INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

What are differences? Comparing DNN and human by their performance and characteristics in speaker age estimation

Yuki Kitagishi, Naohiro Tawara, Atsunori Ogawa, Ryo Masumura, and Taichi Asami

NTT Corporation

{yuki.kitagishi, naohiro.tawara, atsunori.ogawa, ryo.masumura, taichi.asami}@ntt.com

Abstract

We compare speaker age estimation results obtained by human listeners and a latest deep neural network (DNN) model to reveal differences in their estimation characteristics. A DNN model can achieve high speaker age estimation performance and is expected to be utilized in practical applications. Only a few studies compared speaker-age estimation performance between human listeners and machine learning models. However, the differences in their estimation characteristics have yet to be revealed. Our experimental results reveal that the DNN model performs comparable or superior to the listeners but is more sensitive to elderly speech, acoustic characteristics, and lengths of speech samples than the listeners. The results also reveal that the speakers' gender and some specific acoustic features negatively affect the listeners' estimation performance.

Index Terms: speaker age estimation, human versus DNN, estimation characteristics

1. Introduction

A high-performance automatic speaker age estimation has been achieved by using deep neural networks (DNNs) [1–15], and its application is expected in various fields of services, such as forensics, age-dependent advertisements, and supporting call center operations [16–18]. One concrete example of strong interest is the call center. In conventional systems, a customer's age is manually estimated by a call center operator from the customer's telephone call. Automatic speaker age estimation can support the operators by reducing their burden.

Many studies reported speaker age estimation performance and characteristics obtained by human listeners [19–39]. Some of them suggested that the listeners' estimation error is about ten years in a mean absolute error (MAE) [19–24]. Their estimation characteristics based on some speaker- and listener-related factors, such as their age, gender, and language were also reported [19, 23–32].

Only a few studies analyzed the difference in the estimation performance between the listeners and the machine learning models using the same speech dataset [23,30], but the differences in the estimation characteristics are not revealed. Huckvale and Webb compared the estimation performance obtained by their human participants and a support vector regression (SVR)-based model [23]. They revealed that the MAEs obtained by the participants and the SVR-based model were comparable. They also revealed a tendency that the MAEs for elderly speakers obtained by the model are more degraded than the participants. However, the differences in estimation characteristics between the listeners and the machine learning models are not revealed. Because only MAEs obtained by the participants and the model were compared.

The differences in estimation characteristics between the listeners and the machine learning model should be revealed. We expect the DNN model to perform superior to the listeners since modern DNN models estimate the speakers' age more accurately than conventional models; the modern DNN methods archived the estimation performance of about five years in MAE [4, 6, 10–15]. However, there is a possibility that the advantages and disadvantages of the DNN model's and the listeners' estimation are different, as reported in [23]. Also, it is not revealed whether the listeners' estimation characteristics reported in conventional studies [19, 23–32] are particular to the listeners or not.

To reveal these unsolved questions, we create a novel inhouse dataset and analyze the differences in the estimation characteristics between human listening participants and the latest automatic speaker age estimation system. The dataset is designed to contain simulated telephone calls between operators and customers, assuming the practical use of the automatic speaker age estimation at the call center. We employ 21 call center operators as the participants and the latest automatic age estimation system based on a Transformer model [11], and compare their age estimation results. In addition to comparing their estimation performance, we reveal differences in their estimation characteristics by comparing their misestimation tendencies.

Our main findings in this study can be summarized as follows.

(1) The latest DNN model achieves comparable or surpassed the estimation performance demonstrated by the listeners.

(2) However, compared with the listeners, the DNN model tends to be more sensitive to short-time input speech, mismatches between training and testing acoustic characteristics, and elderly speech.

(3) The speaker's gender and some specific age-related acoustic features negatively affect the listeners' estimation performance.

