Tokyo BISH Bash #8

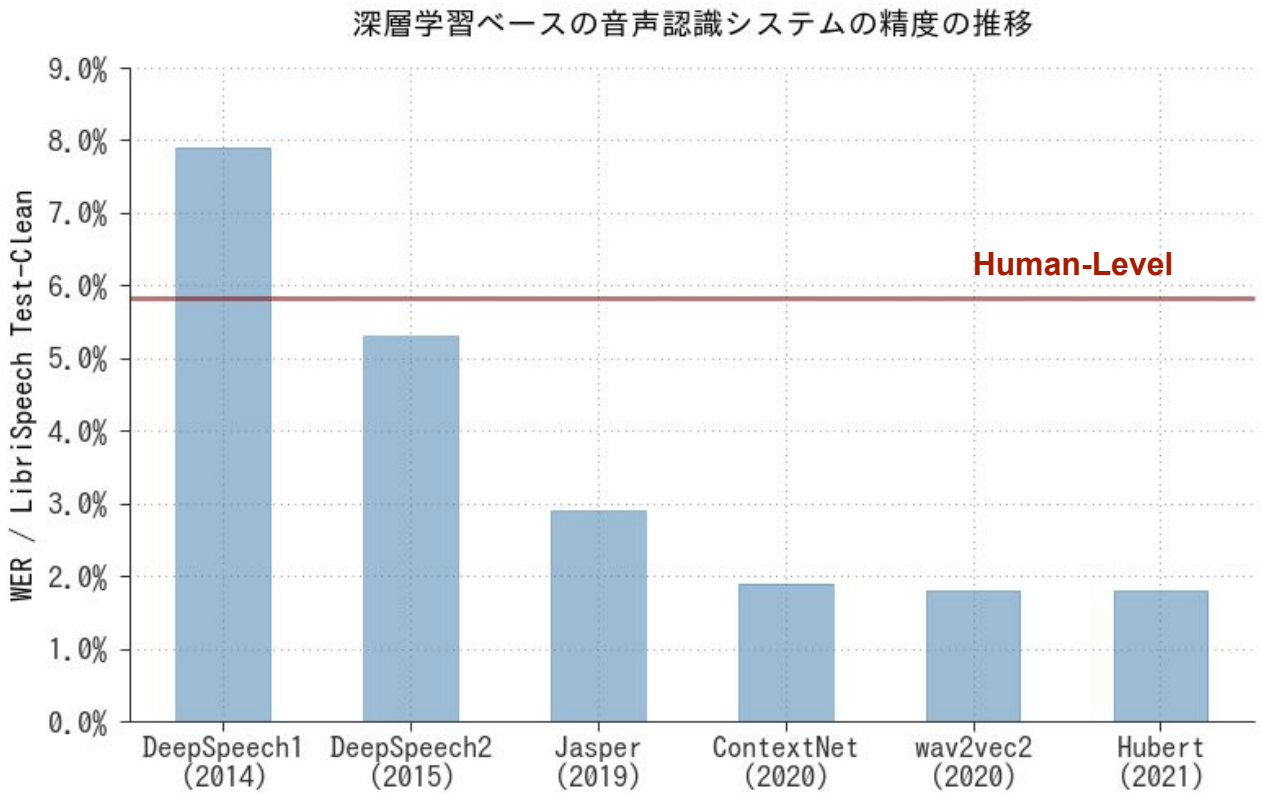
大規模日本語音声コーパスReazonSpeechの現状と展望

レアゾンヒューマンインタラクション研究所　藤本誠二

Reazon
Human Interaction
Lab

目次

はじめに: 深層学習の爆発的な発展



深層学習ベースの音声認識システムは、既に人間よりも低いError-Rateで音声を認識できている。

はじめに: 近年の音声認識モデルの傾向

学習に用いる音声コーパスが巨大化・多言語化している

音声認識エンジン	ラベル付き学習データセット [時間数]	備考
SpeechStew [Chan21]	5,140時間	LibriSpeech・Common Voiceなど7つのソースを統合して作成
OpenAI Whisper [Radford22]	680,000時間	インターネットから多言語のデータを大量に収集
Google USM [Zhang23]	210,000時間	Youtube + 公開データを統合して作成
Massively Multilingual Speech [Pratap23]	44,700時間	インターネットのデータから作成した 1107 言語の大規模多言語コーパス

はじめに: 日本における音声コーパス事情

音声コーパス	サイズ [時間数]	備考
JSUT [Sonobe17]	10時間	単一話者の読み上げ音声を録音
Common Voice	73時間	クラウドソースで作成（時間数は Common Voice Corpus 13.0時点）
CSJ [Maekawa00]	660時間	主に講演音声から作成
LaboroTVSpeech [Ando20]	2,000時間	テレビの放送の音声・字幕から作成
JTubeSpeech [Takamichi22]	1,300時間	Youtube動画の音声・字幕から作成

