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 [1,2,3]　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] 産業技術総合研究所　[4] 早稲田大学
shinnosuke_takamichi@keio.jp, @forthshinji (X (ex: Twitter))

概要

本論文では，音声・音響・音楽信号を対象とするオープンな音基盤モデル構築に向けた，データセットの策定結果を報告する．汎用的な音基盤モデルを構築してその知見を共有するには，構築に資するデータセットを再現可能な形で整備すべきである．本論文では，音基盤モデルの満たすべき入出力条件を整理し，策定したデータセットについて分析する．

1　はじめに

大規模言語モデル（言語基盤モデル） [1, 2, 3] は，伝統的な自然言語処理タスクや zero-shot タスクにおいて顕著な性能を示しており，現代における自然言語処理の基盤技術となっている．これを受け，自然言語以外の様々なメディアデータ，センサデータに対する大規模モデルが期待される [4, 5]．音についても例外でなく，例えば，音声信号（人間の声） [6, 7, 8, 9]，音響信号（環境音，動物音声など） [10, 11]，音楽信号（器楽音など） [12, 13, 14] について，それぞれ基盤モデルの構築が進められている．これらは，各研究分野において，次なる基盤技術として創出されようとしている．

これを踏まえ我々は，音声・音響・音楽を扱う大規模モデル（音基盤モデル[1)]）について，以下のように考える．図 1 はその図示である．

- **音声・音響・音楽を統一的に扱えること**：これらの音は，これまで別の研究分野で扱われてきたが，物理（例えば波動方程式）と情報（例えば心理）の観点で通底するため，統一的な基盤モデルで扱うことが望ましい [15, 16, 17, 18]．また，記号（例えば自然言語，音楽記号）で記述されない音信号も含むべきである．
- **時間長，サンプリング周波数，チャネル数について任意の値を扱えること**：音信号は，どの種類でも可変長信号を前提とする点で共通するが，種類に応じて保存フォーマットが異なる．例えば，音声信号の主周波数帯域を扱うなら 16 kHz サンプリングが用いられるが，音楽信号はより広い 48 kHz サンプリングを要請する．同様に，音声信号はモノラル (1 ch) 信号で扱うことが多いが，空間情報を扱う場合はステレオから数十チャネルまで扱う必要がある．これらを踏まえると，音基盤モデルは，時間長，サンプリング周波数 [19]，チャネル数 [20, 21, 22] について任意の値をとる入出力信号を扱わなければならない．関連して，それ以外の録音条件（例えばマイクロホン配置）についても，任意の条件に対応すべきである．
- **言語学，物理学，心理学，生物学などの隣接分野に接続できること**：音響学は，言語学，物理学，心理学，生物学と隣接する学問（例えば音声言語，物理音響，音響心理，生物音響）である．そのため，音基盤モデルはこれらの学問にも利用できる

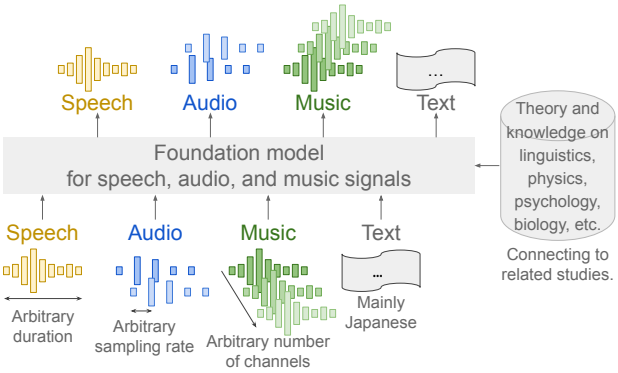

図 1　本研究で目標とする基盤モデル．

1)　本研究で扱う基盤モデルであらゆる音を扱うわけではないが，簡単のためこのように表記する．

ことが望ましい．

- **日本あるいは日本語を扱えること**：画像と同様に，音も土地や文化に依存する．本研究は特に日本語および日本を扱う基盤モデルを目標とする．
- **音研究を多分野に広く開放できること**：基盤モデルの登場により音声・音響・音楽研究のパラダイムが大きく変化し，音声・音響・音楽研究者以外も当該基盤モデルを扱うように変化すると予想される．このような時代を迎えるために，音基盤モデルのデータと開発環境を広く一般に公開しなければならない．
- **基盤モデルを扱える人材を創出できること**：言語基盤モデルプロジェクト LLM.jp [23] と同様に．基盤モデル開発を先導あるいは追従できる人材を育成しなければならない．

