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），○高道 慎之介，佐伯 高明（東京大学），
Chen William（CMU），塩田 さやか（東京都立大学），渡部 晋治（CMU）

1　はじめに

音声データセットの整備により，様々な end-to-end 音声認識合成モデル [1, 2, 3] や自己教師ありモデル [4, 5] を学習できるようになった．特に，多言語の発話内容ラベル付き [6, 7, 8, 9] あるいはラベル無しデータセット [10] の貢献は大きい．しかしながら，パブリックなデータセットの整備は技術発展に遅れており，例えば，Whisper や Google USM では 10 万時間以上のデータが学習に使用される [11, 12] ものの，相応サイズのパブリックなデータセットは現存しない．

これに対し本研究では，YouTube データを基とする多言語音声データセット YODAS (YouTube-oriented dataset of audio and speech) を提案する．本データセットは以下の 3 つのサブセットから成る．

1. manual：手動書き起こしを持つ約 8.6 万時間
2. automatic：手動書き起こしを持つ約 34 万時間
3. unlabeled：書き起こしのない約 14 万時間

上記サイズは 2023 年 7 月時点のものであり，今後増強する予定である．データセットは `https://huggingface.co/espnet` にて公開可能である．

2　データ収集

YouTube からのデータ収集にあたり，以下の 2 条件を設けた．

- 当該動画が Creative Commons でライセンスされていること．
- 手動あるいは自動書き起こしを有すること．ただし書き起こしのない動画も許容する．

この条件を満たす動画を収集するにあたり，jtube-speech ツールキット [16] を改良した．その枠組みは図 1 に示すように，複数のクライアントと単一の主ノードから成る．

2.1　キーワードクライアント

データ収集では，収集対象の動画を効率的に発見することが要請される．本クライアントでは，キーワード検索と YouTube のフィルタリング機能を用いてこれを達成する．

初期フェーズでは，多言語の Wikipedia 記事からキーワードを抽出することで，検索キーワードのリストを作成する．図 2 は，各言語の検索キーワード数の分布である．同図より英語のキーワードが大部分を占めることが分かる．しかしながら本研究ではデータ多様性を担保するために，全キーワードを検索するのではなく，主要言語以外の言語のキーワードの検索を優先する．次のフェーズでは，前述した 2 条件を満たすように検索フラグを立て，キーワード検索を実施した．この実装には，高関連度動画のみがヒットする HTTP リクエストではなく，低関連度動画もヒットする AJAX を用いる．

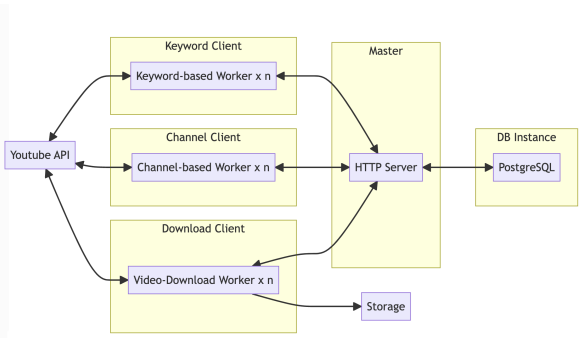

Fig. 1　データ収集．各クライアントは複数のワーカから成り，各ワーカはタスクを実行し主ノードあるいは YouTube と通信する．ダウンロードワーカはダウンロードしたデータを外部ストレージに転送する．

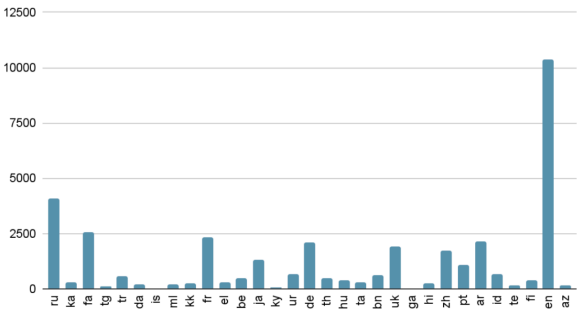

Fig. 2　検索キーワードの言語分布．

2.2　チャンネルクライアント

キーワード検索のみでは十分な数の動画を発見できない．これは，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である．そこで YouTube チャンネルに基づいて新たな動画を発見する．キーワード検索で発見された動画からその YouTube チャンネルを識別し，そのチャンネルに属する動画を新たな候補とする．この枠組みは，一つの YouTube チャンネルに属する動画のライセンスは似通るという仮定に基づく．