はじめに: 小括

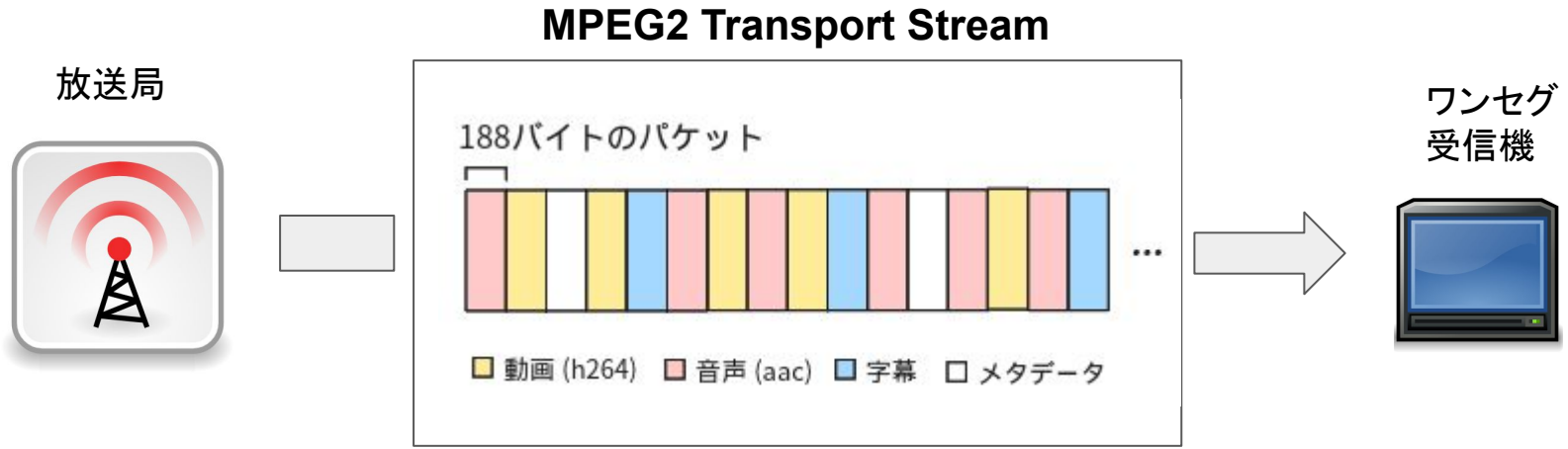
1. 深層学習ベースの手法は、音声認識の領域において人間の水準を凌駕する成績を達成している。
2. 精度の向上とともに、学習に用いる音声コーパスの量は巨大化している。
    - 国外では数万時間のデータセットの活用が標準になりつつある。
3. 一方で、日本語の音声コーパスはまだ整備の途上にある。

ReazonSpeechコーパスとは

- 放送音声（ワンセグ放送）から作成したオープン日本語音声コーパス
- 2023年6月現在、コーパス規模は19000時間。
  - 一般に入手可能な日本語コーパスとしては世界最大
- ツールキットを含めてすべてオープンソースで公開。
  - 後述するように音声認識モデルも公開しています。

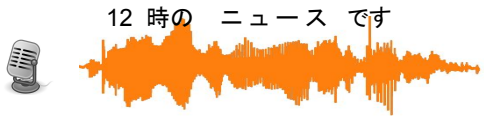
公式サイト	https://research.reazon.jp/
Hugging Face	https://huggingface.co/datasets/reazon-research/reazonspeech
GitHub	https://github.com/reazon-research

