



# Confidence Measure for Automatic Age Estimation From Speech

Amruta Saraf, Ganesh Sivaraman, Elie Khoury

Pindrop, Atlanta, USA

{asaraf,gsivaraman,ekhoury}@pindrop.com

## Abstract

Age estimation from speech is a problem that has wide applications in call centers, virtual assistants and IoT devices. The estimated age is used for various system decisions related to personalization, parental control, and anomaly detection. Performance of speech based automatic age estimation systems is generally stated in the literature using the Mean Absolute Error (MAE) and Pearson Correlation Coefficient (PC). In real-world applications of these systems, MAE and PC provide little insight into the confidence of a point estimate. An MAE of 5 years on a test set provides only the average error across all estimates in the test set and hence does not provide any information about the confidence of each individual estimate. A confidence score of predicted age is essential to know the trustworthiness of the predictions made by the system. This paper formulates age estimation from speech as a label distribution learning problem to come up with a measure of the confidence related to the point estimate being within a desired range from the ground-truth. It further uses it to analyze the age estimation system under conditions of varying speech quality. We show that the proposed measure of confidence is better than a fixed error-margin.

**Index Terms:** age estimation, confidence measure, human-computer interaction, computational paralinguistics

## 1. Introduction

Automatic speaker age prediction (ASAP) enables a wide variety of applications in the call center, IoT devices, age based personalization of ads and services, and parental control. This task was sporadically discussed in the literature since the late 1950s [1], but it got much more attention after the age sub-challenge in the Interspeech 2010 paralinguistic challenge [2].

Research on age estimation from speech has benefited tremendously from the progress in the field of automatic speaker recognition. Therefore, different variants of speaker recognition front-ends have been studied in ASAP systems, such as Gaussian mixture model (GMM) [3] super-vectors used in [4], i-vectors [5] used in [6], or more recently x-vectors variants [7, 8] used in [9, 10]. In our work, we also use x-vector as a front-end extractor. While it is possible to fully train the x-vector for the task of age estimation like in [9], or use high-dimensional statistics pooling outputs of a front-end pre-trained for speaker recognition task, followed by training a back-end specific for age estimation like in [10], we decide to favor increased efficiency and portability, by directly consuming low-dimensional speaker embeddings of a pre-trained x-vector system and solely focusing on the age estimation back-end.

Historically, ASAP methods can be grouped into two main categories, regression based and classification based. Regression based methods aim to predict the exact age of the speaker by penalizing the difference between the true age and the estimated age during the training using simple losses like mean-square error (MSE) [9], root mean squared error (RMSE) [11],

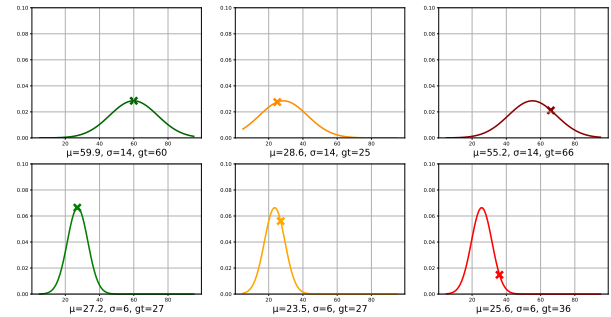


Figure 1: Depiction of the output distribution as a reflection of the input label distribution. In each sub-figure,  $\mu$  is the age estimate and the mean of the distribution,  $\sigma$  is the square-root of variance predicted by the system and  $gt$  is the ground-truth age label. The first and second rows depict examples where the variance is high and low, respectively. The first, second and third columns depict examples where the age estimate is most accurate, satisfactory and not very useful, respectively.

mean-absolute error (MAE) [12], and hinge loss used in Support Vector Regression (SVR) [6]. Classification based methods were also used in existing ASAP systems, either to determine the age group of the subjects [2], or to estimate their exact age [9, 10]. It is worth noting that both [9, 10] also investigate the fusion of classification and regression approaches. One drawback of both regression and classification approaches is that they do not incorporate well the scarcity of the training data. Another drawback is that they do not account for the uncertainty of age estimation during either training or testing.

