YODAS: YOUTUBE-ORIENTED DATASET FOR AUDIO AND SPEECH

Xinjian Li[1], Shinnosuke Takamichi[2],Takaaki Saeki[1], William Chen[1], Sayaka Shiota[3], Shinji Watanabe[1]

[1]Carnegie Mellon University, USA
[2]The University of Tokyo, Japan
[3]Tokyo Metropolitan University, Japan

ABSTRACT

In this study, we introduce YODAS (YouTube-Oriented Dataset for Audio and Speech), a large-scale, multilingual dataset comprising currently over 500k hours of speech data in more than 100 languages, sourced from both labeled and unlabeled YouTube speech datasets. The labeled subsets, including manual or automatic subtitles, facilitate supervised model training. Conversely, the unlabeled subsets are apt for self-supervised learning applications. YODAS is distinctive as the first publicly available dataset of its scale, and it is distributed under a Creative Commons license[1]. We introduce the collection methodology utilized for YODAS, which contributes to the large-scale speech dataset construction. Subsequently, we provide a comprehensive analysis of speech, text contained within the dataset. Finally, we describe the speech recognition baselines over the top-15 languages.

Index Terms— multilingual speech processing, speech recognition, large-scale speech dataset

1. INTRODUCTION

In recent years, significant advancements have been achieved in the field of speech recognition. With a sufficiently large speech dataset, it becomes feasible to train various end-to-end models using objectives such as CTC, ASG, seq2seq, RNN Transducer, and others [1, 2, 3, 4, 5]. We also observe the trend of using self-supervised learning models such as HuBERT and wav2vec2 to take advantage of unlabeled datasets [6, 7, 8]. Those improvements have been realized primarily through the utilization of large-scale multilingual speech datasets. For example, the BABEL project was a pioneering endeavor that scaled multilingual capabilities significantly [9]. The Common Voice project, facilitates an online speech collection interface, offering speech datasets in over 100 languages and encompassing 18,000 hours validated recording hours [10]. The MLS, a multilingual dataset, was derived from Librispeech [11]. Concerning linguistic diversity, the CMU Wilderness and MMS-Lab dataset, originating from the religious domain, covers nearly 1,000 languages [12,
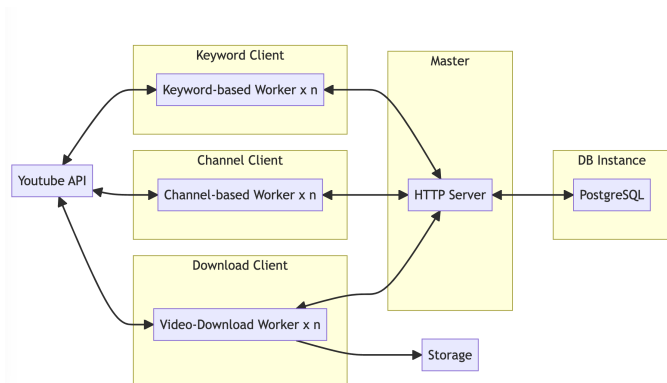
Fig. 1. Diagram of our data collection architecture: It incorporates three types of clients: Keyword-based, Channel-based, and Download workers. Each worker fulfills specific tasks and interacts with both the master node and the YouTube platform. Additionally, the Download worker also handles the transfer of downloaded data to external storage.

13]. Unlabeled dataset such as Libri-light has also been applied successfully to train self-supervised models such as HuBERT, wav2vec2, and WavLM [14, 7, 6, 15].

Despite the achievements with large-scale datasets, most public speech datasets available do not exceed 100,000 hours. In contrast, industry-utilized speech models are typically much more extensive. For instance, Whisper and Google's USM, have been trained with over 100,000 hours and up to 1,000,000 hours of data, respectively [16, 17]. However, the details of the datasets used to train these models remain undisclosed, which makes it difficult to reproduce those models. Addressing the limitation of the lack of industry-scale large dataset, this paper presents YODAS (YouTube-Oriented Dataset of Audio and Speech)—a large-scale multilingual dataset, which comprises the following three subsets:

1. The *manual* subset encompasses 86,400 hours of audio data paired with manual transcriptions.

2. The *automatic* subset, containing 335,845 hours of audio data, is supplemented with automatic transcrip-

tions.