ReazonSpeechコーパスの構築の流れ

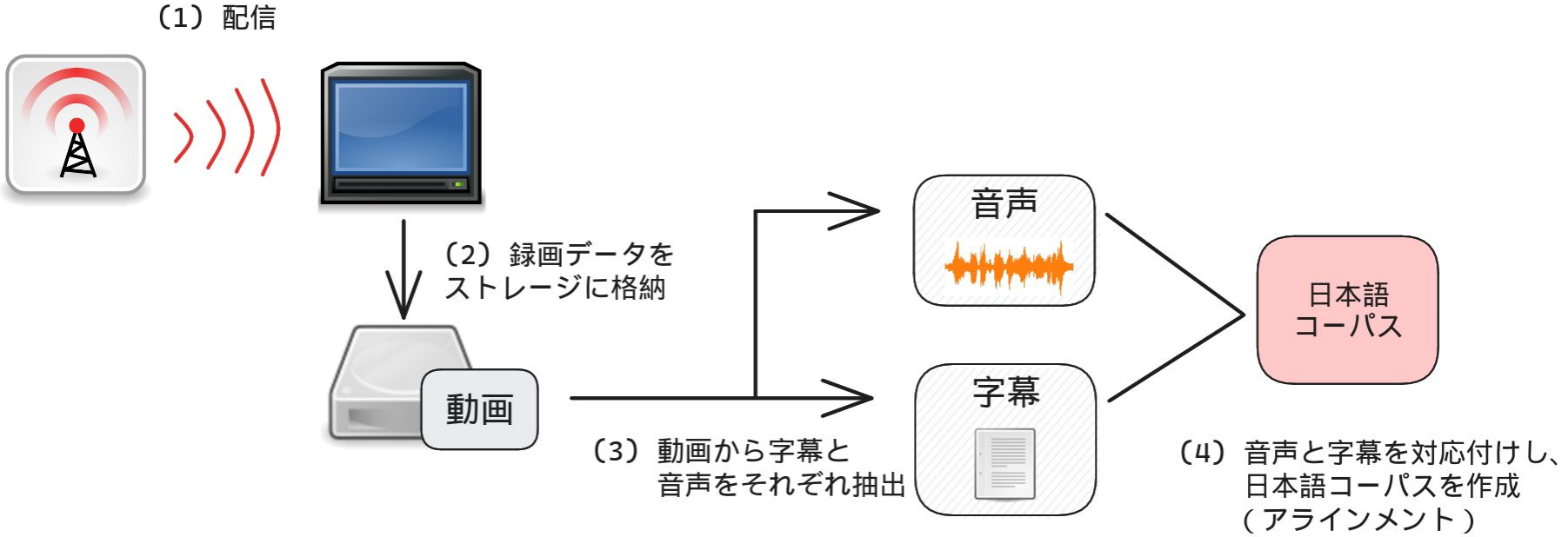


放送のストリームデータから情報を抽出し、発話音声と対応する書き起こしラベルのペアを生成する。

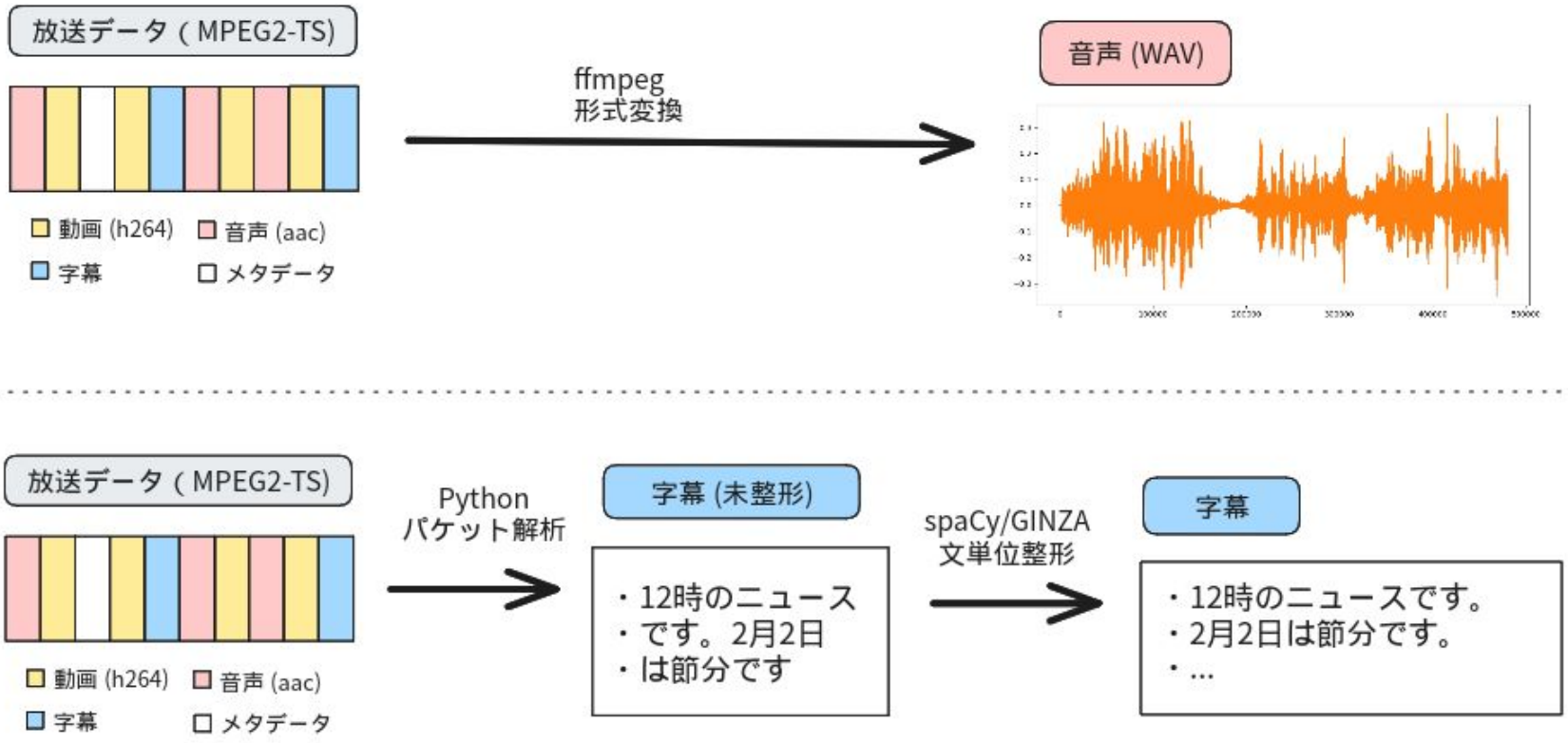
12時のニュースです。



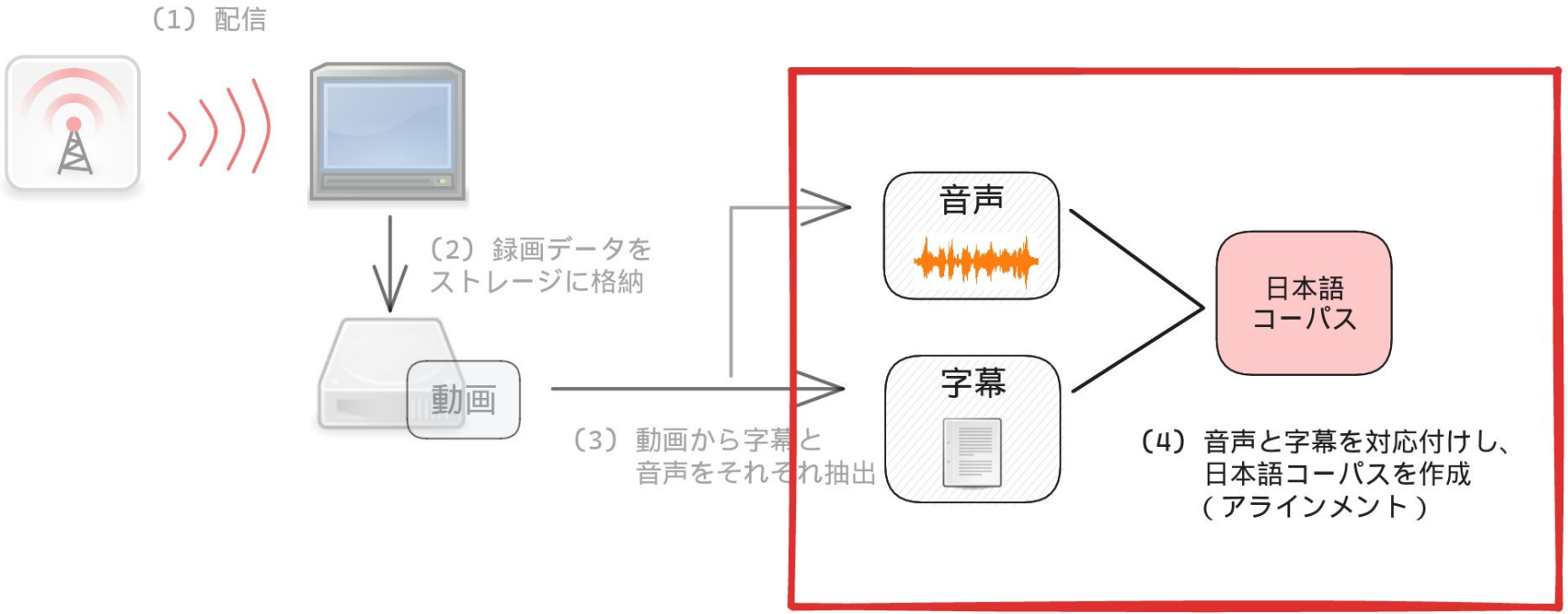
ReazonSpeechコーパスの構築の流れ（承前）



ReazonSpeechコーパスの構築の流れ（承前）



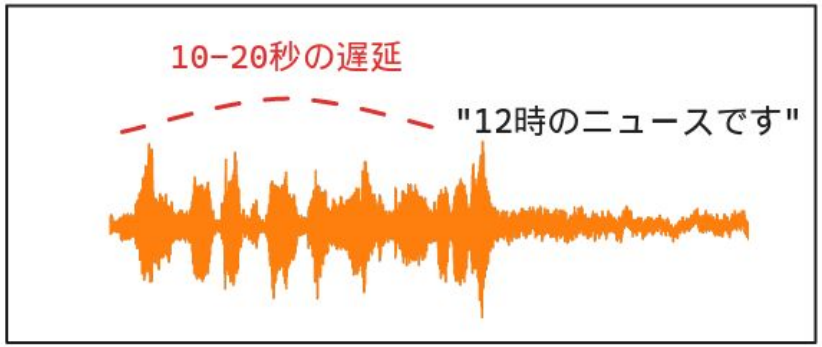
ReazonSpeechコーパスの構築の流れ（承前）



ReazonSpeechコーパスの構築の流れ（承前）

- 音声と字幕は正確に一対一に対応しているわけではない。
- 様々なノイズを処理して、正確に発話のタイミングを特定する。

音声のタイミング ≠ 字幕のタイミング

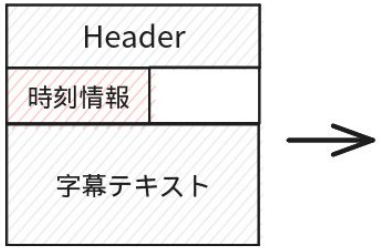


