INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Age Estimation with Speech-age Model for Heterogeneous Speech Datasets

Ryu Takeda, Kazunori Komatani

The Institute of Scientific and Industrial Research (SANKEN), Osaka University, Japan

rtakeda@sanken.osaka-u.ac.jp, komatani@sanken.osaka-u.ac.jp

Abstract

This paper describes an age estimation method from speech signals for heterogeneous datasets. Although previous studies in the speech field evaluate age prediction models with held-out testing data within the same dataset recorded in a consistent setting, such evaluation does not measure real performance. The difficulty of heterogeneous datasets is overfitting caused by the corpus-specific properties: transfer function of the recording environment and distributions of age and speaker. We propose a speech-age model and its integration with sequence neural networks (NNs). The speech-age model represents the ambiguity of age as a probability distribution, which also virtually extends the limited range of age distribution of each corpus. A Bayesian generative model successfully integrates the speech-age model and the NNs. We also applied mean normalization technique to cope with the transfer function problem. Experiments showed that our proposed method outperformed the baseline neural classifier for completely open test sets in the age distribution and recording setting.

Index Terms: age estimation, heterogeneous datasets, speech signal

1. Introduction

1.1. Background and Motivation

Age estimation from speech signals [1] can be help for a spoken dialogue system. The system can change its behavior in accordance with a user's age once the system estimates it correctly. Machine learning, in particular neural networks (NNs), is a standard approach for age estimation because 1) the reliability or probability of the estimation is usually provided and 2) multi-modal integration using speech and vision is also possible. The main issue is the construction of an age estimation model using training data.

Our focus is to construct an age estimation model for heterogeneous training and non-held-out test datasets. This is a practical situation; an acoustic environment of real application is often different from that of held-out data. This problem has been seldom handled in speech community, while a number of studies in the vision field have discussed this issue [2, 3]. Almost all studies in the speech field use held-out testing data from the same training datasets [4, 5, 6]. Evaluating this environment does not measure real performance of models and is not realistic for many actual use cases.

Important properties for age estimation, such as age distribution, annotated age labels, and transfer function of the recording environment, are usually different among corpora. This is mainly because most of the speech corpora were collected for the acoustic model training of automatic speech recognition (ASR). One example is a corpus of senior citizens (over 60 years old) with phonemically balanced or isolated dictionary word utterances. The age label is sometimes given as a range, such as 60–69 years old. The transfer function of microphone and signal processing filter also affects the spectral envelope of the recorded speech signals. The real age distribution and spectral envlope of three corpora for training are shown in Fig. 3, and we can understand how much different these properties are among corpora. The details are described in Section 4.1.

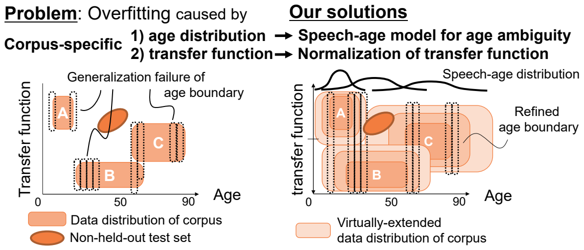
Problem: Overfitting caused by
Corpus-specific
1) age distribution → Speech-age model for age ambiguity
2) transfer function → Normalization of transfer function
Our solutions
Generalization failure of
age boundary
Transfer function
A
B
C
0
50
90
Age
Data distribution of corpus
Non-held-out test set
Speech-age distribution
Refined
age boundary
transfer function
A
B
C
0
50
90
Age
Virtually-extended
data distribution of corpus

Figure 1: *Problem of heterogeneous datasets and our solution*

The problem with heterogeneous datasets is the overfitting of the models to training data, which is mainly caused by the corpus-specific properties: the age distributions and the transfer function of recording environment. The overfitting results in seriously catastrophic and unreliable age estimation for non-held-out test sets because such models perform well only with data similar to the training set. Figure 1 depicts an example of the corpus distribution along the age and transfer function axes. The age distribution of each corpus is not complete and non-uniform. If classifiers are trained with these corpora, the age boundaries of classifiers will be fitted to a specific corpus.