2. Related works

Only a few studies analyzed the differences in speaker age estimation performance between the participants and the machine learning models, by using same speech datasets for the participants and the model. Schötz compared human participants' age estimation performance with automatic age estimation based on a classification and regression trees (CART) [30]. Schötz indicated that the participants' age estimation performance outperformed the CART-based model, achieving 8.89 and 14.45 in MAEs, respectively. However, there is a significant estimation performance gap between the CART-based and modern DNN models. On the other hand, Huckvale and Webb also compared

10.21437/Interspeech.2023-1507

their participants' performance with the SVR [23]. They indicated that the SVR-based model slightly outperformed the participants, achieving 8.63 and 9.79 in MAEs, respectively. They also showed that the MAEs were more degraded for elderly speakers than younger speakers by comparing the MAEs per age group. Especially the MAE for the elderly speakers obtained by the SVR-based model was more degraded than the participants' MAE. However, more detailed differences in the estimation characteristics between the participants and the models were not revealed. Besides, their experiments were conducted using speech datasets with restricted speaking styles, such as isolated word utterances [30] and read speech [23].

In this paper, we conduct the speaker age estimation experiments with the participants and the latest DNN model, and analyze their estimation characteristics in addition to comparing their estimation performance. Besides, we analyze the estimation characteristics in practical environments by using simulated telephone calls at the call center and employing the participants who were experienced in the call center operation.

3. Speaker age estimation

We define three kinds of speaker age; the speaker's chronological age called "CA" [30], the speaker's age perceived by human listeners called "PA" [30], and the speaker age estimated automatically by the machine learning model, which we call "EA".

3.1. Speaker age estimation by human listeners

Speaker age estimation by human listeners is categorized into two types based on the range of speaker age; numerical and categorical estimations. In the numerical estimation [19–24], the participants estimate the speaker's age as one-year-step age value. In the categorical age estimation [32,36], the participants estimate the ages as categorical age labels, such as a five-year-step age group [32] or more coarse age label (e.g., child, young adult, middle-age, retired, and senior [36]).

We employ a five-year-step age class for our experiments. According to conventional studies [19–24], human listeners' estimation performance is about ten years in MAE. There is a possibility that the numerical age estimation is too difficult for human listeners, and it badly affects the listeners' age estimation performance. We consider the five-year-step age classes a reasonable age range due to the humans' MAE.

3.2. Automatic speaker age estimation by DNNs

Recent automatic speaker age estimation emploies numerical age estimation [1–15] and is categorized into two types; regression [1–7] and classification [8–15].

We use a classification model that estimates the posterior probabilities for 0 to 100 years old based on the state-of-the-art method [11]. It is formulated as the estimation of numerical CA y_i from input acoustic feature sequence X_i of i^{th} utterance; it is defined as, $\hat{y}_i = \sum_{k=0}^{100} P(k|X_i, \Omega)k$, where $\hat{y}_i$ is the EA, $P(k|X_i, \Omega)$ is the posterior probability for each age class $k \in [0, 100]$, and Ω is a set of DNN parameters. $\hat{y}_i$ is computed by the expected value of the posterior [12–14].

The Ω is optimized with a stochastic gradient descent algorithm on training speech samples with actual speaker-age labels. Some recent studies trained the DNN model based on the method of label distribution learning [8–14] by using a soft target like a normal distribution $\mathcal{N}(k|y_i, \sigma^2)$ as an actual target. We also use such a soft target and define the loss function as a soft-target cross-entropy