字幕に対応する発話があるとは限らない

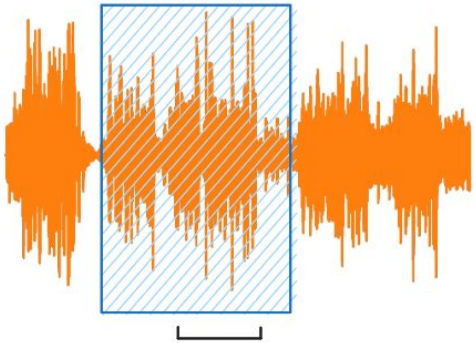
ReazonSpeechコーパスの構築の流れ（承前）

- 字幕の時刻情報の活用で「抽出精度」と「処理速度」の一挙両得！

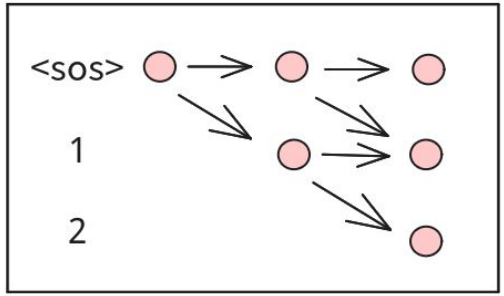
(1) 字幕パケットから
表示タイミングを抽出



(2) 表示タイミングの周辺を
マージンを設けて切り取り

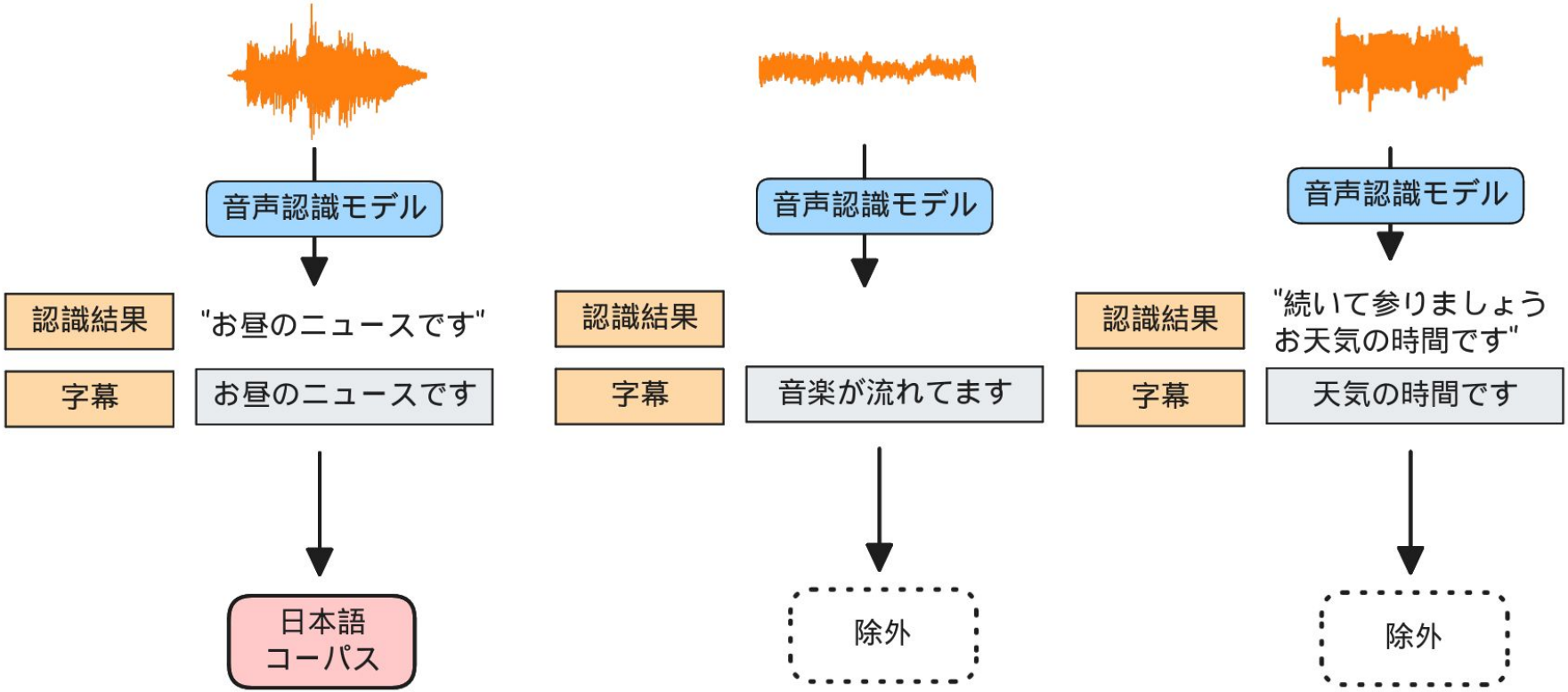


(3) 音響モデルを元に遷移確率行列を
計算し、最も確率の高い区間を特定



※ Kürzinger, Ludwig, et al. "CTC-segmentation of large corpora for german end-to-end speech recognition." arXiv:2007.09127

ReazonSpeechコーパスの構築の流れ（承前）



