LABORO-ASV: A SMALL YET EFFECTIVE ADD-ON DATASET FOR JAPANESE SPEAKER VERIFICATION

Xinyi Zhao, Hiromasa Fujihara

Laboro.AI, Inc., Tokyo, Japan
{zhao,fujihara}@laboro.ai

ABSTRACT

Japanese speaker verification struggles due to the lack of high-quality datasets in Japanese. We attempt to find a solution by collecting and exploiting Laboro-ASV, a small but high-quality dataset, to circumvent the difficulties in producing a large-scale dataset in Japanese. We combine large English datasets and small Japanese datasets as the training data, aiming to improve the performance on the Japanese speaker verification task. Our dataset Laboro-ASV, as an add-on dataset, reaches the best performance among all evaluated datasets. Insights into what makes an add-on dataset more effective are also provided.

Index Terms— speaker verification, dataset, small-scale, add-on, cross-lingual

1. INTRODUCTION

Speaker recognition has received a lot of interest, particularly after the release of large-scale open-source datasets in recent years. There are VoxCeleb datasets [1, 2] for English and CN-Celeb [3] dataset for Mandarin Chinese. It is possible to use these datasets for Japanese automatic speaker verification (ASV), but unfortunately, speaker recognition is not completely language-independent. The language mismatch in training and testing data frequently leads to unsatisfying performance. Although numerous studies have attempted to address this cross-lingual issue by language adaptation [4, 5], having a dataset in the target language unquestionably makes things easier and more straightforward.

While it is possible to copy the data collection pipeline from VoxCeleb or CN-Celeb and apply it to other languages, it is also worth noting that there are limited resources for languages that are not as widely spoken as English or Mandarin Chinese. Instead of creating a large-scale Japanese dataset equivalent to VoxCeleb, we would like to look into the possibility of smaller datasets. The performance would be incomparable due to the huge disparity in size, hence our main purpose is to investigate how, if possible, to exploit the potential of smaller datasets and gain satisfactory performance on the Japanese speaker verification task.

In this study, we make two contributions. First, we create the Laboro-ASV dataset and make it open-source. It is a Japanese ASV dataset collected from Japanese terrestrial television programs. By using data from traditional media, we aim to provide a universal method so that good ASV datasets in more languages can be collected in the same fashion. Second, we suggest adding Laboro-ASV as a supplemental dataset to other datasets used for training in the target language, and comparing the outcomes. In the Japanese speaker verification task, the combination of VoxCeleb 1 and Laboro-ASV outperforms all the other datasets and yields the best performance. Furthermore, we also discuss the key attributes of an effective add-on dataset.

2. RELATED WORKS

Large-scale speaker recognition datasets have been studied in research. For English, there are VoxCeleb 1 [1] and 2 [2], and for Mandarin Chinese, there is CN-Celeb [3]. Focused on the real-world environment, both projects download videos from online resources and extract spoken utterances for selected persons of interest (POIs) with the help of face detection and verification. To implement such a pipeline for collecting a Japanese speaker recognition dataset, a curated list of Japanese speakers and a face image dataset of the chosen POIs are necessary, both of which are, unfortunately, not available.

To the best of our knowledge, JTubeSpeech-ASV (JTube-ASV) [6] is the only large-scale Japanese speaker recognition dataset collected "in the wild". Single-speaker videos are first collected from YouTube, and all videos from the same YouTube channel are labeled as the same speaker. By manually checking the JTube-ASV, we find that this dataset might not be the best for Japanese speaker verification, because occasionally multiple speakers share the same speaker ID and not all utterances are in Japanese.

Unlike the existing works, we make public a small yet high-quality Japanese speaker recognition dataset, and use it as an add-on dataset for Japanese speaker verification. We focus on the collection of a high-quality speaker recognition dataset using constrained resources, as well as the impact of a small dataset in the target language. We also provide insights into what makes an add-on dataset more effective.

3. DATASET COLLECTION

This section first describes the pipeline we design and apply to obtain a speaker recognition dataset, following

with the comparison of our dataset with other existing speaker recognition datasets.