このような音基盤モデルを開発するための第一歩として本研究では，基盤モデル構築に用いるデータセット一覧を策定する．本論文では作成した一覧とメタ情報について分析した結果を報告する．これらの情報は https://github.com/sarulab-speech/audio-foundation-model-dataset にて入手可能である．

2 データセット策定

音信号において，自然言語における Common Crawl [2] のような，ウェブクロール済み共通データは多くない．これは，音信号の多くが著作権や財産権に強く関係するからである．そこで，既存の整備済み音データセットの情報を取集し，それに対しメタ情報を付与する．

データセットについて，以下の 2 種類を考える．

1. **学習用データ**：基盤モデルの学習に直接的に使用する，可聴音とそのメタ情報を含むデータ．現時点ではメタ情報として自然言語を主とするが，将来の発展を見据えて，画像を含んでもよいものとする．
2. **拡張用データ**：学習用データの拡張に使用する，可聴音とは限らない信号とそのメタ情報を含むデータ．例えば音響的な伝達関数（空間的な音の聞こえや，頭の形による音の聞こえ）と物理的なメタ情報（例えば部屋の大きさ，頭の形）．データ駆動でない拡張（例えば SpecAugment [24]）は，データ拡張に用いる可能性があるがデータセットではないため，ここには含めない．

前者の学習用データについて，各データセットに対して以下の情報を付与する．

- **音の種類**：音声，音響，音楽の別．それ以外の音（例えば超音波）は対象外とする．動物音声は音響に分別するものとする．
- **チャネル数**：モノラル (1 チャネル)，ステレオ (2 チャネル) などの別．3 チャネル以上のデータセットは “3 チャネル以上” とする．
- **サンプリング周波数**：サンプリング周波数．可聴音帯域を中心として 8 から 48 kHz を採る．可聴音帯域を超えるものは “48 kHz 以上” とする．
- **圧縮音源か否か**：可逆 (wav, flac など) あるいは非可逆（mp3 など）のファイルフォーマットの別．後者は，人間の心理聴覚特性に合わせて情報が欠落している場合がある．なお，非可逆フォーマットで保存した後に可逆フォーマットにした場合は，“非可逆” に分別する．
- **信号対雑音比 (SNR)**：目的音（例えば音声）に対する雑音（例えば車の音）のパワー比．スタジオなどの理想環境においては “雑音なし”，一般的な環境においては “雑音あり”，SNR の記載がある場合は数値とした．目的音を定義できない場合は “N/A” とする．
- **残響時間 (RT60)**：空間による響きの時間長．SNR と同様に，“残響あり”，“残響なし”，数値 (RT60 値) とする．
- **時間長**：音の時間長の合計値．この値には無音区間を含む．
- **言語**：音声の場合は発話言語，音響・音楽の場合はメタ情報の言語の別．“日本語”，“英語”，“中国語”，“それ以外” とする．
- **ライセンス**：コーパスのライセンス．本研究では当該コーパスの利用範囲が商用・非商用を問わず収集する．

後者の拡張用データについては，チャネル数，サンプリング周波数，信号対雑音比，残響時間，ライセンスを学習用データと同様に付与し[3]，さらに以下の情報を付与する．

- **伝達関数の種類**：室内伝達関数，頭部伝達関数，その他（室外など）の伝達関数の別．
- **マイクロホンアレイ数，インパルス応答数（伝達関数の数）**：観測に用いたマイクロホンアレイの

2） https://commoncrawl.org/

3） 可聴音とは限らないため，音の種類は付与しない

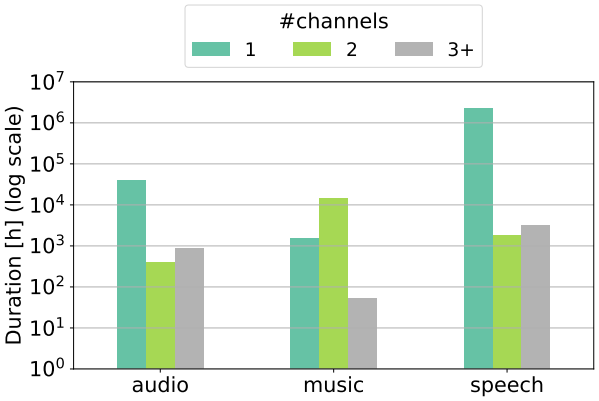

図 2　チャネル数ごとに時間数を集計した図.

