Speaker Age Estimation Using Age-Dependent Insensitive Loss

Yuki Kitagishi*, Hosana Kamiyama*, Atsushi Ando*, Naohiro Tawara†,
Takeshi Mori* and Satoshi Kobashikawa*
* NTT Media Intelligence Laboratories, Japan
E-mail: yuki.kitagishi.wp@hco.ntt.co.jp Tel: +81-46-859-5160
† NTT Communication Science Laboratories, Japan

***Abstract*—This paper proposes a new speaker age estimation method that uses an age-dependent insensitive loss. Most conventional speaker age estimation frameworks ignore the ambiguity of a perceptual speaker age. These "over-sensitive" frameworks can cause critical errors far from the actual age. We propose an age-dependent insensitive loss for speaker age estimation. The key idea of the proposed method is that the age estimator should allow some ambiguity of actual age and this ambiguity should depend on age. The age-dependent insensitivity is learned by ε-MAE (mean absolute error) loss and soft target cross entropy loss in regression and classification problems. Experimental results showed that the proposed method improves the mean absolute error and the ratio of critical error by 5.2% and 5.7% for the regression problem and 9.6% and 31.5% for the classification problem.**

Index Terms: speaker age estimation, ε-insensitive loss, soft target cross-entropy loss, age-dependent insensitive loss, speaker attribution estimation

I. Introduction

Speech contains not only text information but also speaker attribute information such as gender and age. The rapid development of novel speech applications is demanding technologies that can estimate such speaker attribute information. For example, speaker age estimation could be used to personalize the advertisements to suit the customer's age, or prioritize suspects in forensic cases [1], [2].

The methods that have been proposed to estimate speaker age fall into two types; group and individual estimation. The first, group estimation, classifies speaker's age into several classes such as child, young, adult and senior [3]–[7]. Though it is robust in limited training data, it is not practical for some applications due to the roughness of the estimation. The other methods estimate the speaker's age directly by solving regression or classification problems. The estimation task is formulated as either the regression [7]–[12] or the classification problems [13]. Such methods can estimate the speaker age in detail. However, more training data are needed to achieve high performance.

Most conventional works defined the actual age as the ground-truth without any ambiguity. However, it is known that there is a high ambiguity in age perception from speech [14]. For this problem, some conventional works introduced some frameworks to allow ambiguity in age labels. For example, some methods use a soft target that spreads like a Gaussian distribution centered on the actual age as the actual target for cross-entropy loss in facial age estimation [15], [16]. Also, there is a method that reduces the variance of the posterior probability by the softmax function in facial age estimation [17]

In this research, we propose an age-dependent insensitive loss for learning the speaker age estimation model. The age-dependent insensitive loss regards the estimated error below a threshold as the correct like the ε-insensitive loss [18]. The threshold is not the constant decided manually but age-dependent value. Our evaluations showed that the estimation performances that are mean absolute error (MAE) and the number of critical error are improved by age-independent insensitive loss as was adopted in previous works on both regression and classification problems.

II. Speaker age estimation

This section describes two conventional formulations of the speaker age estimation task; the regression and the classification.

A. Regression Task

Generally, speaker age is estimated directly as a type of regression problem [7]–[12]. The regression task is formulated as an estimation for the actual age of the speaker y from the input speech feature x; it is defined as,

$$\hat{y} = f(x, \mathbf{\Theta}_{\mathrm{r}}), \tag{1}$$

where $\hat{y}$ is the estimated age, $f(\cdot)$ is the projection function such as SVR or neural networks, and $\mathbf{\Theta}_{\mathrm{r}}$ is a set of parameters of $f(\cdot)$. x is a series of acoustic features or a speech representation such as speaker embedding vector [7]–[10]. In conventional works, $\mathbf{\Theta}_{\mathrm{r}}$ is optimized by stochastic gradient descent using MAE or mean squared error (MSE) loss. In this paper, we use MAE loss $\mathcal{L}_{\mathrm{MAE}}$ that is defined as,

$$\mathcal{L}_{\mathrm{MAE}} = |\hat{y} - y|. \tag{2}$$

B. Classification Task

Recently, some methods that estimate age based on face images or speech as classification problems have been proposed [15], [19]. In these studies, each age value is treated as one