3. The *unlabeled* subset consists of 144,174 hours of raw audio data, devoid of any transcription.

When used conjointly, the *manual* and *automatic* subsets offer a comprehensive resource of 480k hours for supervised model training. Meanwhile, all three subsets may be used in conjunction with the application of self-supervised learning techniques. It's worth noting that the combined subsets will result in an extensive corpus of over 560k hours by July 2023, and this amount will continue to grow over time. This marks the first time a dataset of this scale with the Creative Commons license has been made publicly available. We plan to release it from the Huggingface datasets.

2. DATA COLLECTION

We designed our data collection architecture to fulfill two specific requirements. Firstly, the video content must be accompanied by a Creative Commons license. Secondly, the video should possess either an automatic subtitle or a manual subtitle as much as possible, although we also allow unlabeled videos. In order to meet these criteria, we devised a framework by improving upon an existing toolkit [20]. Our framework is depicted in Figure 1, which utilizes three distinct clients and a master node which we will discuss next. Our data collection software will be open to the public for individual use or collaborative efforts.

2.1. Keyword-based Client

The principal challenge within our data gathering pipeline lies in the efficient identification of proper videos to download. This is achieved by implementing keyword-based crawling and combining it with YouTube's native filtering feature, allowing us to pinpoint a subset of relevant videos.

In the first phase, we construct a list of keywords by extracting unique keywords from the dumps of multilingual Wikipedia articles. Figure 2 presents the language distribution of unique query keywords within one of our data shards. As depicted by the figure, English commands the majority share. Rather than querying every keyword from this distribution, we prioritize those derived from less prominent languages trying to enhance the diversity of our video dataset. Subsequently, we initiate a keyword-based query on YouTube by appending appropriate flags, enforcing that the videos listed in the search results should (mostly) possess subtitles and adhere to the Creative Commons license. The naive HTTP GET request tends to yield a subset of videos that have the highest relevance to the supplied keywords. In order to capitalize on the number of available videos, we use AJAX to dynamically crawl lower-ranked videos, mimicking the process of scrolling down the search result page.
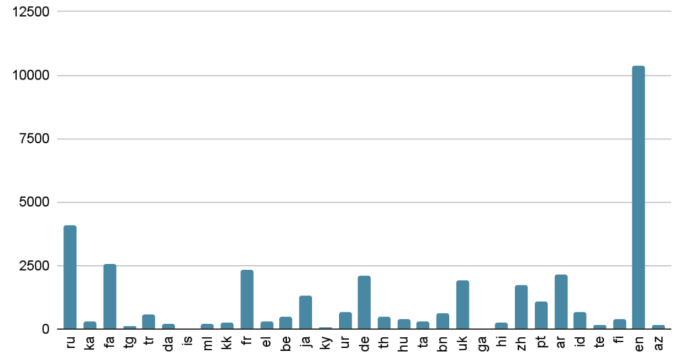
Fig. 2. language distribution of unique query keywords used in one of our shards (i.e., worker).

2.2. Channel-based Client

The keyword-based crawling approach alone, unfortunately, is insufficient in scaling our dataset to the substantial size that we desire. This is due to the tendency of YouTube to repetitively present the same popular video subset in its search results rather than proposing new content.

To mitigate this challenge and broaden our video exploration, we combine keyword-based crawling with a channel-based crawling strategy. For each video discovered during the keyword-based search, we extract the corresponding channel. This channel then serves as a means to identify all affiliated videos with ease. This method significantly aids in the discovery of videos that are otherwise less likely to appear using the keyword-centric approach. Most crucially, videos hosted within the same channel typically share similar licensing and subtitle characteristics, simplifying the process of identifying new videos that meet our specific criteria.