We propose a speech-age model and its integration with sequence NNs to solve the age distribution problem. Our speech-age model expresses such age ambiguity of speech as a probabilistic distribution, which results in a smoothing of real age distribution. From a macro perspective, the age distribution of each corpus is virtually extended by this model as shown in Figure 1. The transfer function of each corpus is also normalized by signal processing technique, and the gap among their recording settings is alleviated. As a result, the age boundaries of classifiers will be generalized more because the virtually-extended corpora are used for training. Age-dependent distribution to capture aging feature are also investigated as a speech-age model. A Bayesian generative model is used to integrate the speech-age model and sequence NNs.

1.2. Related Work

Studies on age estimation in the speech field usually use a specific corpus, not heterogeneous datasets. Here, the age estimation problem is formulated as a classification problem, e.g. 100 classes (each class represents each age), or a regression problem. There is no large difference between them [4].

Several neural architectures and speech features for age estimation in the speech and vision fields have been investigated. For example, convolutional NNs (CNNs) [4, 7] and fully-connected networks are standard architectures. Studies have applied a sequence neural model, such as long short-term memory

http://dx.doi.org/10.21437/Interspeech.2021-861

(LSTM) [5, 8] to these architectures.

Speech features and multi-task learning in the speech field have also been investigated. The major input features of NNs are x-vectors, mel-filter bank spectra and mel-frequency cepstral coefficients (MFCC) [4, 5, 6]. Multi-task learning is also used to estimate not only age but also other speaker properties. In [6], the age and height of a speaker can be estimated simultaneously because they are correlated in the growth process.

A number of studies use the KL-divergence of a Gaussian distribution as a cost function in the vision field [9, 10], but the usage of such a distribution is not discussed in the speech field. There has been no comprehensive discussion regarding the age-dependent distribution nor their impact on heterogeneous datasets even in vision field to the best of our knowledge.

2. Baseline Method

2.1. Problem Statement

Age estimation from a speech signal is formulated as the maximum posterior estimation given from speech features. We denote the t-th frame N_x-dimension speech feature as $\mathbf{x}_t \in \mathbb{R}^{N_x}$, and its sequence from 1 to T as $\mathbf{x}_{1:T}$. The age class of a target person, $\hat{z}$, is estimated by determining an argument that maximizes the posterior probability as

$$\hat{z} = \arg\max_z p(z|\mathbf{x}_{1:T}; \mathbf{\Theta}), \tag{1}$$

where $\mathbf{\Theta}$ is a model parameter set. We assume that the number of classes is N_c. Note that *a raw speech signal is not appropriate as input in the case of heterogeneous datasets* because the corpus-specific transfer function of a microphone, room, or low-pass filter easily causes overfitting.

The posterior probability $p(z|\mathbf{x}_{1:T}; \mathbf{\Theta})$ is built using K training corpora, $D = \{D_k\}_{k=1,...,K}$. The supervised training is usually applied because each corpus includes the pair of z and $\mathbf{x}_{1:T}$. The model parameter set $\mathbf{\Theta}$ is usually estimated by minimizing the cross-entropy cost function

$$J(\mathbf{\Theta}) = \textstyle\sum_{n \in D} \sum_{z_n} - q(z_n) \log p(z_n|\mathbf{x}_{n,1:T}; \mathbf{\Theta}), \tag{2}$$

where z_n and $\mathbf{x}_{n,1:T}$ are the age and speech features of n-th data, respectively. $q(z)$ is an age-label distribution, and it is usually set to a delta function that takes a value of 1 only if z matches the age label z_n, otherwise, it takes a value of 0. If the age label represents the range z_l–z_r, such as 20–29 years old, we can set $q(z)$ to a uniform distribution in the region.

2.2. Neural Sequential Model

The posterior probability is usually modelled by NNs that can deal with sequential data. Two major models are BLSTM [11] and self-attention mechanism (ATTN) [12, 13, 14].

