Common Crawl を用いた大規模音声音響データセットの構築 *

◎淺井 航平 * (東大/NII LLMC)，杉浦 一瑳 * (京大/NII LLMC)，中田 亘 * (東大)，
栗田 修平 (NII LLMC)，高道 慎之介（慶大/東大），小川 哲司 (早大)，
東中 竜一郎（名大/NII LLMC）

1 はじめに

Web 上のコンテンツを収集し公開するプロジェクトである Common Crawl[1]から大量のデータセットを構築する試みは，自然言語処理 [1] や画像処理 [2] 分野において広く行われている一方で，音声音響分野では存在しない．本稿では，日本語キャプション[2]のついた音声音響データを日本語音声音響データと称し，Common Crawl からの大規模データセット構築を試みる．

ウェブなどをデータソースとしてデータセットを構築する試みがいくつかある．外国語の音声音響データ構築の先行研究としては，LibriVox に基づく LibriSpeech [3]（約 0.1 万時間），Internet Archive に基づく The People's Speech [4]（約 1 万時間），YouTube に基づく YODAS [5]（約 50 万時間），Freesound に基づく LAION-Audio-630K（約 4 千時間） [6] がある．日本語については音声データに限れば，YouTube や PodcastIndex に基づく J-CHAT [7]（約 7 万時間），ワンセグ放送に基づく ReazonSpeech[3]（約 3.5 万時間）などが存在する．

しかし，オープンな音声音響データセットの量は昨今の大規模音基盤モデルを学習するのに十分でない．実際，近年の音声基盤モデル開発は非常に多くのデータセットを使って学習している．例えば，音声認識モデルの Whisper [8] は 68 万時間，full-duplex 対話モデルの Moshi [9] は 700 万時間，音声言語モデルの Kimi-Audio [10] は 1300 万時間以上のデータで学習されている．特に日本語において，この学習データ量とデータセット量の乖離を緩和するために，新たなデータソースからデータセットを構築する方法を模索することが重要である．

本研究では，Common Crawl から音声音響データの URL を抽出し，データをダウンロードすることで大規模音声音響データセットを構築する方法を検討する．この方法は，Common Crawl を新たなデータソースに出来ること，さらに，Common Crawl の定期更新に基づいて継続的にデータセットを更新できるメリットがある．一方で，Common Crawl に存在する音データはキャプションの言語，音の種別（音声，音響，音楽），音質，ファイルフォーマットが様々であるため，言語と音の両面からフィルタリングの方法を検討する必要がある．本稿ではそのパイプラインについて述べるとともに，得られる音声音響データ URL と音データの統計情報を報告する．

2 データセット構築パイプライン

本研究において, データセット構築を以下のパイプラインで行なう.

1. Common Crawl から音 URL の抽出
2. 音データ URL をもとに音データのダウンロードおよびフィルタリング

Common Crawl は 1ヶ月に 1 回程度の頻度で Web 上のデータをクロールしてスナップショットとして公開している．ここでは 2025-18 のスナップショット[4]に含まれる 100,000 個の WARC ファイル（Web アーカイブの保存用ファイル形式）を用いる．Common Crawl から音 URL を抽出する部分については, HTML ファイルをもとに抽出する方法と, RSS ファイルをもとに抽出する方法の 2 つを試す.

2.1 HTML ベースのアプローチ

WARC ファイルには, Web 上からクロールした大量の HTML が含まれている．ここでは, 日本語の HTML のみを高速にフィルタリングするために, chardet[5]を用いた文字コードのデコーディングに成功した HTML について, HTML のメタタグとして lang=“ja” がついている HTML のみを残す． さらに, trafilatura[6] の extract() を用いて text が抽出でき, title タグが存在し, 言語判定ライブラリの lingua で日本語判定されたもののみを残す．この様にして得られた日本語 HTML について, 音 URL を beautifulsoup4[7]ライブラリで HTML 中のタグを探索することで取り出す. その結果 2025-18 のスナップショットから, 1,029,780 個の音 URL を抽出した．さらに抽出できた音 URL の完全重複削除をした結果, 724,787 個の音 URL が残った．図 1 に重複削除後の音 URL のトップレベルドメインの頻度トップ 50 を示す.