2.3. Download Client

The aforementioned workers solely undertake the task of identifying the correct video; however, the responsibility of downloading the video or its subtitles falls to the download worker. This worker perpetually monitors the database to ascertain whether a new video has been discovered by its predecessors. Upon identifying new videos, The download worker first retrieves the video, converting it into an audio format encoded at 1 channel and 24 kHz. Subsequently, it downloads the list of all available subtitles. Each video may either have multiple subtitles or none at all. The subtitles can be either manually uploaded by the user (manual subtitle) or automatically recognized by YouTube as enabled by the user (automatic subtitle). One significant challenge in this endeavor is determining the "correct" subtitle and the language of each video. Surprisingly, many videos possess multiple subtitles across diverse languages. We employ a heuristic method to identify the language and choose which subtitle to download. If the target video only has a unique

Dataset	# Languages	Total Hours	Speech Type	Labeled	Public	License
BABEL [9]	17	1k hours	Spontaneous	Yes	Yes	IARPA Babel License
Common Voice [10]	112	18k hours	Read	Yes	Yes	CC-0
MLS [11]	8	50.5k hours	Read	Yes	Yes	CC BY 4.0
FLEURS [18]	102	1.4k hours	Read	Yes	Yes	CC BY 2.5
CMU Wilderness [12]	700	14k hours	Read	Yes	Yes	-
MMS-Lab [13]	1,107	44.7k hours	Read	Yes	No	-
VoxLingua107 [19]	107	6.6k hours	Spontaneous	Yes	Yes	CC BY 4.0
Librilight [14]	1	60k hours	Read	No	Yes	CC BY 4.0
Whisper[16]	97	680k hours	Unknown	Yes/No	No	-
USM [17]	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 1. A comparison of YODAS dataset with a few other large-scale multilingual datasets. Our YODAS dataset is the first public dataset to reach a scale of over 500k hours.

manual subtitle with no other subtitles, it's highly probable this singular subtitle accurately reflects the language. We then proceed to download this subtitle and assign the language ID. Similarly, if the target video only has one automatic subtitle, we consider it to be accurate and proceed to download this subtitle. However, when a video has more than one manual or automatic subtitle with conflicting language IDs, the task of identifying the correct language becomes complicated. In such cases, we forego downloading the subtitles and leaving the video unlabeled. Although we attempt to identify the language using Language Identification (LID) tools, the results are not significantly successful. Therefore, we earmark this identification task for future work.

2.4. Master Node

In addition to those clients, we deploy a master node to monitor the overall progress. This node is connected to a PostgreSQL database and hosts an HTTP server that accepts GET/POST HTTP requests from each worker. The master node manages all resources—be it keywords, channels, or videos—each marked with one of three states: not-started, being processed, or done. The 'being processed' state serves to prevent simultaneous downloads of the same resource by separate workers. All the workers typically function as HTTP clients, querying the master node to request the next available resource. Once the resource has been resolved (i.e., downloaded), these workers mark it as done.

3. ANALYSIS

3.1. Overview

As described in the previous section, the YODAS dataset was segmented into three subsets, namely: *Manual*, *Automatic*,

and *Unlabeled*. Videos within the manual subset are characterized by user-uploaded, (possibly) manually generated subtitles, while those included in the automatic subset have associated automated subtitles. Conversely, the unlabeled subset consists of videos devoid of subtitles, primarily due to our current inability to accurately identify the language.

3.2. Speech Analysis

Table 2 presents the key statistics concerning the distribution of raw video durations and utterance durations (shown in parentheses) within our datasets. Notably, the raw video duration of the automatic subset exhibits a notably higher mean duration and standard deviation compared to the other two datasets. Conversely, the average duration of utterances and its standard deviation are notably lower in the automatic subset as compared to the manual subset. This is because YouTube tends to chunk speech into small segments as we will discuss further in the next subsection. In total, we have compiled an extensive dataset comprising 86,000 hours for the manual subset, 336,000 hours for the automatic subset, and an additional 144,000 hours for the unlabeled subset.

	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 (-)

Table 2. Descriptive statistics of raw video duration distribution (measured in hours) and utterance duration distribution in parentheses (measured in seconds) in three subsets.

