日本音響学会 2024年 春季研究発表会

YODAS：YouTube 動画から構築される多言語大規模音声データセット

Li Xinjian(CMU)，〇高道 慎之介，佐伯 高明(東大)，
Chen William(CMU)，塩田 さやか(都立大学)，渡部 晋治(CMU)

MOONSHOT RESEARCH & DEVELOPMENT PROGRAM
AVATAR SYMBIOTIC SOCIETY

概要

- **オープンなコーパスは研究を加速する**
  - Whisper [Radford22] などは10万時間以上を利用
  - 一方でオープンコーパスは最大でも数万時間程度
  - YouTube は豊富なデータ資源になりうる

- **新たなオープンコーパス YODAS を構築**
  - **Y**ouTube-**o**riented **d**ataset of **a**udio and **s**peech
  - YouTube をクロールして構築
    - jtubespeech [Takamichi21] の構築アルゴリズムをベースに
  - Huggingface datasets に対応

- **本研究では，その構築結果と実験評価を報告**

YouTubeを基にした，日本語を含むコーパス比較

	jtubespeech [Takamichi21]	**YODAS (本研究)**
元動画	ライセンスを問わず 発見した動画すべて	Creative Commons license 動画のみ
公開物	動画IDのみ	音声，テキスト
言語	30言語	140言語
サイズ	1万時間〜 (日本語)	42万時間 (全言語)
用途例	音声データ自体は配布せず 大規模モデルを作る	音声データの再現性も担保する