BLSTM is based on forward and backward recurrent NNs (RNNs). We denote the forward and backward context vectors at the t-th frame as $\mathbf{h}_{\mathrm{f},t} \in \mathbb{R}^{N_h}$ and $\mathbf{h}_{\mathrm{f},t} \in \mathbb{R}^{N_h}$, respectively. Their sequences of context vectors, $\mathbf{h}_{\mathrm{f},1:T} = [\mathbf{h}_{\mathrm{f},1}, ..., \mathbf{h}_{\mathrm{f},T}]$ and $\mathbf{h}_{\mathrm{b},1:T} = [\mathbf{h}_{\mathrm{b},1}, ..., \mathbf{h}_{\mathrm{b},T}]$, are calculated from input features as

$$[\mathbf{h}_{\mathrm{f},1:T}, \mathbf{h}_{\mathrm{b},1:T}] = \mathrm{BLSTM}(\mathbf{x}_{1:T}). \tag{3}$$

The attention network is expected to select the effective feature frame for age estimation. The relative weights $a_{1:T} = [a_1, ..., a_T]$ for context vectors from frames 1 to T are also estimated from the same context vectors,

$$a_{1:T} = \mathrm{ATTN}(\mathbf{h}_{\mathrm{f},1:T}, \mathbf{h}_{\mathrm{b},1:T}) \ \ (\textstyle\sum_{t=1}^{T} a_t = 1). \tag{4}$$

The ATTN network consists of linear projection, tanh function with N_a-dimension, linear projection and softmax function.

Finally, the posterior probability is calculated by a classification network (CLS). The context vectors averaged by the relative weights are used for the input as

$$q(z|\mathbf{x}_{1:T}) = \mathrm{CLS}(\mathbf{g}), \quad \mathbf{g} = \textstyle\sum_{t=1}^{T} a_t [\mathbf{h}_{\mathrm{f},t}, \mathbf{h}_{\mathrm{b},t}]. \tag{5}$$

The CLS consists of linear projection and softmax function (final output). If we assume the uniform weights, the mean of the context vectors can be used like [6].

3. Proposed Method

We first explain the generative model to solve the age estimation problem by introducing our speech-age model. The sequence NNs is logically embedded into the top-down model. We also discuss several distributions of speech-age models and cost function for stochastic gradient descent (SGD) training.

3.1. Probabilistic Generative Model

We introduce three kinds of latent variables to represent the observed speech features. The first one is a *speech-age* y, which represents an age of speech, not the real age of a person. The second one is the hidden states $s_{1:T} = [s_1, ..., s_T]$, which correspond to the language and pronunciation context similar to those in a hidden Markov model (HMM) [15] used in the ASR field. The last one is an *age-mask* $m_{1:T} = [m_1, ..., m_T], m_i \in [0\ 1] (i = 1, ..., T)$, which represents the degree of how the person's speech-age appears at each speech feature $\mathbf{x}_t$.

The posterior probability consists of a likelihood and prior computed using Bayes' theorem. We derived the following hybrid model that partly includes the discriminative model as

$$p(z|\mathbf{x}_{1:T}) = \textstyle\sum_y p(\mathbf{x}_{1:T}|y) p(y|z) p(z) / p(\mathbf{x}_{1:T}) \tag{6}$$

$$= \textstyle\sum_y \dfrac{p(y|\mathbf{x}_{1:T})}{p(y)} p(y|z) p(z), \tag{7}$$

$$p(y|\mathbf{x}_{1:T}) = \textstyle\sum_{m_{1:T}, s_{1:T}} p(y|\mathbf{x}_{1:T}, m_{1:T}, s_{1:T}) \\ p(m_{1:T}|\mathbf{x}_{1:T}, s_{1:T}) p(s_{1:T}|\mathbf{x}_{1:T}), \tag{8}$$

where $p(\mathbf{x}_{1:T}|y)$ is a likelihood given from a speech-age, and $p(y|z)$ is a speech-age probability given from the person's real age z. The posterior $p(y|\mathbf{x}_{1:T})$ is modelled by NNs with parameter set $\mathbf{\Theta}$. $p(z)$ is a prior distribution of age, and it is assumed to be a uniform distribution in the estimation phase. The oversample technique [16] is applied to avoid learning this prior probability of the training datasets in the training phase.