ReazonSpeechコーパスの構築の流れ（まとめ）

- 放送局から配信されるワンセグデータを蓄積する。
- 録画から音声と字幕をそれぞれ抽出する。
  - 音声：ffmpeg
  - 字幕：Pythonパーサ + spaCy/GINZAでセンテンス整形
- 音声と字幕をアラインメントして日本語コーパスを作成。
  - 字幕の時刻情報を上手く活用する。
  - 音声認識モデルで文字起こしすることで品質を担保。

観点：音声コーパス構築の循環問題

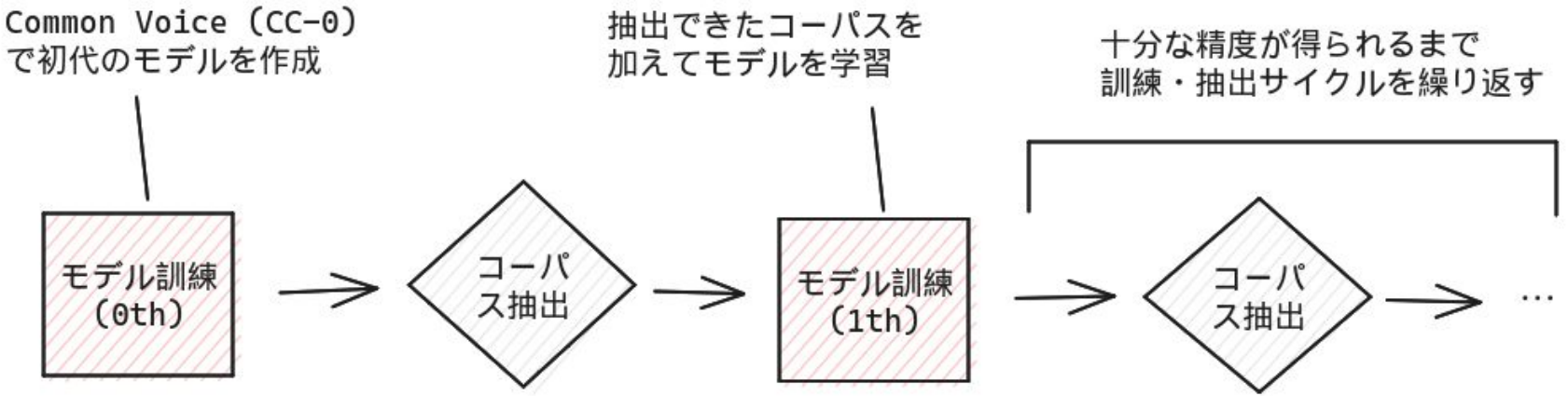
- **大規模な音声コーパスを構築するには、高精度な音声認識モデルが必要**
  - 音声認識モデル無しにはアラインメントが正しく計算できない
  - 最後の確認の工程においても認識精度が必要。
- **高精度な音声認識モデルを構築するには、大規模な音声コーパスが必要**
  - 冒頭の論点。深層学習ベースの音声認識モデルではデータ量が必須。
  - モデルとコーパスがお互いに前提条件となっている！

