手の松任谷由実氏が50年前にデビューした当時の歌声を人工再現する技術**を開発しました。(中略)、当時の声色と歌唱表現を忠実に再現することに成功しました。”

https://www.i.u-tokyo.ac.jp/news/files/ist_pressrelease_20221003_takamichi_s.pdf

信号処理ってそもそも何?

コンセプト動画



IEEE Signal Processing Societyによる信号処理の紹介動画
https://www.youtube.com/watch?v=EErkgr1MWw0

カバーする学会

- **IEEE (米国電気電子学会)**
  - アメリカ合衆国に本部を置く, 電気・情報工学分野における世界最大規模の学術研究団体
- **IEEE Signal Processing Society (SPS)**
  - IEEEの信号処理分野のコミュニティ
  - 最初 (1948年) に設立され, 現在は4番目に大きいソサイエティ (Computer, Power and Energy, Communications, Signal Processing)

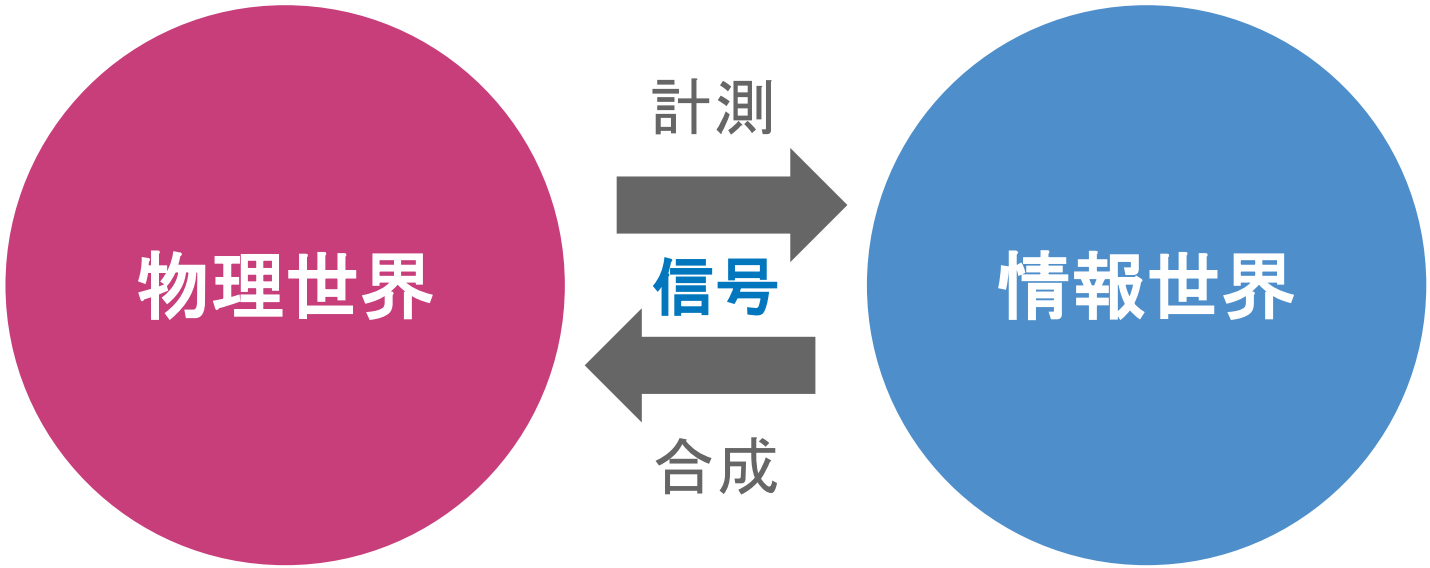
https://signalprocessingsociety.org/

多岐にわたる応用分野
(太字はシステム1研が関連する分野)

- **IEEE Signal Processing Society – Technical Committees**
  - **Audio and Acoustic Signal Processing**
  - Bio Imaging and Signal Processing
  - Computational Imaging
  - Design and Implementation of Signal Processing Systems
  - Image, Video, and Multidimensional Signal Processing
  - Industry DSP Technology
  - Information Forensics and Security
  - **Machine Learning for Signal Processing**
  - Multimedia Signal Processing
  - **Sensor Array and Multichannel**
  - Signal Processing for Communications and Networking
  - Signal Processing Theory and Methods
  - **Speech and Language Processing**

信号とは世界を行き来するもの

- **信号とはセンサ等で計測した物理量の時間的/空間的な変動，あるいはそれを記号として表したもの**
  - 音声, 音楽, 画像, 動画, 超音波ソナー, 電波, 脳波, 筋電位, 地震波, 株価, etc…



信号処理とは,
信号を数理的な手法で分析/加工/合成する技術

入力に対して写像によってなんらかの処理を行い,
出力を生成する.

信号処理の具体例: 雑音除去

入力　　**雑音除去**　　出力

雑音を含む音声信号 → $\mathcal{L}$ → 雑音が除去された音声信号

音声や雑音の性質を利用して
音声のみを抽出するような処理を行う

信号処理の具体例: 音声認識

音声信号のパターンに対して
対応するテキストを出力するような関数を
教師データを用いて学習する

音メディアの重要性

- **コミュニケーション手段としての音**
  - 音声は, 人間同士の最も原始的なコミュニケーションの1つ
  - 遠隔コミュニケーション手段としての電話は, 形を変えながら現在でも広く利用される (収音・再生, 符号化, エコーキャンセラ)
  - ラジオ, テレビ, インターネット放送と変遷する放送メディア (収音・再生, 符号化)
  - 人間とコンピュータ・ロボットがインタラクションするためのインターフェース (音声認識, 理解, 合成, 対話)
- **芸術表現手段としての音**
  - 音楽を始めとして, 音を使った多様な芸術表現が可能
  - 音を記録・再生する技術 (デバイス, 記録メディア, 符号化)
  - 技術発展が新たな芸術表現や文化を生む (楽音・歌声合成・変換)

音楽の創作活動 × 最新の AI

歌声合成
Synthesizer V (2018)

楽曲制作
Suno AI (2023)

音メディアに関する信号処理研究

どのような基準で最適化?
最適化アルゴリズムは?

どのようにセンシングする?
入力信号のモデリングは?

どのような出力デバイス?
出力をどう評価する?

物理音響, 統計モデリング, 機械学習, 聴覚などを総合的に用いて,
音声・音響・音楽信号の性質を上手く取り入れ, 所望のシステムを実現

音源分離 (計算機による選択的聴取)

猿渡 洋
(教授)

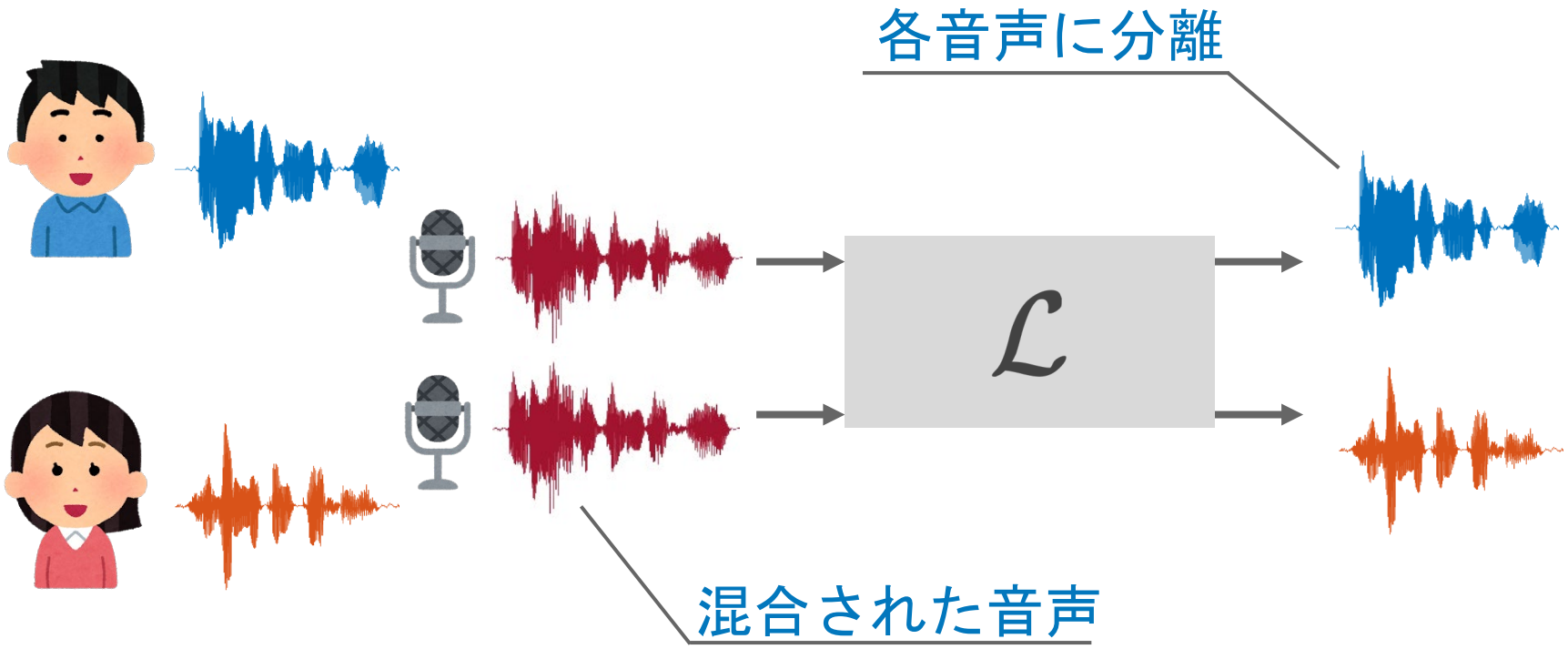
山岡 洸瑛
(助教)