Next, the language distribution of the manual subset and
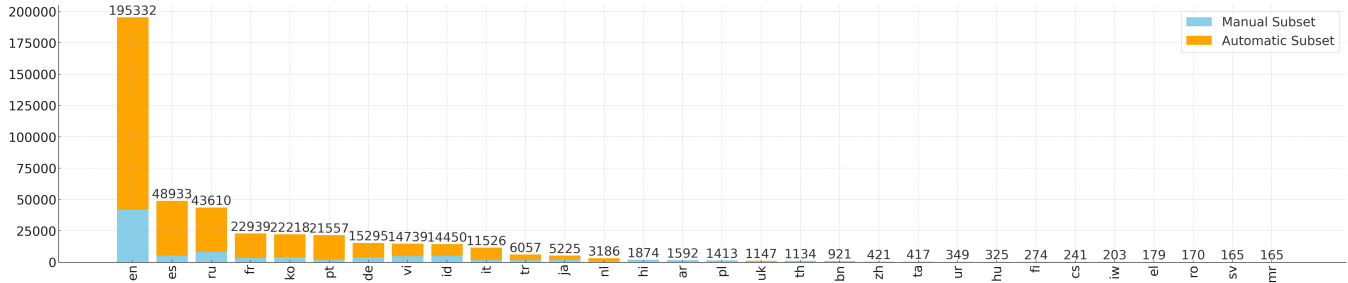
Fig. 3. Total duration (measured in hours) in the manual and automatic subset. The lower-blue bar shows the duration of the manual subset, the top-orange bar indicates the automatic subset. The combined duration is illustrated on top of each bar.

the automatic subset is illustrated in Figure 3. This figure presents the distribution of the top 30 languages out of a total of 140 evaluated. As anticipated, English (en) emerges as the most prevalently used language, with Spanish (es) and Russian (ru) occupying the second and third positions, respectively, when assessed based on duration. Although the language distribution trend appears similar between the automatic and manual subsets, the automatic subset has only a very limited number of languages (14 languages) compared with the manual subset (140 languages).

3.3. Text Analysis

The top 10 writing systems in our dataset are Latin, Cyrillic, CJK, Hiragana, Greek, Devanagari, Hangul, Malayalam, Katakana, and Arabic. It largely corresponds to the aforementioned language distribution, with the Latin alphabet appearing as the most frequently used writing system. The Cyrillic script, originating from Russian language videos, also features prominently within our dataset.

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

Table 3. Descriptive statistics of subtitle transcription measured in the number of characters in two subsets.

Table 3 is the comparison of character length distribution from the manual subset and automatic subset. The manual subset tends to have a larger number of characters per utterance and have a larger variance. Conversely, the automatic subset has an even distribution where most utterances are short and have little variance. This is because the automatic subtitle frequently divides long utterances into small chunks as Table 4 indicates, this splitting might be a feature to help viewers to follow subtitles easier.

utt id	automatic transcription
00682	if you're trying to do something in your
00683	community and you're spending your money
00684	public money or somebody's money to
00685	really do something that makes a

Table 4. A sample of the automatic transcriptions, one individual utterance is usually divided into multiple small chunks.

4. EXPERIMENT

The YODAS dataset offers a versatile resource that can be employed for a variety of tasks including supervised training, weakly-labeled supervised training, and self-supervised training. In this work, we focus specifically on the use of the dataset for the monolingual speech recognition task.

4.1. Speech-Text Alignment

The raw dataset, as collected, presents considerable noise with regard to speech-text alignment, suggesting that its unfiltered usage might be inappropriate. There are instances where neither the manual nor automatic subtitles accurately represent the underlying spoken discourse. For instance, subtitles occasionally serve as concise descriptors of the current scene in a video, annotating elements such as laughter or musical segments, rather than transcribing the actual spoken dialogue. Our heuristics to decide the language based on the list of subtitles might also introduce errors, those errors possibly arising from user inaccuracies or misidentifications in YouTube's language detection system.

To filter the dataset, we first apply the speech-text alignment. In particular, we use a pre-trained acoustic model to score every utterance in the audio and only consider using the high-scoring utterance pairs [21]. The score is obtained from the CTC loss where a lower value (loss) implies a better alignment [22, 23]. Figure 4 presents a scatter plot depicting the relationship between the duration and alignment score of 1,000 utterances randomly sampled from the *manual* subset. From the plot, it is evident that while there are occasional