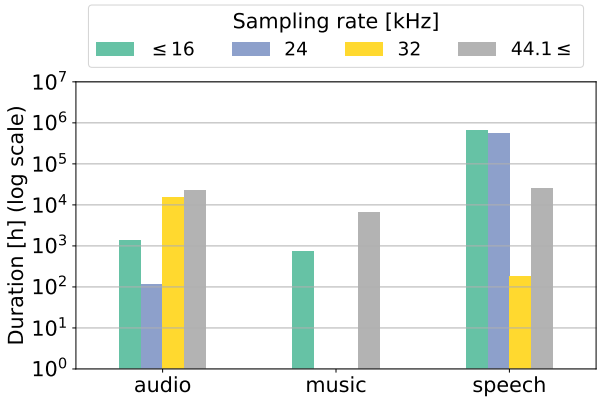

図 3　サンプリング周波数ごとに時間数を集計した図.

数およびそのチャネル数と，インパルス応答の総数.

- **観測場情報**：観測場に関する情報．例えば部屋のサイズ.

3　データセットの分析

作成期間は 2024 年 10 月から 12 月であり，それまでに公開されているデータセットを対象とした．学習用データの総時間数は約 220 万時間，データセット数は約 450 であった．各データセットに付与したメタ情報のうち約 2.5%については，投稿時点で正確な情報を付与できなかったため，以降の分析には含めていない.

3.1　学習用データの分析

学習用データに付与したメタ情報を分析した.

3.1.1　チャネル数

図 2 に，データセットの総時間数を，音種類ごとチャネル数ごとに集計した結果を示す．この図より，本研究で収集した学習用データの大半は，モノラルの音声信号であることがわかる．音響信号はモノラルが，音楽信号はステレオが主たるチャネル数だが，いずれも音声信号の 1 %程度の量である．また，音声信号においてもステレオ以上の時間数は極端に小さい．以上のことから，基盤モデルの構築時には，モノラル音声信号の処理と比較して，音響信号・音楽信号・ステレオ以上の信号の処理の性能を調査する必要がある.

3.1.2　サンプリング周波数

図 3 は，サンプリング周波数ごとの総時間数である．音声信号において主たるサンプリング周波数は

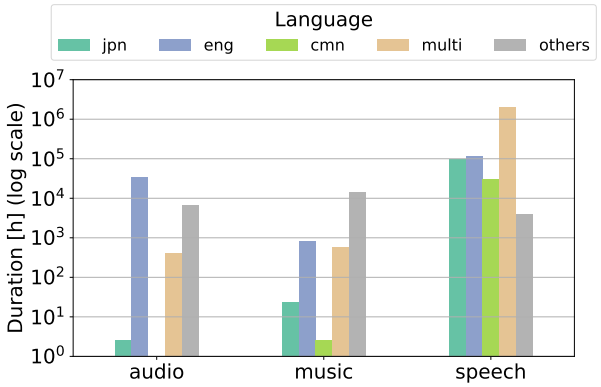

図 4　言語ごとに時間数を集計した図．“multi” は複数言語.

16, 24 kHz であり，それ以上のサンプリング周波数のデータは非常に限定的である．一方，人間の可聴周波数帯域を考慮すれば 44.1 kHz 以上が好ましい．そのため，基盤モデルの構築では高いサンプリング周波数の音声信号の処理について調査する必要がある．他方で，音響信号と音楽信号については，比較的高いサンプリング周波数のデータセットが多い．ただし，これらのデータセットの殆どは圧縮音源であることに注意されたい.

3.1.3　言語

言語ごとの総時間数を図 4 に表示している．興味深いことに，日本語音声は，英語音声および中国語音声と同程度の時間数である．ただし，これは単言語音声のデータセットに限った比較であり，複数言語の音声データセット (“Multi”) を集計に含めていない．複数言語の音声データセットの殆どは英語が主であるため，実際には，英語と日本語で数十倍程度の時間数の差がある．他方，音響信号と音楽信号において，日本語データセットはごくわずかしか含まれない．非日本語の音響・音楽データセットを翻