選択的聴取

- **カクテルパーティー効果 (選択的聴取)**
  - カクテルパーティのような騒がしい人混みの中でも, 自分の名前や自分が興味のある会話は, 自然と聞き取ることができる現象

人間の持つこの機能を計算機で表現できるだろうか?
→ (ブラインド) 音源分離

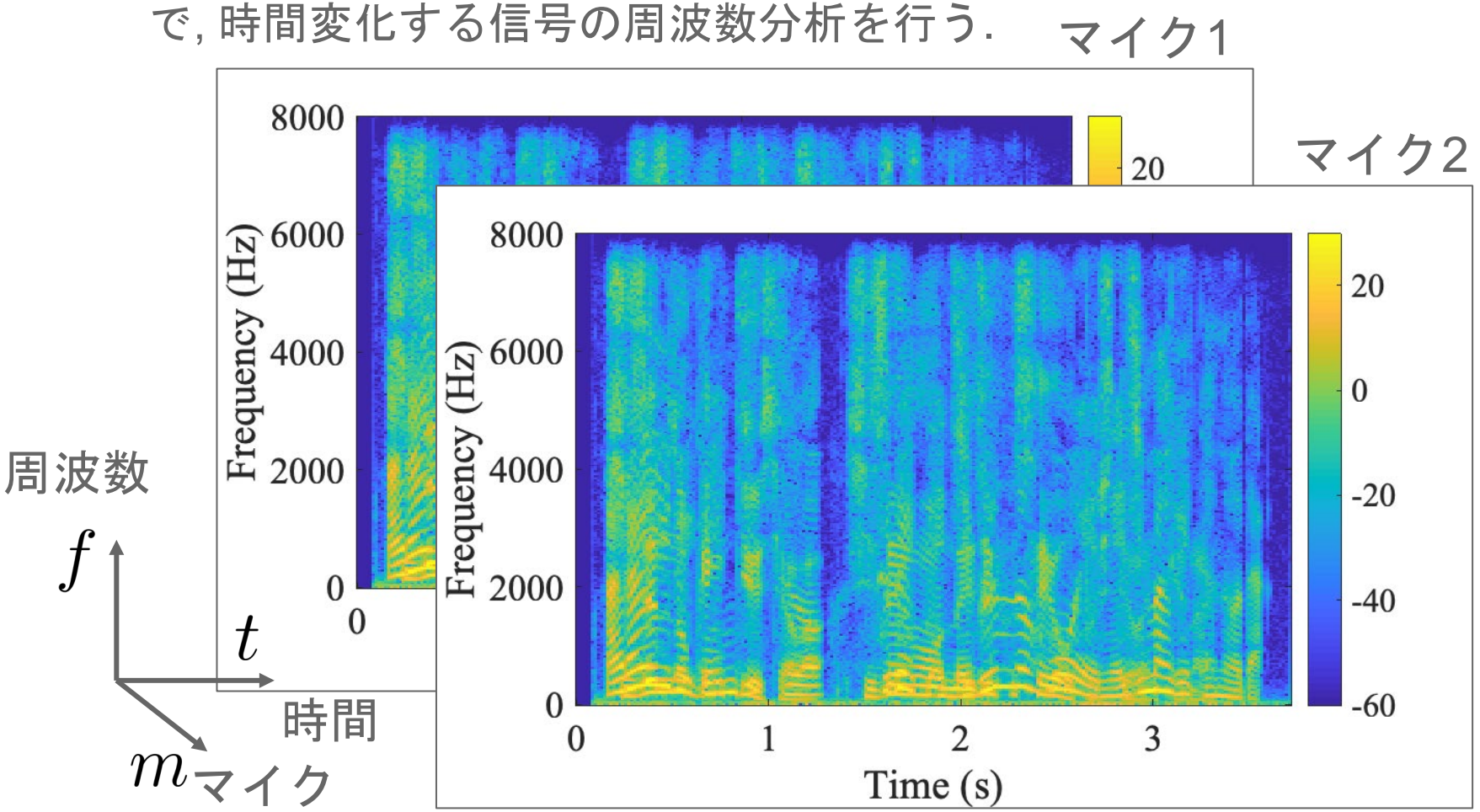
ブラインド音源分離 (Blind Source Separation)



音源やマイクの位置関係などの情報が未知の状態で
(複数の) マイク信号のみから各音源の信号を分離

前準備：
時間周波数領域表現

- **短時間フーリエ変換 (Short-Time Fourier Transform)**
  - 時間信号を短い時間フレームに区切ってフーリエ変換を行うことで, 時間変化する信号の周波数分析を行う.



問題設定:
音源信号の混合 → 混合信号

- J 個の音源信号と M 個のマイクによる観測信号が, 時間周波数領域で以下のように関係付けられるとする.

$$x_{ft,m} = \sum_{j=1}^{J} a_{f,mj} s_{ft,j}$$

音源信号

観測信号

伝達関数

$$\begin{cases} m : \text{マイク} \\ j : \text{音源} \\ t : \text{時間} \\ f : \text{周波数} \end{cases}$$

- 行列形式で書けば,

$$\boldsymbol{x}_{ft} = \boldsymbol{A}_f \boldsymbol{s}_{ft}$$

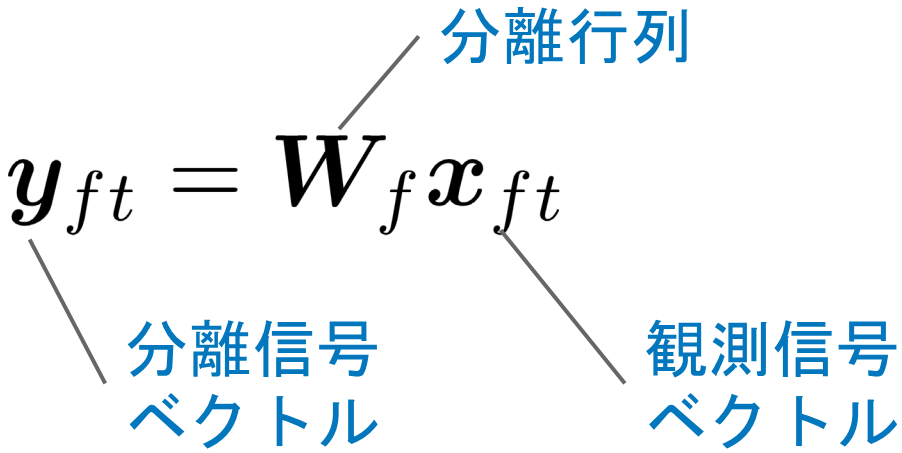
伝達関数行列 (混合行列)

観測信号ベクトル

音源信号ベクトル

問題設定:
混合信号の分離 → 分離信号

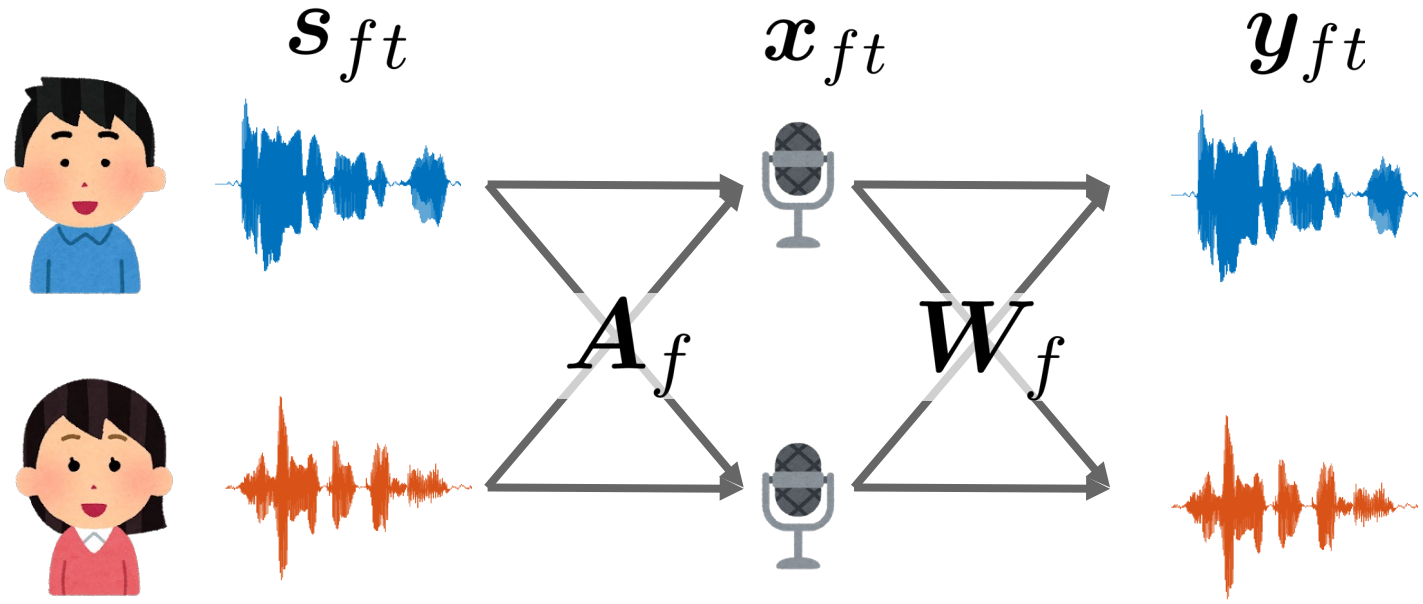
- 混合した観測信号 $\boldsymbol{x}_{ft} \in \mathbb{C}^M$ を各音源の信号 $\boldsymbol{y}_{ft} \in \mathbb{C}^J$ に分離するための分離行列 $\boldsymbol{W}_f \in \mathbb{C}^{J \times M}$ を求めたい.