---

* Building large-scale speech and audio dataset from Common Crawl. By Kohei Asai (UTokyo/NII LLMC), Issa Sugiura (Kyoto Univ/NII LLMC), Wataru Nakata (UTokyo), Shuhei Kurita (NII LLMC), Shinnosuke Takamichi (Keio Univ./UTokyo), Tetsuji Ogawa (Waseda Univ.), and Ryuichiro Higashinaka (Nagoya Univ./NII LLMC). ∗ indicates equal contribution.

[1] https://commoncrawl.org/

[2] 音声の書き起こしに限らず，環境音や楽音の説明文も含む.

[3] https://huggingface.co/datasets/reazon-research/reazonspeech

[4] https://data.commoncrawl.org/crawl-data/CC-MAIN-2025-18/index.html

[5] https://pypi.org/project/chardet/

[6] https://trafilatura.readthedocs.io

[7] https://pypi.org/project/beautifulsoup4/

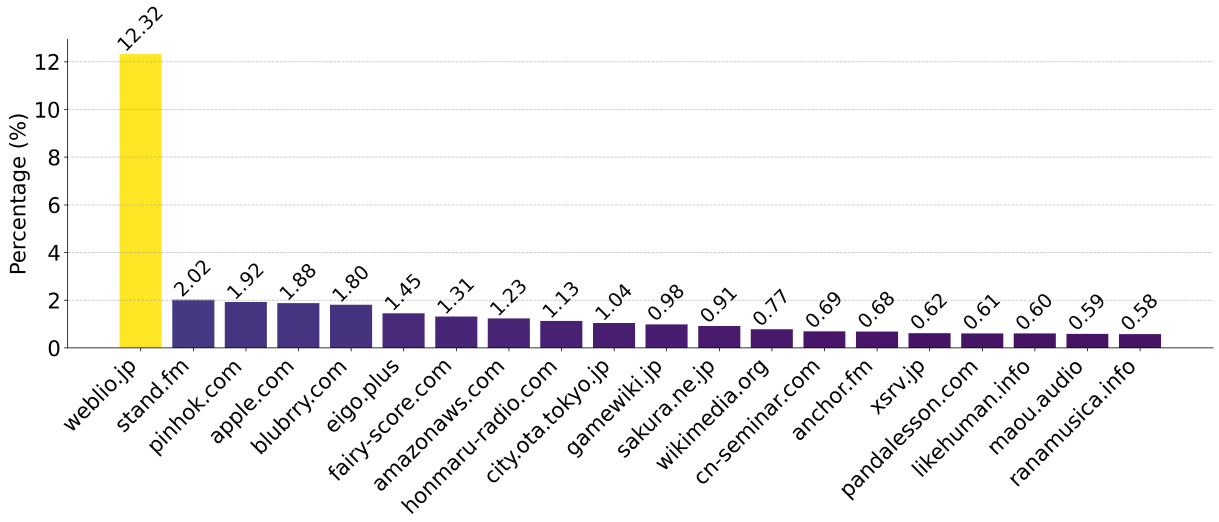

Fig. 1　HTML から抽出した音 URL のトップレベルドメインの頻度トップ 20.

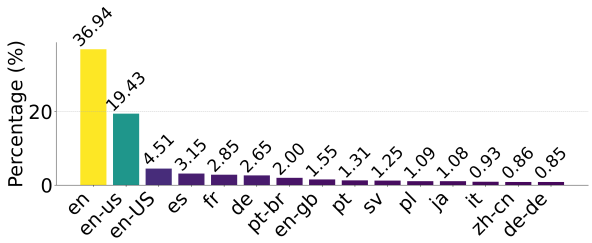

Fig. 2　RSS ベースの音 URL の language field のトップ 15.

なお, 音 URL の付近を調べても，ライセンスタグは 1 件も得られなかった．このことから, Web 上に添付される多くの音 URL にはライセンス情報が紐づけられていないことが示唆される.

2.2　RSS ベースのアプローチ

HTML の方法で得られた音 URL には, 非音声の音データも多く含まれていた．ここでは, ポッドキャストの音 URL がインデックスされた XML 形式の RSS フィードを収集し, そこから音 URL を抽出し, 音 URL ベースの重複削除を行う．その結果, 1 つの snapshot から, 5,060,098 事例の音 URL が取得できた．そのうち, XML の language field が日本語の事例は 56,728 事例あった．Language field のトップ 10 を図 2 に示す．英語の音 URL が圧倒的に多いことがわかる．Language field が日本語の音 URL のトップレベルドメインの頻度トップ 50 を図 3 に示す．HTML ベースの音 URL の場合の頻度トップ 50 と異なり，ポッドキャスト関連のドメインが多いことがわかる.