outliers with poor alignment scores such as 18.0, the majority of utterances exhibit a duration of less than 10 seconds and possess an alignment score superior to 5. Conversely, Figure 5 portrays a scatter plot derived from the *automatic* subset. This plot reveals a significant proportion of misaligned utterances. It should be noted that for analytical purposes, all scores exceeding 20 have been capped at 20. These misaligned utterances typically display a considerable duration, often extending to as long as 50 seconds, and are predominantly attributed to music or background noise.
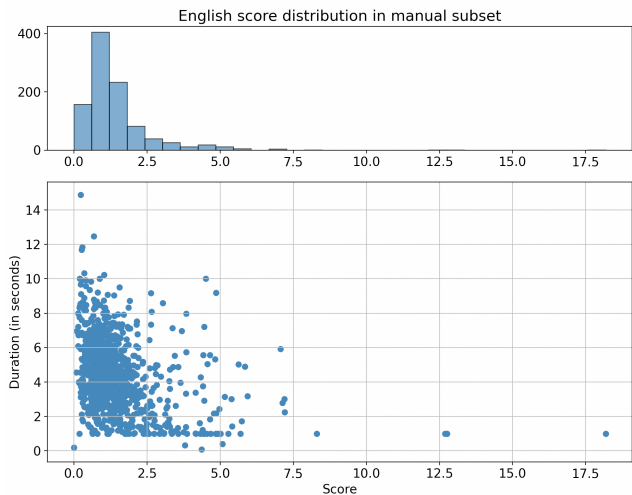
Fig. 4. the score histogram and scatter plot of the relationship between the duration and the alignment in the manual subset.
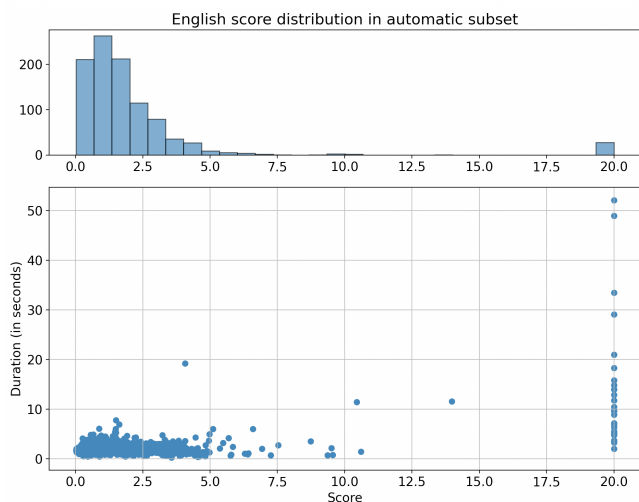
Fig. 5. the score histogram and scatter plot between the duration and the alignment score in the automatic subset.

In the subsequent experiment, we employ an alignment threshold to exclude utterances with scores worse than **2.0**. From the refined subset, a sample of 1,000,000 utterances (at most) is randomly selected to constitute the training set for each language, while a separate, smaller selection of 1,000 utterances is assigned to the testing dataset.

4.2. Baseline

We build simple monolingual baseline models for the top-25 languages in the manual subset. Our model is based on the pre-trained XLSR representations [24], where we have a linear layer randomly initialized on top of the pre-trained representations, which is then optimized with the CTC loss [22]. The preparation is done by using ESPnet [25] and s3prl [26]. The subword vocabulary is prepared with BPE using SentencePiece [27, 28], where we use 300 as the vocabulary size for most languages except for CJK languages where we use 5000 for Mandarin and 3000 for Japanese. For simplicity, we do not perform speech augmentation such as SpecAugment [29] and Speed Perturbation [30]. The acoustic model is optimized with the AdamW optimizer with a fixed learning rate of 0.0001 [31]. The decoding is done greedily without any language models.

4.3. Results

Table.5 displays the testing outcomes for the top-15 languages, measured by the Character Error Rate (CER). The respective CER for each language spans from 6 to 15. The best performance is recorded for Hungarian, with a CER of 6.2, while Japanese exhibits the least performance with a CER of 14.7. The average CER across all languages is 9.97. We observe that languages possessing a larger BPE vocabulary size, such as Mandarin (cmn) and Japanese (jpn), tend to correspond with higher character error rates (Mandarin has a CER of 12.5 and Japanese has a CER of 14.7). Conversely, languages that adhere to more straightforward spelling rules generally exhibit lower character error rates. For example, the writing system in Hungarian is mostly phonemic and achieves the lowest CER 6.2 in our experiment.
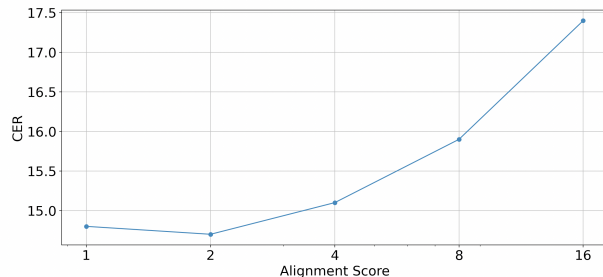
Fig. 6. the relationship between the alignment score and its performance on the speech recognition task.