- 簡単のため, $J = M$ の場合のみを考えることとし, 分離行列 $\boldsymbol{W}_f$ は正方行列とする.

問題設定:
音源分離で推定するもの

- 混合行列 $\boldsymbol{A}_f$ が未知の状況でマイク信号 $\boldsymbol{x}_{ft}$ のみから各音源を分離する分離行列 $\boldsymbol{W}_f$ を推定する.



どうやって$\boldsymbol{W}_f$を推定する?

独立成分分析 (ICA):
分離行列を推定するために

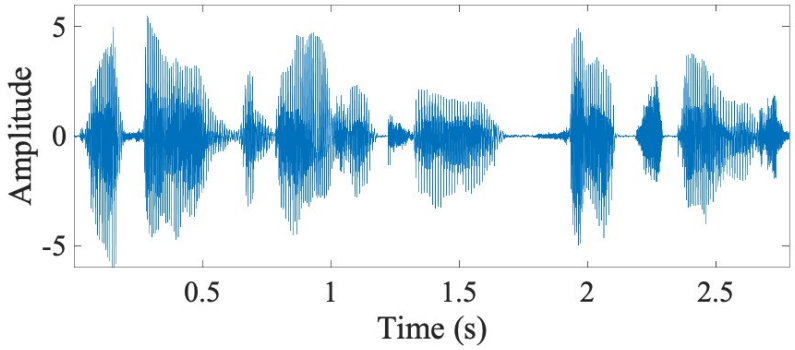
- **独立成分分析 (I**ndependent **C**omponent **A**nalysis**)**
  - 混合された音源が統計的に独立であるという仮定の下で分離行列を推定する.

$$p(\boldsymbol{y}_{ft}) = p(y_{ft,1}, \ldots, y_{ft,J}) = \prod_{j=1}^{J} p(y_{ft,j})$$

- **統計的な独立性**
  - 無相関性/白色性よりも強い仮定であり,
    優ガウスな分布に従う信号源の混合を分離することを可能にする.
  - ICA に基づく BSS は様々な評価尺度を用いて実現されているが,
    ここでは音源分離においてよく用いられる,
    最尤推定による ICA を紹介する.

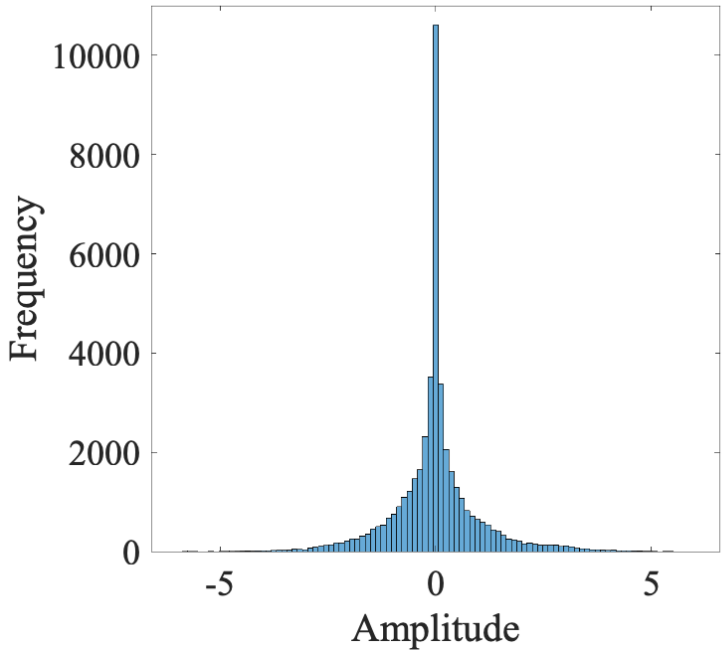
なぜ統計的独立性を使う?:
音源数が1の場合

音声波形

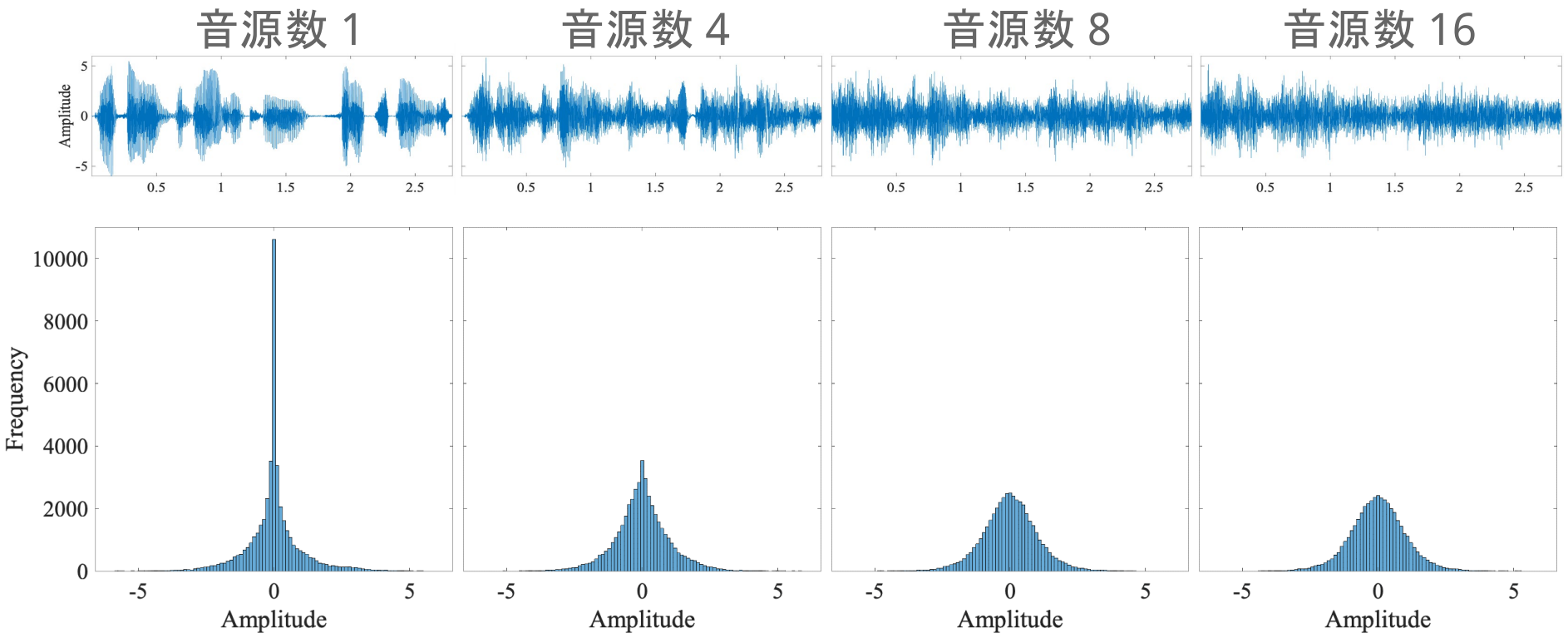


振幅値 (縦軸)の
ヒストグラムを
計算すると...

振幅値の
ヒストグラム



なぜ統計的独立性を使う?:
音源数が増える (= 複数の音声を混合する) と…



混合数を増やすと正規分布に近づく

分離を実現するにはこの逆を行うことが必要

最尤推定による ICA

- 分離行列 $\boldsymbol{W}_f$ の尤度関数を考えると，

$$\begin{aligned}\mathcal{L}(\boldsymbol{W}_f) &= p(\boldsymbol{x}_{f,1}, \ldots, \boldsymbol{x}_{f,T} | \boldsymbol{W}_f) \\ &= \prod_{t=1}^{T} p(\boldsymbol{x}_{ft} | \boldsymbol{W}_f) \\ &= \prod_{t=1}^{T} |\det \boldsymbol{W}_f|^2 \, p(\boldsymbol{y}_{ft})\end{aligned}$$

- ここで線形変換 $\boldsymbol{y}_{ft} = \boldsymbol{W}_f \boldsymbol{x}_{ft}$ の確率密度に関する以下の関係式を用いた．

$$p(\boldsymbol{x}_{ft} | \boldsymbol{W}_f) = |\det \boldsymbol{W}_f|^2 \, p(\boldsymbol{y}_{ft})$$

最尤推定による ICA

- 音源信号が互いに独立であると仮定すれば,

$$p(\boldsymbol{y}_{ft}) = \prod_{j=1}^{J} p(y_{ft,j})$$

- $\boldsymbol{W}_f$ の負の対数尤度を考えると,

$$\begin{aligned}
\mathcal{J}(\boldsymbol{W}_f) &= -\log \mathcal{L}(\boldsymbol{W}_f) \\
&= -\log \left\{ \prod_{t=1}^{T} |\det \boldsymbol{W}_f|^2 \prod_{j=1}^{J} p(y_{ft,j}) \right\} \\
&= \sum_{t=1}^{T} \sum_{j=1}^{J} G(y_{ft,j}) - 2T \log |\det \boldsymbol{W}_f|
\end{aligned}$$

$$G(y_{ft,j}) = -\log p(y_{ft,j})$$

最尤推定による ICA

- $G(y_{ft,j})$ はコントラスト関数と呼ばれ, 音源信号が従うと仮定できる確率密度関数に基づいて設定する必要がある.