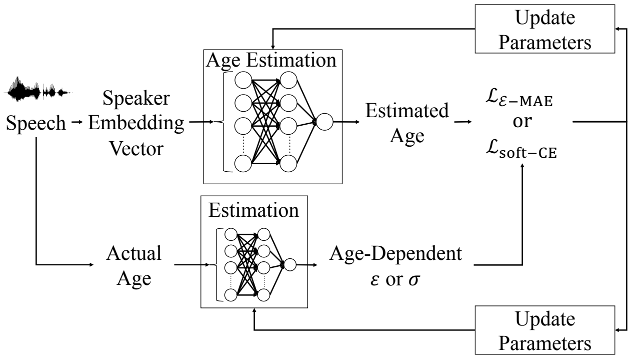

Update Parameters
Age Estimation
Speech
Speaker Embedding Vector
Estimated Age
$\mathcal{L}_{\varepsilon\text{-MAE}}$ or $\mathcal{L}_{\text{soft-CE}}$
Estimation
Actual Age
Age-Dependent ε or σ
Update Parameters

Fig. 1. Block diagram of proposed method for speaker age estimation

class. The classification task is formulated as estimation for y from x; it is defined as,

$$\hat{y} = \sum_{y_n=\min(n)}^{\max(n)} P(y_n|x)y_n, \tag{3}$$

where y_n is a specific age value, n is a set of age values; it is defined as $n \in (0 \cdots N)$, N is a maximum value of age values, and $P(y_n|x)$ is the posterior probability for each age value. Generally, $\hat{y}$ is estimated based on maximum posterior probability as $\hat{y} = \arg\max P(y_n|x)$ [13], however, we calculate $\hat{y}$ as the expected value from $P(y_n|x)$ [15], [19]. In conventional works, $P(y_n|x, \mathbf{\Theta}_\mathrm{c})$ is calculated using the age estimation model with a set of parameters ($\mathbf{\Theta}_\mathrm{c}$), and $\mathbf{\Theta}_\mathrm{c}$ is optimized by stochastic gradient descent using cross-entropy loss $\mathcal{L}_\mathrm{CE}$; it is defined as,

$$\mathcal{L}_\mathrm{CE} = -\sum_{y_n=\min(n)}^{\max(n)} T(y)\log[P(y_n|x, \mathbf{\Theta}_\mathrm{c})], \tag{4}$$

where $T(y)$ is the actual target of y. Generally, the 1-hot target is used as the actual target; it is defined as,

$$T(y) = \begin{cases} 1 & (y_n = y) \\ 0 & (\text{otherwise}). \end{cases} \tag{5}$$

III. Age-dependent insensitive loss for speaker age estimation

Most conventional speaker age estimation models were trained to estimate the ages more accurately. However, we consider that it is reasonable to assume that the appropriate variance varies depending on age because of the difference in the standard deviation (SD) of the acoustic features [14]. In this paper, we proposed that 1) the model regards the estimation error below a threshold as correct when training the model, 2) the thresholds are varied according to the actual age. Figure 1 shows the proposed method.

A. Age-Dependent Insensitive Loss for Regression

For the regression model, we use the ε-MAE loss that treats the estimation error as being zero if it lies within the threshold range. The ε-MAE loss ($\mathcal{L}_{\varepsilon\text{-MAE}}$) is defined as,

$$\mathcal{L}_{\varepsilon\text{-MAE}} = \max(|\hat{y} - y| - \varepsilon, 0), \tag{6}$$

where ε is the range value of the threshold for regression. Figure 2 shows the loss value by the conventional and ε-MAE with $\varepsilon = 6$.

Here, age-dependent ε is calculated by a multilayer perceptron (MLP) based on actual age; as shown in below of Figure 1. Age-dependent ε is defined as,

$$\varepsilon_y = \max(f_\mathrm{r}(y, \mathbf{\Omega}_\mathrm{r}) + b_\mathrm{r}, 0), \tag{7}$$

where f_r is the MLP for estimation the age-dependent ε, $\mathbf{\Omega}_\mathrm{r}$ is the parameters of $f_\mathrm{r}(\cdot)$, and b_r is the bias. $\mathbf{\Omega}_\mathrm{r}$ were trained simultaneously with the model for speaker age estimation using $\mathcal{L}_{\varepsilon\text{-MAE}}$.