3.1. Pipeline

All the source data in Laboro-ASV dataset comes from the recordings of Japanese terrestrial television programs from February to July 2022. The source data for each TV show contains its audio and video, as well as additional information such as the title, the genres, and the performers of the show. From such source data, we create a speaker recognition dataset by using the pipeline described below.

Stage 1. POIs and TV Programs Selection

Having approximately 750 performers and 25,000 TV programs to begin with, it is important to select a list of POIs and TV programs that are appropriate for building a speaker recognition dataset. An ideal POI is expected to have as many utterances as possible, which translates to the preference for POIs that make more appearances on TV. As for TV programs, not all genres are suitable for our purpose. Voice activity detection (VAD) and speaker change detection (SCD) are especially challenging to implement, for instance, on variety shows due to the constant background noises, as well as on music and sports shows due to the lack of speech. We limit the genres of the TV programs to news shows, tabloid shows, and TV series, and then rank all the performers by their appearances on these shows. The top 200 performers are included in our POI list, and all of the TV programs that (1) involve these POIs and (2) fall under the appropriate genres previously stated are included in the TV program list.

Stage 2. Manual Seed Annotation

"Seed" utterances are required for speaker verification as the ground truth for each POI. Instead of using face verification, the pipeline is kept simple by manually annotating audio recordings to obtain seed utterances. For each POI, five TV programs are selected, and programs with multiple POIs are preferable because it reduces the total number of shows that need to be annotated. Annotators are asked to annotate the timestamps when the POI speaks in the programs without interfering with other speakers. Using the annotation tool ELAN [7], it took two part-time Japanese-speaking annotators approximately 2 months to complete the annotation in our scenario, where 284 TV programs were chosen for the 200 POIs.

According to the timestamps, long segments of POIs speaking are retrieved; then, they are further divided into shorter utterances as the "seeds" using VAD, which is covered in Stage 3 below. The POI will be removed from our POI list if its seed utterances' combined duration length is too short or their similarity scores are too low. Eventually, we obtained a seed utterances dataset involving 142 POIs.

Stage 3. Segmentation

Dataset	Lang	# POI	# Utter	Dur	Hours
VoxCeleb 1	En	1,211	116	8.2	340
VoxCeleb 2	En	6,112	185	7.8	2,442
CN-Celeb	Zh	1,000	130	7.8	274
JTube-ASV	Ja	1,574	102	4.8	214
Laboro-ASV	Ja	142	478	5.0	95

Table 1. *Datasets statistics.* ***Lang****: the language of the dataset;* ***# POI****: the number of POIs;* ***# Utter****: the average number of utterances per POI;* ***Dur****: the average length of utterances in seconds;* ***Hours****: the total length of all utterances in the dataset in hours.*

The audio data of the TV programs are first divided into shorter segments by performing VAD, so that only segments with speech remain. We employ two techniques in a cascade for VAD: (1) an acoustic-feature-based method [8] and (2) a hybrid CNN-BiLSTM model pretrained on the AVA-Speech dataset [9, 10].

The segments acquired from VAD are then divided into utterances by performing SCD. An LSTM model pretrained on Estonian broadcast data is applied for this purpose [11].

Stage 4. Speaker Verification and Classification

To identify the unknown speaker from each utterance obtained from the SCD step, binary speaker verification is done first. For each utterance, the speaker embedding is calculated using a x-vector TDNN model pretrained on VoxCeleb 1 and VoxCeleb 2 datasets [12, 13]. Every unknown utterance from a certain TV show is paired with seed utterances of all POIs that appear in the same show. Similarity score is generated for each utterance pair by applying Probabilistic LDA (PLDA) [14].

Using the similarity scores, a prediction of the speaker is made for each utterance using a k-nearest neighbors (k-NN) voting algorithm. We find that it works best for our case when k = 50, meaning the top 50 seed utterances with the highest similarity scores are taken into consideration. We also introduce two thresholds to make the classification stricter. One threshold is for the similarity score. All utterance pairs with scores lower than 0.5 are disregarded. The other threshold requires the voting percentage to be above a certain number; otherwise, the prediction is invalid. We set the voting percentage threshold as 30%, meaning a valid prediction can be made only when there are 15 (= 50 × 30%) seed utterances of one POI reaching the similarity score of 0.5. If a valid prediction cannot be made, the utterance will be classified as "spoken by a non-POI".