3　音声データの統計

本節では，データ収集プロセスの統計的性質と，構築されたコーパスの概要について報告する．以降の分析は，2.1 節あるいは 2.2 節で抽出した，日本語キャプションを持つ音 URL に基づいて実施している.

3.1　HTML ベースのアプローチ

以下に示す統計は，収集した音 URL の 1/1000 の無作為サンプルに対して分析を行った結果である.

データセット全体の時間長　1/1000 のサブセットの時間長は 38 時間であったので，データセット全体では 38000 時間程度と推定できる.

音ファイル形式　収集した音ファイルの形式を分析したところ，図 4 に示すように，ファイルの拡張子は.mp3 が 83.3% を占め，.m4a が 8.0% であった．この結果から，本コーパスの大部分は，非可逆のファイル形式で保存されており，圧縮された音データで構成されていることがわかる.

Whisper-AT によるタグ付け処理　収集した音ファイルのオーディオ種別と言語を分析するため，Whisper-AT [11] を用いたタグ付けを行った．Whisper-AT を用いたタグ付けには，音の種別の推定と，言語の推定が含まれる．このサブセットの中で，音声のダウンロードとタグ付け処理がどちらも成功したサンプルの割合は 88.4 %であった．以下では，Whisper-AT による処理に成功したデータセットについて，その特性を詳述する.

音の種別の分布　Whisper-AT によるタグ付け結果をもとに収集した音の種別を分析したところ，図 5 に示すように，全体の時間長のうち 58.9 %が Speech となった．よって，このデータセットはとりわけ音声に関するタスクにおいて有効に活用することができる．また，Conversation の割合は 2.9 %にとどまった.

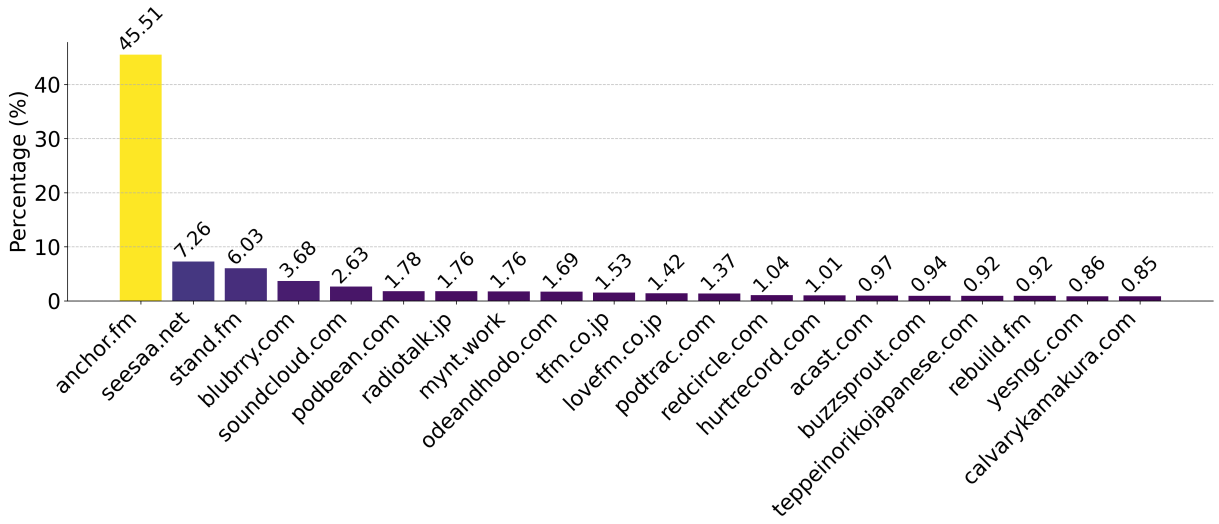

Fig. 3　language field が日本語の RSS から抽出した音 URL のトップレベルドメインの頻度トップ 20.