Therefore, in a recent work [12], we propose the use of a ranking based approach for the first time for age estimation from speech. This approach coined label distributed learning was borrowed from facial age estimation [13], but it was adapted to ASAP [12]. Our paper builds upon this previous work, but we propose to use a simple mean loss instead of previous losses used for label distribution learning, namely Kullback-Leibler divergence (KLD) [14], Generalized Jeffries-Matusita (GJM) distance [15] and mean and variance loss (MVL) [16]. The proposed loss does not control the variance, neither by fixing its value as in KLD and GJM training, nor by reducing its loss as in MVL training. The main advantages of this loss is that it empirically outperforms previous losses, and it enables the variance to be more naturally used as a measure of confidence during inference. We argue that the estimation of the confidence margin or interval during inference is an important measure to operationalize speaker age estimation in real-world applications. Figure 1 shows several types of constellations of the estimated age or  $\mu$  and the estimated confidence margin or  $\sigma$ .

ground-truth

In some use cases that look at age ranges, like personalized ads, it might be better off ignoring the predicted age if the confidence margin is very large, instead of recommending ir-

relevant ads to the wrong age range. We hypothesize that the confidence margin may vary from one sample to another due to the health conditions of the speaker (such as smoking), age range, and quality of audio signal (e.g. duration of net speech, and noise conditions).

While MAE is an important metric to quantify the overall performance of the system, it is just an average error and does not guarantee that any predicted age would fall within the MAE error bounds. For instance if the MAE of the system is 5 years, this does not imply that any given sample will have a predicted age that falls within  $\pm 5$  years of the ground-truth age.

Therefore, we define a new accuracy metric that accounts for both the point estimate as well as the confidence margin. Using this new metric, we empirically show that the variable confidence margin estimated by the network provides a higher accuracy than all fixed confidence margins that are below its average value.

The rest of the paper is organized as follows: Section 2 talks about the dataset used in the study. Section 3 describes the metrics prevalent in literature for the description of age estimation systems. It also motivates and defines the accuracy metric and the confidence measure associated with the age estimate from a system. In Section 4, the features, training algorithms and inference mechanisms are described. Section 5 shows the results of the inference trends of the metrics, while Section 6 draws conclusions from these trends.

## 2. Datasets

Two datasets are used in our analysis. The first one is a subset of the NIST SRE data that has age labels. All utterances from the SRE 2008 dataset [17] with age labels are included in the train set, while all the utterances from SRE 2010 [18] are included in the test set. The train set is augmented with four degraded copies using music, babble, noise and reverberation. The background noise files for augmentation are from the MUSAN noise corpus [19]. The room impulse responses (RIRs) for reverberation are randomly selected from the publicity available RIR dataset [20, 21, 22]. After augmentation, we get 53,190 utterances in the train set spanning 1,224 speakers. The test set consists of 6,953 from 445 speakers. There is no overlap of speakers between the train and test set. The age range in the SRE data is from 18 to 91 years. The utterances in the SRE data span 2 seconds to 232 seconds of estimated net speech duration.

The second dataset is the AgeVoxCeleb dataset [23] which has age labels assigned to utterances from Voxceleb2 [24] that was downsampled to 8kHz. This dataset includes 151,890 utterances from 4,471 speakers in the train set, and 16,050 utterances from 497 speakers in the test set. There is no overlap of speakers between the train and test set. The age range is from 5 to 95 years. The utterances are shorter than NIST SRE and have a maximum of 94 seconds of estimated net speech duration.