コーパスとモデルをブートストラップする

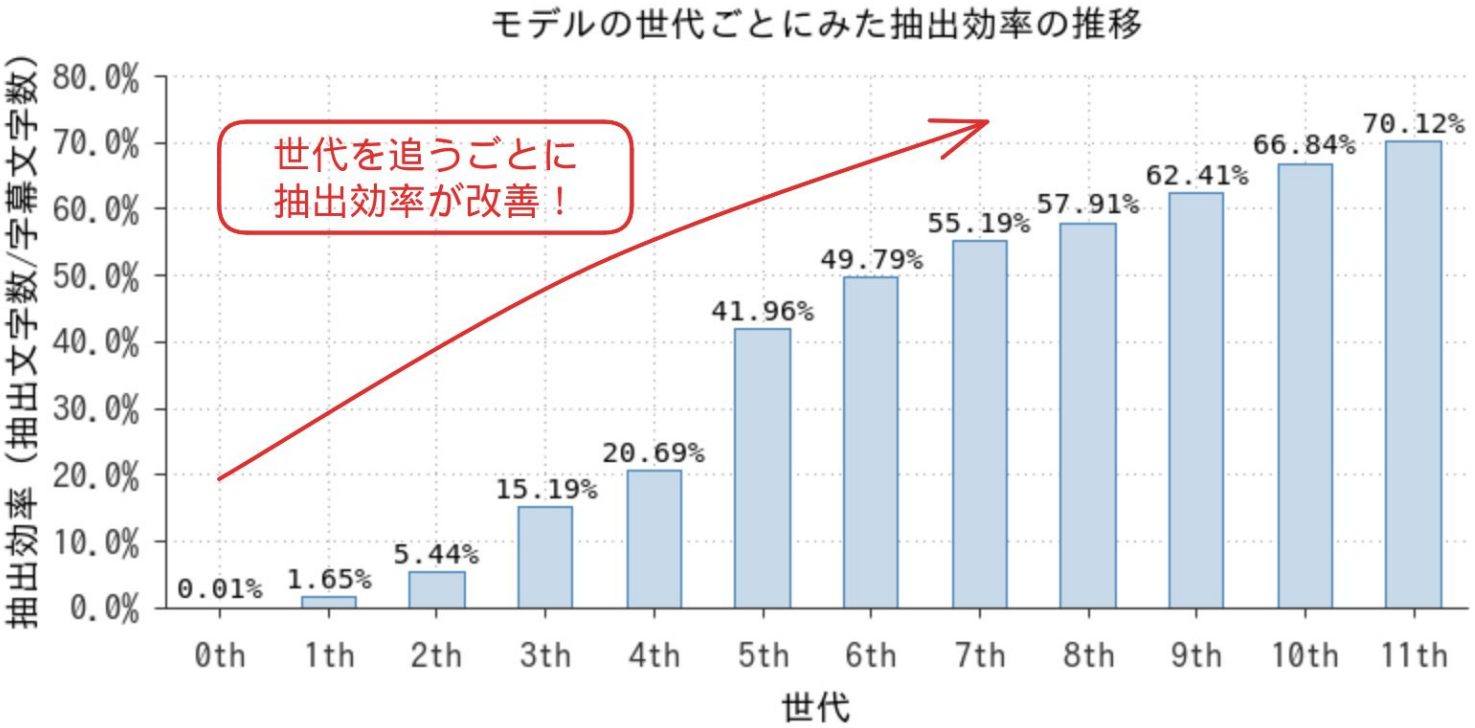
目標

高精度な音声認識モデルと、大規模日本語コーパスを一から同時に構築する

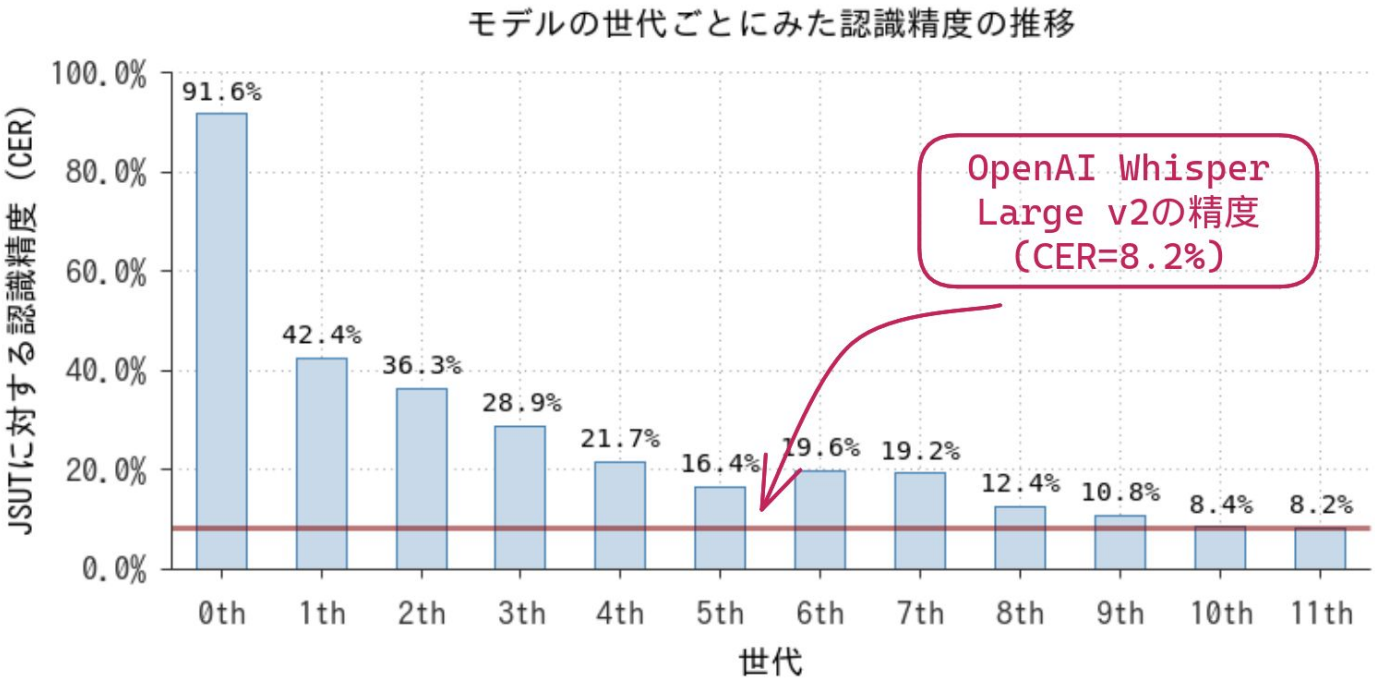
手法



コーパスとモデルをブートストラップする (承前)



