DISTRIBUTION LEARNING FOR AGE ESTIMATION FROM SPEECH

Amruta Saraf, Elie Khoury

Pindrop, Atlanta, USA

ABSTRACT

Age estimation from speech is becoming important with increasing usage of the voice channel. Call centers can use age estimates to influence call routing or to provide security by comparison with the speaker's age on-file. Voice assistants can use it for parental control applications. The problem of age estimation from speech has been often viewed as a regression or classification problem. However these methods do not explicitly incorporate ordinal ranking or uncertainty in age estimation that humans often do. In this work, we hypothesize that the age follows a normal distribution centered around the real age with a particular confidence interval. We investigate three different distribution learning losses, namely KL divergence, GJM distance and mean-and-variance loss. Cross-dataset experiments were conducted on the NIST SRE08/10 and AgeVoxCeleb data, and their results show that the distribution learning methods are very competitive and in most cases better than traditional approaches.

Index Terms— age estimation, x-vectors, distribution learning, regression, classification

1. INTRODUCTION

Voice is becoming a predominant channel for executing important actions such as transactions over the telephone in call centers. Short commands over personal voice activated devices can help one to place orders or do other transactions over the voice channel. This is especially useful in a hands-free environment like cars or in the living room or kitchen while engaged in another activity. With such proliferation of voice everywhere, automatic speaker recognition has become important for the security and privacy of the speakers and it has been actively researched in the last three decades. Along with speaker identity, other peripheral information about the speaker, including the gender, age, language, accent, dialect, emotional state, health and many others are also garnering interest in the research community.

Although the aspect of determining age from speech has been discussed since the late 1950s [1], the automatic age estimation research got more eyes on it in the 2010s, especially with the age sub-challenge in the Interspeech 2010 paralinguistic challenge [2]. Work on automatic speech-based age estimation has ranged from researching age-informative features such as pitch and speech rate [3] to exploring fully end-to-end systems [4], including the use of front-end speaker representations such as Gaussian mixture model (GMM) [5] super-vectors used in [6], i-vectors [7] used in [8], or more recently x-vectors variants [9, 10] used in [4, 11]. In our work, we also use x-vector as a front-end extractor, and we focus on exploring different back-end methods for age estimation.

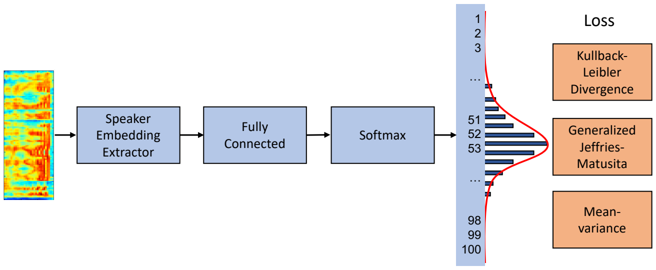

Speaker Embedding Extractor
Fully Connected
Softmax
1
2
3
...
51
52
53
...
98
99
100
Loss
Kullback-Leibler Divergence
Generalized Jeffries-Matusita
Mean-variance

Fig. 1. Proposed age estimation: The front-end extracts speaker embeddings from speech utterances. The back-end is trained using Gaussian-based distribution learning losses.

In the parallel field of face recognition, age estimation from face images has been well researched for more than two decades [12], and accuracy has been drastically improving in recent years to achieve less than three years in mean absolute error (MAE) between estimated age and apparent age [13, 14]. This success is mainly attributed to the use of deep architecture such as ConvNets [15] or VGGs [16], but also to the exploration of a wide variety of methods to solve the problem of age estimation. Those methods could be grouped into three categories: regression based, classification based, ranking based methods.

Regression based methods have been widely used by both computer-vision and speech communities [8, 17, 18]. The goal of these methods is to predict the exact age of the subject (face or voice) by penalizing the difference betweeen the true age and the estimated age during the training using simple losses like mean-square error (MSE) [18], hinge loss used in SVR [19] or more sophisticated ones like locally adjusted regression [17]. Regression based methods are the most intuitive among the three categories, however they do not incorporate well both the scarcity of the training data and the uncertainty in age estimation that is dependent on the subject, age range, and the quality of the image or audio signal.

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

Classification based methods were also explored by both research communities for the task of age estimation, traditionally to determine the age group of the subjects [20, 2], but more recently to estimate their exact age [4, 11]. While results in those recent studies [4, 11] are encouraging, classification approaches ultimately suffer from ignoring the rank ordering and the correlation between neighboring ages during training. It is worth noting that both [4, 11] investigate the fusion of classification and regression approaches.

Ranking based methods were recently introduced for facial age estimation [21, 22, 23, 14]. In contrast to regression methods, these methods explicitly make use of the ordinal relationship between different ages at training stage. These methods, particularly the ones that explore label distribution learning [22, 23, 14], were found very successful especially in the scenario where each face image is labeled by multiple individuals to estimate the apparent age.