2.3　ダウンロードクライアント

本クライアントは動画と字幕のダウンロードを担う．重複ダウンロードを避けるため，本クライアントはダウンロード済み動画を格納するデータベースを監視する．動画の音フォーマットを 24 kHz モノラルに統一し，入手可能なすべての言語の字幕をダウンロードする．字幕のダウンロードでは，動画アップロードユーザの付与した手動字幕と，YouTube が自動付与した自動字幕を区別して保存する．

*YODAS: YouTube-oriented multi-lingual speech dataset, by Xinjian Li (CMU), Shinnosuke Takamichi, Takaaki Saeki (UTokyo), William Chen (CMU), Sayaka Shiota (Tokyo Metropolitan University), Shinji Watanabe (CMU)

Table 1　多言語大規模コーパスとの比較．YODAS は 50 万時間を超える初めてのオープンコーパスである．

Dataset	# Languages	Total Hours	Speech Type	Labeled	Public	License
BABEL [13]	17	1k hours	Spontaneous	Yes	Yes	IARPA Babel License
Common Voice [6]	112	18k hours	Read	Yes	Yes	CC-0
MLS [7]	8	50.5k hours	Read	Yes	Yes	CC BY 4.0
FLEURS [14]	102	1.4k hours	Read	Yes	Yes	CC BY 2.5
CMU Wilderness [8]	700	14k hours	Read	Yes	Yes	-
MMS-Lab [9]	1,107	44.7k hours	Read	Yes	No	-
VoxLingua107 [15]	107	6.6k hours	Spontaneous	Yes	Yes	CC BY 4.0
Librilight [10]	1	60k hours	Read	No	Yes	CC BY 4.0
Whisper[11]	97	680k hours	Unknown	Yes/No	No	-
USM [12]	300	12M hours	Spontaneous	Yes/No	No	-
YODAS (manual)	140	86k hours	Spontaneous	Yes	Yes	CC BY 3.0
YODAS (automatic)	20	336k hours	Spontaneous	Yes	Yes	CC BY 3.0
YODAS (unlabelled)	-	144k hours	Spontaneous	No	Yes	CC BY 3.0

Table 2　動画時間長 [hour] と発話時間長 [秒] の統計量．発話時間長はカッコ内の数値．

	Manual	Automatic	Unlabeled
Mean	0.15h (5.6s)	0.23h (3.2s)	0.15h (-)
Std	0.35h (8.9s)	0.37h (1.6s)	0.25h (-)
Min	0.00h (0.0s)	0.00h (0.1s)	0.00h (-)
Max	24.9h (42.1s)	24.9h (87.7s)	24.9h (-)

字幕が常に正しいとは限らないため，正しい字幕と言語 ID を識別する必要がある．本研究では，ヒューリスティックな方法で言語とダウンロード対象の字幕を決定する．もし当該動画が単一言語の手動字幕のみを有するならば，その言語 ID は正しいとみなす．単一言語の自動字幕のみを有する場合も同様に扱う．当該動画が複数言語の字幕を有する場合には，字幕をダウンロードせず，当該動画をラベル無し動画として扱う．なお，言語 ID 識別器による識別を試みたがその精度は不十分であった．

2.4　主ノード

全体を管理する主ノードを設ける．本ノードは，PostgreSQL データベースに接続され各ワーカからの GET/POST HTTP リクエストを受け付ける（すべてのワーカは HTTP クライアントとして機能する）．主ノードは各ワーカの状態を not-started, being processed, done のいずれかに設定する．being processed 状態は，別ワーカが同じリソース（例えば動画）を処理するのを防ぐためにある．各リソースの処理（例えばダウンロード）が完了したのち，当該ワーカは done 状態となる．

3　分析

3.1　データ量

手動字幕動画は manual サブセット，自動字幕動画は automatic サブセット，字幕なし動画は unlabeled サブセットとして扱う．表 2 は時間長の統計量である．automatic サブセットの動画時間長の平均は，他サブセットの平均よりも顕著に長いことが分かる．逆に，同サブセットの発話時間長の平均と標準偏差は，他サブセットよりも顕著に短い．

図 3 は，140 言語中で時間長の大きい 30 言語における時間長である．英語 (en) が首位であり，次点にスペイン語 (es) とロシア語 (ru) が続く．manual サブセットと automatic サブセットの量の傾向は言語間でおよそ変わらないが，14 言語のみに対応する automatic サブセットに比べ manual サブセットは 140 言語に対応する．