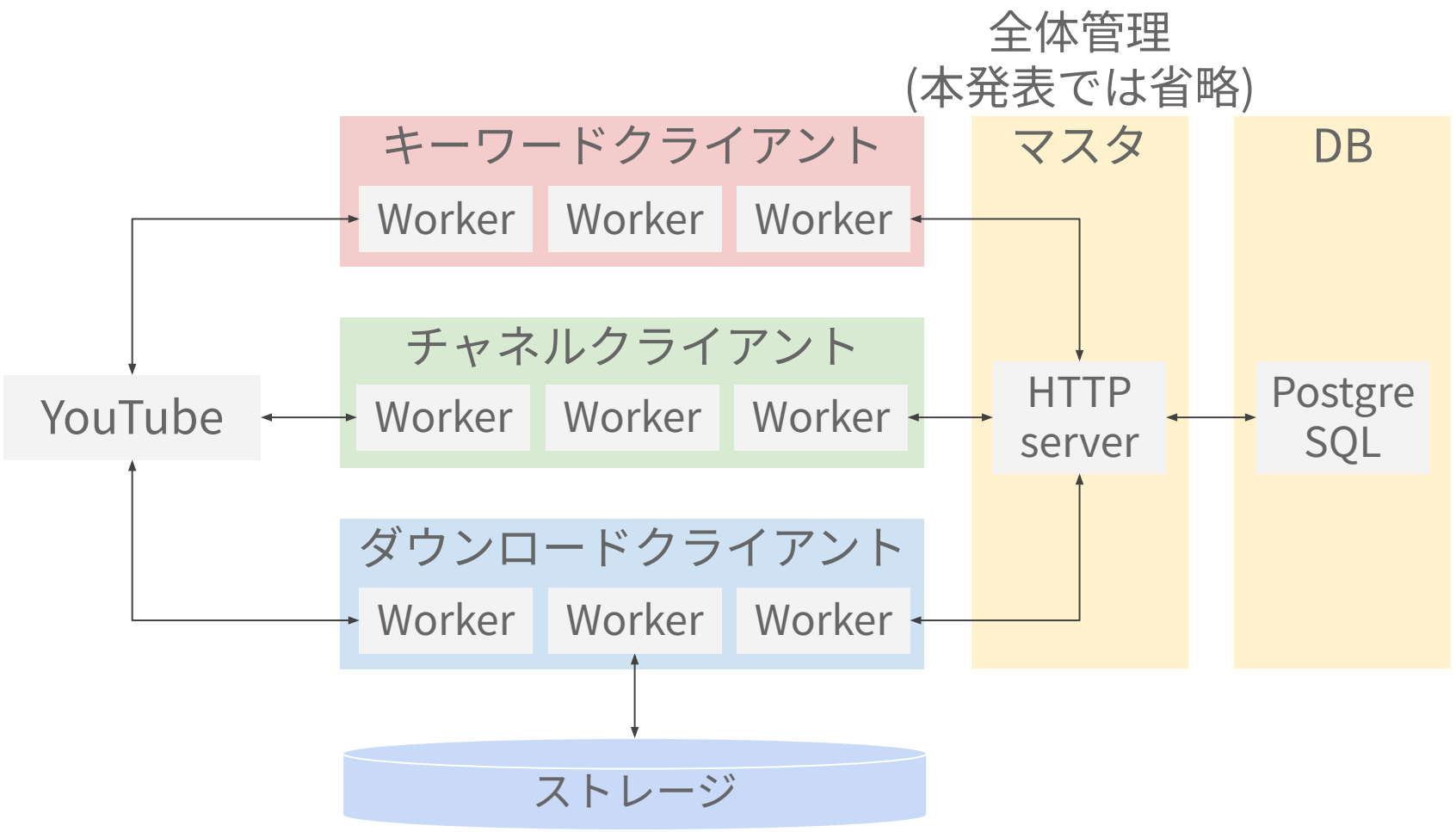
データ収集

収集における 2 つのポリシー

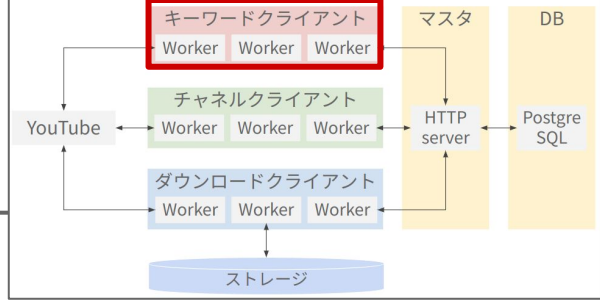
1. **当該動画が Creative Commons でライセンスされていること**
   a. データの再配布を可能にするため

2. **手動あるいは自動字幕を有すること．ただし字幕のない動画も許容する**
   a. 字幕あり動画は，教師あり学習などに利用可能
   b. 字幕なし動画は，自己教師あり学習などに利用可能

データ収集の全体像


全体管理
(本発表では省略)
キーワードクライアント
Worker
Worker
Worker
マスタ
DB
チャネルクライアント
YouTube
Worker
Worker
Worker
HTTP
server
Postgre
SQL
ダウンロードクライアント
Worker
Worker
Worker
ストレージ

キーワードクライアント

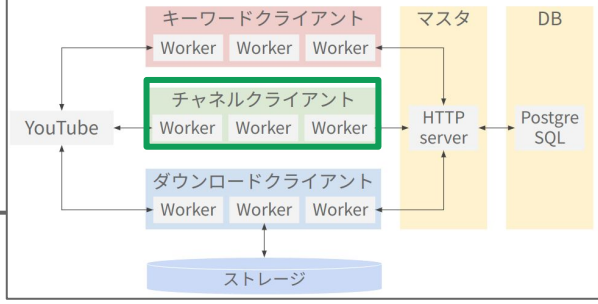


- **目的**
  - 収集対象の動画IDを効率的に検索すること

- **手順**
  - 多言語の Wikipedia 記事からキーワード抽出 [Takamichi21]
    - ハイパーリンクのあるフレーズをキーワードとして利用
  - YouTube動画検索
    - 前述のポリシーを満たすよう検索フラグを立てる
    - 言語多様性を担保するため，rich-resource 言語以外のキーワードの使用を優先

- **Tips**
  - HTTPリクエスト [Takamichi22] → AJAX に変更
    - 低関連度動画もヒットするように

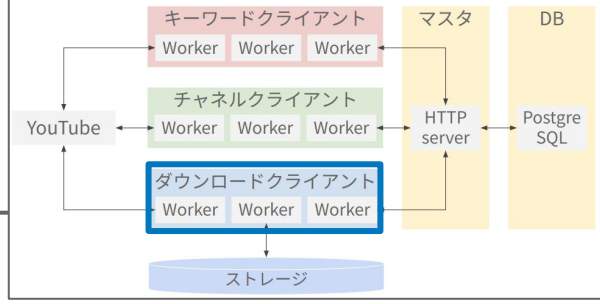
チャンネルクライアント



- **目的**
  - キーワードクライアントだけでは集めきれない動画IDを取得
    - YouTube検索が未視聴の動画よりも人気の動画を表示するため

- **手順**
  - キーワード検索で発見された動画から YouTube チャンネルを識別
  - そのチャンネルに属する動画を新たな候補とする
    - ポリシーに反しないかを別途判定
    - “Creative Commons ライセンス動画を持つチャンネルには，他にも同様のライセンスの動画があるだろう” という仮説