3.2. Datasets Description

Table 1 summarizes some existing speaker identification datasets. Laboro-ASV is a relatively small dataset. When compared with JTube-ASV, the data source guarantees that the majority of utterances are spoken in Japanese, and the collection pipeline ensures a lower error rate for POI classification.

4. EXPERIMENTS

This section explains the experimental setup for the Japanese speaker verification task. We suggest adding small datasets to the training sets for speaker embedding models, and assessing the performance of Laboro-ASV in comparison with other datasets. We also discuss what makes an effective add-on dataset by improving the existing Japanese dataset JTube-ASV.

4.1. Experiments Setup

All datasets are evaluated by speaker verification experiments. The trial set of JTube-ASV dataset is adopted as the testing set for all the experiments, and there are no duplicate POIs between the training and trial sets of JTube-ASV. Equal error rate (EER) is used as the performance metric.

We focus on the evaluation and comparison of datasets; therefore, all experiments are based on the same x-vector/PLDA method. For each dataset, an x-vector TDNN model is trained to extract the utterance-level speaker embeddings, and PLDA is used to perform speaker verification.

4.2. Speaker Verification Experiments

We first evaluate datasets individually on the Japanese speaker verification task. This includes two groups of English datasets (1) VoxCeleb 1 (2) VoxCeleb 1 + VoxCeleb 2, and 2 Japanese datasets (1) Laboro-ASV (2) JTube-ASV. To the best of our knowledge, JTube-ASV dataset was the only large-scale open source Japanese ASV dataset. Therefore, it is adopted as the baseline for our experiments. The results are given in Table 2.

Being aware that the disparity in the size of the datasets will inevitably make the performance incomparable, we also use Laboro-ASV as an add-on dataset to train alongside VoxCeleb 1. We further minimize the size of Laboro-ASV so that it could be more easily used as an add-on dataset. After ranking all POIs by the number of utterances, only the top 50 POIs are selected to create the Laboro-ASV-50 dataset. For comparison, another model is trained with the combination of VoxCeleb 1 and JTube-ASV. The results are given in Table 2.

The VoxCeleb 1 + Laboro-ASV gives the best performance among all the experiments, and even after minimizing the size, Laboro-ASV-50 still serves the purpose of a powerful add-on dataset.

4.3. To Create an Effective Add-On Dataset

JTube-ASV alone gives a decent performance on Japanese speaker verification task, but when used as the add-on dataset, the performance is not improved much. We perform more tests on JTube-ASV to explain this behavior and determine what makes an add-on dataset useful. There are several potential decisive factors for the quality of an add-on dataset:

Dataset	**EER (%)**
VoxCeleb 1	6.08
VoxCeleb 1+2	4.39
JTube-ASV	4.82
Laboro-ASV	9.21
VoxCeleb 1 + JTube-ASV	4.82
VoxCeleb 1 + Laboro-ASV	**3.95**
VoxCeleb 1 + Laboro-ASV-50	**3.95**

Table 2. *Results for Japanese speaker verification.*

(1) the number of POIs, (2) the average length of utterances, (3) the average number of utterances per POI, (4) the POI purity, and (5) the Japanese language purity.

The first two factors can be easily excluded because JTube-ASV has much more POIs than 50, and the average length of utterances is very similar to Laboro-ASV. However, the other three factors require more in-depth research, and three JTube-ASV subsets have been created for this purpose. Table 3 provides a statistical summary of all add-on datasets used in our experiments.

To increase the average number of utterances per POI, all the POIs in JTube-ASV are ranked by the number of utterances, and the top 184 POIs are selected to create the new JTube-More-Utterances (JTube-MU) dataset. JTube-MU has an average of 478 utterances for each POI, which is the same as Laboro-ASV. Table 4 presents the results. JTube-MU as the add-on dataset has a big improvement compared to the VoxCeleb 1 + JTube experiment. This demonstrates that a vital attribute of an effective add-on dataset is having more utterances per POI.

In JTube-ASV, only single-speaker videos are included, and videos from the same YouTube channel are labeled as the same POI. However, the channels are not manually verified and one speaker ID may be shared by multiple speakers. For each POI, we select one video that contains the most utterances to form the JTube-Pure-POI (JTube-PP) dataset. As the results are shown in Table 4, the performance of JTube-PP becomes worse than using the original dataset, and it suggests that the POI purity is not the major issue in JTube-ASV.

