2-Q-27 LABORO-ASV: A SMALL YET EFFECTIVE ADD-ON DATASET FOR JAPANESE SPEAKER VERIFICATION

Xinyi Zhao, Hiromasa Fujihara
Laboro.AI, Inc., Tokyo, Japan

Laboro

INTRODUCTION

Speaker recognition datasets in existence

- VoxCeleb 1 & 2 for English
- CN-Celeb for Mandarin Chinese
- ___??___ for Japanese

Motivations

- In the scene of Japanese speaker recognition, language mismatch in training (English, Chinese, etc.) and testing (Japanese) data leads to unsatisfying performance
- Instead of creating a large-scale Japanese dataset equivalent to VoxCeleb, we investigated the possibility of smaller datasets

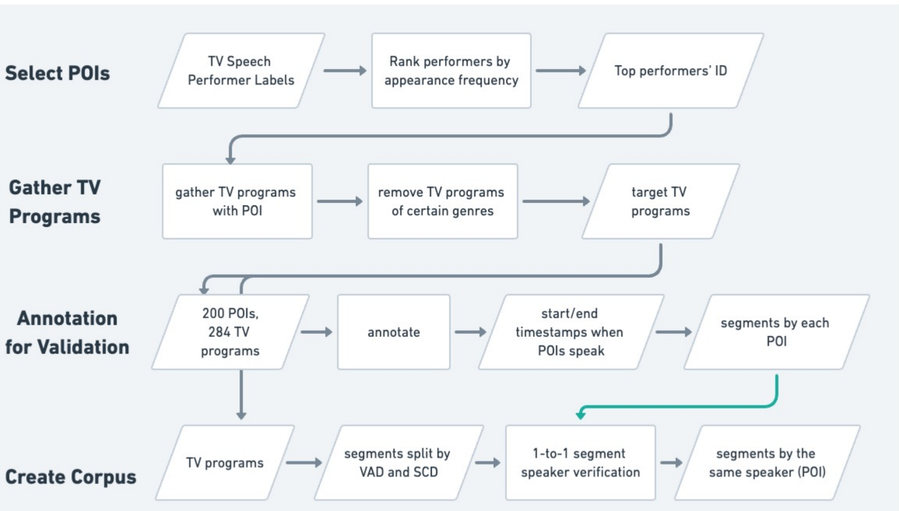
Contributions

- We create a open-source Japanese speaker recognition dataset **Laboro-ASV**
- We discuss the key attributes of a small yet effective ASV dataset

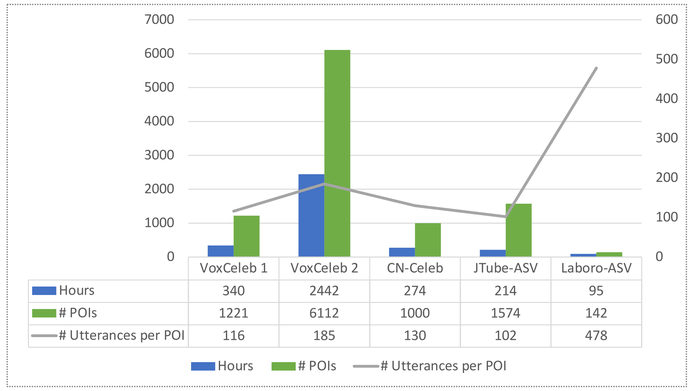
DATA COLLECTION

Source data

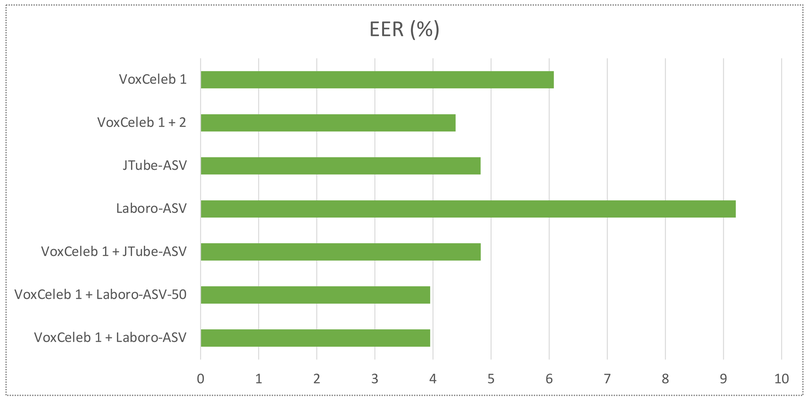
- 25,000 Japanese terrestrial TV programs from February to July 2022
- Additional information including the genre and the performers, etc.



DATA STATISTICS



EXPERIMENTS: SETUP AND RESULTS



Setup

- All experiments are based on the same setup for comparison
- X-vector/PLDA are used for speaker embedding extraction and speaker verification. The evaluation metric is EER. The testing set is the the trial set of JTubeSpeech ASV

Datasets involved

- VoxCeleb 1 & 2
- JTubeSpeech ASV (JTube-ASV)
  A large-scale Japanese speaker recognition dataset collected "in the wild"
- Laboro-ASV
  The ASV dataset we created
- Laboro-ASV-50
  A subset of Laboro-ASV including only the top 50 POIs by the number of utterances

DATA ABLATION ANALYSIS

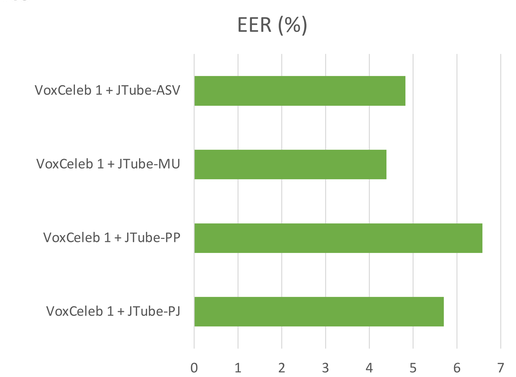
To see how the following factors effect the results

1) The number of total POIs
2) The average length of utterances
3) The average number of utterances per POI
4) The POI purity
5) The Japanese language purity

We curated and compared the following datasets

- JTube-ASV
- JTube-More-Utterances (JTube-MU)
  A subset of JTube-ASV that includes 184 POIs and have 478 utterances per POI
- JTube-Pure-POI (JTube-PP)
  A subset of JTube-ASV that has as more accurate POI information for each utterance as possible
- JTube-Pure-Japanese ((JTube-PJ)
  A subset of JTube-ASV that includes exclusively Japanese utterances

Results



CONCLUSIONS

- Laboro-ASV dataset serves as an effective supplemental dataset for Japanese speaker verification
- Laboro-ASV and its subset Laboro-ASV-50 give the best performance in our experiments
- The average number of utterances per POI is a crucial attribute for a high-quality add-on dataset