- 音声・音響信号では, 優ガウスな分布として以下のような分布を用いる場合が多い.
  - α , β は非負のパラメータ

$$p(y_{ft,j}) \propto \exp\left(-\frac{\sqrt{|y_{ft,j}|^2+\alpha}}{\beta}\right)$$

- $\mathcal{J}(\boldsymbol{W}_f)$ を最小化するように $\boldsymbol{W}_f$ を求める.
  - 例えば勾配法などの反復法

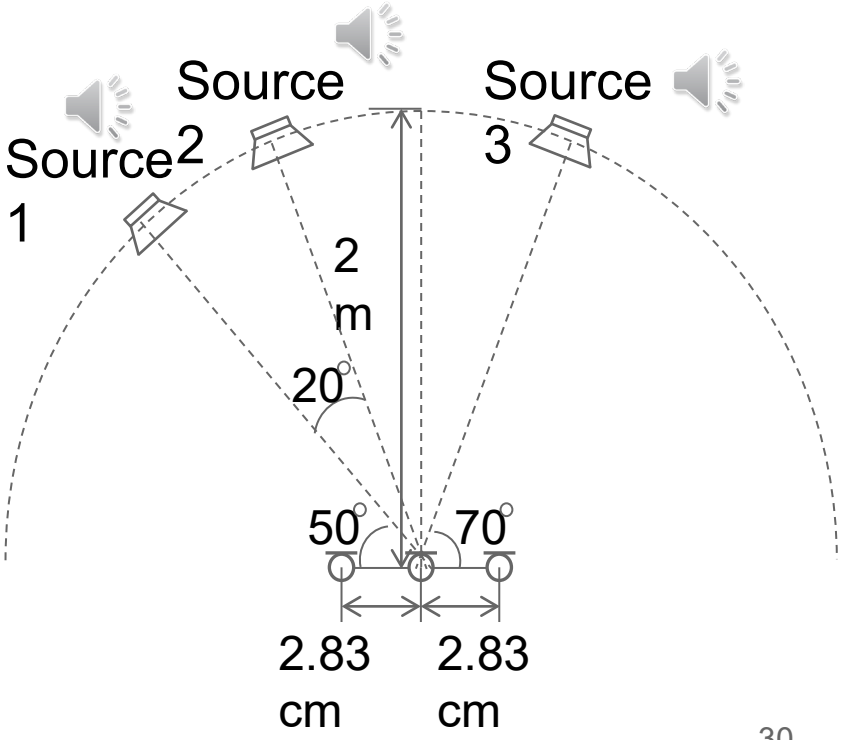
より発展的な内容: 独立ベクトル分析, 補助関数法

- **パーミュテーション問題**
  - コスト関数 $\mathcal{J}(\boldsymbol{W}_f)$ を最小化する $\boldsymbol{W}_f$ を求めることで分離が達成できると考えられるが, 分離行列 $\boldsymbol{W}_f$ は周波数ごとに別々に求まるため, 音源の順序に関する任意性が残る.
    - パーミュテーション問題を解決する方法の一つとして, 音源信号の各要素ではなく，全周波数の要素を並べたベクトルの独立性を仮定する, 独立ベクトル分析 (**I**ndependent **V**ector **A**nalysis) がある
- **最適化**
  - 勾配法はコスト関数の減少を保証しない
    (ハイパーパラメータの設定によってはそもそも収束しない)
  - → **M**ajorization-**M**inimization アルゴリズム
    (非線形最適化問題を効率的に解く手法) がよく用いられる

音源分離デモ

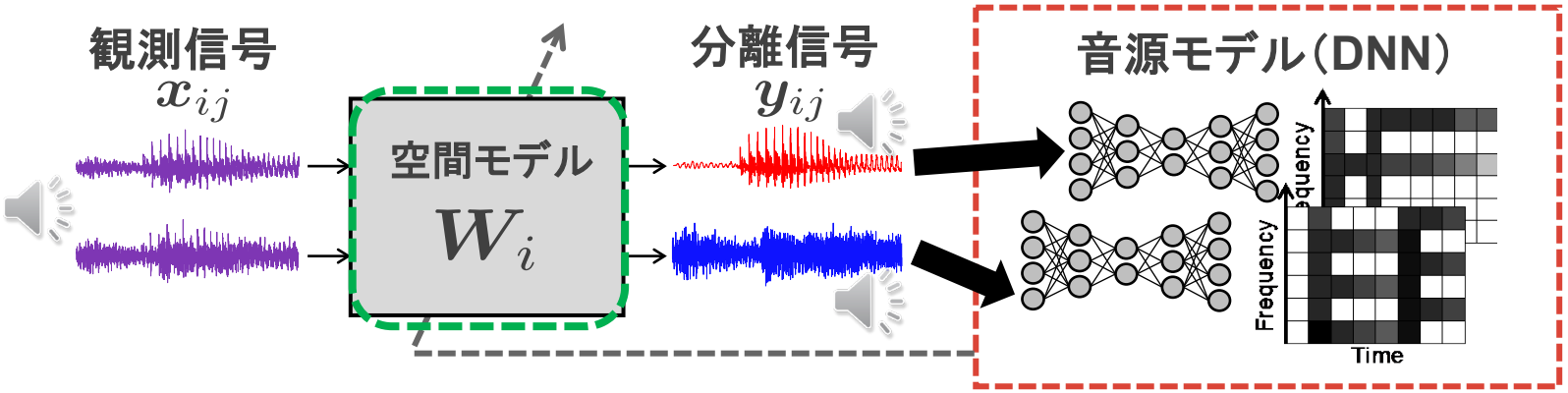
高速 ICA による
リアルタイム聞き分け

楽音 (ドラム, 弦楽器, 歌声) の
音源分離



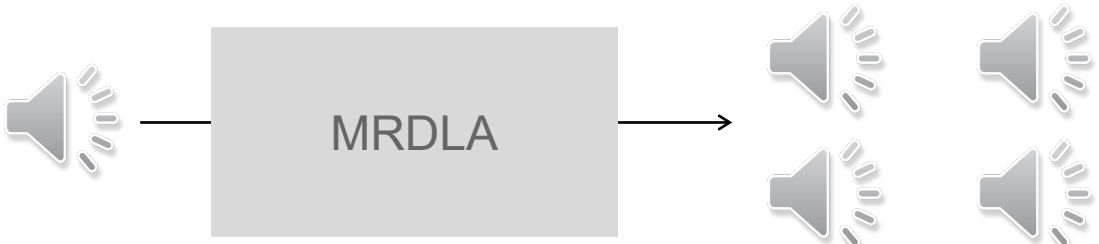
より発展的な内容: 深層学習ベースの音源分離

- **独立深層学習行列分析 (Independent Deeply-Learned Matrix Analysis)**
  - 空間情報はブラインドに推定し, 音源情報は深層学習でモデル化



Makishima: “Independent Deeply Learned Matrix Analysis for Determined Audio Source Separation,” IEEE/ACM TASLP, 2019.

- **多重解像度深層分析 (MultiResolution Deep Layered Analysis)**
  - 最先端の深層学習と古典的な信号処理 (離散ウェーブレット変換) を融合させ, 高性能な音源分離を実現



Nakamura: “Time-Domain Audio Source Separation With Neural Networks Based on Multiresolution Analysis,” IEEE/ACM TASLP, 2021.

音メディアの合成・変換 (計算機による音生成)

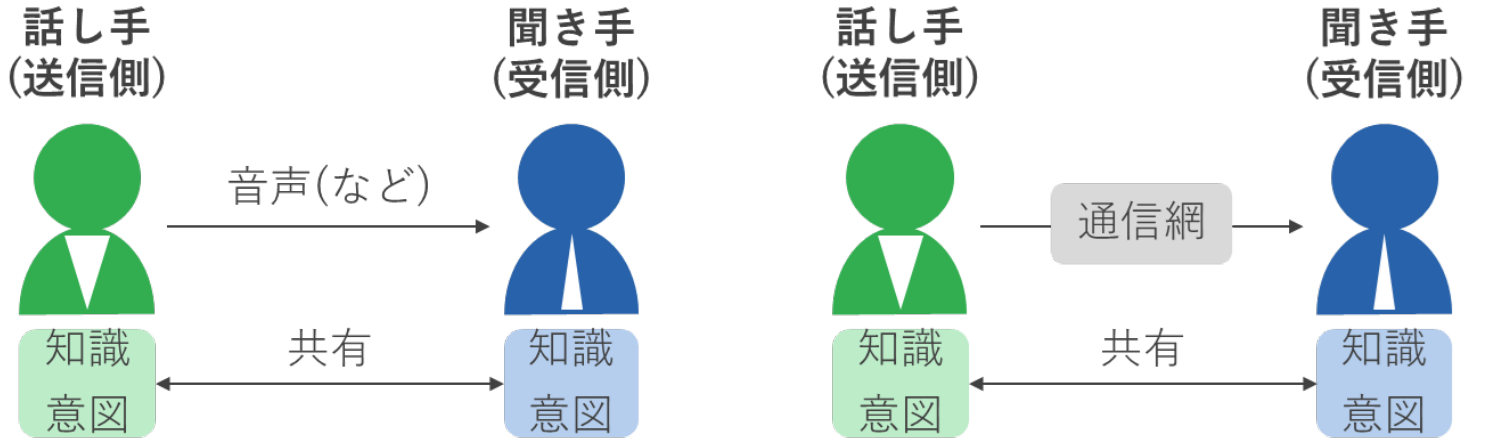
齋藤 佑樹
(講師)

岡本 悠希
(特任助教)

高道 慎之介
(特任准教授)

人間の音声コミュニケーション

- **人間の音声コミュニケーション**
  - 話し手は, 自らの意図, 感情, 話者属性などを音声波形に変換する
  - 聞き手は, 話し手の意図, 感情, 話者属性を音声波形から推定する