Table 3　テキストの文字数の統計量．

	Manual	Automatic
Mean	58.2	33.5
Std	27.6	8.6
Min	0.0	0.0
Max	588	44

3.2　テキスト分析

トップ 10 言語の書記体系は，ラテン，キリル，CJK 統合漢字，平仮名，ギリシャ，デーヴァナーガリー，ハングル，マラヤーラム，片仮名，アラビア文字である．ラテンアルファベットの頻度が最も大きく，次いでキリル文字が多い．

表 3 は，テキストの文字数の統計量である．manual サブセットはその平均も標準偏差も大きい傾向にあり，automatic サブセットはこの逆の傾向にある．

4　実験

YODAS は教師あり学習，弱教師あり学習，自己教師あり学習など多くの利用が可能だが，本論文では単言語の音声認識性能のみを評価する．

4.1　音声とテキストのアライメント

データセットのアライメントはノイジーであるため前処理を施す．学習済みの音響モデルを用いてアライメントを実施 [17, 18] し，CTC スコアの高い音声–テキスト対を残す [19]．

図 4 は，発話時間長とスコアの散布図である．図のファイルサイズを抑えるため，manual サブセットからランダムに抽出した 1000 発話のみの結果を載せている．スコアの低い外れ値（18.0）が一部含まれるものの，大部分は良好なスコア（5.0）である．図 5 は，automatic サブセットから同様にサンプリングした結果である．一定の割合で不良のスコア（20.0）が含まれることがわかる．これらの発話は長い時間長（例えば 50 秒以上）であり，主に音楽や背景雑音に起因する．

以降では，スコアのしきい値を 2.0 に定め，それより良い発話のみを使用する．その発話セットのうち，

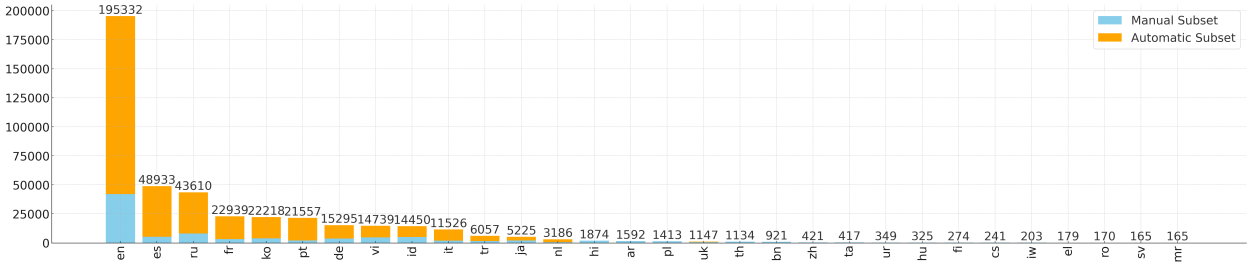

Fig. 3　言語ごとの時間長.

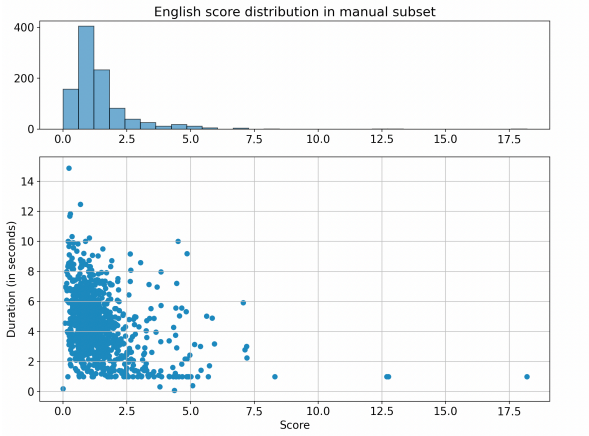

Fig. 4　発話時間長とアライメントスコアのヒストグラムと散布図（manual サブセット）.

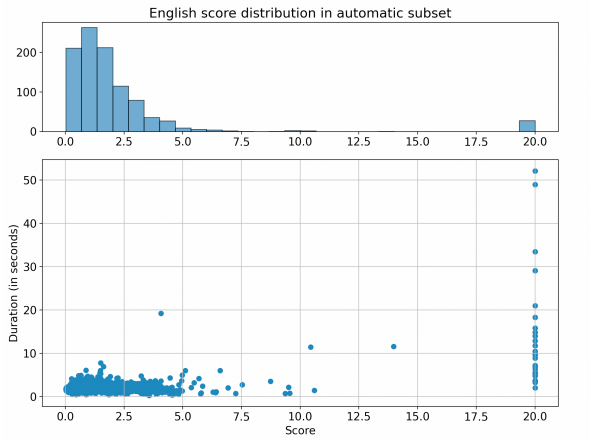

Fig. 5　発話時間長とアライメントスコアのヒストグラムと散布図（automatic サブセット）.