ダウンロードクライアント

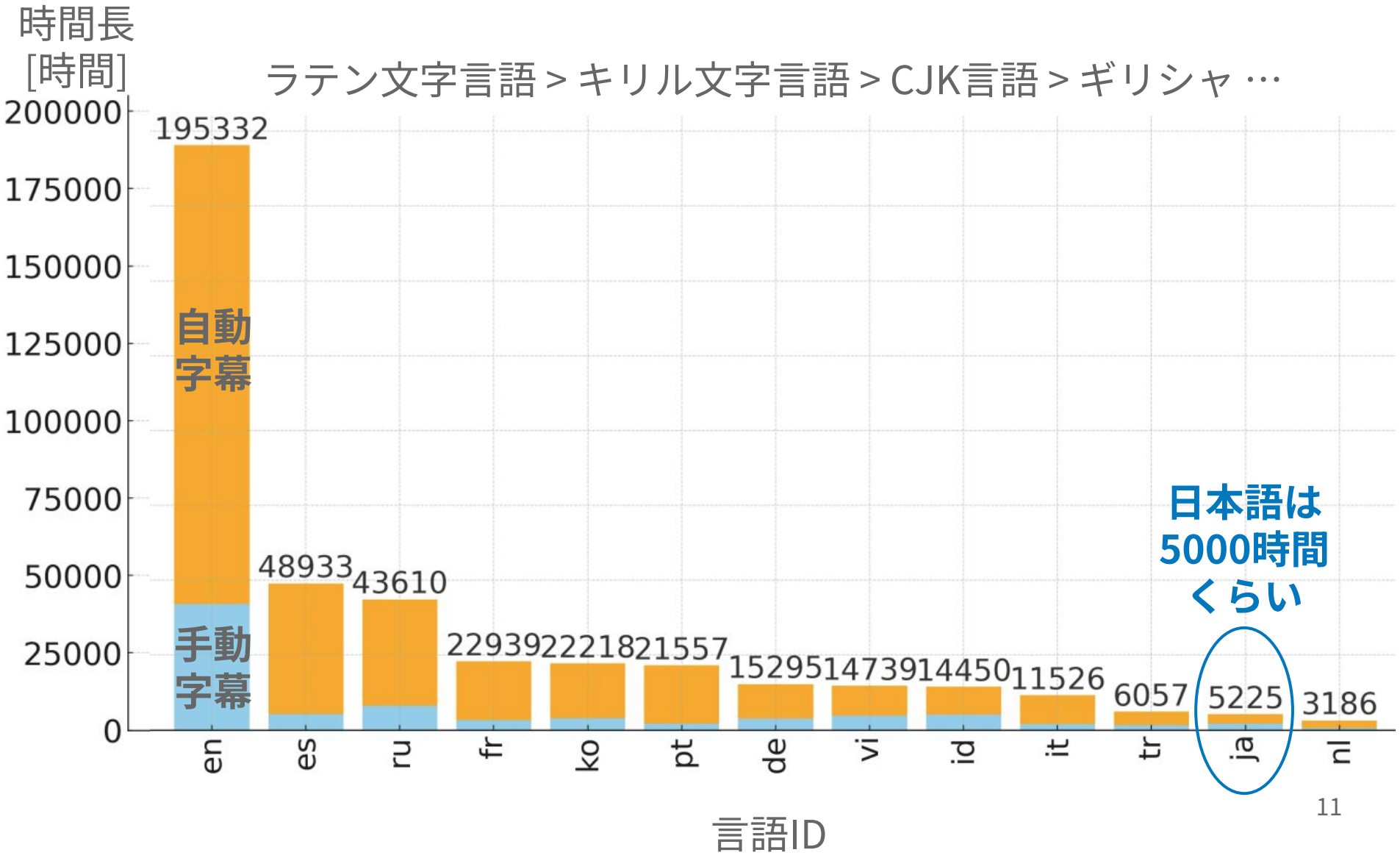


- **目的**
  - 動画と字幕のダウンロード

- **手順**
  - 取得した動画IDを用いてダウンロード
  - 保存
    - 音ファイルフォーマットを 24 kHz モノラルに統一
    - 手動字幕と自動字幕を分けて保存
  - 言語識別（字幕の言語IDが常に正しいとは限らないため）
    - 単一言語の字幕のみを有するなら，字幕の言語IDは正しいとする

データ分析

言語ごとの時間長 (字幕あり動画のみ)