To evaluate the age estimation across different conditions of audio quality, we prepare variations of the test sets with varying net speech and SNRs. We add noise (randomly selected from the MUSAN noise corpus [19]) to the audio in a controlled fashion, limiting the SNR to a desired level. We prepare five sets of these noise infused test utterances, at SNR values of 5dB, 10dB, 15dB, 20dB, and 25dB. We also prepare variations of the test sets having nine different levels of net speech ranging from 2s to 30s. The net speech variations are simulated by simply trimming the audio files when the required level of net-speech is reached.

## 3. Metrics

The baseline metrics for the performance of ASAP systems are the MAE and PC as they are a standard comparison provided in much of the related literature. However, there is a limitation in using this metric. Consider the case where an age estimation system has an MAE of 5 years. This is a bulk estimate, applying to the test data as a whole. It means that if you use this system in an application that sees data similar in volume and age distribution to the test dataset, the mean of the absolute error between the ground-truth and estimated age would be 5 years. If the speaker age of a particular speech utterance is estimated to be, say 45 years, there is no knowing how far from the ground-truth it is.

To overcome this limitation, we propose a new measure of accuracy at a given margin. This accuracy is defined as the percentage of test utterances for which the absolute error between the ground-truth and age estimate of that utterances is less than that margin:

$$acc@margin = \frac{N_{correct}^{margin}}{N_{total}^{margin}} \quad (1)$$

where  $N_{correct}^{margin}$  is the number of utterances such that  $|y_i - \hat{y}_i| \leq margin$  (where  $y_i$  is the ground-truth age and  $\hat{y}_i$  is the predicted age), and  $N_{total}^{margin}$  is the total number of utterances for this given margin. A higher margin allows more instances to be deemed as correctly classified, implying higher  $acc@margin$ . Naturally, as margin increases, it should follow that the  $acc@margin$  would show a monotonic increase. Therefore, when comparing two systems, the system with a higher  $acc@margin$  for the same margin is a better system. The  $acc@margin$  also serves as a measure of confidence. For instance, if the system predicts for an utterance a speaker age of 45 and a margin of 5 years, and we know that the  $acc@5years$  is 90% on a representative test data, then we can trust the system with 90% confidence that the true age lies between 40 and 50 years.

## 4. Automatic Speaker Age Prediction

The proposed ASAP system has two main components: a front-end speaker embedding extractor and a back-end age prediction.

The speaker embedding is a standard CNN-based x-vector. Details of this system can be found in [12]. This front end is trained on a combination of Switchboard, NIST SRE 2004-2012 and Voxceleb, comprising a total of 348k audio files from about 12.8k speakers. The training of the x-vector consists of two phases: The first phase uses categorical cross-entropy loss on chunks of 2 seconds of speech. The second phase consist of freezing all convolutional layers and train a 256-dimensional embedding layers using large margin cosine loss (LMCL) [25] on full-length utterances.

The back-end age prediction consumes directly speaker embeddings. It consists of a shallow neural network that uses a dense layer of 256 units with Relu activation, followed by a Softmax layer where the number of units is equal to the age range. The outputs are treated as posterior probabilities of age labels. In contrast to previous work where Kullback-Leibler Divergence (KLD) [14], Generalized Jeffries-Matusita (GJM) distance [15] and Mean and Variance Loss (MVL) [16], this work uses Mean Loss (ML), which tries to minimize the difference in the means of the output Softmax distribution and the label distribution without explicitly requiring to know the distribution of the ground-truth. Therefore, there is no need to set a fixed sigma

Train	Test	MSE	MAE	CCE	KLD	GJM	MVL	ML
SRE08/10	SRE08/10	7.0/0.80	7.2/0.79	7.8/0.72	7.4/0.8	7.4/0.83	7.3/0.77	<b>5.8/0.84</b>
SRE08/10	AgeVoxCeleb	<b>10.4/0.42</b>	<b>10.4/0.42</b>	14.1/0.29	11.4/0.41	11.7/ <b>0.43</b>	11.7/0.37	10.8/0.42
AgeVoxCeleb	AgeVoxCeleb	8.5/0.65	8.5/0.65	9.7/0.58	8.4/0.65	9.9/0.62	8.3/0.66	<b>8.1/0.67</b>
AgeVoxCeleb	SRE08/10	8.2/0.72	7.9/0.73	<b>7.5/0.70</b>	7.7/0.74	13.3/0.74	<b>8.2/0.78</b>	<b>7.5/0.77</b>