The influence of the transfer function of speech signal can be reduced from speech features with signal processing techniques. Because the transfer function appears as a constant additive noise in log-power spectrum domain, the constant term can be removed by subtracting the mean log-power spectrum of all frames in the utterance from log-power spectrum at each frame. The mean subtraction in the speech feature domain is well-known as a cepstral normalization in ASR area [17].

3.2. Design of Speech-age Probability

The new probability $p(y|z)$ is essential for the design of our method. We propose to use three types of distribution, Gaussian, Poisson and generalized Gaussian distribution (GGD). Although the variance of the Gaussian probability density function (PDF) is independent to age z, those of the other two distributions are dependent on z, in which a child tends to be easier to

Table 1: *Corpus statistics: phonemically balanced sentence (PBS).*

	Corpus	Number of speakers	Number of men / women	Age range	Age label resolution	Utterance Content	Amount (hours)
Train	SLC-3	86	43 / 43	[6 12]	1 (exact)	Travel dialogue, PBS	35
	APP+APPDIC	3700	1300 / 2400	[14 65]	1 (exact)	Travel dialogue, Word	89
	S-JNAS	360	181 / 179	[60 91]	1 (exact)	Newspaper, PBS	169
Test	JNAS	79	31 / 38	[18 70]	10 (e.g. 20–29)	Newspaper, PBS	0.5
	CSJ	107	54 / 53	[18 80]	5 (e.g. 20–24)	Lecture, Dialogue	0.8

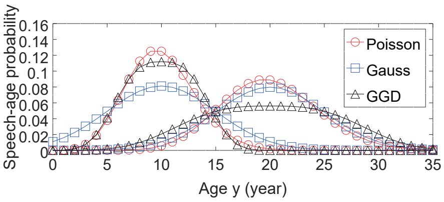

Figure 2: *Comparison of discretized $p(y|z)$ at $z = 10, 20$*

idenfity than elder people.

$$p(y|z) = \mathcal{N}(z, \sigma^2), \quad p(y|z) = z^y e^{-z}/y!, \text{ and} \tag{9}$$

$$p(y|z) = \frac{\beta}{2z\alpha\Gamma(1/\beta)} \exp\left(-\frac{|y-z|^\beta}{(z\alpha)^\beta}\right). \tag{10}$$

Because y and z are discrete variable, the actual PDF is discretized and satisfies $\sum_y p(y|z) = 1$ in the fixed range.

Another difference between Gaussian and GGD is the sharpness of the distribution peak. The peak shape of GGD PDF around the mean is flatter than that of the Gaussian PDF. Figure 2 shows a comparison of the shape of these three distributions at $z = 10, 20$ under certain parameters. This property of GGD is expected to relax the overfitting to the real age and may be close to our human hearing senses.

3.3. Integration with NNs and Cost Function

We can consider the correspondence relationship between neural layers and $p(y|\mathbf{x}_{1:T}; \boldsymbol{\Theta})$ in the generative model because the difference between them is a deterministic or stochastic process. The missing aspect of NN design for age estimation also becomes clear through this process. By considering the role of each function, we can determine that $p(s_{1:T}|\mathbf{x}_{1:T})$, $p(m_{1:T}|\mathbf{x}_{1:T}, s_{1:T})$ and $p(y|\mathbf{x}_{1:T}, m_{1:T}, s_{1:T})$ correspond to the BLSTM, ATTN, and CLS layers, respectively.

This model relationship indicates the exact inputs of the ATTN and CLS layers. The modified functions become

$$a_{1:T} = \text{mATTN}(\mathbf{h}_{\text{f},1:T}, \mathbf{h}_{\text{b},1:T}, \mathbf{x}_{1:T}), \text{and} \tag{11}$$

$$p(y|\mathbf{x}_{1:T}) = \text{mCLS}(\mathbf{g}), \quad \mathbf{g} = \textstyle\sum_t a_t[\mathbf{h}_{\text{f},t}, \mathbf{h}_{\text{b},t}, \mathbf{x}_{1:T}]. \tag{12}$$

Note that the speech feature $\mathbf{x}_{1:T}$ is used in both functions and is expected to behave as a kind of constraint.