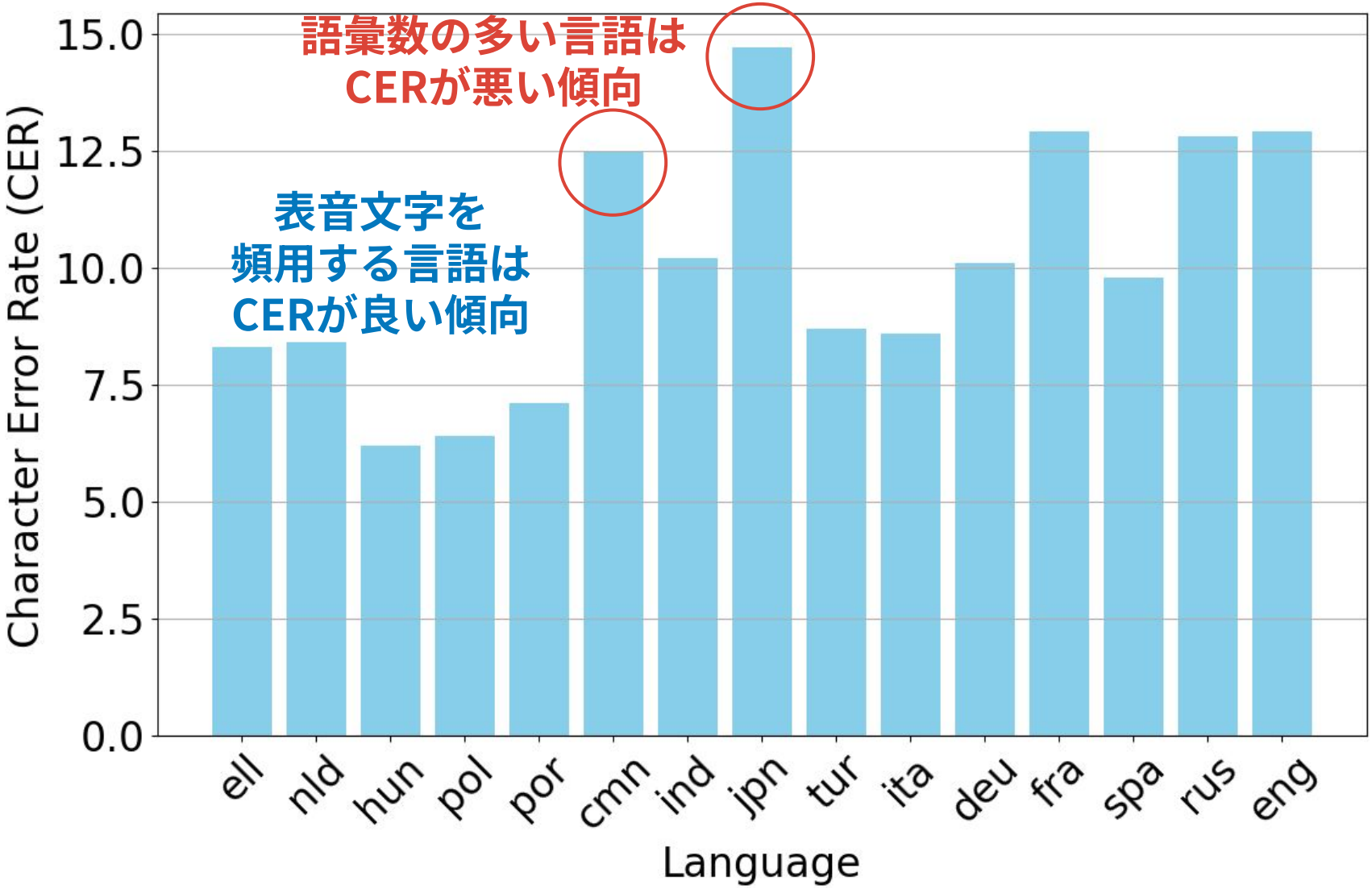


音声認識実験 (YODASは他のタスクにも使えるが本論文では音声認識に限定)

実験条件

使用する字幕	手動字幕のみ
言語数	手動字幕の多いトップ25言語
学習データ	100万発話 / 言語 (CTCスコアによる足切りを事前に実施)
評価データ	1,000発話 / 言語
音響モデル	XLSR [Babu21] + 線形層
語彙数	5,000 (中国語)，3,000 (日本語)，300 (それ以外の言語)
データ拡張	なし

音声認識結果


語彙数の多い言語は
CERが悪い傾向
表音文字を
頻用する言語は
CERが良い傾向
Character Error Rate (CER)
15.0
12.5
10.0
7.5
5.0
2.5
0.0
ell
nld
hun
pol
por
cmn
ind
jpn
tur
ita
deu
fra
spa
rus
eng
Language

コーパスのサンプル

https://huggingface.co/datasets/espnet/yodas

まとめ

まとめ

- **目的**
  - 配布可能な大規模音声コーパスを作る！ -> YODAS！
- **手段**
  - 字幕付きの Creative Commons License 動画を効率的に収集
- **結果**
  - 数十万時間規模のコーパスを構築
- **今後の予定**
  - コーパス規模の拡大，アノテーションなど