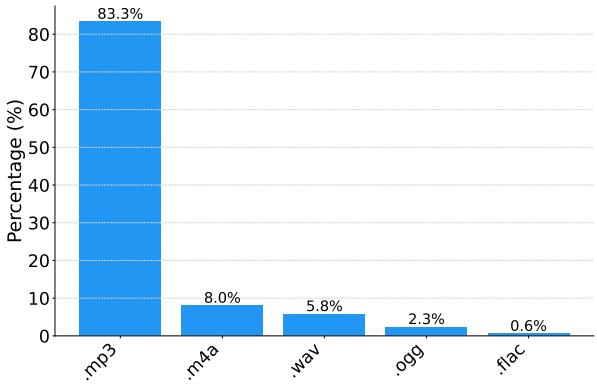

Fig. 4　HTML から抽出した音の音ファイル形式の分布.

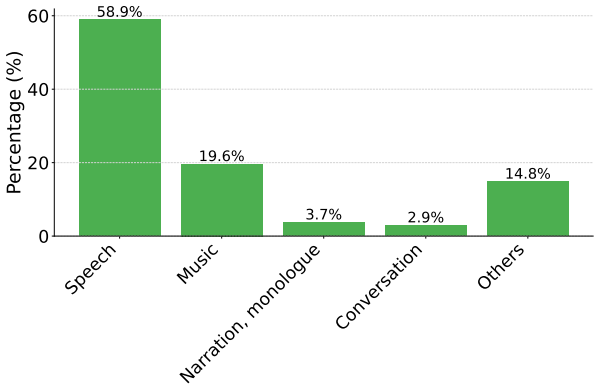

Fig. 5　HTML から抽出した音の種別の分布.

音声の言語の分布　Whisper-AT によるタグ付け結果をもとに，音声データに絞って言語を分析したところ，図 6 に示すように，94.7 %が日本語となった．よって，日本語音声の合計のデータ量は 32000 時間程度となる．

3.2　RSS ベースのアプローチ

以下に示す統計は，収集した音 URL の 1/10000 の無作為サンプルに対して分析を行った結果である．

データセット全体の時間長　1/10000 のサブセットの時間長は 45 時間であったので，データセット全体では 45 万時間程度と推定できる．

音ファイル形式　収集した音ファイルの形式を分析したところ，図 7 に示すように，ファイルの拡張子は.mp3 が 87.4% を占め，.m4a が 12.5% であった．この結果から，HTML ベースのアプローチと同様に，ほとんどが非可逆圧縮された音ファイルで構成されることがわかった．

Whisper-AT によるタグ付け処理　3.1 節と同様に Whisper-AT を用いたタグ付け処理を行ったところ，成功率は 39.8 %と HTML ベースのアプローチに比べて大きく下がる結果となった．この原因としては，音ファイルの URL が壊れている，無音区間が長すぎるなどの原因が考えられる．以下では，Whisper-AT による処理に成功したのデータセットについて，その特性を詳述する．

音の種別の分布　3.1 節と同様に音の種別を分析したところ，図 8 に示すように，全体の時間長のうち 63.9 %が Speech となった．よって，このデータセットはとりわけ音声に関するタスクにおいて有効に活用することができる．

音声の言語の分布　3.1 と同様に音声データに絞って言語を分析したところ，図 9 に示すように，日本語はわずか 4.0 %となった．よって，日本語音声の合計のデータ量は約 18000 時間程度となる．日本語音声の割合が低いのは，RSS ベースの音収集において language field を絞らずに収集していることを考えると自然な