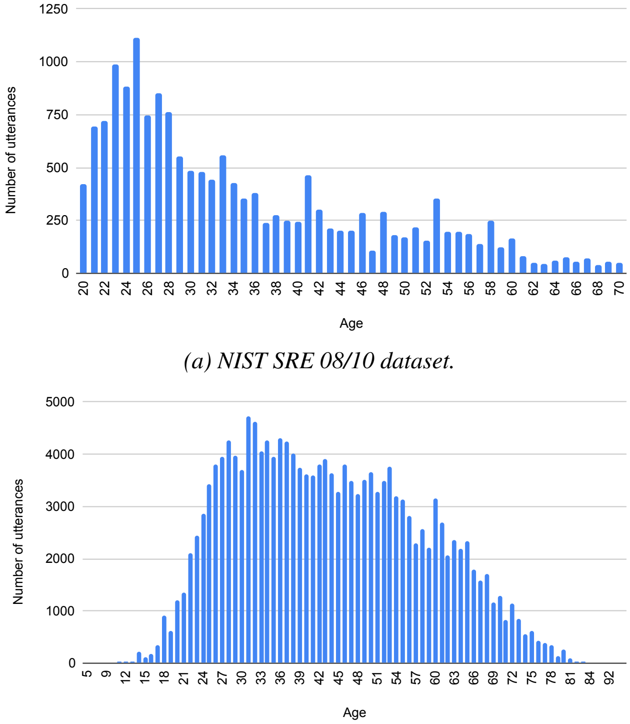
In this paper we introduce distribution learning for speech age estimation. While speech research lacks datasets with labeled "apparent" age, the promises of distribution learning that were validated for facial age estimation still hold for speech, that is, when humans estimate someone's age (from image or speech), it is relatively easy for them to give an age estimate with a particular confidence interval. This could translate into a normal distribution centered around the estimated age with a particular standard deviation. This assumption of Gaussian distribution led us to investigate three distribution learning losses: Kullback-Leibler divergence (KLD) [13], Generalized Jeffries-Matusita (GJM) distance [14] and mean and variance loss [23]. The proposed age estimation system is illustrated in Figure 1.

The remainder of the paper is organized as follows: Section 2 details the datasets and metrics used in our study. Section 3 describes the proposed age estimation system. Section 4 presents the different losses including KLD and GJM. Section 5 shows the experimental results of the difference losses including cross-dataset conditions. Section 6 concludes the paper.

2. DATASETS AND METRICS

In this set of experiments, we have used two different datasets. The first one is the subset of NIST SRE[1] 2008 [24] and 2010 [25] datasets having age labels. This is an 8 kHz audio corpus. To have a comparison against state-of-the-art, we tried to emulate the benchmark conditions as given in [26]. This is why, we restricted to only those utterances between 20 and 70 years. With that, we get a total of 16,830 utterances and those span 1,588 speakers. We also replicated the 15 folds, no speaker overlap conditions and recorded average metrics over the 15 folds as the SRE performance. We augmented the training data of each fold by four noise types including noise, babble, music and reverb. We restricted the test data in each fold to have more than 12s of net speech. The age distribution for the dataset is in Figure 2(a).

The second dataset is the AgeVoxCeleb dataset which is the contribution of the work by Naohiro *et al.* [11]. It is a subset of the VoxCeleb2 dataset, but with annotated age labels. It is a 16 kHz audio corpus, so we downsampled it to 8 kHz to compare with the first dataset. The age range is from 5 to 95 years. It has 167,940 utterances and 4,968 speakers. Again to emulate the same benchmark for comparison, we used the train and test specified in the dataset. It spans ages from 5 years to 95 years. The age distribution is shown in Figure 2(b).



(a) NIST SRE 08/10 dataset.

(b) AgeVoxCeleb dataset.

Fig. 2. Age distribution of utterances in the datasets.

The performance of the age estimation system is measured using both the Mean Absolute Error (MAE) and the Pearson Coefficient.

3. AGE ESTIMATION SYSTEM

The age estimation system has two main components: the front-end speaker embedding extractor and the back-end age prediction.

[1] https://www.nist.gov/itl/iad/mig/speaker-recognition

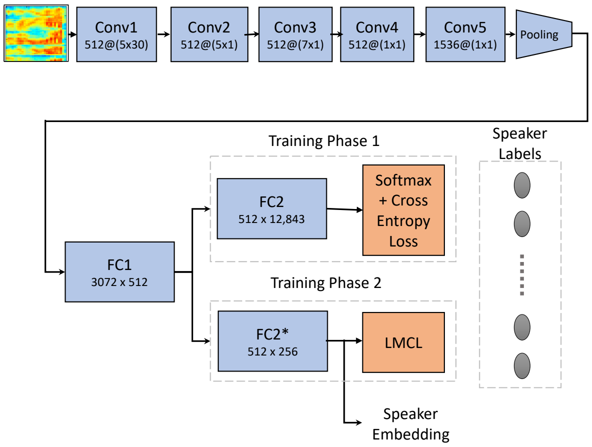


Fig. 3. Block diagram of the Speaker embedding network.