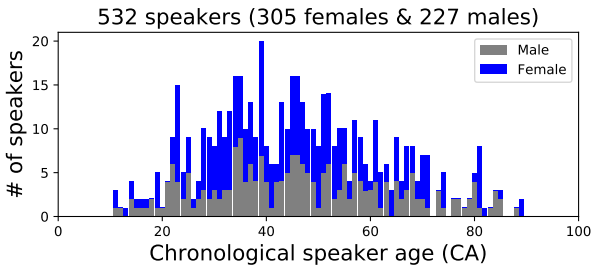


Figure 1: *Gender and age distribution of Call center dataset*

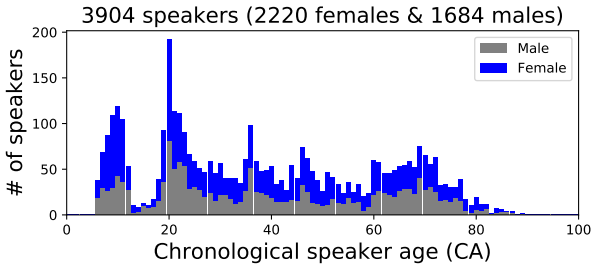


Figure 2: *Gender and age distribution of Short sentence dataset*

$\mathcal{L}(P(X_i, \Omega), y_i)$. The loss is defined as, $\mathcal{L}(P(k|X_i, \Omega), y_i) = -\sum_{k=0}^{100} \mathcal{N}(k|y_i, \sigma^2) \log P(k|X_i, \Omega)$, where σ is a standard deviation (SD) of the normal distribution.

4. Speaker age estimation experiments

4.1. Dataset

We created two different domain in-house speech datasets since the DNN model sometimes is trained under mismatched conditions between training and practical-use environments. Figure 1 and 2 show gender and age distributions of each dataset.

One is the Call center dataset comprises telephone calls spoken by 532 Japanese speakers (305 females and 227 males) from 11 to 89 years old. The speakers recorded according to five scripted scenarios at the call center. One of the scenarios is a telephone conversation between a customer and an operator. The others are utterances for an interactive voice response system. This dataset contains five single-channel calls per speaker that were recorded indoors using her/his smartphone device, and only the speakers' voices were recorded. All calls were downsampled by quantizing at 16 bits and converting a sampling frequency of 8k Hz. This dataset was randomly split into training, validation, and test subsets without speaker duplication; there are 289, 76, and 170 speakers, respectively. The average speech length after removing non-voiced duration using the voice activity detection (VAD) [40] is 35.25 seconds and its SD is 31.85.

The other is the Short sentence dataset comprises the utterances spoken by 3,904 Japanese speakers (2,220 females and 1,684 males) from 6 to 91 years old. This dataset contains 20 single-channel utterances per speaker that were recorded indoors. All utterances were downsampled by quantizing at 16 bits and converting a sampling frequency of 8k Hz. This dataset was randomly split into training, validation, and test subsets without speaker duplication; there are 3,297, 355, and 352 speakers, respectively. The average speech length after removing non-voiced duration using the VAD is 6.73 seconds and its SD is 6.61.

Table 1: *Automatic speaker age experiment performance obtained by DNN; MAE (years old)/ρ*

	Short sentence dataset		Call center dataset	
	Female	Male	Female	Male
OD	**4.87/0.94**	**5.15/0.94**	8.56/0.78	8.27/0.81
ID	4.98/**0.94**	5.31/**0.94**	**5.39/0.87**	**6.38/0.88**

Table 2: *Effect of segment length to MAE for Call center dataset*

		Test segment length (sec.)				
		5	10	20	30	full
Female	*OD*	10.00	9.37	9.70	9.44	**8.56**
	ID	6.45	5.61	5.86	5.90	**5.39**
Male	*OD*	9.13	8.47	8.89	8.91	**8.27**
	ID	7.23	6.40	6.52	6.54	**6.38**

4.2. Speaker age estimation experiment using DNNs

Method: We conducted automatic speaker age estimation experiments using the transformer model. The model architecture was determined referring to [11]. That is, the model has 13 WavLM transformer encoder layers of "WavLM Base +" [41], five ECAPA-TDNN layers [42], an attentive statistics pooling layer, two dense layers, and an output layer with a softmax function. The hyper-parameters of each layer were the same in [11]. We used RAdam optimizer [43] with settings β_1 and β_2 of 0.9 and 0.999, respectively, to optimize the model parameters. The σ of the loss was 3.6, the mini-batch size was 64, and the weight of L2 regularization was 1e$-$4. The training speech samples were randomly clipped as the segment of five continuous seconds after removing non-voiced durations using the VAD [40] at each epoch to make the mini-batches. Model training procedure was split into two phases referring to [11]; first, the pre-trained parameters of the WavLM model were frozen, and the other parameters were trained with the initial and minimum learning rates were 1e$-$4 and 1e$-$6, respectively. Then, all of the model parameters were fine-tuned with the initial and minimum learning rates were 1e$-$5 and 1e$-$7, respectively. We used the same learning rate decreasing strategy and the method of data augmentation to the training subset (adding noises and revering using simulated room impulse responses) as the same in [15] All hyper-parameters were determined by using the results of the validation subset.