B. Age-Depend Insensitive Loss for Classification

For the classification model, we use a soft target cross-entropy loss that covers the region from the actual age to the range of the threshold as the actual target. In this paper, we use the Gaussian distribution for the soft target ($T(y)_\mathrm{gauss}$); it is defined as,

$$T(y)_\mathrm{gauss} = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y_n - y)^2}{2\sigma^2}\right), \tag{8}$$

where σ is the SD of the Gaussian distribution that means the range of the threshold. The soft target cross-entropy loss using $T(y)_\mathrm{gauss}$ as the actual target ($\mathcal{L}_\text{soft-CE}$) is defined as,

$$\mathcal{L}_\text{soft-CE} = -\sum_{y_n=\min(n)}^{\max(n)} T(y)_\mathrm{gauss}\log[P(y_n|x)]. \tag{9}$$

Figure 3 shows the soft target by Equation (8) with the actual age is 29 and $\sigma = 3$.

Here, age-dependent σ is calculated using an MLP based on actual age; to see below of Figure 1: age-dependent σ is estimated from actual age using the MLP. Age-dependent σ is defined as,

$$\sigma_y = \max(f_\mathrm{c}(y, \mathbf{\Omega}_\mathrm{c}) + b_\mathrm{c}, 0), \tag{10}$$

where f_c is the model for estimation of the age-dependent σ, $\mathbf{\Omega}_\mathrm{c}$ is the parameters of $\mathbf{\Omega}_\mathrm{c}$ and b_c is the bias. $\mathbf{\Omega}_\mathrm{c}$ were trained simultaneously with the model for speaker age estimation using $\mathcal{L}_\text{soft-CE}$.

IV. Speaker age estimation experiments

In order to show the effectiveness of the proposed age-dependent intransitive loss over the conventional loss, we conducted speaker age estimation experiment.

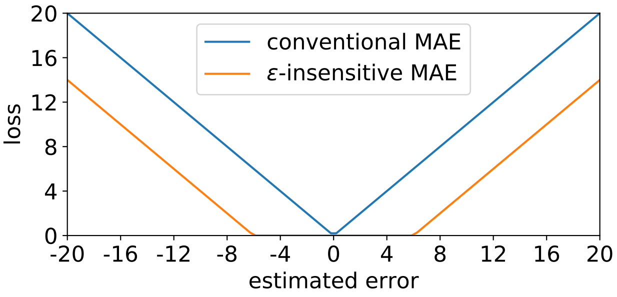


Fig. 2. Conventional and ε-MAE ($\varepsilon = 6$).

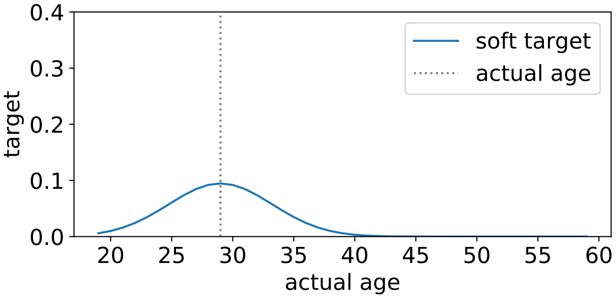
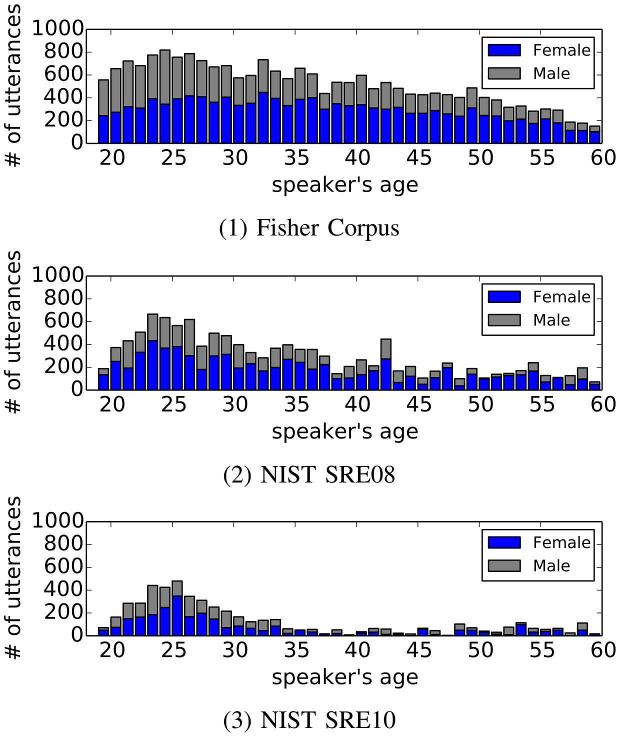


Fig. 3. Soft target for cross-entropy loss (actual age is 29 and $\sigma = 3$).