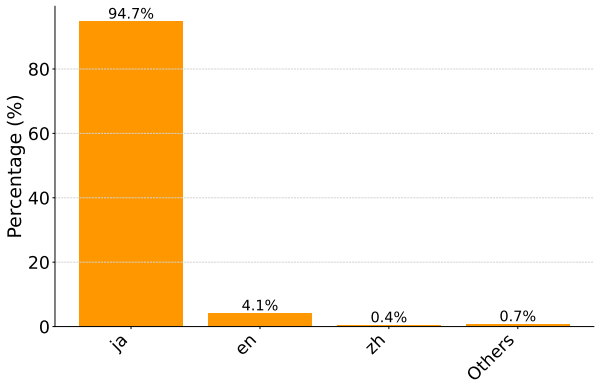

Fig. 6　HTML から抽出した音声の言語の分布．

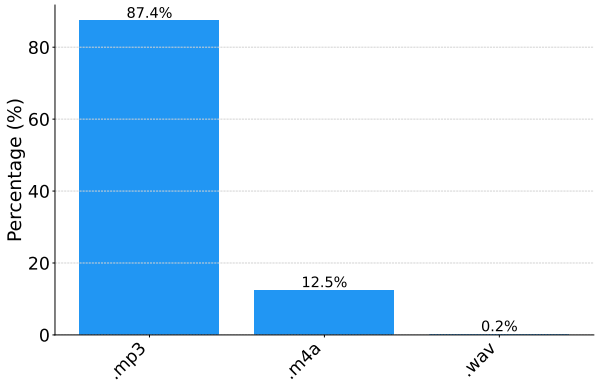

Fig. 7　RSS から抽出した音の音ファイル形式の分布．

結果である．

4　おわりに

研究では，Common Crawl を用いた大規模音コーパスの構築方法について検討した．

本手法では，HTML ベースの手法でも RSS ベースの手法でも日本語音声データは数万時間程度にとどまったため，大規模音声モデルの開発に必要な大量のデータを収集するためには，対象とする HTML ファイルの範囲を広げるだけでなく，Common Crawl 以外のデータソースを活用することも視野に入れて検

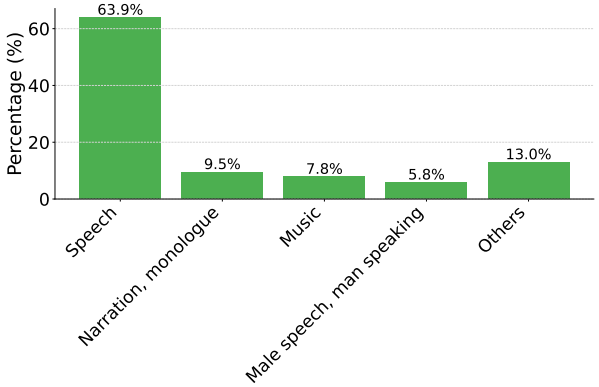

Fig. 8　RSS から抽出した音の種別の分布．

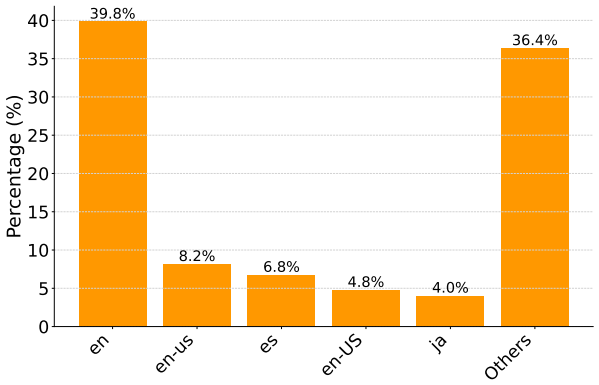

Fig. 9　RSS から抽出した音声の言語の分布．

討する必要があると考えられる．また，他の言語の音声も合わせて pre-training を行うことで，日本語にも活用できる可能性がある．

謝辞：本研究は，JSPS 科研費 23K24895，JST 創発的研究支援事業 JPMJFR226V の支援を受けて実施した．

参考文献

[1] G. Penedo et al., “The fineweb datasets: Decanting the web for the finest text data at scale,” in *Advances in Neural Information Processing Systems*, A. Globerson et al., Eds., vol. 37. Curran Associates, Inc., 2024, pp. 30 811–30 849.

[2] C. Schuhmann et al., “Laion-5b: An open large-scale dataset for training next generation image-text models,” 2022.

[3] V. Panayotov et al., “Librispeech: An asr corpus based on public domain audio books,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[4] D. Galvez et al., “The people’s speech: A large-scale diverse english speech recognition dataset for commercial usage,” in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

[5] X. Li et al., “Yodas: Youtube-oriented dataset for audio and speech,” 2024.

[6] Y. Wu et al., “Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation,” 2024.

[7] W. Nakata et al., “J-CHAT: Japanese large-scale spoken dialogue corpus for spoken dialogue language modeling,” 2024.

[8] A. Radford et al., “Robust speech recognition via large-scale weak supervision,” 2022.

[9] A. Défossez et al., “Moshi: a speech-text foundation model for real-time dialogue,” 2024.

[10] KimiTeam et al., “Kimi-audio technical report,” 2025.

[11] Y. Gong et al., “Whisper-at: Noise-robust automatic speech recognizers are also strong general audio event taggers,” in *Interspeech 2023*, 2023, pp. 2798–2802.