We set up two kinds of training conditions; the *out-of-domain* (*OD*) and *in-domain* (*ID*) conditions. In the *OD* condition, the model training procedure used only the Short sentence dataset. It was the out-of-domain condition of the Call center dataset. In the *ID* condition, the procedure used both datasets. It was the in-domain condition of the Call center dataset.

We evaluate the experimental results using the MAE and Pearson's correlation coefficient ρ between CA and EA.

Results: The results are shown in Table 1. Both models yielded good estimation performance for the Short sentence dataset. However, the performance of the *OD* condition did not better than the *ID* condition for the Call center dataset. These performance degradations were due to the domain mismatch between both datasets. It is known that differences in the acoustic characteristics such as speaking style can degrade the performance [8, 9].

Table 2 shows the effect of test segment length to the MAEs for the Call center dataset. Some studies reported that the estimation performance is degraded in the case of the short-time input samples [7, 11, 14, 15]. Our results showed the same trends. In particular, the MAEs were more degraded when the length of test speech samples was five seconds; the MAEs were degraded more 10% than the use of full length.

Table 3: *Experimental results of PA/EA for Call center dataset*

	C-MAE / MSD / ρ	
Estimator	Female Speaker	Male Speaker
Participant #01 (20s F)	1.58 / 1.49 / **0.79**	**1.61** / 1.31 / 0.82
Participant #02 (30s F)	1.38 / 1.21 / 0.76	1.62 / **1.21** / 0.81
Participant #03 (30s F)	**1.37** / **1.12** / 0.76	1.69 / 1.19 / 0.80
Participant #04 (30s F)	1.99 / 1.60 / 0.75	1.70 / 1.45 / 0.77
Participant #05 (40s F)	1.72 / 1.35 / 0.70	1.74 / 1.52 / 0.79
Participant #06 (40s F)	1.79 / 1.47 / 0.76	1.72 / 1.38 / 0.78
Participant #07 (40s F)	1.86 / 1.59 / 0.72	1.68 / 1.30 / 0.78
Participant #08 (40s F)	1.94 / 1.46 / 0.74	1.67 / 1.30 / **0.83**
Participant #09 (40s F)	2.11 / 1.61 / 0.72	1.77 / 1.39 / 0.76
Participant #10 (40s F)	2.21 / 1.65 / 0.75	1.96 / 1.49 / 0.78
Participant #11 (50s F)	2.06 / 1.53 / 0.71	1.65 / 1.42 / 0.79
Participant #12 (50s F)	2.04 / 1.59 / 0.69	1.80 / 1.41 / 0.73
Participant #13 (50s F)	2.15 / 1.67 / 0.73	1.80 / 1.42 / 0.77
Participant #14 (50s F)	2.15 / 1.69 / 0.59	2.06 / 1.65 / 0.66
Participant #15 (50s F)	3.34 / 1.82 / 0.76	2.10 / 1.49 / **0.83**
Participant #16 (60s F)	2.14 / 1.67 / 0.72	2.51 / 2.02 / 0.75
Participant #17 (30s M)	3.43 / 1.95 / 0.61	2.17 / 1.72 / 0.65
Participant #18 (50s M)	1.89 / 1.45 / 0.69	1.90 / 1.53 / 0.72
Participant #19 (50s M)	2.38 / 1.71 / 0.66	1.86 / 1.55 / 0.73
Participant #20 (50s M)	2.30 / 1.75 / 0.75	2.19 / 1.58 / 0.80
Participant #21 (60s M)	2.03 / 1.66 / 0.68	2.35 / 2.07 / 0.67
Participants' Avg.	1.63 / 1.34 / 0.81	**1.27** / **1.04** / **0.86**
DNN (*OD* condition)	1.99 / 1.50 / 0.69	1.86 / 1.50 / 0.74
DNN (*ID* condition)	**1.30** / **1.13** / **0.84**	1.38 / 1.07 / 0.85