The cost function for back-propagation is also modified in accordance with the probabilistic model. The new cost function for a specific paired data becomes

$$J(\boldsymbol{\Theta}) = \sum_z -q(z)[\log \sum_y \frac{p(y|\mathbf{x}_{1:T}; \boldsymbol{\Theta})}{p(y)} p(y|z)] + \text{const.} \tag{13}$$

We can simplify the cost function by applying Jensen's inequality to Eq. (13) with $\sum_y p(y|z) = 1$. The upper bound of the cost function becomes

$$J(\boldsymbol{\Theta}) \leq \textstyle\sum_{z,y} -q(z)p(y|z) \log p(y|\mathbf{x}_{1:T}; \boldsymbol{\Theta})) + \text{const.} \tag{14}$$

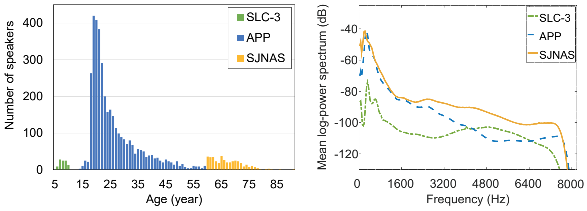

Figure 3: *Age distribution (left) and mean log-power spectrum (right) of each training corpus*

Only the first term is used to calculate the gradient for the SGD learning of NNs $p(y|\mathbf{x}_{1:T}; \boldsymbol{\Theta})$. The prior $p(y)$ is assumed as uniform in this paper. Our model includes *the convolution of real age and speech-age distributions*, and $p(y|z)$ works as a probabilistic filter.

4. Experiments

4.1. Data for Training and Test Set

We used speech data from four types of Japanese corpus as a training set. The set consists of four corpora, SLC-3[1], APP, APPDIC[2] and S-JNAS[3]. The speaker age of these four corpora is annotated by the exact age. The ranges of age of each corpus are quite different; the speakers of SLC-3, APP/APPDIC and S-JNAS are 6–12 years old, 14–64 years old and 60–91 years old, respectively. The utterance contents are also different among corpora. Speakers of some corpora uttered phonemically balanced sentences and newspaper articles, and those of other corpora uttered dictionary words or travel dialogue scripts. The data amount of SLC-3, APP/APPDIC and S-JNAS were 35, 89, and 169 hours, respectively. This information is summarized in the upper three rows of Table 1.

Figure 3 shows the age histogram and the mean log-power spectrum of these three corpus. The numbers of the speakers are quite different among corpora; the APP corpus includes over 3500 speakers. The spectrum shapes of each corpus are also different each other. Such difference can be caused not only by aging but also by the transfer function of microphone and low-pass filter.

We used speech data from two kinds of Japanese corpus as a *non-held-out* test set: JNAS[4] and Corpus of Spontaneous Japanese (CSJ) [18]. The speakers for the test set were uniformly selected over age and gender distributions. The total number of speakers of JNAS and CSJ were 79 and 107, respectively. Five utterances per person were also selected, and the duration of each utterance ranged from 3 to 5 seconds. The ages of these corpora were labelled with a range, and the range resolutions were 5 (e.g. 10–14 years old) for CSJ and 10 (10–19 years old) for JNAS. Speakers of JNAS uttered phonemically

[1] https://alaginrc.nict.go.jp/slc-outline.html#3

[2] https://www.atr-p.com/products/sdb.html#503

[3] http://research.nii.ac.jp/src/en/S-JNAS.html

[4] http://research.nii.ac.jp/src/en/JNAS.html

Table 2: *Age estimation accuracy: mean absolute error (MAE).*

	Configuration		MAE of test corpus		
	Speech-age distribution	Mod. NNs	JNAS	CSJ	Ave.
Baseline	None (Delta)		20.3	15.7	18.0
Proposed	Poisson		16.2	11.8	14.0
	Gaussian		16.1	**<u>10.9</u>**	13.5
	GGD		14.8	12.2	13.5
	Poisson	✓	17.0	11.9	14.5
	Gaussian	✓	16.4	11.1	13.8
	GGD	✓	**<u>12.7</u>**	11.8	**<u>12.3</u>**
Proposed w/o OS	Poisson		19.3	11.6	15.5
Proposed w/o MN	Poisson		23.0	13.5	18.2

balanced sentences and newspaper articles, and those of CSJ uttered lecture presentations and dialogue. Detailed information is summarized in the lower two rows of Table 1.