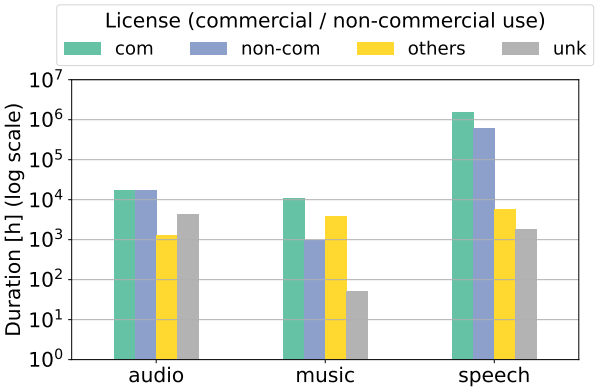

図 5　ライセンスごとに時間数を集計した図．“unk” はライセンス表記なし．

訳する方法も取りうるが，土地や文化に依存する文脈を基盤モデルが捉えられるかを，慎重に議論すべきである．

3.1.4　ライセンス

最後に，各データセットのライセンスを商用利用可能か否かに分類して，それぞれの総時間数を計算した．商用利用可能のライセンスは，“NC” を含まない Creative Commons License，MIT License, Apache License v2.0 などである．商用利用不可のライセンスは，“NC” を含む Creative Commons License, “non-commercial” や “research only” の表記のあるライセンスなどである．結果を 図 5 に示す．

いずれの音の種類においても半分以上の時間数は商用利用可能であることから，商用利用可能な基盤モデルの構築についても本リストの一定の貢献が期待される．音楽信号については，曲ごとにライセンスが異なるデータセットが多く，others に多く分類されている．音楽信号の時間数を増やすためにも，ライセンスの詳細な確認が必要である．

3.2　拡張用データの分析

学習用データと同様に分析した．

3.2.1　サンプリング周波数

図 6 に，伝達関数の種類ごとにインパルス応答の数を集計した結果を示す．多くの場合において 32 kHz 以上のサンプリング周波数であり，可聴周波数帯域を広くカバーできていることがわかる．ただし，現時点の伝達関数の種類ごとの総数については，室内伝達関数（room）が顕著に多く，それ以外は，室内伝達関数の 1/100 程度の量である．また，

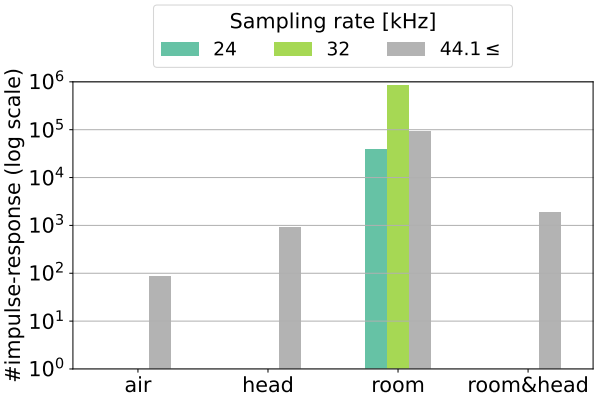

図 6　サンプリング周波数ごとにインパルス応答の数を集計した図．

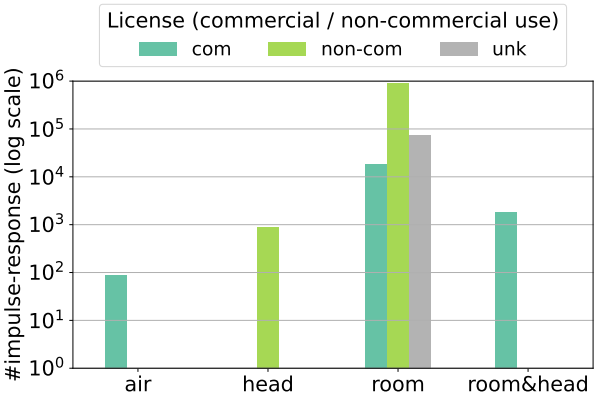

図 7　ライセンスごとにインパルス応答の数を集計した図．“unk” はライセンス表記なし．

単一の屋内環境の多数の点で測定されたインパルス応答を提供するデータセットも多く，環境の多様性も限られている．

3.2.2　ライセンス

図 5 と同様のライセンス分類で，拡張用データを分類した．その結果を図 7 に示す．データリストに含まれるほとんどの伝達関数が商用利用不可であり，商用利用可能なものは全体の数% 程度である．人工データ生成（例えば鏡像法 [25]）などにより，この欠落を補う方法の検討が必要である．

4　おわりに

本稿では，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を策定し，その分析を実施した．今後はデータローダの作成などを実施予定である．本論文のデータセットを用いた研究プロジェクトはオープンソースプロジェクトとして扱う．公開した際には，第一著者の連絡先や GitHub で contributor を募集する．

謝辞： 本研究は，創発的研究支援事業 JPMJFR226V，JSPS 科研費 23H03418，22H03639，JST ムーンショット型研究開発事業 JPMJMS2011 助成を受けたものです．

参考文献

[1] T. Brown et al., "Language models are few-shot learners," in **Advances in Neural Information Processing Systems**, H. Larochelle et al., Eds., vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901.