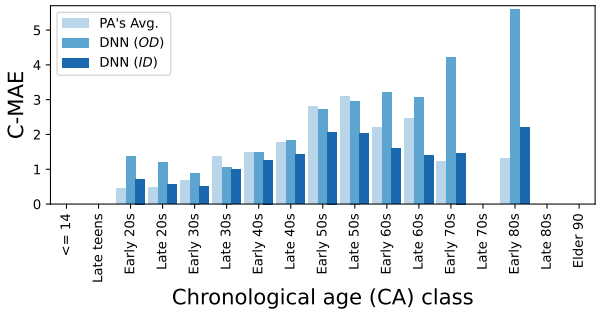
4.3. Speaker age estimation experiment by human listeners

Method: We conducted a speaker age estimation experiment with 21 Japanese listening participants who were experienced in call center operation (16 females and 5 males in their 20s–60s); the details are shown in Table 3. They listened to the calls included in the test subset of the Call center dataset using headphones and labeled the speakers' age-class and gender.

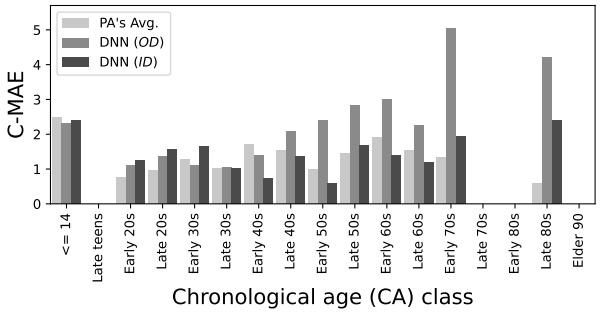
We defined 17 five-year-step age classes with age-class numbers as "1. *early teens or younger*", "2. *late teens*", "3. *early 20s*", $\cdots$, "16. *late 80s*", and "17. *90 years old or older*".

To evaluate the estimation results of these age classes, we define the "Categorical MAE (C-MAE)", which is the MAE between the age-class number of PA and CA. It is formulated as, C-MAE $= \frac{1}{N}\sum_{i=1}^{N}|c_i - \hat{c}_i$, where c_i and $\hat{c}_i \in [1, 17]$ are the age-class number of CA and PA, and N is the number of testing samples. One point in C-MAE approximates a five-year period in the MAE. We also use the ρ value between the age-class numbers of CA and PA. In addition, we evaluate the mean SD of estimated error per the speaker (MSD) as intra-estimator variability. The C-MAE per the spekaer (C-MAE$_s$) is defined as, C-MAE$_s$ $= \frac{1}{5}\sum_{j=1}^{5}|y_{s,j} - \hat{y}_{s,j}|$, where s is the speaker's ID and j is the scenario's ID. The MSD is defined as, MSD $= \frac{1}{S}\sum_{s=1}^{S}\sqrt{\frac{1}{5}\sum_{j=1}^{5}(\text{C-MAE}_s - (y_{s,j} - \hat{y}_{s,j}))^2}$, where S is the number of speakers.

Results: The results are shown in Table 3. The participants estimated the speakers' age with errors of 1.37–3.43 points in C-MAE for the female speakers and 1.62–2.51 points in C-MAE for the male speakers. The performance of PAs' average also are shown in Table 3, and there were better than each participant's performance. This is due to a phenomenon similar to ensemble estimation in machine learning. Also, the MSDs were 1.12–



(a) *Female speakers (teens or younger/late 70s/late 80s or elder are not existed)*



(b) *Male speakers (late teens/late 70s/early 80s/elder 90 are not existed)*

Figure 3: *C-MAEs per each age class for Call center dataset*

1.95 points for the female speakers and 1.30–2.07 points for the male speakers. We consider there are no outlier performance.