(a) 人と人の対面コミュニケーション　(b) 人と人の遠隔コミュニケーション

- **計算機はこれを再現できるだろうか?**
  - 音声認識 … 相手の音声波形から内容 (言語内容に限らない) を推定
  - 音声言語理解 … 相手の内容を理解して, 自分の内容を計画
  - 音声合成 … 自分の内容から音声波形を生成

発声メカニズムのモデル: ソース・フィルタモデル

声道 (フィルタ)

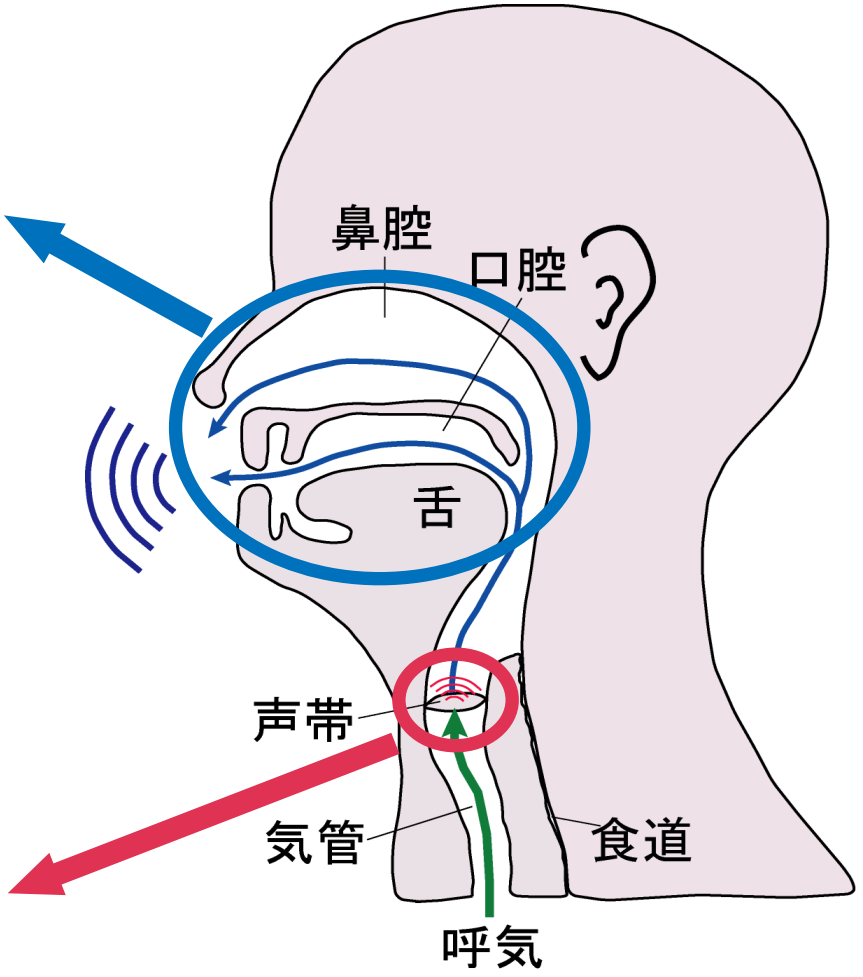
口や喉を動かして
音源信号を音色付け

音声波形

畳み込み

音源 (ソース)

肺からの呼気を
声帯の開閉で制御



音声波形を数式で表すと

音声生成の式 (周波数領域における掛け算 = 時間領域なら畳み込み演算)

音源		フィルタ		(チャネル)		音声
$S(f,\ t)$	$\times$	$F(f,\ t)$	$\times$	$C(f,t)$	$=$	$X(f,\ t)$
声の高さ		声色		空間伝搬・マイク特性		周波数 時刻

注意: 次のページで体内を映した映像が出てきます.
苦手な方は音だけを聞いて下さい.

音声波形を数式で表すと

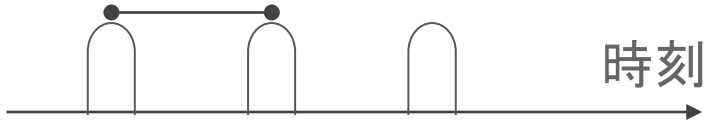
音声生成の式 (周波数領域における掛け算 = 時間領域なら畳み込み演算)

音源 $S(f,\ t)$ × フィルタ $F(f,\ t)$ × (チャネル) $C(f,t)$ = 音声 $X(f,\ t)$

声の高さ　声色　空間伝搬・マイク特性　周波数　時刻

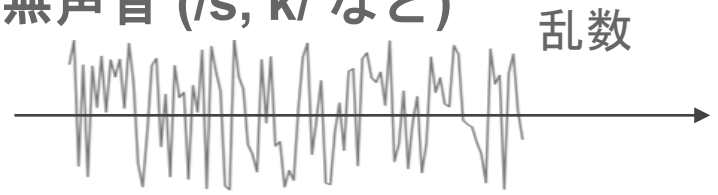
有声音 (/a, i/ など)

声門の開閉周期に対応



振幅スペクトルは櫛形 (スパース)

無声音 (/s, k/ など)



振幅スペクトルは連続的 (デンス)

音声波形を数式で表すと

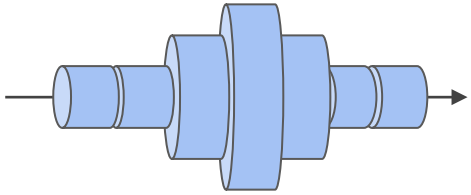
音声生成の式 (周波数領域における掛け算 = 時間領域なら畳み込み演算)

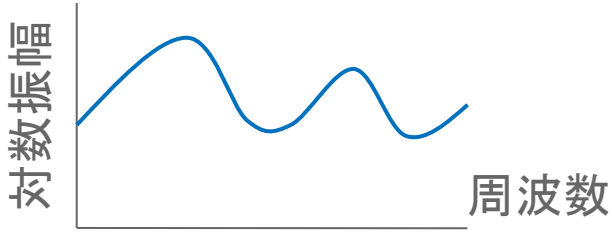
$$\underbrace{S(f,t)}_{\text{音源}} \times \underbrace{F(f,t)}_{\text{フィルタ}} \times \underbrace{C(f,t)}_{\text{(チャネル)}} = \underbrace{X(f,t)}_{\text{音声}}$$

声の高さ　　**声色**　　空間伝搬・マイク特性　　周波数　時刻

音響管の連接でモデル化可能

自己回帰過程を仮定



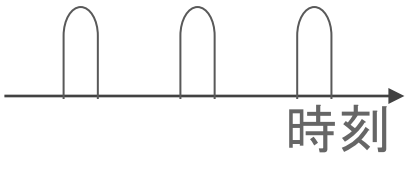


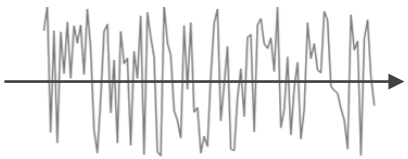
振幅スペクトルは連続的 (デンス)

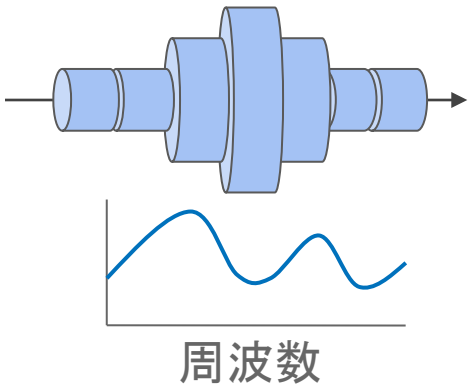
まとめると

音声生成の式 (周波数領域における掛け算 = 時間領域なら畳み込み演算)

音源		フィルタ		(チャネル)		音声
$S(f,\ t)$	×	$F(f,\ t)$	×	$C(f,t)$	=	$X(f,\ t)$
声の高さ		声色		空間伝搬・マイク特性		周波数　時刻

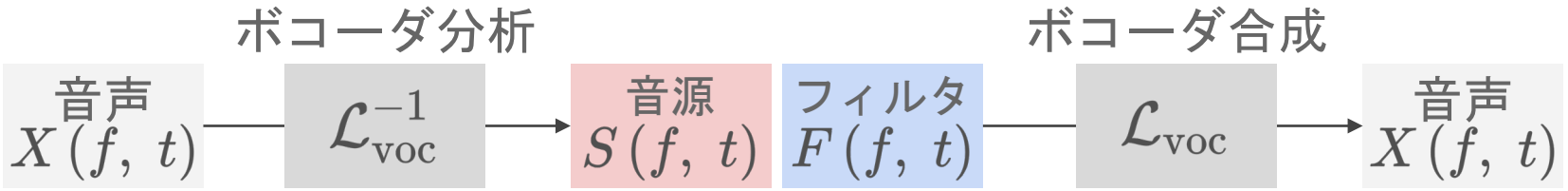






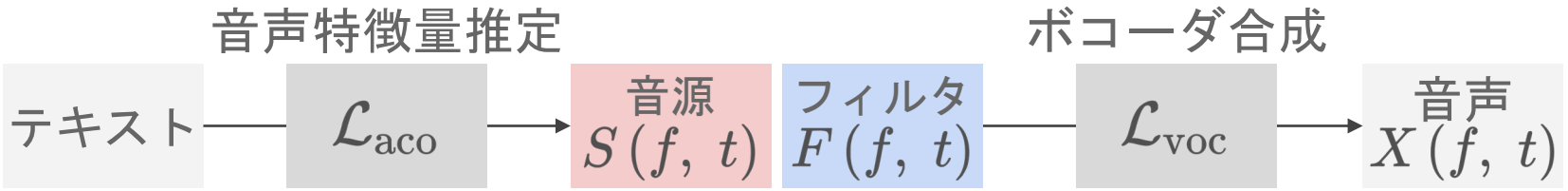
音源とマイクが離れているなら，距離の2乗で減衰

音声分析・合成 (ボコーダとも呼ばれる)