およそ 100 万発話をランダムに選択し，各言語の学習セットする．同様に，1000 発話をテストセットとする．

4.2　ベースライン

manual サブセットのトップ 25 言語について単言語音声認識モデルを構築した．モデルは学習済み XLSR モデル [20] の後にランダム初期化した線形層を追加したもの，目的関数は CTC loss [17] とした．実装には ESPnet [21] と s3prl [22] を用いた．サブワード語彙は，SentencePiece [23, 24] を用いて BPE 基準で構築した．CJK 統合漢字の言語を除いて，語彙数は 300 とした．中国語と日本語の語彙数はそれぞれ 5000, 3000 とした．実装の簡易化のため，データ拡張は実施しなかった．オプティマイザは，学習率を 0.0001 とした AdanW [25] である．

4.3　結果

表 4 にトップ 15 言語の結果を示す．文字誤り率（CER）は 6 から 15 までの値をとっている．最良結果の言語は 6.2 を記録したハンガリー語であり，最低結果の言語は 14.7 を記録した日本語だった．全言語の平均 CER は 9.97 であった．全体傾向として，語彙数が大きいほど CER が悪く，表音文字を頻用する言語では CER が良い傾向にある．

次に，manual サブセットと automatic サブセットの品質を比較した．公正な実験のため，10 万発話を各サブセットからランダムに抽出した．実験は英語のみで実施した．表 5 はその結果である．automatic

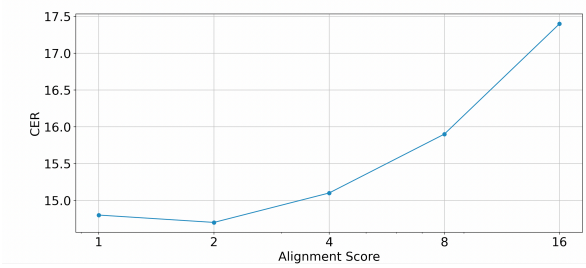

Fig. 6　アライメントスコアと音声認識性能の比較.

サブセットと比較して manual サブセットが顕著に優れており，この差は特に削除誤りに起因していることがわかる．自動書き起こしを学習に使用して性能が劣化する傾向は，既存研究 [11, 26] に一致する．

これまでの実験ではスコアの閾値を 2.0 に設定した．この閾値の影響を調査するために様々な閾値において学習データセットを設定した．学習データセットは，manual サブセットからランダムに抽出した 10 万発話（約 160 時間）とした．テストデータセットは全設定で共通である．図 6 より，閾値 16 から 2 にかけては単調に減少するものの，2 から 1 にかけて僅かに増加することがわかる．この僅かな増加は，スコアの高い発話は比較的短くなる傾向にあるためだと思われる．

最後に，学習データセットのサイズの影響を調査した．発話は英語の manual サブセットからランダムに抽出した．スコア閾値は 2.0 である．表 7 より，

Table 4　manual サブセットにおけるトップ 15 言語の音声認識結果.

language	ell	nld	hun	pol	por	cmn	ind	jpn	tur	ita	deu	fra	spa	rus	eng	average
CER (↓)	8.3	8.4	6.2	6.4	7.1	12.5	10.2	14.7	8.7	8.6	10.1	12.9	9.8	12.8	12.9	9.97

Table 5　音声認識結果における挿入, 削除, 置換誤り.

	CER (↓)	Add (↓)	Del (↓)	Sub (↓)
Manual	14.9	3.2	7.5	4.2
Automatic	32.3	1.6	26.6	4.1

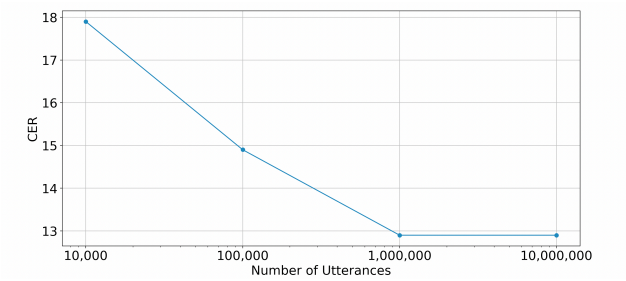

Fig. 7　学習データサイズと音声認識性能の比較.