We subsequently analyzed the quality of the dataset across both the manual and automatic subsets by training independent models solely on each subset. For a balanced compar-

language	ell	nld	hun	pol	por	cmn	ind	jpn	tur	ita	deu	fra	spa	rus	eng	average
CER ($\downarrow$)	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. monolingual speech recognition performance on the top-15 languages (ordered by the duration) from the manual subset. The evaluation is done by using character error rate (CER) where a lower number indicates a better performance.

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

Table 6. A comparison of the speech models trained with the manual subset and the automatic subset. We demonstrate both the CER and its error decomposition of Addition (Add), Deletion (Del), and Substitution (Sub).

ison, 100,000 utterances were randomly selected from each subset, post application of the 2.0 score filter, as introduced in Section 4.1. This comparison was solely conducted within the English subset. The findings, as displayed in Table 6, reveal that models trained on the manual subset yield significantly superior performance compared to those trained on the automatic subset. Further analysis indicates the primary cause of this discrepancy was the deletion error. The automatic subset presented a notably high deletion error rate of 26.6, whereas the manual subset recorded a markedly lower rate of 7.0. These findings align with previous research, which has indicated that the use of automatically-generated transcripts tends to undermine system performance [16, 32]. Consequently, these results underscore the importance of prioritizing the utilization of the manual subset over the automatic subset in the training of models.
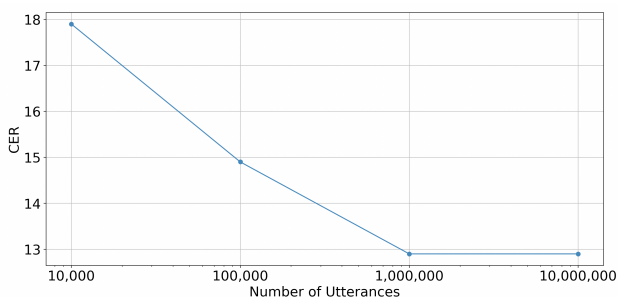
Fig. 7. the relationship between the size of the training dataset and its performance on the speech recognition task.

In the previous experiments, an alignment score threshold of 2.0 was implemented as a cut-off. To investigate how altering this threshold might impact performance, we performed an experiment with varied threshold values, ranging from 1 to 16. For analytical purposes, scores exceeding 16 were normalized to 16. We conducted this analysis using 100,000 ut-

terances (160 hours) from the manual subset with the same testing set. The outcome of this exercise is illustrated in Figure 6. It is evident that the setting of the alignment threshold is critical to model performance where raising the threshold from 16 to 2 results in consistent performance improvement. However, further tightening of the threshold from 2.0 to 1.0 caused a minor degradation in performance, from a CER of 14.7 to 14.8. This slight decrease may be attributed to the fact that utterances in the training set with an alignment score of 1.0 tend to be shorter and comprise fewer words than those within the subset with a score of 2.0. These findings imply the significance of judiciously selecting the alignment threshold when training models. Although a lower threshold might seem intuitively beneficial, it may inadvertently exclude longer, richer utterances, thereby potentially impacting performance.

Finally, we investigated the impact of modifying the size of the training dataset, ranging from 10,000 to 10 million utterances (16 to 16k hours). All utterances were randomly selected from the English manual subset, adhering to an alignment threshold of 2.0. The result is depicted in Table 7. The results demonstrate that increasing the quantity of utterances consistently enhances model performance. Interestingly, the model trained with 1 million utterances and the model trained with 10 million utterances exhibit a similar CER. This phenomenon could potentially be attributed to the simplistic architecture we employed - namely a linear model built upon pre-trained features. Such an architecture may not be fully equipped to leverage the expanded dataset.