Table 1: Performance (MAE in years/PC) of age estimation systems employing different losses. Loss names and abbreviations are given in Section 4.

such as in KLD and GJM training. The Mean Loss is defined as follows:

$$L_m = \frac{1}{2N} \sum_{i=1}^N \left( \sum_{j=1}^K j * p_{i,j} - y_i \right)^2 \quad (2)$$

where  $N$  is the total number of utterances,  $K$  is the maximum age in the dataset,  $p_{i,j}$  is the posterior probability of the  $i^{th}$  utterance for the  $j^{th}$  age bin, and  $y_i$  is the ground-truth age.

We compare the performance of all the systems in Table 1. The other losses are Mean-Squared Error (MSE), Mean Absolute Error (MAE) and Categorical Cross-Entropy (CCE). We observe that the performance of the model trained using Mean Loss performs well under different matched and mismatched conditions. We use only the Mean Loss trained models for further experiments in the next section.

Finally, since the variance was not constrained during training, we found it to be suitable as a measure of the confidence margin or interval during inference. Confidence margin is an important measure to operationalize speaker age estimation in real-world applications. If the system estimates a confidence margin that is very small, this indirectly indicates that the system has high confidence in estimating the age, and thus can be used to enable some personalization or security features in the product. Contrarily, if the system estimates a large confidence margin, it might be better off disabling those extra features.

We assume that different speaker and audio aspects could impact the confidence margin. While the speaker-related information such as the health of the speaker are difficult to find in existing datasets, the audio quality aspects can be simulated, mainly varying the duration of net speech and signal-to-noise ratio. This will be studied in the experimental section.

## 5. Experimental Results

In the previous section, we explore the possibility to estimate both the age and its associated confidence margin, using the Mean loss that provides low MAEs and good Pearson Coefficient scores as shown in Table 1. All remaining experiments use Mean loss and the focus is shifted toward computing the Accuracy as described in Section 3.

Before the use of the sample-dependent estimated margin, the user of the ASAP system was intuitively selecting a fixed confidence margin for all predicted ages, e.g.  $X + / - 5$  or  $X + / - 10$ . A more reasonable selection of the fixed margin is based on the MAE value computed on the test or development data. During the selection, there is trade-off assessment that is considered. The higher the confidence margin, the higher the accuracy defined in Eq. 1, however, the prediction is less desirable. A prediction with 30 years margin, i.e.  $X + / - 30$  will result in very high accuracy but it is not practical. Therefore, when comparing between systems, the higher the accuracy at the same given margin, the better the system.

Our proposed ASAP system allows to estimate both the age and the margin for every given speech utterance. The objective of the first experiment is to assess if the estimated margin is any better than the fixed margin.

SNR (dB)	MAE (yrs)	PC	Acc@estimated-margin (%)	Mean margin (yrs)	Acc@mean-margin (%)
<b>5</b>	8.7	0.62	80.6	13.8	78.7
<b>10</b>	8.4	0.65	81.9	13.8	80.3
<b>15</b>	8.2	0.67	82.9	13.8	81.6
<b>20</b>	8.1	0.68	83.3	13.8	81.8
<b>25</b>	8.1	0.68	83.1	13.8	81.8

Table 2: Performance of AgeVoxCeleb data with respect to SNR. All metrics improve with higher SNR, but acc@estimated-margin always outperforms acc@mean-margin.

