15 Can you guess the age from this voice?

Deep speaker attribute estimation with speaker clustering

Abstract

Estimating speaker-attributes such as age and gender is an important task with a wide range of applications. While the recent proposed deep neural network models have been achieving high performance, the estimated results tend to be less reliable because of the overfitting problem. In order to solve this problem, we propose a general framework for correcting the unreliable results of the arbitrary speaker-attribute estimation models. The proposed algorithm first applies speaker clustering to the target utterances to detect similar speakers of target utterances. Then, the speaker-attribute class of each cluster is determined by voting on the utterances assigned to the cluster. Finally, we can correct the result of unreliable utterances by replacing their result with the clusters' speaker-attribute class. Our approach is evaluated on age-gender classification and gender regression tasks, yielding significant improvements in classification accuracy and mean absolute error.

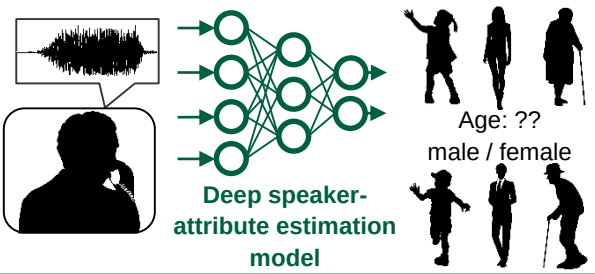
Deep speaker attribute estimation

Problem definition

Estimate speakers' age and gender from their voices with deep learning model

Applications

Call center response decision, marketing support, realization of a voice dialogue system that changes behavior according to user attributes, etc.



Difficulties of task

- There is **large deviation in amount of training data for each age-group** (Fig.1)
- **Models tend to be overfitting to specific speaker / age** (Fig. 2)

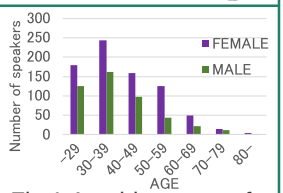


Fig.1 Age histogram of NIST-SRE08

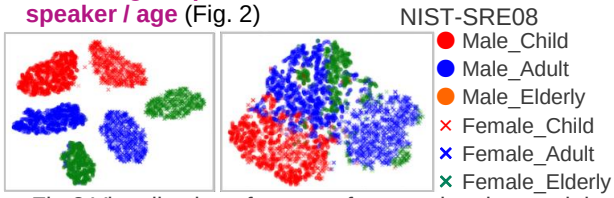
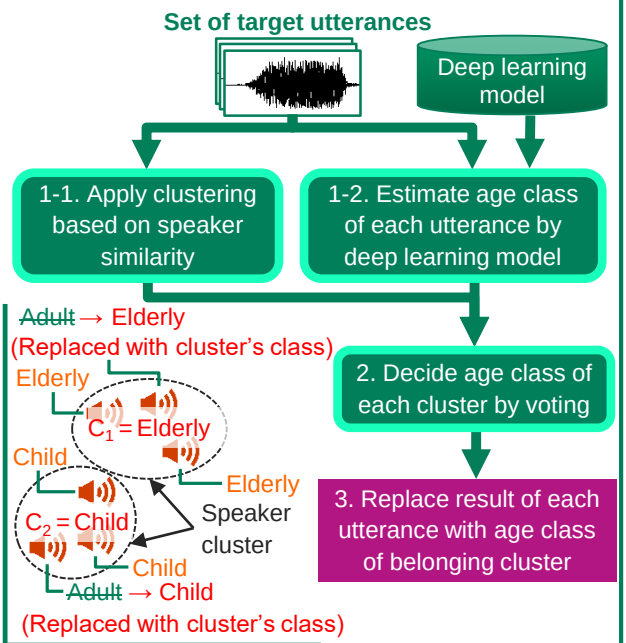


Fig.2 Visualization of output of age-estimation model (cited from [1]) (left: training data, right: evaluation data)

Error correction based on speaker clustering

Correct estimation results of deep learning model by majority voting of results of similar speakers



: Target speech and estimated class

C.: Class decided by voting for each cluster

Evaluation criteria	Conventional	Proposed
Classification accuracy (Child, Adult, Elderly)	59 %	**72 %**
Mean absolute error	± 10.9 years	**± 8.7 years**

References

[1] N. Tawara, H. Kamiyama, S. Kobashikawa, A. Ogawa, "Improving speaker-attribute estimation by voting based on speaker cluster information," *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 6594-598, May 2020.

[2] N. Tawara, H. Kamiyama, S. Kobashikawa, A. Ogawa, "Frame-level phoneme-invariant speaker feature extraction for text-independent speaker recognition on extremely short utterances," *Reports of the autumn meeting the Acoustical Society of Japan.*, pp. 815-816, Sept. 2019.

Contact

Naohiro Tawara Email: cs-openhouse-ml@hco.ntt.co.jp

Signal Processing Research Group, Media Information Laboratory

オープンハウス 2020