[2] A. Chowdhery et al., "PaLM: Scaling language modeling with pathways," **Journal of Machine Learning Research**, vol. 24, no. 240, pp. 1–113, 2023.

[3] H. Touvron et al., "LLaMA: Open and efficient foundation language models," 2023. [Online]. Available: https://arxiv.org/abs/2302.13971

[4] Gemini-Team-Google, "Gemini: A family of highly capable multimodal models," 2024. [Online]. Available: https://arxiv.org/abs/2312.11805

[5] OpenAI, "GPT-4 technical report," 2024. [Online]. Available: https://arxiv.org/abs/2303.08774

[6] D. Zhang et al., "SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities," in **Findings of the Association for Computational Linguistics: EMNLP 2023**, Dec. 2023, pp. 15 757–15 773.

[7] J. Wu et al., "On decoder-only architecture for speech-to-text and large language model integration," 2023. [Online]. Available: https://arxiv.org/abs/2307.03917

[8] S. Arora et al., "UniverSLU: Universal spoken language understanding for diverse tasks with natural language instructions," in **Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, K. Duh et al., Eds., Jun. 2024, pp. 2754–2774.

[9] P. K. Rubenstein et al., "AudioPaLM: A large language model that can speak and listen," 2023. [Online]. Available: https://arxiv.org/abs/2306.12925

[10] S. Deshmukh et al., "Pengi: An audio language model for audio tasks," in **Advances in Neural Information Processing Systems**, vol. 36, 2023, pp. 18 090–18 108.

[11] M. Hagiwara, "Aves: Animal vocalization encoder based on self-supervision," in **ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**, 2023, pp. 1–5.

[12] Y. LI et al., "MERT: Acoustic music understanding model with large-scale self-supervised training," in **The Twelfth International Conference on Learning Representations**, 2024.

[13] S. Liu et al., "Music understanding LLaMA: Advancing text-to-music generation with question answering and captioning," in **ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**, 2024, pp. 286–290.

[14] W. Liao et al., "Music foundation model as generic booster for music downstream tasks," 2024. [Online]. Available: https://arxiv.org/abs/2411.01135

[15] Y. Chu et al., "Qwen-Audio: Advancing universal audio understanding via unified large-scale audio-language models," **arXiv preprint arXiv:2311.07919**, 2023.

[16] R. Huang et al., "AudioGPT: Understanding and generating speech, music, sound, and talking head," 2023. [Online]. Available: https://arxiv.org/abs/2304.12995

[17] C. Tang et al., "SALMONN: Towards generic hearing abilities for large language models," in **The Twelfth International Conference on Learning Representations**, 2024. [Online]. Available: https://openreview.net/forum?id=14rn7HpKVk

[18] D. Yang et al., "UniAudio: An audio foundation model toward universal audio generation," 2024. [Online]. Available: https://arxiv.org/abs/2310.00704

[19] K. Saito et al., "Sampling-frequency-independent convolutional layer and its application to audio source separation," **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, vol. 30, pp. 2928–2943, 2022.

[20] Z. Zheng et al., "BAT: Learning to reason about spatial sounds with large language models," in **Proceedings of the 41st International Conference on Machine Learning**, ser. Proceedings of Machine Learning Research, R. Salakhutdinov et al., Eds., vol. 235. PMLR, 21–27 Jul 2024, pp. 61 454–61 469. [Online]. Available: https://proceedings.mlr.press/v235/zheng24i.html

[21] Anonymous, "Both ears wide open: Towards language-driven spatial audio generation," in **Submitted to The Thirteenth International Conference on Learning Representations**, 2024, under review. [Online]. Available: https://openreview.net/forum?id=qPx3i9sMxv

[22] C. Tang et al., "Can large language models understand spatial audio?" in **Interspeech 2024**, 2024, pp. 4149–4153.

[23] 河原大輔 et al., "LLM-jp: 日本語に強い大規模言語モデルの研究開発を行う組織横断プロジェクト," **自然言語処理**, vol. 31, no. 1, pp. 266–279, 2024.

[24] D. S. Park et al., "SpecAugment: A simple data augmentation method for automatic speech recognition," in **Interspeech 2019**, 2019, pp. 2613–2617.

[25] R. Scheibler et al., "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in **ICASSP 2018 - 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**, 2018, pp. 351–355.