### 5.1. Estimated Margin versus Fixed Margin

In this experiment, we first compute the accuracy of the whole test set as the number of age estimates that have the ground-truth falling within the associated margin bounds. This accuracy is coined **acc@estimated-margin**. We then compute the mean of the estimated margins over the test set and call it **mean\_margin**. We also compute the group-wise accuracies of the different margin values (they are acc@margin according to Equation 1). The significance of this is that, if we have an utterance with an estimated age of 40 years and an associated margin of 8 years, there is a 90% confidence (Figure 2) that the true age is between 32 and 48. As another example, if an age estimate is 50 years and the associated margin is 5 years, then, there is an 84.5% confidence that the true age is between 45 and 55 years.

The merit of using the network predicted margin is also shown in Figure 2. The mean-margin is 9.28. If we compute the acc@margin for margin fixed at mean\_margin, we call it **acc@mean-margin**. This acc@mean-margin = 81.5%, which is lower than acc@estimated-margin, which is 83.6% for our data. Moreover, the acc@margin for the estimated margin case is greater than all acc@margin for fixed margin values less than mean-margin. Thus, the network’s margin estimate provides more confidence in the estimate than using a fixed margin having an equivalent overall value.

### 5.2. Impact of Audio Quality

In an effort to understand the impact of audio quality on the performance of age estimation systems, we performed the following experiments:

#### 5.2.1. Signal-to-Noise Ratio

We extended the AgeVoxCeleb test data to include the quality measure of SNR, wherein we add noise to the audio in a controlled fashion, limiting the SNR to [5, 10, 15, 20, 25] dB. Table 2 shows the impact of SNR on AgeVoxCeleb performance. A zoomed in look at the differences in the accuracies at margins, that were either fixed or derived from the network, is seen in Figure 3. We observe that as expected, all the metrics, including MAE, PC, acc@estimated-margin, acc@mean-margin improve with higher SNR (better audio quality). However, it is a pronounced effect that at all SNR levels, the acc@estimated-margin is higher than acc@mean-margin. We also did the same for SRE data and found that the same trend applies.

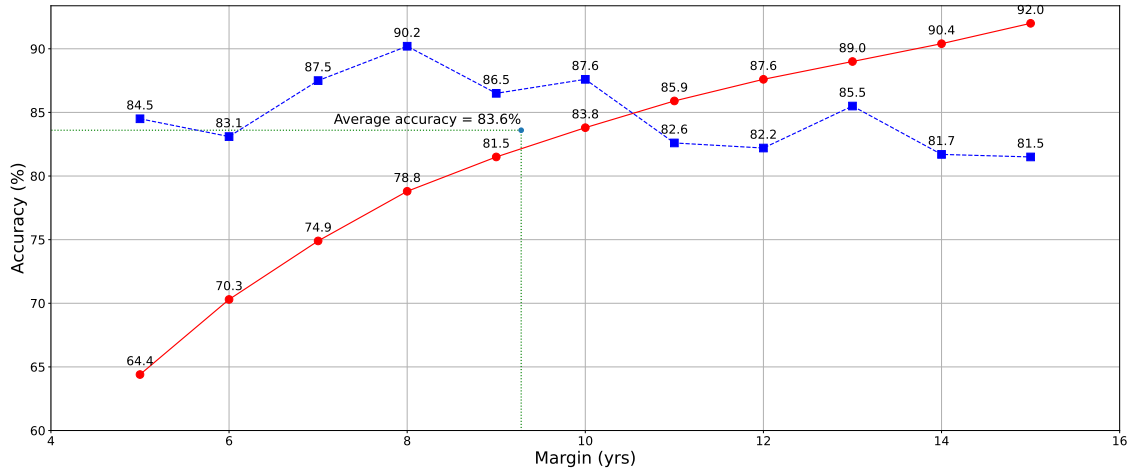


Figure 2: SRE dataset: Accuracies as a function of the margin:  $acc@margin$  for estimated margins (blue),  $acc@margin$  for fixed margins (red) and the  $acc@mean-margin$  (green). At the point of mean-margin,  $acc@estimated-margin$  is better than  $acc@mean-margin$ . For margins lower than that, using estimated margins gives better performance than using fixed margins.