- **分析は逆問題の一種**
  - １つの観測 (音声波形) からその内部要素 (音源・フィルタ) を推定
- **音声波形の統計的性質や人間の聴覚特性を利用する. 例えば,**
  - 音声波形はミクロ的 (20 ~ 25 ms) では定常
    - 時間解像度を落としたパラメータ表現
  - 音源信号は, 周期信号と非周期信号の重み和でモデル化可能
    - 周期, 重みのみで音源信号を表現
  - 人間の聴覚は, 音声波形の位相に鈍感
    - 振幅スペクトル (各周波数の成分の大きさ) のみのモデル化

テキスト音声合成



- **音声を制御するテキストの要因は?**
  - 音素 … 言語音の最小単位 (/a, i, u, e, o/)
  - アクセント … 橋, 端, 箸は高さのパターン (アクセント) が異なる
  - 音調 … 発話のニュアンスなど. 例えば同じ疑問符でも抑揚が異なる
- **音声特徴量推定**
  - 文字列から音声特徴量列への変換 (sequence-to-sequence 変換)
  - 昨今の深層学習が大活躍. 特に系列変換モデル, 生成モデル
    - Attention, Transformer, GAN, Diffusion, etc.

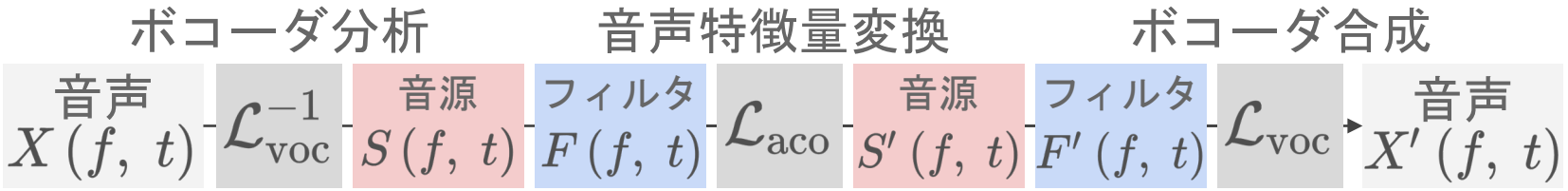
“ざっくりいうと, 先ほど少しお話ししましたけども, 戦後のそういうサブカルチャーのイメージという…”

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…

人間のように非流暢になる音声合成

Matsunaga: “Empirical study incorporating linguistic knowledge on filled pauses for personalized spontaneous speech synthesis,” APSIPA ASC, 2022.

声質変換

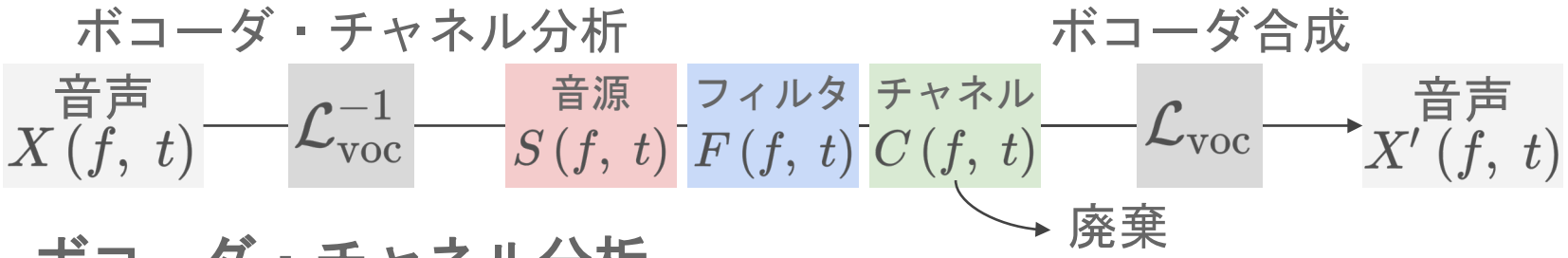


- **音声特徴量推定**
  - 例えば話者変換ならば, 話者Aの特徴を話者Bの特徴に変換
  - 深層学習が大活躍. 特に系列変換モデル, 生成モデル
    - 話者やニュアンスの変換に, 深層学習の非線形性変換が合う

松任谷由実 & AI 荒井由実
Call me back (2022紅白)

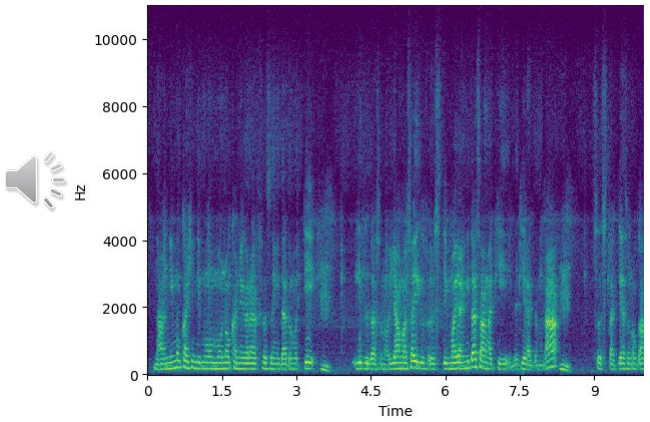
リアルタイム音声変換ソフト
Paravo (2024)

音声復元

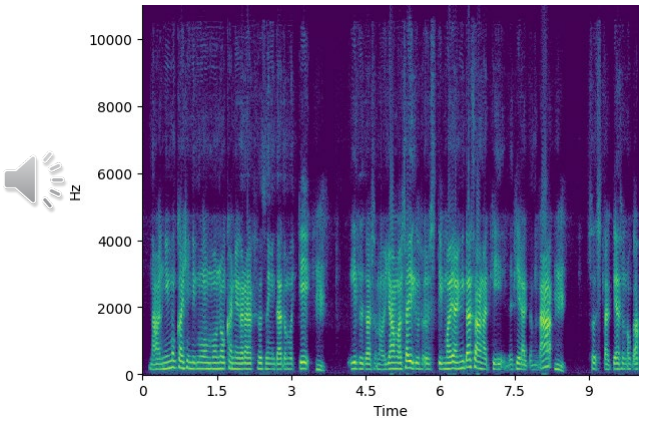


- **ボコーダ・チャネル分析**
  - ボコーダ分析よりも, 更に多くのパラメータを含む逆問題
  - 音声特徴量 (音源・フィルタ), 音響特徴量 (チャネル) の統計的性質の違いを利用

1960年代に録音された昔話音声



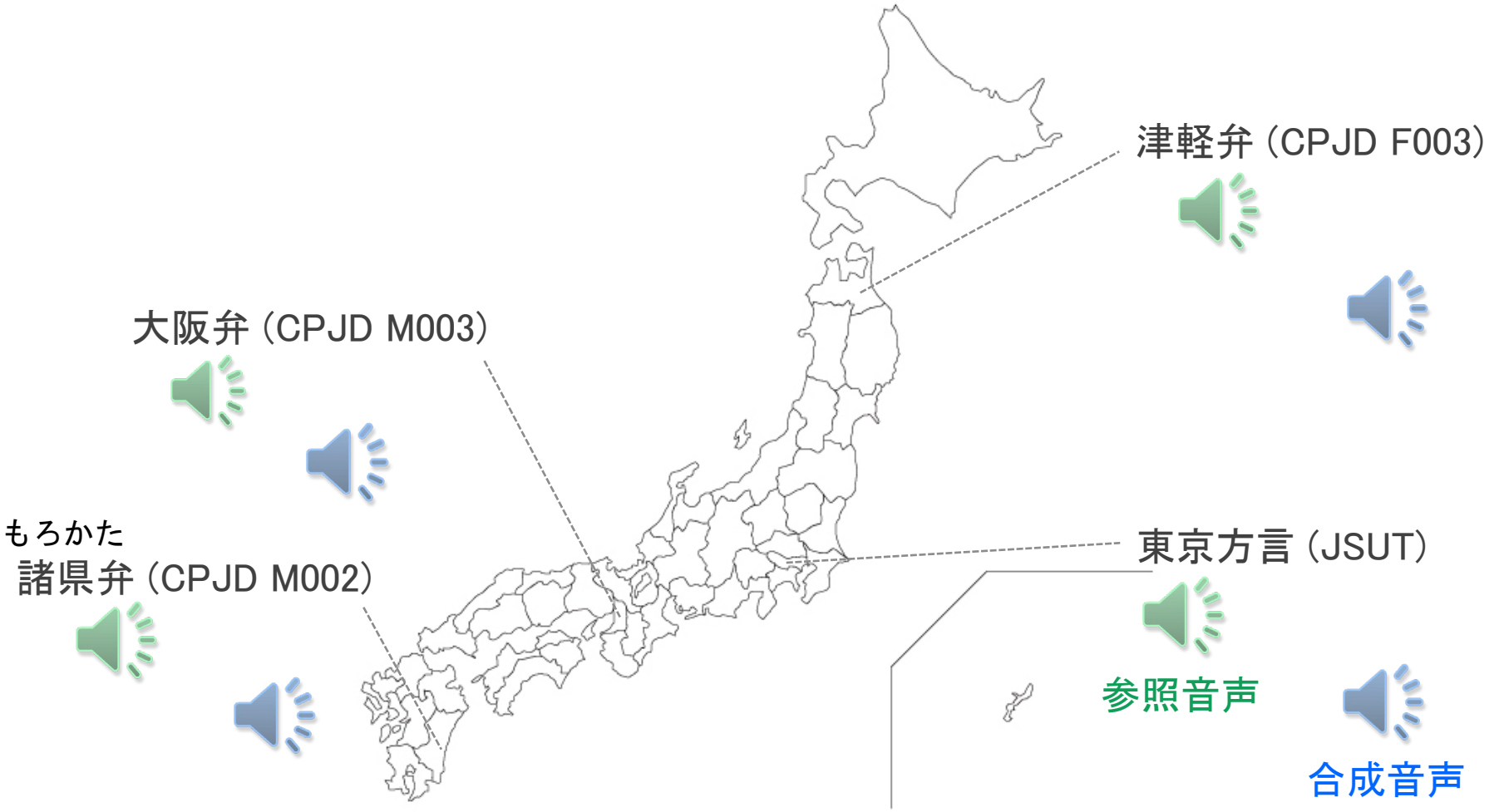
機械学習で復元した音声



本質的に「正解」がない問いに対する機械学習論的アプローチ

Saeki: “SelfRemaster: Self-Supervised Speech Restoration for Historical Audio Resources,” IEEE Access, 2024.