4.4. Comparison of experimental results by DNN model and human listeners

The EAs were converted from numerical age value to the five-year-step age class to compare with the participants' categorical age estimation performance. The C-MAEs, MSDs, and ρ values between the age-class number of EA and CA are shown in the bottom two rows of Table 3.

4.4.1. Comparison of experimental results

The DNN models showed comparable or superior performance to the participants. The DNN model trained under the *OD* condition outperformed by 57% of the participants (12/21) for the female speakers and 38% of the participants (8/21) for the male speakers. The DNN model trained under the *ID* condition outperformed by 86% of the participants (18/21) for the female speakers and all participants for the male speakers.

Figure 3-(a) and (b) show the C-MAEs for female and male speakers, respectively, on each CA-class. Each bar indicates the C-MAE of PA's average and the DNN models trained under the *ID* and *OD* conditions. The C-MAEs by the DNN models for elderly speakers were more degraded than the younger speakers as the same in [23, 24, 32]. Especially, the C-MAEs were more degraded in the case of the *OD* condition than the *ID* condition. These degradations were due to the small number of elderly speakers in the training subset and the difficulty of estimating elderly speakers' age [23].

4.4.2. Comparison of estimation characteristics between human listers and DNN model

We analyzed the differences in the estimation characteristics of the DNN model trained under the *ID* condition and PAs' average. Here, we defined large error as misestimation by more than three age classes and small error as misestimation within two age classes.

Difficult age estimation for human listeners: We found three large misestimation characteristics of the listeners' age estimation. The participants tended to (1) underestimate the late 40s to the late 60s female speakers' age as the early 20s to the late 40s; this result agrees with the findings of [27] and was confirmed in Figure 3-(a), (2) overestimate the male speakers' age as elderly if the participants captured age-related features such as hoarse voice, and (3) underestimate the late 30s to the late 40s male speakers' age as the early and late 20s if they did not capture distinct age-related features. On the other hand, 70 % of the large error samples obtained by the participants were estimated as small errors obtained by the DNN model. There is a possibility that such characteristics are particular to human listeners.

Difficult age estimation for DNN model: The DNN model tended to be sensitive to the difference in acoustic characteristics between training and testing samples. The mismatches of the acoustic characteristics between training and testing samples degrade the estimation performance as shown in Table 1 and conventional studies [8,9]. Human listeners also need to become familiar with the speaker age estimation, however, there is a possibility that they are not so sensitive to the difference in the acoustic characteristics, like the DNN model.

As discussed in Section 4.2, the MAEs obtained by the DNN model tended to be degraded on short utterances. Although the borderline length depends on the experimental setup and dataset, the estimation performance by the DNN model is degraded when the length of the test sample is 5–15 seconds or shorter. However, we received a report in our experiment that the participants could label the speakers' age and gender by listening to about the first 10 to 15 seconds of the samples. This suggests that human listeners can better estimate the PA even if the speech samples are short compared to the DNN models.

5. Conclusion

We compared speaker-age estimation performance and characteristics obtained by human listening participants and the latest DNN model. A practical use of automatic speaker age estimation is expected to be achieving high estimation performance by using DNNs. We assumed the concrete example of a call center application, in which the DNN model substituted for a call center operator to estimate a customer's age from her/his telephone call. Using our in-house simulated call center dataset, we compared the speaker age estimation results obtained by the latest DNN model and the participants who were experienced in the call center operation. Our experiments yielded the following findings; (1) the latest DNN model achieved comparable or superior estimation performance to the listeners. (2) However, compared with the listeners, the DNN model tended to be more sensitive to elderly speech, mismatches the acoustic characteristics between training and testing, and lengths of speech samples. (3) The speakers' gender and some specific age-related acoustic features badly affected the listeners' estimation performance, In future works, we intend to compare the estimation performance and characteristics between the call center operators and non-operators and analyze the full detail.