4.2. Configurations

The speech features were MFCCs with and without utterance-wise cepstral mean normalization (MN) to confirm the influence of corpus-specific transfer function. OpenSMILE [19] was used to extract these features from utterance speech signals. We used the configuration of 39-dimenson MFCC12_E_D_A and MFCC12_E_D_A_Z set for with and without MN, respectively.

All neural models were implemented by PyTorch [20]. The number of dimensions of the context vector of BLSTM (N_h) was 128, and the intermediate number of dimensions of ATTN (N_a) was 24. The number of age classes (N_c) was 100 (from 0 to 99 years old). Additional linear projections to the input and the output layers would cause overfitting. Adam [21] was used for the SGD learning, and its learning rate was set to $5.0 \cdot 10^{-4}$ and decayed at each epoch by multiples of 0.75. The number of epochs was 24. All other parameters were set to default.

The parameters of the speech age distribution were selected from several trials heuristically. The σ^2 of Gaussian PDF was 100 because it performed best among 25, 100, and 400. The α and β of GGD were set to 0.5 and 3, which were not tuned.

Oversampling (OS) was conducted on the basis of the age histogram of the training set. The utterances with age label z were duplicated in accordance with the ratio of the frequency of age z to the maximum frequency among all ages. The fractions were rounded by the floor operation.

4.3. Results

We used a mean absolute error (MAE) to measure age estimation performance. MAE is calculated by the mean absolute distance between predicted age $\hat{z}$ and ground truth age z [22]. If the age label is expressed by range z_l–z_r, such as 10–14, the ground truth age was set to the central value $(z_l + z_r)/2$, in this case, 12. **The MAEs for presudo held-out set of training data were ranged from 7 to 9, and these scores were similar to those of other studies** [5, 6].

The MAEs of Baseline and our methods are summarized in Table 2. The *Speech-age distribution* column denotes the type of distribution $p(y|z)$. Our proposed methods with speech-age model are evaluated with and without network modification (Mod. NNs) by Eqs. (11)-(12). The *JNAS* and *CSJ* columns show the results for each open corpus. The *Ave.* column means the average of MAE over JNAS and CSJ, and our aim is to improve this averaged value. The OS and MN were applied both to Baseline and proposed methods. The results of the proposed method without OS or MN are also shown for comparison.

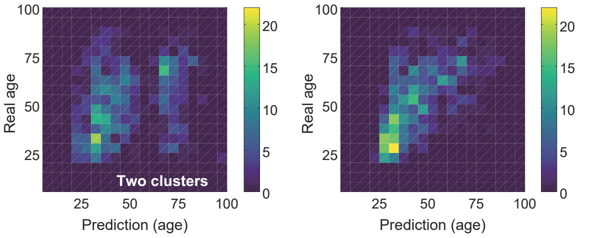

Figure 4: *Confusion matrices of Baseline (left) and Proposed with GGD and modified NNs (right) for CSJ set*

We found that the speech-age model is effective for heterogeneous datasets by comparing Baseline and proposed method. Three types of distribution improved the averaged MAE by 2–4 compared with those of Baseline. The tendencies of GGD and the other two distributions are different; GGD is good for JNAS and the others are good for CSJ. This indicates that the flatness around the distribution peak is important to reduce large estimation errors. Such a property seems to match the ambiguity of identifying age from speech. The age dependency of distribution is not so effective individually by the results of the Poisson and Gaussian distributions.

Our network modification with GGD speech-age model performed best. However, the results with the Poisson and Gaussian distributions were almost the same with those without modification. The combination of GGD and modified networks might match actual data.

We also found that MN was also effective for heterogeneous datasets. When MN was not applied to the proposed method, its performances degraded severely by an average of 4.2. The performance without OS also degraded by an average of 1.5, which shows the importance of the fundamental method. These two methods should be applied to improve performance stability.