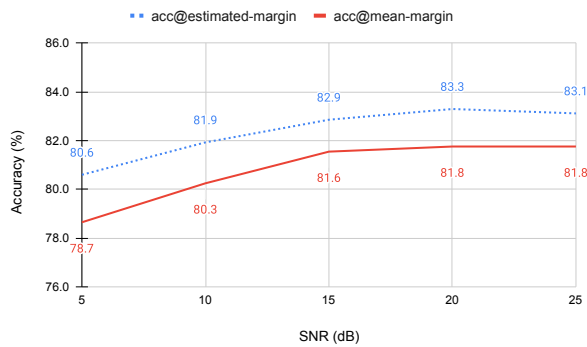


Figure 3: Accuracy for AgeVoxCeleb as a function of Signal-to-Noise Ratio. Using estimated margins makes  $acc@margin$  more robust to noise level than using fixed margins.

### 5.2.2. Duration of Net Speech

The SRE dataset, has longer utterances, so we perform this analysis only on SRE data. We extract speaker embeddings from the first T seconds of net speech for T in [2, 4, 6, 8, 10, 15, 20, 25, 30]. Table 3 shows how the performance varies when the net speech restriction of the utterance is varied. A zoomed in look at the differences in the accuracies is seen in figure 4. When there is less net speech available, the speaker recognition abilities are generally poorer. We observe that this holds true even for age estimation metrics including MAE and PC. Of significance, is the observation that for various durations of net speech up to 25 seconds, the  $acc@estimated-margin$  is higher than  $acc@mean-margin$ , and subsequently levels.

## 6. Conclusions

The  $acc@margin$  provides a better confidence about a point estimate of a speech based age estimation system than system-wide metrics like MAE and PC. Further, the accuracy of the point estimate improves if the confidence margin used is estimated by

ns (s)	MAE (yrs)	PC	Acc@estimated-margin (%)	Mean margin (yrs)	Acc@mean-margin (%)
2	8.7	0.62	73.7	11.7	68.8
4	7.4	0.72	76.9	10.7	72.7
6	6.9	0.76	77.6	10.1	75.9
8	6.6	0.79	78.0	9.7	75.4
10	6.4	0.81	77.6	9.5	76.7
15	6.1	0.82	77.6	9.0	75.6
20	6.0	0.83	77.4	8.7	76.3
25	5.9	0.84	76.9	8.6	76.7
30	5.8	0.84	77.1	8.5	77.3

Table 3: Performance of SRE data with respect to net speech. MAE, PC and the mean-margin improve with increasing net speech.  $acc@estimated-margin$  outperforms  $acc@mean-margin$  at least until 25 seconds of net speech.

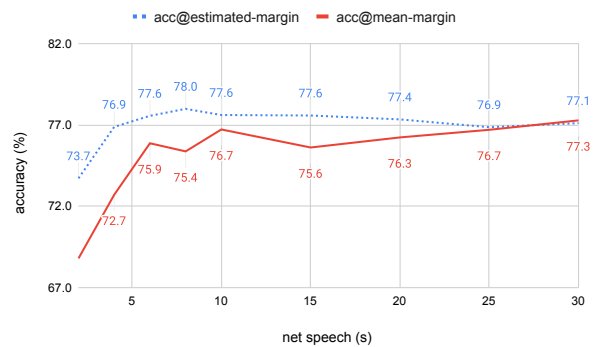


Figure 4: Accuracy for SRE as a function of net speech duration. Using estimated margins makes the  $acc@margin$  more robust to shorter utterances than using fixed margins.

the network itself, rather than by using an equivalent fixed margin (=mean-margin). While the accuracy of the ASAP system under low SNR and low net speech decreases, the accuracy using estimated margin is always higher than the accuracy computed at an equivalent fixed margin. Future work will focus on lowering of confidence margin estimated by the network, while maintaining low MAE, high PC and high  $acc@margin$ .