6. References

[1] A. Fedorova, O. Glembek, T. Kinnunen, and P. Matějka, "Exploring ANN back-ends for i-vector based speaker age estimation," in *INTERSPEECH*, 2015, pp. 3036–3040.

[2] J. Rownicka and S. Kacprzak, "Speaker age classification and regression using i-vectors," in *INTERSPEECH*, 2016, pp. 1402–1406.

[3] 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 *ICASSP. IEEE*, 2019, pp. 6580–6584.

[4] D. Kwasny and D. Hemmerling, "Gender and age estimation methods based on speech using deep neural networks," *Sensors*, vol. 21, no. 14, 2021.

[5] K. Hechmi, T. N. Trong, V. Hautamäki, and T. Kinnunen, "Voxceleb enrichment for age and gender recognition," in *ASRU. IEEE*, 2021, pp. 687–693.

[6] T. Gupta, D.-T. Truong, T. T. Anh, and C. E. Siong, "Estimation of speaker age and height from speech signal using bi-encoder transformer mixture model," in *INTERSPEECH*, 2022, pp. 1978–1982.

[7] 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.

[8] R. Takeda and K. Komatani, "Age estimation with speech-age model for heterogeneous speech datasets," in *INTERSPEECH*, 2021, pp. 4164–4167.

[9] N. Tawara, A. Ogawa, Y. Kitagishi, H. Kamiyama, and Y. Ijima, "Robust speech-age estimation using local maximum mean discrepancy under mismatched recording conditions," in *ASRU. IEEE*, 2021.

[10] N. Tawara, A. Ogawa, Y. Kitagishi, and H. Kamiyama, "Age-vox-celeb: multi-modal corpus for facial and speech estimation," in *ICASSP. IEEE*, 2021, pp. 6963–6967.

[11] Z. Kang, J. Wang, J. Peng, and J. Xiao, "SVLDL: Improved speaker age estimation using selective variance label distribution learning," in *SLT. IEEE*, 2022, pp. 1037–1044.

[12] Y. Kitagishi, H. Kamiyama, A. Ando, N. Tawara, T. Mori, and S. Kobashikawa, "Speaker age estimation using age-dependent insensitive loss," in *APSIPA. IEEE*, 2020, pp. 319–324.

[13] A. Saraf, G. Sivaraman, and E. Khoury, "Confidence measure for automtic age estimation from speech," in *INTERSPEECH*, 2022, pp. 2033–2037.

[14] S. Si, J. Wang, J. Peng, and J. Xiao, "Towards speaker age estimation with label distribution learning," in *ICASSP. IEEE*, 2022, pp. 4618–4622.

[15] 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 *INTERSPEECH*, 2018, pp. 277–281.

[16] 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," *Comput. Speech Lang.*, pp. 4–39, 2013.

[17] D. C. Tanner and M. E. Tanner, *Forensic Aspects of Speech Patterns: Voice Prints, Speaker Profiling, Lie and Intoxication Detection*. Lawers & Judges Publishing, 2004.

[18] T. Pellegrini, V. Hedayati, I. Trancoso, A. Hämäläinen, and M. S. Dias, "Speaker age estimation for eldery speech recognition in european portuguese," in *INTERSPEECH*, 2014, pp. 2962–2966.

[19] A. Braun and L. Cerrato, "Estimating speaker age across languages," in *ICPhS*, 1999, pp. 1369–1372.

[20] R. M. Krauss, R. Freyberg, and E. Morsella, "Inferring speakers' physical attributes from their voices," *Journal of Experimental Social Psychology*, vol. 38, no. 6, pp. 618–625, 2002.

[21] K. Amilon, J. van de Weijer, and S. Schötz, *The Impact of Visual and Auditory Cues in Age Estimation*. Springer Berlin Heidelberg, 2007, pp. 10–21.