方言音声合成



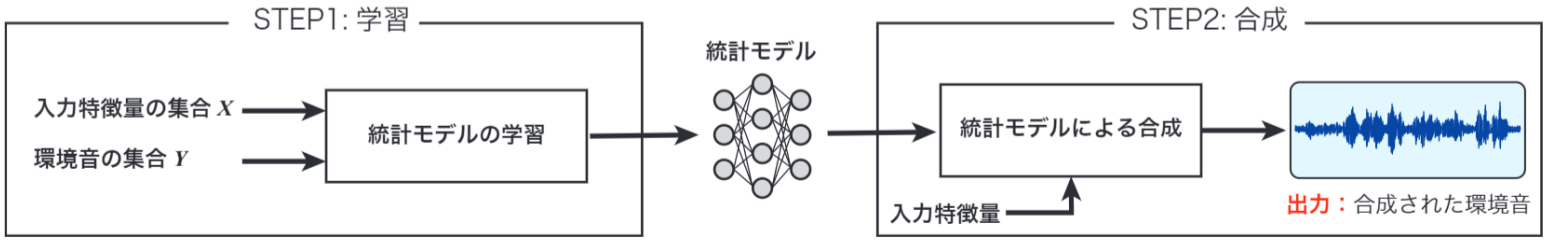
話者の方言の違いを超えた音声コミュニケーションを実現

「音声」の合成を超えて… 環境音の合成

- **環境音とは: 我々の身の周りに存在するあらゆる音**
  - 「車の音」「目覚まし時計の音」「雨の音」など
  - 「音声」や「楽音」も含む
- **環境音は多くの場面において重要な役割を持つ**
  - メディアコンテンツ作品への活用
    - アニメや映画などで場面・状況を説明 & 登場人物の心象を表現
- **環境音合成: あらゆる環境音を人工的に生成する技術**
  - メディアコンテンツにおける効果音, 背景音の生成や,
    高度な音響 VR・メタバースサービスの実現

環境音合成の問題設定

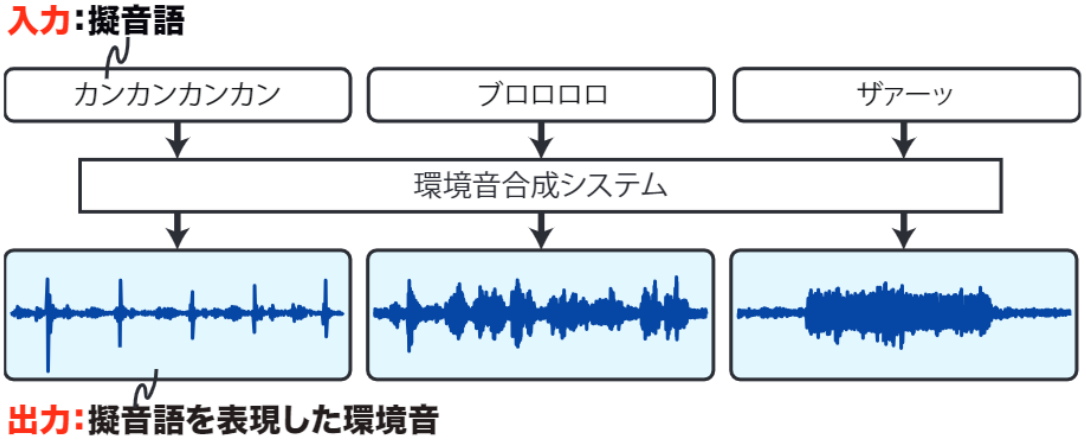
- **機械学習に基づく環境音合成**
  - 何らかの入力特徴量から，それに対応する環境音を合成する統計モデルを学習
  - 近年では，大量の環境音データを用いた深層学習が主流に

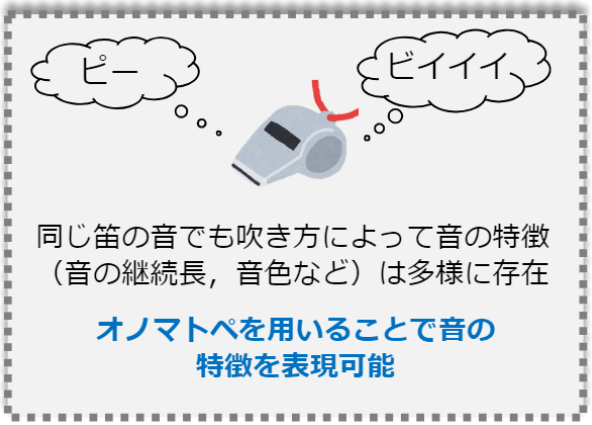


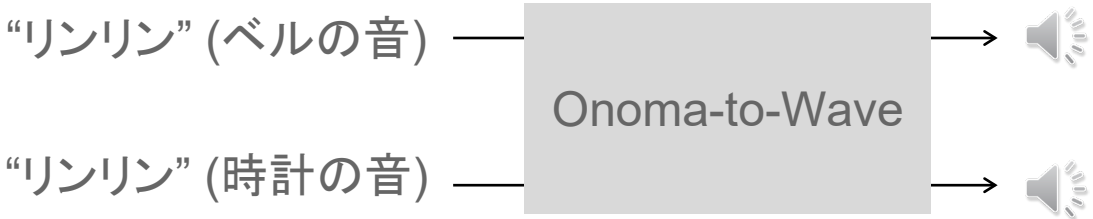
特徴量	音高	リズム	時間的な音の変化	音の種類
音響イベントラベル [Kong+ 2021]				✅
画像 [Zhou+ 2018]				✅
オノマトペ [Okamoto+ 2022]			✅	
音の説明文 [Liu+ 2023]			✅	✅
音声 [Okamoto+ 2024]	✅	✅	✅	

オノマトペ文字列からの環境音合成 (Onoma-to-Wave)

- **合成したい環境音の特徴をオノマトペで表現**
  - オノマトペ: 音の特徴を自然言語で記述したもの
- **オノマトペを入力特徴量とすることで, 多様な音の特徴を表現**
  - 時間構造 (音の繰り返し etc.) や音色などの制御を期待