The confusion matrices in Fig. 4 show the influence of overfitting and its relaxation by proposed methods. Since the age label of CSJ is annotated with a range of 5 years, such as 5–10 years old, the 20-class confusion matrices are summarized from the raw 100-class confusion matrices. Undesirable two clusters exist in the Baseline's matrix, which is obviously affected by the age distribution of APP and SJNAS in the training set. Our speech-age model was no suffered by such bias and performed well. The estimated age of our method is correlated with the real age, and the estimation tends to be younger than the real age.

5. Conclusion

We proposed a speech-age model and its integration with sequence NNs for age estimation from heterogeneous speech datasets. The speech-age model is utilized as a smoothing function to relax the corpus-specific age distribution. The generative model logically integrates the speech-age model and sequence NNs. The utterance-wise mean normalization was confirmed to be effective to normalize the differences in the transfer functions of each corpus. Experiments showed that our proposed method outperformed the baseline by 5.7 in terms of the MAE of age.

6. Acknowledgements

This work was supported by JST, PRESTO Grant Number JPMJPR1857, Japan and JSPS KAKENHI Grant Number JP19H05692.

7. References

[1] N. Minematsu, M. Sekiguchi, and K. Hirose, "Automatic estimation of one's age with his/her speech based upon acoustic modeling techniques of speakers," in *Prof. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 2002, pp. I–137–I–140.

[2] Q. Tian and S. Chen, "Cross-heterogeneous-database age estimation through correlation representation learning," *Neurocomputing*, vol. 238, pp. 286–295, 2017.

[3] A. Akbari, M. Awais, Z. Feng, A. Farooq, and J. Kittler, "Distribution cognisant loss for cross-database facial age estimation with sensitivity analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence (early access)*, 2020.

[4] P. Ghahremani, P. S. Nidadavolu, N. Chen, J. Villalba, D. Povey, S. Khudanpur, and N. Dehak, "End-to-end deep neural network age estimation," in *Proc. of Interspeech*, 2018, pp. 277–281.

[5] R. Zazo, P. Sankar Nidadavolu, N. Chen, J. Gonzalez-Rodriguez, and N. Dehak, "Age estimation in short speech utterances based on LSTM recurrent neural networks," *IEEE Access*, vol. 6, pp. 22 524–22 530, 2018.

[6] S. B. Kalluri, D. Vijayasenan, and S. Ganapathy, "A deep neural network based end to end model for joint height and age estimation from short duration speech," in *Prof. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2019, pp. 6580–6584.

[7] G. Levi and T. Hassncer, "Age and gender classification using convolutional neural networks," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 34–42.

[8] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[9] H. Pan, H. Han, S. Shan, and X. Chen, "Mean-variance loss for deep age estimation from a face," in *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5285–5294.

[10] A. Othmani, A. R. Taleb, H. Abdelkawy, and A. Hadid, "Age estimation from faces using deep learning: A comparative analysis," *Computer Vision and Image Understanding*, vol. 196, p. 102961, 2020.

[11] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[12] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. of International Conference on Learning Representations*, 2014.

[13] Y. Kim, C. Denton, L. Hoang, and A. M. Rush, "Structured attention networks," in *Proc. of International Conference on Learning Representations*, 2017.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. of Advances in Neural Information Processing Systems*, vol. 30, 2017.

[15] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer-Verlag New York, 2006.

[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, no. 1, pp. 321–357, 2002.

[17] A. Acero and R.M. Sterm, "Cepstral normalization for robust speech recognition," in *Proc. of the Speech Processing in Adverse Conditions*, 1992, pp. 89–92.

[18] K. Maekawa, "Corpus of spontaneous Japanese: Its design and evaluation," in *Proc. of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.

[19] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proc. of the 18th ACM International Conference on Multimedia*, 2010, pp. 1459–1462.

[20] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. of Advances in Neural Information Processing Systems*, 2019, pp. 8024–8035.

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. of International Conference on Learning Representations*, 2015.

[22] Y. Fu, G. Guo, and T. S. Huang, "Age synthesis and estimation via faces: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 11, pp. 1955–1976, 2010.