[22] E. Moyse, A. Beaufort, and S. Brédart, "Evidence for an own-age bias in age estimation from voices in older persons," *European Journal of Ageing*, vol. 11, no. 3, pp. 214–247, 2014.

[23] M. Huckvale and A. Webb, "A comparison of human and machine estimation of speaker age," in *SLSP*, 2015, pp. 111–122.

[24] S. S. 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 Psycholagy*, vol. 6, p. 978, 2015.

[25] T. Shipp and H. Hollien, "Perception of the aging male voice," *Journal of Speeech and Hearing Research*, vol. 12, no. 4, pp. 703–710, 1969.

[26] R. Huntley, H. Hollien, and T. Shipp, "Influences of listener characteristics on perceived age estimations," *Journal of Voice*, vol. 1, no. 1, pp. 49–52, 1987.

[27] R. D. Jacques and M. P. Rastatter, "Recognition of speaker age from selected acoustic features as perceived by normal young and older listeners," *Folia Phoniatr (Basel)*, vol. 42, no. 3, pp. 118–124.

[28] M. Bruckl and W. Sendlmeier, "Aging female voices: an acoustic and perceptive analysis," in *VOQUAL*, 2003, pp. 163–168.

[29] N. Minematsu, K. Yamauchi, and K. Hirose, "Automatic estimation of perceptual age using speaker modeling techniques," in *EUROSPEECH*, 2003, pp. 3005–3008.

[30] S. Schötz, *Perception, Analysis and Synthesis of Speaker Age*. Linguistics and Phonetics, 2006.

[31] H. Kasuya, H. Yoshida, S. Ebihara, and H. Mori, "Longitudinal changes of selected voice source parameters," in *INTERSPEECH*, 2010, pp. 2570–2573.

[32] H. Goy, M. K. Pichora-Fuller, and P. van Lieshout, "Effects of age on speech and voice quality ratings," *The Journal of the Acoustical Society of America*, vol. 139, no. 4, pp. 1648–1659, 2016.

[33] J. D. Harnsberger, R. Shrivastav, and W. S. B. Jr., "Modelling perceived vocal age in American English," in *INTERSPEECH*, 2010, pp. 466–469.

[34] M. Pettorino and A. Giannini, "The speaker's age: A perceptual study," in *International Congress of Phonetic Sciences*, 2011, pp. 1582–1585.

[35] C. L. Lortie, M. Thibeault, M. J. Guitton, and P. Tremblay, "Effects of age on the amplitude, frequency and perceived quality of voice," *Journal of the American Aging Association*, vol. 37, no. 6, p. 24 pages, 2015.

[36] R. G. Hautamäki, A. Kanervisto, V. Hautamäki, and T. Kinnunen, "Perceptual evaluation of the effectiveness of voice disguise by age modification," in *Odyssey*, 2018, pp. 320–326.

[37] D. E. Hartman, "The perceptual identity and characteristics of aging in normal male adult speakers," *Journal of Communication Disorders*, vol. 12, no. 1, pp. 53–61, 1979.

[38] J. Kreiman and G. Papçun, "Voice discrimation by two listener populations," *The Journal of the Acoustical Society of America*, no. S1, p. S9, 1985.

[39] S. E. Linville, "Acoustic-perceptual studies of aging voice in women," *Journal of Voice*, vol. 1, no. 1, pp. 44–48, 1987.

[40] Z.-H. Tan, A. kr. Aarkar, and N. Dehak, "rVAD: An unsupervised segment-based robust voice activety detection method," *Computer Speech and Language*, vol. 59, pp. 1–21, 2020.

[41] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[42] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *INTERSPEECH*, 2020, pp. 3380–3384.

[43] L. Liu, H. Jiang, P. He, C. Weizhu, X. Liu, J. Gao, and J. han, "On the variance of the adaptive learning rate and beyond," in *ICLR*, 2020, 13 pages.