5. CONCLUSION

In this study, we presented the YODAS dataset, a comprehensive, multilingual dataset compiled from the YouTube platform. We delineated our data collection pipeline and provided preliminary analyses and baseline models based on the dataset. We anticipate that the YODAS dataset will serve as a valuable resource for the speech research community

6. ACKNOWLEDGEMENTS

7. REFERENCES

[1] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "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] Ronan Collobert, Christian Puhrsch, and Gabriel Synnaeve, "Wav2letter: an end-to-end convnet-based speech recognition system," *arXiv preprint arXiv:1609.03193*, 2016.

[3] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*. Ieee, 2013, pp. 6645–6649.

[4] Ilya Sutskever, Oriol Vinyals, and Quoc V Le, "Sequence to sequence learning with neural networks," *Advances in neural information processing systems*, vol. 27, 2014.

[5] Rohit Prabhavalkar, Takaaki Hori, Tara N Sainath, Ralf Schlüter, and Shinji Watanabe, "End-to-end speech recognition: A survey," *arXiv preprint arXiv:2303.03329*, 2023.

[6] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12449–12460, 2020.

[7] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "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.

[8] Abdelrahman Mohamed, Hung-yi Lee, Lasse Borgholt, Jakob D Havtorn, Joakim Edin, Christian Igel, Katrin Kirchhoff, Shang-Wen Li, Karen Livescu, Lars Maaløe, et al., "Self-supervised speech representation learning: A review," *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[9] Mark JF Gales, Kate M Knill, Anton Ragni, and Shakti P Rath, "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.

[10] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[11] Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert, "Mls: A large-scale multilingual dataset for speech research," *arXiv preprint arXiv:2012.03411*, 2020.

[12] Alan 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.

[13] Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al., "Scaling speech technology to 1,000+ languages," *arXiv preprint arXiv:2305.13516*, 2023.

[14] Jacob Kahn, Morgane Rivière, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, 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.

[15] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al., "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[16] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28492–28518.

[17] Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, et al., "Google usm: Scaling automatic speech recognition beyond 100 languages," *arXiv preprint arXiv:2303.01037*, 2023.

[18] Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna, "Fleurs: Few-shot learning evaluation of universal representations of speech," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 798–805.

[19] Jörgen Valk and Tanel Alumäe, "Voxlingua107: a dataset for spoken language recognition," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 652–658.

[20] Shinnosuke Takamichi, Ludwig Kürzinger, Takaaki Saeki, Sayaka Shiota, and Shinji Watanabe, "Jtube-speech: corpus of japanese speech collected from youtube for speech recognition and speaker verification," *arXiv preprint arXiv:2112.09323*, 2021.

[21] Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R Mortensen, Graham Neubig, Alan W Black, 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.

[22] Alex Graves and Navdeep Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International conference on machine learning*. PMLR, 2014, pp. 1764–1772.

[23] Ludwig Kürzinger, Dominik Winkelbauer, Lujun Li, Tobias Watzel, and Gerhard Rigoll, "Ctc-segmentation of large corpora for german end-to-end speech recognition," in *International Conference on Speech and Computer*. Springer, 2020, pp. 267–278.

[24] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al., "Xls-r: Self-supervised cross-lingual speech representation learning at scale," *arXiv preprint arXiv:2111.09296*, 2021.

[25] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai, "ESPnet: End-to-end speech processing toolkit," in *Proceedings of Interspeech*, 2018, pp. 2207–2211.

[26] Shu wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech 2021*, 2021, pp. 1194–1198.

[27] Rico Sennrich, Barry Haddow, and Alexandra Birch, "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, Aug. 2016, pp. 1715–1725, Association for Computational Linguistics.

[28] Taku Kudo, "Subword regularization: Improving neural network translation models with multiple subword candidates," *arXiv preprint arXiv:1804.10959*, 2018.

[29] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 2613–2617.

[30] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "Audio augmentation for speech recognition," in *Sixteenth annual conference of the international speech communication association*, 2015.

[31] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[32] Behrooz Ghorbani, Orhan Firat, Markus Freitag, Ankur Bapna, Maxim Krikun, Xavier Garcia, Ciprian Chelba, and Colin Cherry, "Scaling laws for neural machine translation," *arXiv preprint arXiv:2109.07740*, 2021.