(1) Fisher Corpus

(2) NIST SRE08

(3) NIST SRE10

Fig. 4. The number of utterances of each dataset

A. Dataset

We used the Fisher Corpus [20] as the training set, NIST SRE08 [21] as the validation set, and NIST SRE10 [22] as the evaluation set. Each dataset contains channel-separated telephone conversations of two speakers quantized at 16 bits with a sampling frequency of 8 kHz. Figure 4 shows the number of utterances of each dataset. Table I shows the number of speakers, the number of utterances and the average duration of each dataset. In each dataset, some age values had few speakers. For example, very few speakers were over 60 years old. Accordingly, to avoid model overfitting, we used utterances whose speakers ranged from 19 to 59 years old for training and evaluation. That is, n was defined as $n \in (19, .20 \cdots, 58, 59)$, and there were 41-class classification task.

B. Experimental Setup

We created an MLP with two fully-connected layers. The features were 512-dim x-vectors [23] extracted from 23-dim Mel frequency cepstral coefficient (MFCC) with 25-millisecond windows offset by 10-milliseconds. Energy-based voice activity detection was used to remove non-speech frames. MFCCs were normalized by short-time mean determination using a three-second window. The x-vector extractor was trained using Kaldi's SRE16 recipe [24] without NIST SRE10. Each hidden layer had 256 neurons with rectified linear unit (ReLU) non-linearity as the activation function. Batch normalization [25] was used before the non-linearity. The dimension of the output layer was one for the regression problem, or 41 for the classification problem. In the output layer of the classification problem, the softmax function was used as the activate function. We used adaptive moment estimation (Adam) [26] with a learning rate of 0.001, β_1 of 0.9, and β_2 of 0.999 to update the model parameters. We decreased the learning rate by a factor of two when the validation loss did not improve for two successive epochs. The validation loss was computed without using insensitive loss. The minimum learning rate was 1e$-$05. The training was stopped if the loss did not improve for three consecutive epochs when the learning rate was minimum. The mini-batch size was 64. All hyper-parameters were decided using the loss of the validation set.

As the baselines, we defined two and three methods for regression and classification problems, respectively. For the regression, one was using conventional MAE loss as shown in Equation 2. Another was using MAE loss by expected value calculated the posterior probability by the softmax function, and KL loss of the posterior probability with σ and λ as the hyper-parameter were 3 and 1.0, respectively [16]. For, the classification, the first was conventional 1-hot cross-entropy loss as shown in Equation (4). The second was multi-task learning of classification and regression problems with λ as the hyper-parameter was 0.5 [13]. The third was using mean-variance loss that contains the soft target cross-entropy loss, the squared error between the expected value and the actual age, and the variance of the posterior probability with λ_1 and λ_2 as the hyper-parameter were 0.3 and 0.1, respectively [17].

The age-independent ε was 6, and the age-independent σ

TABLE I
DATASET DETAILS

	# of speakers			# of utterances			duration
	overall	female	male	overall	female	male	[sec.]
Fisher Corpus [20]	11,320	6,637	4,683	21,252	12,179	9,037	602.7
NIST SRE08 [21]	1,086	686	400	12,017	7,419	4,598	283.8
NIST SRE10 [22]	399	207	192	5,173	2,808	2,365	303.8

