国立国語研究所 異分野融合型共同研究
2024年度 ワークショップ　(2025/02/27)

歴史的音源アーカイブに向けたオープンコーパスの整備とAI 音声復元技術の開発

歴史的音源のみから学習可能な音声復元技術

高道 慎之介[1,2], 佐伯 高明[1], 中田 亘[1]
(1: 慶大, 2: 東大)

Takamichi Laboratory
慶應義塾大学 高道研究室

プロジェクト目標

- **目標1：AI 音声復元技術の開発← 本内容**
  - 実際の歴史的音源に耐えうる音声復元技術の開発
  - オープンソース化

- **目標2：東北方言昔話コーパスの開発**
  - 東北方言昔話音源のディジタル化
  - ディジタル化音源のアノテーション
  - オープンソース化

問題設定

- やりたいこと：歴史的音源を現代音質に復元
  - 音のうち，音声はそのままに音の歪みを除去したい
    - 音の歪み＝ ”悪い音っぽさ”

- 考え方は２つ.
  - 歴史的音源を，音声成分と歪み成分に分離 (昨年度に発表)

  - 歴史的音源から，音声成分を抽出 (今回新しく試行)

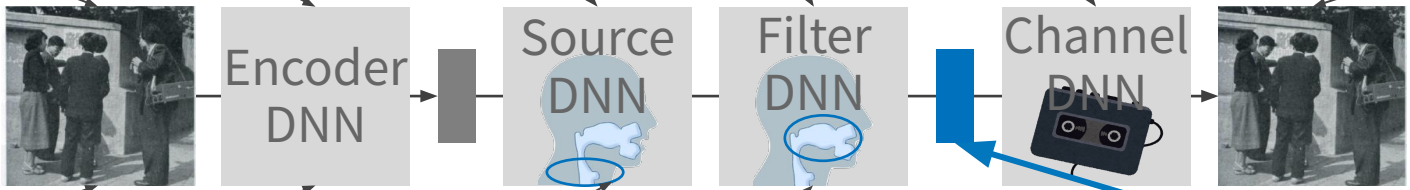
方法1：歴史的音源を，音声成分と歪み成分に分離 (昨年度に発表)

歴史的音源  音声   歪み

音声復元のための自己教師あり学習 (深層ニューラルネット [DNN] を使った機械学習)

- ソース・フィルタ・チャネル分解に基づく自己教師あり学習
  - 自己教師あり学習：データ（歴史的音源）自身から，意味の有る特徴を取り出す機械学習

(学習時)
劣化音声を要素分解して，ソース・フィルタ・チャネル過程を通して再構成

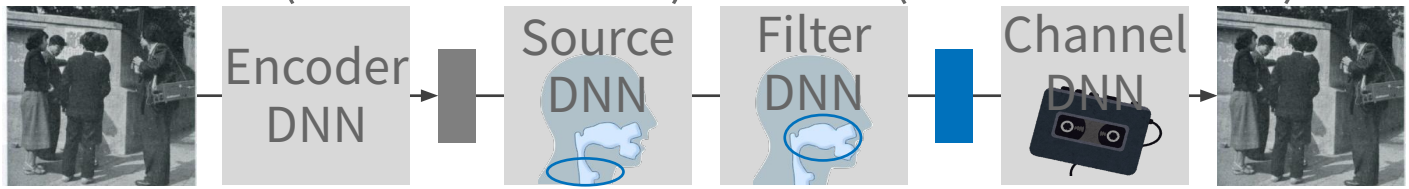


劣化音声を要素分解して，ソース・フィルタ過程だけを通し高品質音声を復元
(復元時)

提案法の動作

ソース・フィルタ・チャネルの
特徴量を抽出 (データから学習)

人間の音声が歪んだ音に
なるように劣化
(データから学習)



ソース・フィルタを通って
人間の音声を復元(データから学習)

これは逆問題なので，各モジュールが狙い通りに動く保証はない！
→ 種々の方法を使って，ソース・フィルタ・チャネルに条件を付ける

方法2：歴史的音源から，音声成分を抽出 (今回新しく試行)