コーパスとモデルをブートストラップする (承前)



1. Takamichi et al. 2017, "JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis," arXiv preprint 711.00354

ReazonSpeechコーパスとは：小括

- 放送音声から作成した**世界最大のオープン日本語音声コーパス**
- **字幕のタイムスタンプの活用した効率的な抽出プロセス**
  - 品質と処理速度を両立し、大規模な構築を可能に
- **コーパスとモデルをブートストラップする**
  - 最高水準の音声認識モデルと、世界最大の日本語コーパスを一挙に構築！

言語処理学会年次大会
(NLP2023) 優秀賞

優秀賞（対象579件中11件）

A5-3 **ReazonSpeech: A Free and Massive Corpus for Japanese ASR**

Yue Yin , Daijiro Mori (Reazon), Seiji Fujimoto (Clear Code)

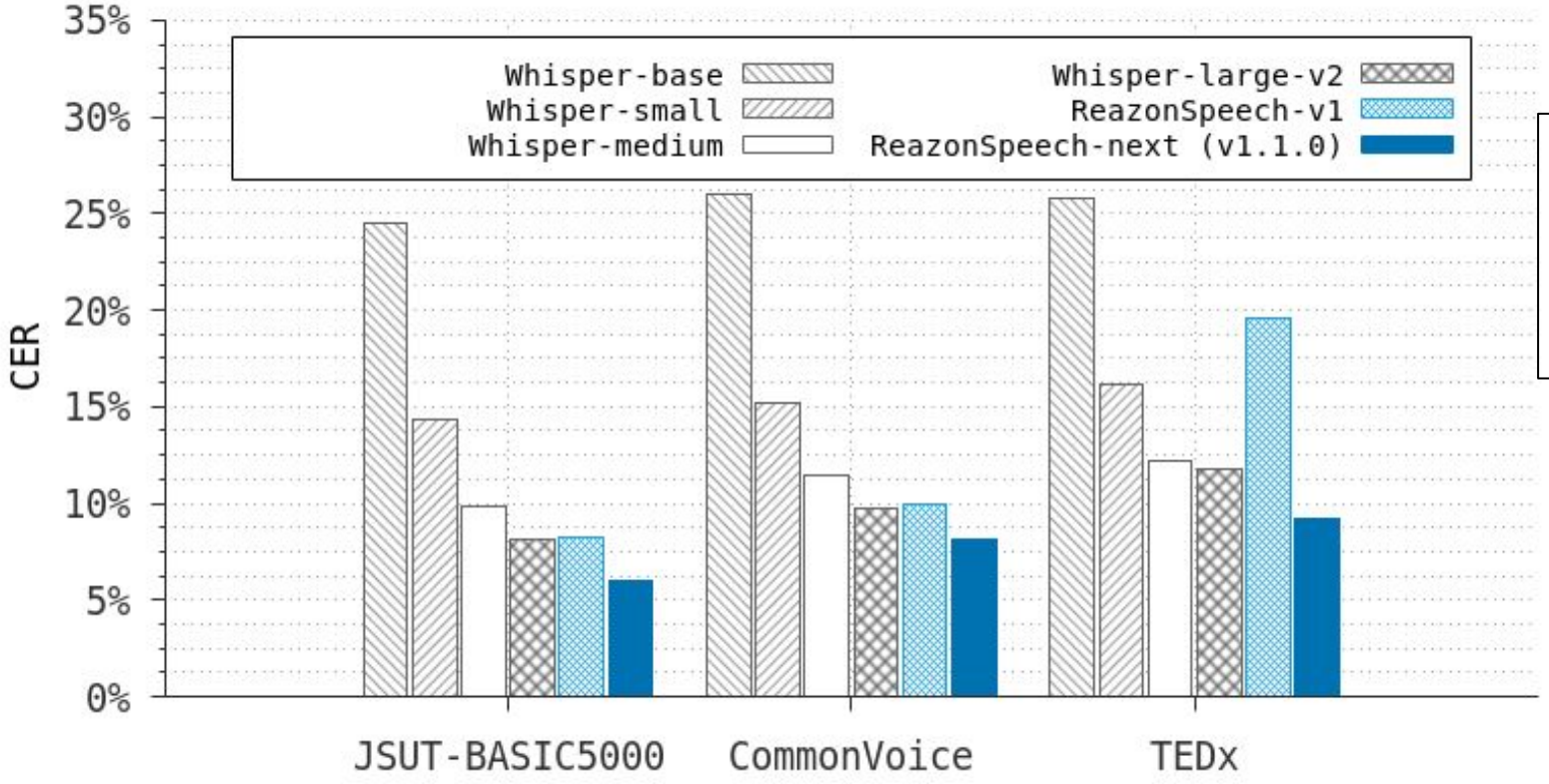
本論文は，日本語の音声認識用コーパスを大規模に自動構築する手法について報告しています．本手法は，テレビ番組から音声と字幕を抽出し，音声認識モデルと音声認識用コーパスをブートストラップ的手続きによって徐々に改良するものです．音声認識用コーパス，その構築ツール，学習した音声認識モデルを無償で公開しており，商用にも用いることができます．コーパスは日本語では最大規模であり，モデルは高い性能を達成しています．これらの貢献から優秀賞にふさわしいと判断しました．

将来の展望

コーパス規模のさらなる拡大

- 「時間とともにコーパス量が増加する」
  - 解析対象のデータは毎日刻々と増えている。
  - 計算リソースを投下すれば、継続的に抽出可能。
- 現在の抽出効率だと…
  - 現行の手法では、年間で約12000時間の抽出が可能。
  - 2024年には、コーパス量は3万時間超に拡大する見込み。