Not all videos in JTube-ASV are in Japanese. We perform utterance-level language identification and all utterances predicted to be spoken in Japanese form another dataset, JTube-Pure-Japanese (JTube-PJ). This is the only option we have for creating the new pure Japanese subset, and it unavoidably results in fewer utterances per POI. As the results are shown in Table 4, the performance of JTube-PJ also becomes worse than the original dataset as an add-on dataset, but we cannot justify the significance of the language purity due to the vast decrease in the number of utterances per POI.

5. CONCLUSIONS

Combining large-scale English datasets with smaller-scale Japanese datasets generally outperforms the datasets individually on the Japanese speaker verification task.

Dataset	# POI	# Utter	Dur	Hours
Laboro-50	50	1206	5	84
JTube-MU	184	478	4.6	112
JTube-PP	813	102	4.9	112
JTube-PJ	1558	62	4.9	137

Table 3. *Statistics of all add-on datasets involved in our experiments.* ***# POI****: the number of POIs;* ***# Utter****: the average number of utterances per POI;* ***Dur (s)****: the average length of utterances in seconds;* ***Hours****: the total length of all utterances in the dataset in hours.*

Dataset	EER (%)
VoxCeleb 1 + JTube-ASV	4.82
VoxCeleb 1 + JTube-MU	**4.39**
VoxCeleb 1 + JTube-PP	6.58
VoxCeleb 1 + JTube-PJ	5.70

Table 4. *Results for Japanese speaker verification on modified JTube ASV datasets.*

Among all the datasets evaluated in our experiments, Laboro-ASV and Laboro-ASV-50 give the best result, showing that Laboro-ASV is a high-quality speaker recognition dataset, and it works effectively as an add-on dataset even after the size being reduced. By modifying JTube-ASV dataset, we also discover that the average number of utterances per POI is a crucial attribute for a high-quality add-on dataset.

6. REFERENCES

[1] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," in *INTERSPEECH*, 2017.

[2] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *INTERSPEECH*, 2018.

[3] Y. Fan, J.W. Kang, L.T. Li, K.C. Li, H.L. Chen, S.T. Cheng, P.Y. Zhang, Z.Y. Zhou, Y.Q. Cai, and D. Wang, "Cn-celeb: a challenging chinese speaker recognition dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7604–7608.

[4] X. Qin, N. Li, Y. Lin, Y. Ding, C. Weng, D. Su, and M. Li, "The dku-tencent system for the voxceleb speaker recognition challenge 2022," 2022.

[5] Z.D. Zhao, Z. Li, W.C. Wang, and P.Y. Zhang, "The hccl system for voxceleb speaker recognition challenge 2022," 2022.

[6] S. Takamichi, L. Kürzinger, T. Saeki, S. Shiota, and S. Watanabe, "Jtubespeech: corpus of japanese speech collected from youtube for speech recognition and speaker verification," 2021.

[7] "Elan (version 6.2) [computer software]," Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. Retrieved from `https://archive.mpi.nl/tla/elan`, 2021.

[8] "voice_activity_detection [source code]," `https://github.com/zlzhang1124/voice_activity_detection`, 2020.

[9] N. Wilkinson and T. Niesler, "A hybrid CNN-BiLSTM voice activity detector," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Toronto, Canada, 2021.

[10] S. Chaudhuri, J. Roth, D. P. W. Ellis, A. Gallagher, L. Kaver, R. Marvin, C. Pantofaru, N. Reale, L.G. Reid, K. Wilson, and Z. Xi, "Ava-speech: A densely labeled dataset of speech activity in movies," 2018.

[11] "online_speaker_change_detector [source code]," `https://github.com/alumae/online_speaker_change_detector`, 2021.

[12] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[13] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.C. Chou, S.L. Yeh, S.W. Fu, C.F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. De Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," 2021.

[14] S. Ioffe, "Probabilistic linear discriminant analysis," in *Proceedings of the European Conference on Computer Vision*. 2006, pp. 531–542, Springer Berlin Heidelberg.