歴史的音源  音声 

音声復元のための教師あり学習

劣化音声の特徴量を抽出し，洗練し，復号したのち，波形に戻す

Speech SSL

Feature cleaner

Feature decoder

Neural vocoder

事前学習&固定

学習

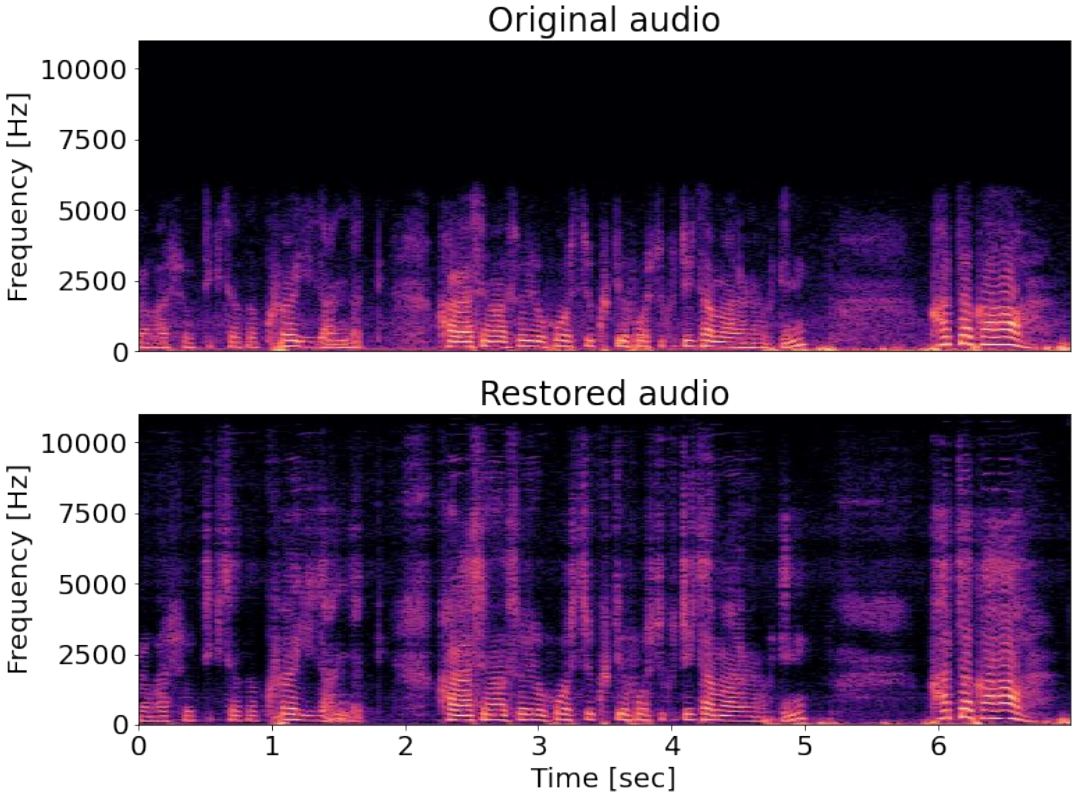
事前学習＆固定

事前学習＆固定

深層学習の事前学習の効果を享受できる．
(査読中なので詳細は省略)