was 3. When we use age-dependent ε or σ, the parameters of the model for estimation the age-depend ε or σ were learned as well. The age-dependent ε was estimated by $f_{\mathrm{r}}(\cdot)$ that had two fully-connected layers, each of which had 32 neurons with ReLU non-linearity as the activate function. The age-dependent σ was estimated by $f_{\mathrm{c}}(\cdot)$ that had three fully-connected layers, each of which had 64 neurons with ReLU non-linearity as the activate function. In each hidden layer, batch normalization was used before ReLU non-linearity, and Adam was used to update the parameters as was done in learning the model for age estimation. The bias for the age-dependent ε (i.e., b_{r}) in Equation 7 was 0 and for the age-dependent σ (i.e., b_{c}) Equation 10 was 4. The initial learning rate was 5e$-$5 and 5e$-$4 for $f_{\mathrm{r}}(\cdot)$ and $f_{\mathrm{c}}(\cdot)$, respectively. The learning rate decreased in the same way as the learning rate of the model for age estimation. The L2 penalty were 0.1 for both $f_{\mathrm{r}}(\cdot)$ and $f_{\mathrm{c}}(\cdot)$. We set the L2 penalty and small learing rate for training $f_{\mathrm{r}}(\cdot)$ and $f_{\mathrm{c}}(\cdot)$ to prevent the aget-dependent ε and σ from diverging to infinity All hyper-parameters were decided using the accuracy of the validation set.

To evaluate the accuracy, we used MAE, Pearson's correlation coefficient, and the ratio of outliers. MAE is defined as,

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|, \tag{11}$$

where N is the number of samples and i is ID of each sample. Pearson's correlation coefficient ρ is defined as follows,

$$\rho = \frac{1}{N-1}\sum_{i=1}^{N}(\frac{\hat{y}_i - m_{\hat{y}}}{s_{\hat{y}}})(\frac{y_i - m_y}{s_y}), \tag{12}$$

where, $m_{\hat{y}}$ and $s_{\hat{y}}$ are the mean and SD of estimated age, and m_y and s_y are the mean and SD of actual age, respectively. In addition, we evaluated the ratio of the number of incorrect age estimates over 10 years old; it is defined as,

$$\textit{outlier-ratio} = \frac{\text{number of }[|\hat{y}_i - y_i| \geqq 10]}{N} \times 100. \tag{13}$$

We considered that incorrect estimated age over 10 years old are critical errors in real world applications.

C. Results

Table II and Table III show the results of the regression and classification problems, respectively. Figure 5 shows the age-dependent ε and σ when early stopping, respectively. We used paired t-test for comparing the MAEs and chi-squared test for comparing the ratio of outliers; the significance level was 0.05.

In the regression problem, our proposed method improved the MAE by 3.1% to 5.2 %, and the outlier ratio by 5.7% to 9.8% then the conventional works. The age-independent ε loss and [16] yielded significantly lower MAEs than those achieved with conventional MAE loss. Using age-dependent ε attained even lower the MAE. The proposed method created fewer outliers, but the differences were not significant.

With regard to the classification problem, our proposed method improved the MAE by 2.0% to 12.8%, and the outlier ration by 9.6% to 31.5% than the conventional works. The 1-hot target yielded significantly higher MAE than the other method. Both two previous works yielded significantly lower MAEs than using the age-independent σ. The age-dependent σ yielded significantly lower MAE than using the age-independent σ, however, the differences between the MAEs of the previous works and using the age-dependent σ were not significant. The use of the 1-hot target created significantly more outliers than the proposed methods and previous works. However, the differences among the proposed methods and previous works were not significant.

D. Discussions

We hypothesized that there is the appropriate value according to the actual age for ε and σ, and proposed a method for considering the hypothesis. We considered that the experimental results showed that our hypothesis was supported; the performances were improved than the conventional works or age-independent insensitive loss.

In the conventional works, the variance of actual age (i.e., age-independent σ and ε) was decided manually. However, there are appropriate variances according to each actual age, and the previous works could not consider it. Therefore, it is considered that the performances became poor by the training with inappropriate variance. We considered that our proposed method could train the model while considering the appropriate variance for each actual age by estimating age-dependent ε and σ from the actual age. Our proposed method could train the model while considering the appropriate variance for each actual age by estimating age-dependent ε and σ from the actual age.

The age-dependent ε and σ show same tendency that was valley type centered on about 35 years old. We consider that this is due to the estimation difficulty to the training set; there were the correlation between the SDs of the estimated error for each actual age and the age-dependent ε ($\rho = 0.83$) or σ ($\rho = 0.62$). Therefore, we believe that the age-dependent ε and σ were trained correctly. However, there is a possibility that the age-dependent ε and σ are affected by the balance

TABLE II
RESULTS: AGE ESTIMATES YIELDED BY REGRESSION PROBLEM

	loss function	ε	MAE	ρ	*outlier-ratio* [%]
baseline	MAE loss	—	4.66	0.83	10.9
	MAE loss + KL loss [16]	—	4.56	0.83	11.4
proposed 1	MAE loss	age-independent	4.54	0.83	**10.3**
proposed 2	MAE loss	age-dependent	**4.42**	**0.84**	10.7