## 7. References

- [1] E. D. Mysak, "Pitch and duration characteristics of older males," *Journal of Speech and Hearing Research*, vol. 2, no. 1, pp. 46–54, 1959.
- [2] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt *et al.*, "The INTERSPEECH 2010 paralinguistic challenge," in *INTERSPEECH*, 2010.
- [3] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.
- [4] T. Bocklet, A. Maier, J. G. Bauer, F. Burkhardt *et al.*, "Age and gender recognition for telephone applications based on GMM supervectors and support vector machines," in *ICASSP*. IEEE, 2008, pp. 1605–1608.
- [5] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel *et al.*, "Front-End Factor Analysis for Speaker Verification," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, pp. 788–798, 2011.
- [6] M. H. Bahari, M. McLaren, H. Van hamme, and D. Van Leeuwen, "Age Estimation from Telephone Speech using i-vectors," in *INTERSPEECH*, 2012.
- [7] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey *et al.*, "X-vectors: Robust DNN embeddings for speaker recognition," in *ICASSP*. IEEE, 2018, pp. 5329–5333.
- [8] J. Villalba, N. Chen, D. Snyder, D. Garcia-Romero *et al.*, "State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and Speakers in the Wild evaluations," *Computer Speech & Language*, vol. 60, p. 101026, 2020.
- [9] P. Ghahremani, P. S. Nidadavolu, N. Chen, J. Villalba *et al.*, "End-to-end Deep Neural Network Age Estimation," in *INTERSPEECH*, 2018.
- [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] D. Kwasny and D. Hemmerling, "Joint gender and age estimation based on speech signals using x-vectors and transfer learning," *arXiv preprint arXiv:2012.01551*, 2020.
- [12] A. Saraf and E. Khoury, "Distribution learning for age estimation from speech," in *ICASSP*. IEEE, 2022, pp. 8552–8556.
- [13] B.-B. Gao, H.-Y. Zhou, J. Wu, and X. Geng, "Age Estimation Using Expectation of Label Distribution Learning," in *IJCAI*, 2018, pp. 712–718.
- [14] B.-B. Gao, C. Xing, C.-W. Xie, J. Wu *et al.*, "Deep label distribution learning with label ambiguity," *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 2825–2838, 2017.
- [15] A. Akbari, M. Awais, M. Bashir, and J. Kittler, "How Does Loss Function Affect Generalization Performance of Deep Learning? Application to Human Age Estimation," in *ICML*. PMLR, 2021.
- [16] H. Pan, H. Han, S. Shan, and X. Chen, "Mean-variance loss for deep age estimation from a face," in *IEEE CVPR*, 2018.
- [17] A. F. Martin and C. S. Greenberg, "NIST 2008 speaker recognition evaluation: Performance across telephone and room microphone channels," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2009, pp. 2579–2582.
- [18] —, "The NIST 2010 speaker recognition evaluation," in *Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010*, 2010, pp. 2726–2729.
- [19] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," oct 2015. [Online]. Available: <http://arxiv.org/abs/1510.08484>
- [20] S. Nakamura, K. Hiyane, F. Asano, and T. Endo, "Sound scene data collection in real acoustical environments," *Journal of the Acoustical Society of Japan*, vol. 20, no. 3, pp. 225–231, 1999.
- [21] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *2009 16th International Conference on Digital Signal Processing*, 2009, pp. 1–5.
- [22] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5220–5224.
- [23] N. Tawara, A. Ogawa, Y. Kitagishi, and H. Kamiyama, "Age-vox-celeb: Multi-modal corpus for facial and speech estimation," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6963–6967.
- [24] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Proc. Interspeech 2018*, 2018, pp. 1086–1090.
- [25] H. Wang, Y. Wang, Z. Zhou, X. Ji *et al.*, "Cosface: Large margin cosine loss for deep face recognition," in *IEEE CVPR*, 2018.