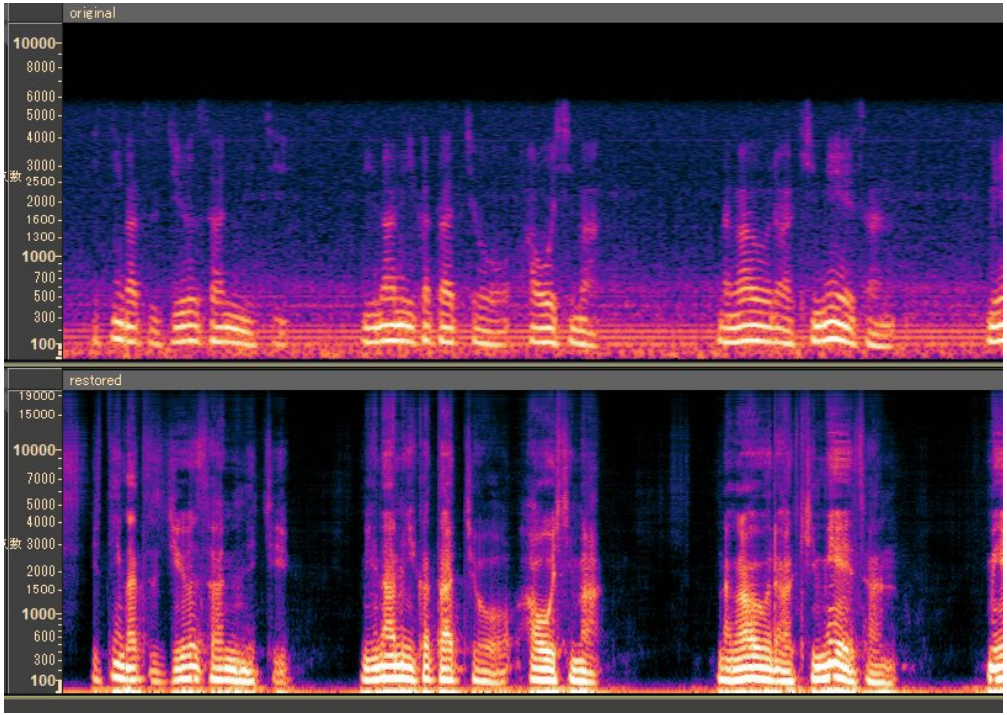
実験結果

方法1の音声サンプル (tape001/03_タコの昼寝)



実際の歴史的音声でも動作を確認 (両音源とも公開済み)

方法2の音声サンプル



実際の歴史的音声でも動作を確認

考察

事前学習モデル (SSL) の効果

	方法1 (音声と歪みを分離)	方法2 (音声を抽出)
チャネル歪み	△ (雑音や乗算ひずみがいまだ残る)	○ (事前学習の効果により，歴史的音源特有の歪みを除去できている)
音声歪み	○ (聴感上の大きな歪みは見られない)	△ (言語音に聞き取れない箇所も多々ある)

音声&歪み→元音源になる
学習の効果

両方を合わせた手法が今後の発展方針

余談：基盤モデルについて

2025年 言語処理学会 年次大会

音声・音響・音楽を扱う オープン基盤モデルの構築に向けた データセット策定

高道 慎之介[1,2,4], 和田 仰[2], 小川 諒[2], 山岡 洸瑛[2], 中田 亘[2], 淺井 航平[2], 関 健太郎[2], 岡本 悠希[2], 齋藤 佑樹[2], 小川 哲司[4,3], 猿渡 洋[2], 中村 友彦[3], 深山 覚[3]

(1: 慶應義塾大学, 2: 東京大学, 3: 早稲田大学, 4: 産業技術総合研究所)