TABLE III
RESULTS: AGE ESTIMATES YIELDED CLASSIFICATION PROBLEM

	loss function	σ	MAE	ρ	*outlier-ratio* [%]
baseline	1-hot target cross-entropy loss	—	5.07	0.80	14.5
	1-hot target cross-entropy loss + MSE [13]	—	4.47	**0.84**	10.1
	mean-variance loss [17]	—	4.44	**0.84**	**10.0**
proposed 1	soft target cross-entropy loss	age-independent	4.58	0.83	11.4
proposed 2	soft target cross-entropy loss	age-dependent	**4.42**	0.83	10.7

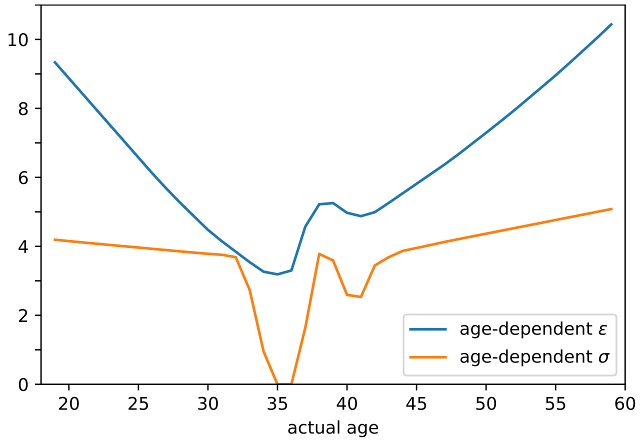


Fig. 5. Age-dependent ε and σ at early stopping

of training set. For example, the age-dependent ε and σ may become larger in the age with few speakers. It is nessesary to prepare the balanced training dataset to prevent such problem.

V. CONCLUSIONS AND FUTURE WORKS

This paper proposed a new method for speaker age estimation using age-dependent insensitive loss. The age-dependent insensitive loss treats the estimation error below the threshold varied according to the actual age as correct. For regression problems, the ε-MAE (mean absolute error) loss is used and ε is varied according to the actual age. For classification problems, the soft target cross-entropy loss is used and the variance of the soft target is varied according to the actual age. In this paper, we evaluated the MAE and the ratio of outliers in estimated ages; both were found to be improved using age-independent insensitive loss as same as conventional works. Moreover, we validated our proposal of using age-dependent insensitive loss.

To validate the method, we conducted experiments using Fisher Corpus as a training set, NIST SRE08 as the validation set, and NIST SRE10 as the evaluation set. In the regression problem, the proposal improved the MAEs by 3.1% to 5.2%, and by 2.0% to 12.8% in the classification problem. Also, the number of estimation errors over 10 years old was decreased by 5.7% to 9.8% in the regression problem, and by 9.6% to 31.5% in the classification problem. As future work, we plan to extend our framework by introducing multi-task loss which estimates other speaker attributions such as gender and individuality. And we plan to incorporate multi-modal features such as face images.

REFERENCES

[1] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "Paralinguistics in Speech and Language—State-of-the-Art and the Challenge," *Computer Speech and Language, Special Issue on Paralinguistics in Naturalistic Speech and Language*, pp. 4–39, Jan. 2013.

[2] S. Shepstone, Z.-H. Tan, and S. Jensen, "Audio-based Age and Gender Identification to Enhance the Recommendation of TV Content," *Consumer Electronics, IEEE Transactions on*, vol. 59, pp. 721–729, Aug. 2013.

[3] 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 *Proc. of International Conference on Acoustics, Speech, and Signal Processing*, ser. ICASSP 2002, May 2002, pp. I–137–I–140.

[4] T. Bocklet, G. Stemmer, V. Zeißler, and E. Nöth, "Age and gender recognition based on multiple systems — early vs. late fusion," in *Proc. Interspeech 2010*, Sep. 2010, pp. 2830–2833.

[5] P. J. Saeid Safavi, Martin Russell, "Identification of Age-Group from Children's Speech by Computers and Humans," in *Proc. Interspeech 2014*, Sep. 2014, pp. 3036–3040.