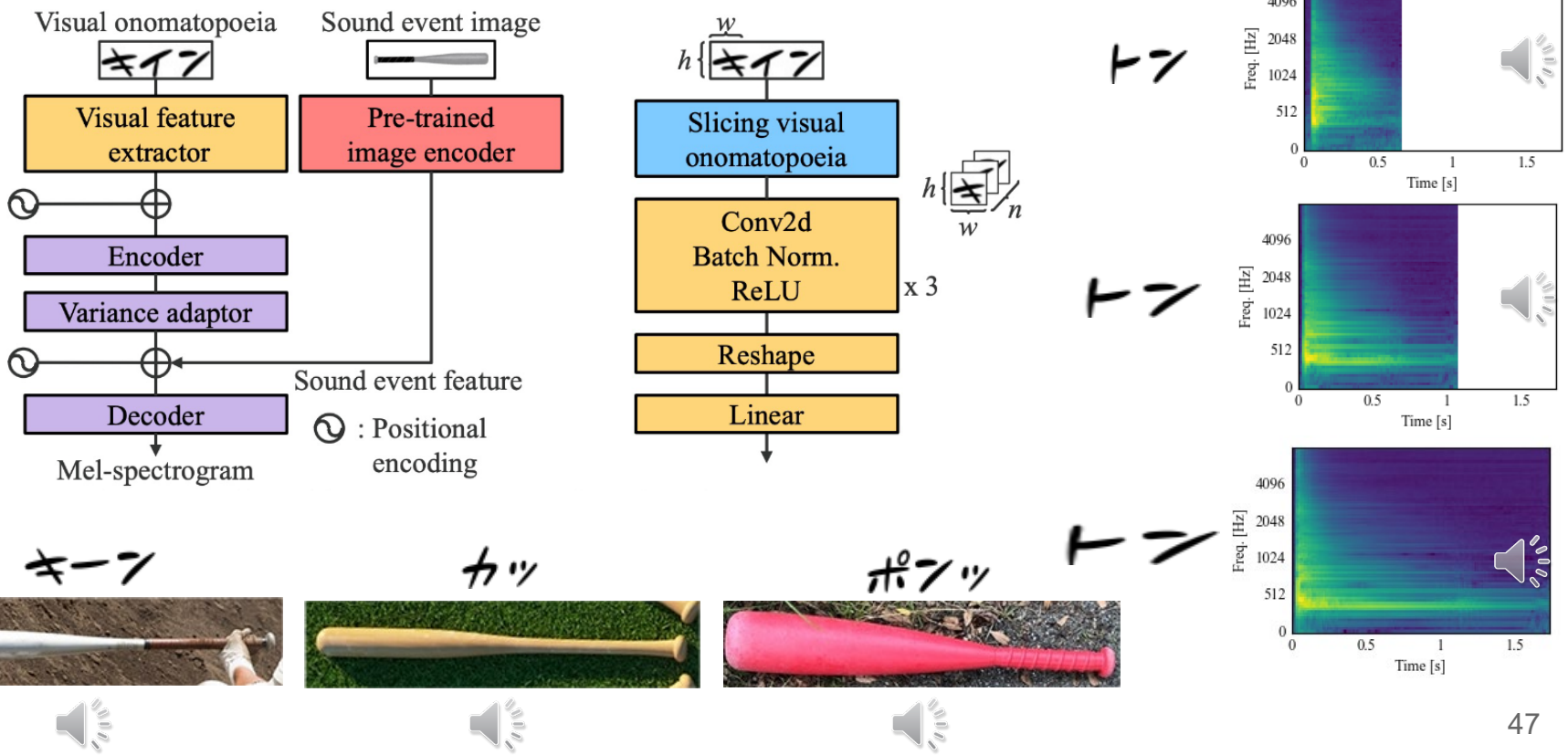




Okamoto: “Onoma-to-wave: Environmental sound synthesis from onomatopoeic words,” APSIPA TSIP, 2022.

オノマトペ画像からの環境音合成 (Visual Onoma-to-Wave)

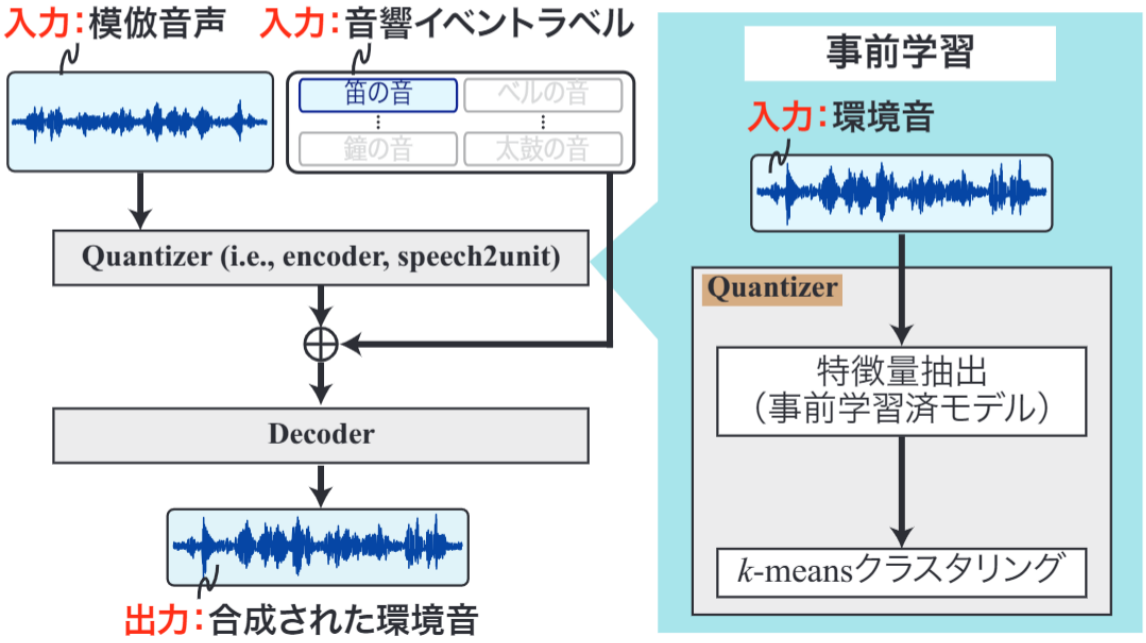
- **文字列ではなく, オノマトペの画像から環境音を直接合成**
  - 音の引き延ばしなどの制御をより直感的に実現可能
  - 環境音を表すイメージ画像での条件付けも可能

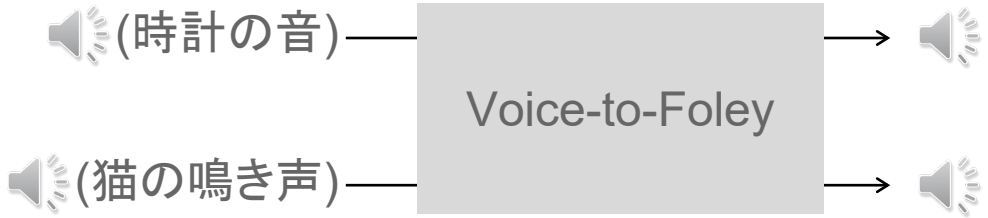


Ohnaka: “Visual onoma-to-wave: environmental sound synthesis from visual onomatopoeias and sound-source images,” ICASSP, 2023.

声真似音声からの環境音合成 (Voice-to-Foley)

- **環境音を人間が声真似した音声から環境音を合成**
  - 音の高さやリズムなどの制御をより直感的に実現可能





Okamoto: “Environmental Sound Synthesis from Vocal Imitations and Sound Event Labels,” ICASSP, 2024.

様々なメディアからの環境音合成

音の説明文 (自然言語記述)

The sound of a steam engine.

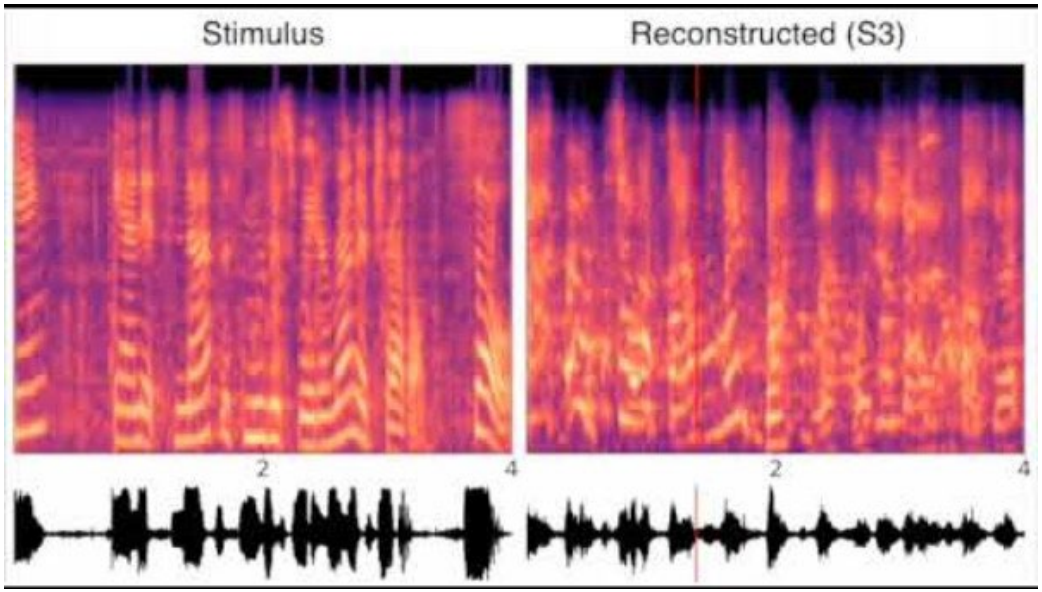
A man is speaking in a huge room.

動画

オリジナル

動画 w/ 合成結果

脳波



本発表のまとめ & これからの音メディア処理

まとめ

- **音を解析・合成するための信号処理**
  - 信号処理 = 信号を数理的な手法で分析/加工/合成する技術
  - 人間のコミュニケーション拡張・創作活動支援など

- **音源分離 = 計算機による選択的聴取**
  - 統計的独立性に基づくブラインド音源分離
  - 深層学習との融合による高精度化

- **音メディアの合成・変換 = 計算機による音生成**
  - 音声合成・変換
  - 環境音合成

これからの音メディア処理

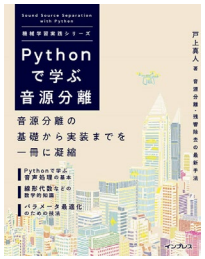
- **計測**
  - 必ずしもビッグデータであるとは限らない (例えば音空間の計測)
  - 一方でセンサ・トランスデューサの数は爆発的に増え, 安価に (cf. トリリオン・センシング)

- **解析**
  - 信号処理, ドメイン知識, 物理音響, 統計的機械学習の高度な融合

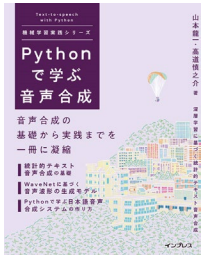
- **合成**
  - 合成された音信号を享受するのは, 多くの場合で人間
  - 人間と音メディア処理の相互作用 (human-in-the-loop, AI-in-the-loop)

この分野を勉強するなら

- 川村, 音声音響信号処理の基礎と実践, 2021
  - 音メディア信号処理の基礎を丁寧に説明
  - 音を題材に信号処理を勉強するのにオススメ



- 戸上, Pythonで学ぶ音源分離, 2020
  - 信号処理に基づく音源分離を基礎から最先端まで説明
  - 理論を学ぶための数学と，実践するための Python コード



- 山本&高道, Pythonで学ぶ音声合成, 2021
  - 機械学習に基づく音源合成を基礎から最先端まで説明
  - 日本語の言語体系解説と，実践するための Python コード

講義の質問, 感想など → yuuki_saito@ipc.i.u-tokyo.ac.jp