より高精度な音声認識モデルの追求



4月4日リリースの最新モデルで、Whisper large-v2を精度で上回る

コーパス作成・利用の輪を広げる

・活用例：HuBERTの事前学習モデルの学習

日本語の音声に特化した事前学習モデルHuBERTを公開

お知らせ　2023.04.28

rinnaが開発した日本語の音声に特化した事前学習モデルHuBERT (Hidden Unit Bidirectional Encoder Representations from Transformers) を、商用利用可能なApache-2.0 ライセンスで公開したことをお知らせします。

rinnaはこれまでに日本語に特化した言語モデルGPT・BERTや言語画像モデルCLIP・Japanese Stable Diffusionなどを公開し、Hugging Faceでのモデルダウンロード数は累計150万を超え、多くの研究・開発者にご利用いただいています。この度、Metaから提案されたHuBERTのモデル構造とレアゾン・ホールディングスが公開した日本語音声コーパスReazonSpeechを用いて、日本語の音声に特化したHuBERTを学習し、Apache-2.0 ライセンスでHugging Faceに公開しました。音声表現が学習された事前学習モデルを公開することで日本語の研究・開発コミュニティに成果を還元し、研究・開発の活発化に繋がることを期待します。

日本語HuBERT (rinna/japanese-hubert-base) : https://huggingface.co/rinna/japanese-hubert-base

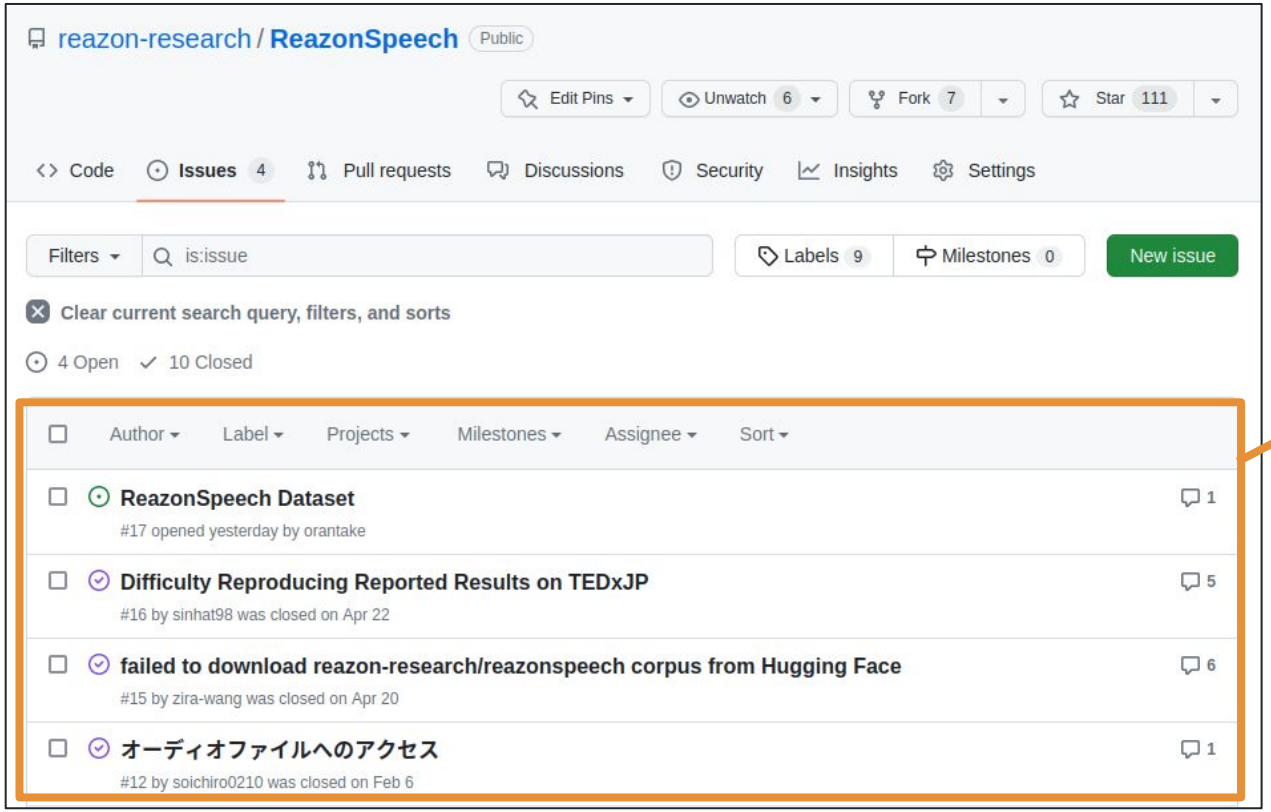
ReazonSpeechは著作権法30条の4の情報解析の範囲であれば誰でも自由に利用可能。

コーパス作成・利用の輪を広げる

公開リソース	ライセンス	URL
学習済みESPnetモデル	Apache-2.0	• reazonspeech-espnet-v1（安定版） • reazonspeech-espnet-next（最新版）
音声処理ライブラリ	Apache-2.0	https://github.com/reazon-research/ReazonSpeech
日本語音声コーパス	CDLA-Sharing-1.0 (ただし利用目的は著作権法30条の4に定める情報解析に限る)	https://huggingface.co/datasets/reazon-research/reazonspeech
研究論文		http://research.reazon.jp/_static/reazonspeech_nlp2023.pdf

ライブラリを活用すれば、誰でもコーパスを作成可能です！

GitHub Issueで気軽に議論しよう！



相談レベルでOKなので気軽に起票ください！

・こういう用途にモデルを使いたい
・データセットを自作したいがわからないことがある
・…

おわりに

みんなでオープン日本語コーパスを盛り上げよう！