データ量の増加により単調に性能が向上する．興味深いことに，100 万発話で学習したモデルの性能は，1000 万発話で学習したモデルと同程度の性能である．この現象は，シンプルなモデル構造に起因すると予想される．

5　まとめ

YouTube から構築された大規模多言語音声データベース YODAS を提案し，そのデータサイズと音声認識実験結果を述べた．

謝辞: 本研究は，アメリカ国立科学財団資金番号 #2138259, #2138286, #2138307, #2137603, #2138296 により支援された，PSC Bridges2 と NCSA Delta via ACCESS allocation CIS210014 を使用した．また本研究は科研費 21H04900, 22H03639, 23H03418, JST 創発的研究支援事業 JP23KJ0828, ムーンショット JPMJPS2011 の助成を受けた．

参考文献

[1] A. Graves et al., “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[2] R. Collobert et al., “Wav2letter: an end-to-end convnet-based speech recognition system,” *arXiv preprint arXiv:1609.03193*, 2016.

[3] R. Prabhavalkar et al., “End-to-end speech recognition: A survey,” *arXiv preprint arXiv:2303.03329*, 2023.

[4] A. Baevski et al., “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[5] W.-N. Hsu et al., “Hubert: Self-supervised speech representation learning by masked prediction of hidden units,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[6] R. Ardila et al., “Common voice: A massively-multilingual speech corpus,” *arXiv preprint arXiv:1912.06670*, 2019.

[7] V. Pratap et al., “Mls: A large-scale multilingual dataset for speech research,” *arXiv preprint arXiv:2012.03411*, 2020.

[8] A. W. Black, “Cmu wilderness multilingual speech dataset,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5971–5975.

[9] V. Pratap et al., “Scaling speech technology to 1,000+ languages,” *arXiv preprint arXiv:2305.13516*, 2023.

[10] J. Kahn et al., “Libri-light: A benchmark for asr with limited or no supervision,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7669–7673.

[11] A. Radford et al., “Robust speech recognition via large-scale weak supervision,” in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.

[12] Y. Zhang et al., “Google usm: Scaling automatic speech recognition beyond 100 languages,” *arXiv preprint arXiv:2303.01037*, 2023.

[13] M. J. Gales et al., “Speech recognition and keyword spotting for low-resource languages: Babel project research at cued,” in *Fourth International workshop on spoken language technologies for under-resourced languages (SLTU-2014)*. International Speech Communication Association (ISCA), 2014, pp. 16–23.

[14] A. Conneau et al., “Fleurs: Few-shot learning evaluation of universal representations of speech,” in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 798–805.

[15] J. Valk, T. Alumäe, “Voxlingua107: a dataset for spoken language recognition,” in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 652–658.

[16] S. Takamichi et al., “Jtubespeech: corpus of japanese speech collected from youtube for speech recognition and speaker verification,” *arXiv preprint arXiv:2112.09323*, 2021.

[17] A. Graves, N. Jaitly, “Towards end-to-end speech recognition with recurrent neural networks,” in *International conference on machine learning*. PMLR, 2014, pp. 1764–1772.

[18] L. Kürzinger et al., “Ctc-segmentation of large corpora for german end-to-end speech recognition,” in *International Conference on Speech and Computer*. Springer, 2020, pp. 267–278.

[19] X. Li et al., “Universal phone recognition with a multilingual allophone system,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 8249–8253.

[20] A. Babu et al., “Xls-r: Self-supervised cross-lingual speech representation learning at scale,” *arXiv preprint arXiv:2111.09296*, 2021.

[21] S. Watanabe et al., “ESPnet: End-to-end speech processing toolkit,” in *Proceedings of Interspeech*, 2018, pp. 2207–2211. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1456

[22] S. wen Yang et al., “SUPERB: Speech Processing Universal PERformance Benchmark,” in *Proc. Interspeech 2021*, 2021, pp. 1194–1198.

[23] R. Sennrich et al., “Neural machine translation of rare words with subword units,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1715–1725. [Online]. Available: https://aclanthology.org/P16-1162

[24] T. Kudo, “Subword regularization: Improving neural network translation models with multiple subword candidates,” *arXiv preprint arXiv:1804.10959*, 2018.

[25] I. Loshchilov, F. Hutter, “Decoupled weight decay regularization,” *arXiv preprint arXiv:1711.05101*, 2017.

[26] B. Ghorbani et al., “Scaling laws for neural machine translation,” *arXiv preprint arXiv:2109.07740*, 2021.