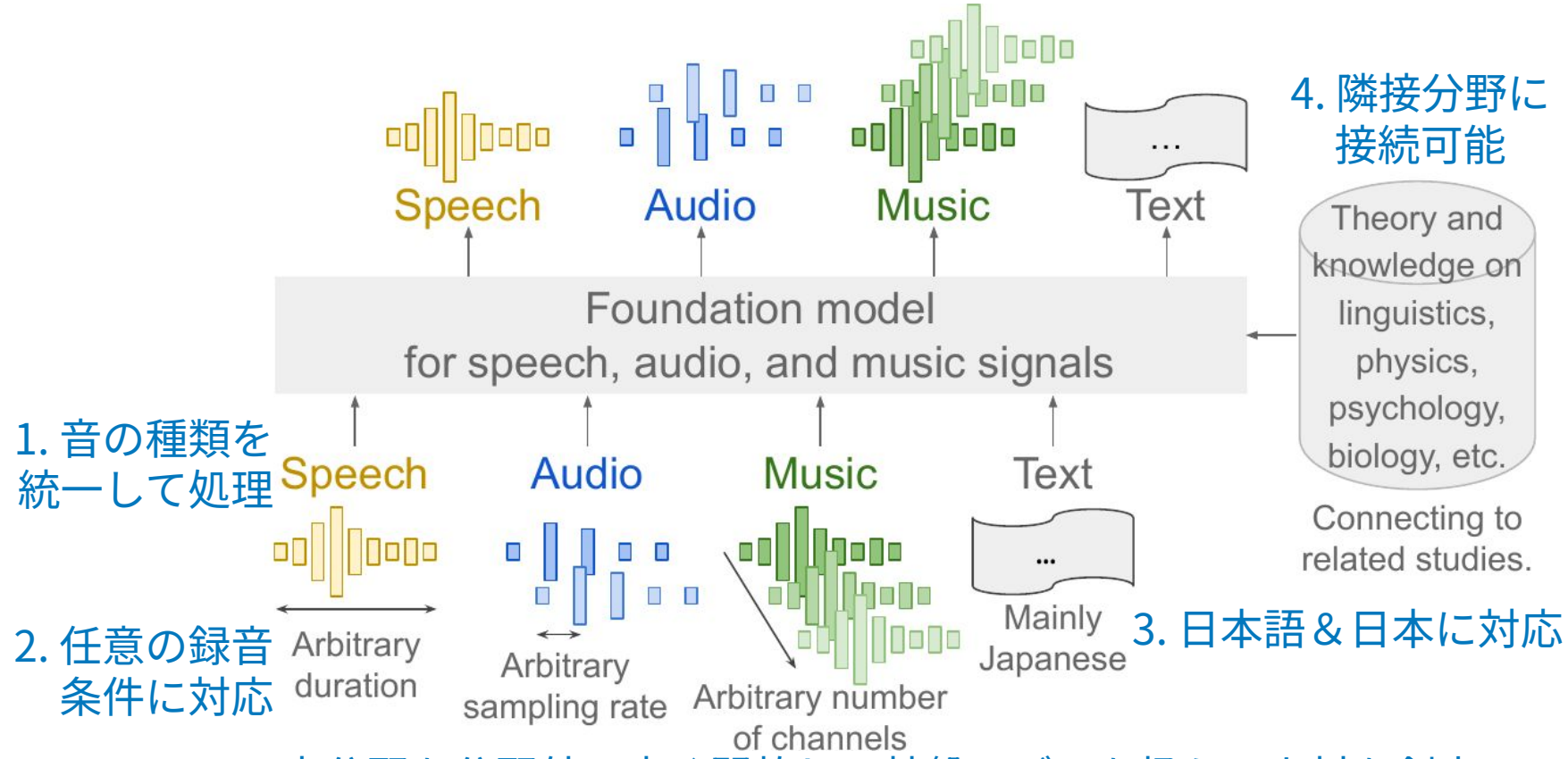
Takamichi Laboratory
慶應義塾大学 高道研究室

背景：音のオープン基盤モデルを作ろう

- 大規模言語モデル (言語基盤モデル) の隆盛 [Brown20][Chowdhery23]
  - 旧来の NLP タスクや zero-shot タスクにおける高い汎用性
  - → 言語以外のメディア・センサデータでも基盤モデルを期待

- 音の基盤モデルの現状
  - それぞれの音の種類で基盤モデルが登場
  - **音声** (人の声)：SpeechGPT [Zhang23], UniverSLU [Arora23]
  - **音響** (環境音，動物音声など)：Pengi [Deshmukh23], AVES [Hagiwara23]
  - **音楽** (器楽音など)： MERT [Li24], Mu-LLaMA [Liu24]

音の基盤モデルに対する我々の狙い



1. 音の種類を統一して処理

2. 任意の録音条件に対応

3. 日本語＆日本に対応

4. 隣接分野に接続可能

5. 音分野を分野外に広く開放し，基盤モデルを扱える人材を創出

概要：基盤モデルに資するデータセット一覧を作る

- 既存のデータセットを収集し，メタ情報を付与
  - 音の分野には，Common Crawl のような，クロール済みダンプデータが殆どないため，整備済みデータセットが中心