[6] H. Kamiyama, A. Ando, S. Kobashikawa, and Y. Aono, "Robust children and adults speech identification and confidence measure based on DNN posteriorgram," in *Proc. of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, ser. APSIPA ASC 2017, Dec. 2017, pp. 502–505.

[7] J. Równicka and S. Kacprzak, "Speaker Age Classification and Regression Using i-Vectors," in *Proc. Interspeech 2016*, Sep. 2016, pp. 1402–1406.

[8] M. Bahari, M. McLaren, H. V. hamme, and D. V. Leeuwen, "Speaker age estimation using i-vectors," *Engineering Applications of Artificial Intelligence*, vol. 34, pp. 99—108, 09 2014.

[9] S. O. Sadjadi, S. Ganapathy, and J. W. Pelecanos, "Speaker age estimation on conversational telephone speech using senone posterior based i-vectors," in *Proc. of International Conference on Acoustics, Speech and Signal Processing*, ser. ICASSP 2016, Mar. 2016, pp. 5040–5044.

[10] A. Fedorova, O. Glembek, T. Kinnunen, and P. Matějka, "Exploring ANN Back-Ends for i-Vector Based Speaker Age Estimation," in *Proc. Interspeech 2015*, Sep. 2015, pp. 3036–3040.

[11] R. Zazo, P. S. Nidadavolu, N. Chen, J. Gonzalez-Rodriguez, and N. Dehak, "Age Estimation in Short Speech Utterances Based on LSTM Recurrent Neural Networks," *IEEE Access*, pp. 22 524–22 530, 2018.

[12] 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 *Proc. of International Conference on Acoustics, Speech and Signal Processing*, ser. ICASSP 2019, 2019, pp. 6580–6584.

[13] 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. Interspeech 2018*, Dec. 2018, pp. 277–281.

[14] S. Skoog Waller, M. Eriksson, and P. Sörqvist, "Can you hear my age? Influences of speech rate and speech spontaneity on estimation of speaker age," *Frontiers in Psychology*, vol. 6, p. 978, 2015.

[15] G. Antipov, M. Baccouche, S.-A. Berrani, and J.-L. Dugelay, "Effective training of convolutional neural networks for face-based gender and age prediction," *Pattern Recognition*, vol. 72, pp. 15–26, 2017.

[16] B.-B. Gao, H.-Y. Zhou, J. Wu, and X. Geng, "Age Estimation Using Expectation of Label Distribution Learning," in *Proc. of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, ser. IJCAI 2018. International Joint Conferences on Artificial Intelligence Organization, Jun. 2018, pp. 712–718.

[17] H. Pan, H. Han, S. Shan, and X. Chen, "Mean-Variance Loss for Deep Age Estimation from a Face," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, ser. CVPR 2018, Jun. 2018, pp. 5285–5294.

[18] A. J. Smola and B. scholköpf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, pp. 199–222, 2004.

[19] R. Rothe, R. Timofte, and L. Van Gool, "DEX: Deep EXpectation of Apparent Age From a Single Image," in *Proc. of The IEEE International Conference on Computer Vision Workshops*, ser. ICCVW 2015, Dec. 2015, pp. 10–15.

[20] C. Cieri, D. Miller, and K. Walker, "The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text," in *Proc. of the Fourth International Conference on Language Resources and Evaluation*, ser. LREC 2014, May 2004.

[21] S. S. Kajarekar, N. Scheffer, M. Graciarena, E. Shriberg, A. Stolcke, L. Ferrer, and T. Bocklet, "THE SRI NIST 2008 speaker recognition evaluation system," in *Proc. of International Conference on Acoustics, Speech and Signal Processing*, ser. ICASSP 2009, Apr. 2009, pp. 4205–4208.

[22] A. F. Martin and C. S. Greenberg, "The NIST 2010 speaker recognition evaluation," in *Proc. of Interspeech 2010*, Sep. 2010, pp. 2726–2729.

[23] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," in *Proc. of International Conference on Acoustics, Speech and Signal Processing*, ser. ICASSP 2018, Apr. 2018, pp. 5329–5333.

[24] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *Proc. of IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, ser. ASRU 2011, Dec. 2011.

[25] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," in *Proc. of International Conference on International Conference on Machine Learning*, ser. ICML 2015, Jul. 2015, pp. 448—-456.

[26] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *Proc. of International Conference for Learning Representations*, ser. ICLR 2015, 2015.