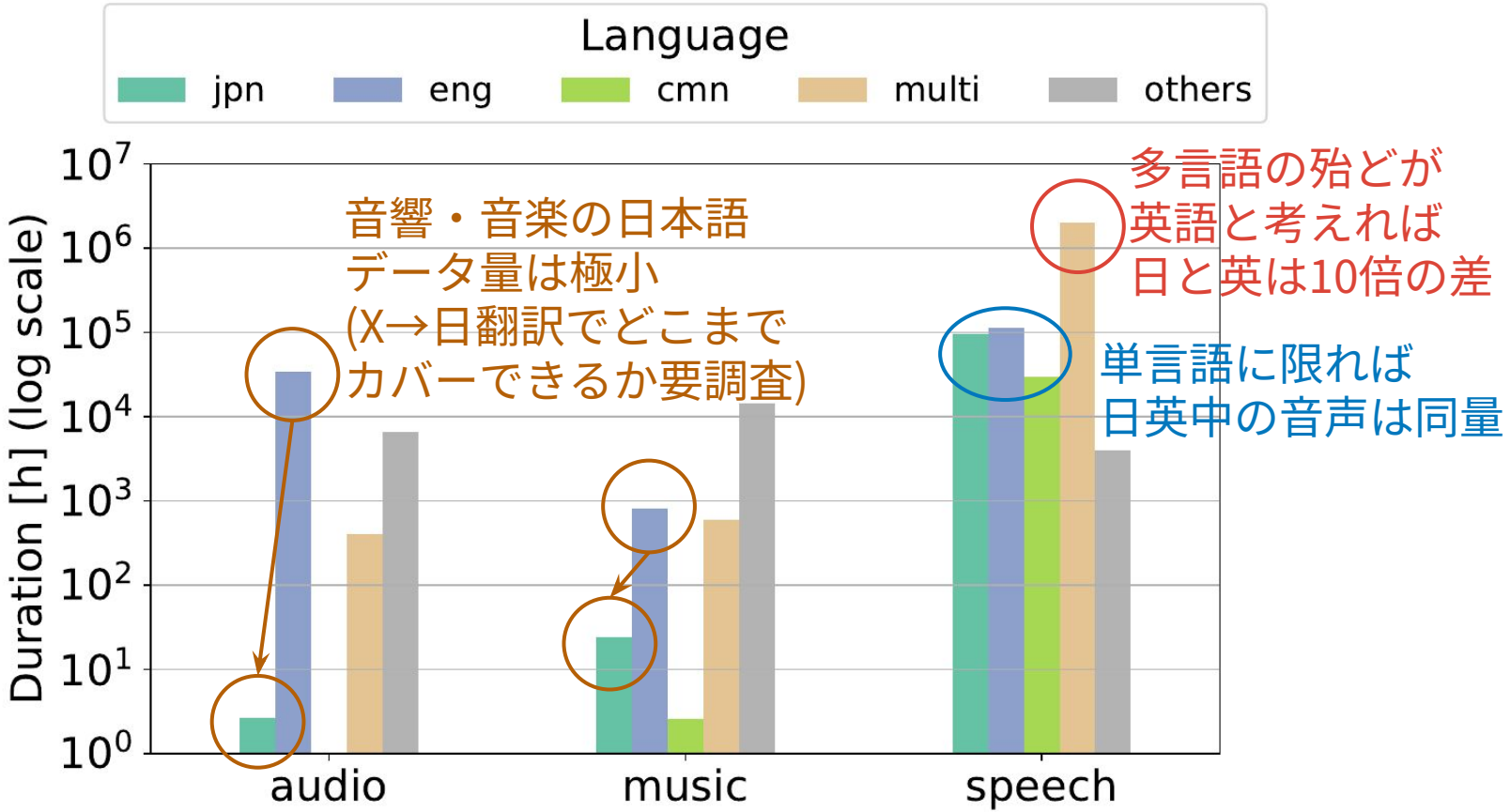
- **データセットの統計を分析し，基盤モデル構築に向けて情報を整理**
  - 約 220 万時間，約 450 個のデータセットについて
  - 一覧とメタ情報は公開済み

収集条件と概要

収集期間	2024/10–12 (3ヶ月間)
主な収集対象言語	日本語，英語，中国語
学習用 / 拡張用データセット数	約 450 個 / 約 80 個
学習用データの総時間数	約 220 万時間 (単純比較はできないが， 現存するオープン基盤モデルの中で最大)
含まれる希少データの例	日本語方言，動物音声，体内音，ディープフェイク，水中音，幼児音声など

次ページ以降で，学習用データセットの時間長の統計について調査
(拡張用データセットについては原稿を参照)

学習用データの分析３：言語
(日本語，英語，中国語，多言語，それ以外)



方言への基盤モデルの応用可能性について

- 諸方言をカバーするデータセットの作成およびその半自動化
  - 現在、東京方言以外は東京方言の 0.05% 程度
- 報酬及び評価のスキーム
  - なにをもって「よい方言処理結果である」とするか
- 時間と空間へのグラウンディング
  - 基盤モデルで{時間, 空間}と対応させて{文字言語, 音声言語} を処理可能？
  - (昨年に競争的資金に応募したが落ちた…)
- 地域コミュニティのインセンティブ
  - 研究者が直接絡まなくてもプロジェクトが進むようになってほしい