3.1. Front-End Speaker Embedding

In this work, we designed a fairly simple CNN-based x-vector. After processing the audio utterances through an energy-based voice activity detection, 24-dimensional Mel frequency cepstrum coefficients (MFCC) are computed on 25 ms windows with an overlap of 10 ms, and then normalized using zero-mean and unit-variance normalization and then fed into the CNN. The training is done in two phases as illustrated in Figure 3. The first phase trains the network using categorical cross-entropy, on 348k audio files from Switchboard, NIST SRE 2004-2012 datasets and AgeVox-Celeb, totalling about 12.8k speakers. The network includes five convolutional layers, a statistics pooling layer, two fully-connected layers and a softmax output layer. The second phase consists of removing both the second fully-connected layer (FC2) and the softmax layer, freezing the remaining layers, and then adding a fully connected layer (FC2*) that is trained using large margin cosine loss (LMCL) [27].

3.2. Back-End Age Estimation

Once speaker embeddings are extracted, we feed them into a shallow neural network whose architecture depends on the approach being used:

- For regression, we use a dense layer with Relu activation composed of 256 units, followed by a second dense layer with Linear activation composed of 8 units, and finally a dense layer with a single neuron that outputs the estimated age.
- For either classification or distribution learning approaches, we also use a dense layer with Relu activation composed of 256 units, followed by a softmax layer where the number of units is equal to the age range. The difference between the two approaches lies in the loss function being used.

The following section details the different losses used to train the back-end age estimation network.

4. DISTRIBUTION LEARNING LOSSES

As discussed in section 1, regression-based methods that use losses like mean square error (MSE) or mean absolute error (MAE) offer the advantage of low error contribution of estimates closer to the ground truth, but do not allow any wiggle room in the neighborhood of the ground truth labels during training. Classification-based methods, where each year is a class, have the drawback that misclassified estimates closer to ground truth and those further from ground truth contribute equally to the error, regardless of rank ordering at training or inference. Several factors are at play while estimating the age of a person. In fact, the chronological age and biological age can be different for a person [28]. Therefore, an approach that can give some room for ambiguity in the ground truth and also considers rank ordering in some form could be more optimal for the problem of age estimation.

We incorporate the ambiguity in the ground truth by replacing each single label by a normal probablity distribution centered around it, similar to [13]. The standard deviation σ of the probability distribution is either in our control or learned by the system. As σ of the assumed normal distribution decreases, it becomes sharper and at the limit $\sigma \to 0$, this method asymptotically approaches the classification method.

The softmax output from the neural network, trained with the x-vector front end and these distributions as targets, is assumed to be the probability distribution of the estimated age. There are several loss functions that measure the similarity between two distributions, of which we have tried the ones that have shown success in facial age estimation.

Kullback-Leibler Divergence (KLD) or the relative entropy is defined as

$$KLD = -\sum_{i=1}^{K} p_i * \log \frac{\hat{p}_i}{p_i} \tag{1}$$

The KLD loss requires the μ and σ of the groundtruth age to be known. For facial age estimation, this could be an easy problem because there are training datasets that have been labeled by multiple annotators. However such datasets are not available for speech. Hence, we assume a constant σ for all ages for simplicity.

Similar to KLD, a new loss function proposed in [14] is called the Generalized Jeffries-Matusita (GJM) and is defined as

$$GJM = \sum_{i=1}^{K} p_i \mid 1 - \left(\frac{\hat{p}_i}{p_i}\right)^{\alpha} \mid^{\frac{1}{\alpha}} \tag{2}$$

This loss also requires the distribution of the groundtruth to be known. Here too, we empircally set σ to 25. Additionally, we varied the α and found that α=0.5 worked best similar to [14].

Train	Test	B1 [26]	B2 [11]	MSE	MAE	CCE	KLD	GJM	MVL
SRE08/10	SRE08/10	6.06/0.75	5.64/0.81	6.70/0.73	7.07/0.74	7.34/0.75	7.04/0.68	**6.12/0.76**	6.41/0.73
SRE08/10	AgeVoxCeleb	-/-	9.65/0.54	10.28/0.43	**9.65/0.50**	10.52/0.47	11.57/0.36	10.91/0.44	11.25/0.42
AgeVoxCeleb	AgeVoxCeleb	-/-	7.43/0.74	8.28/0.68	8.37/0.66	13.24/0.60	8.00/**0.69**	**7.94/0.69**	8.05/0. 68
AgeVoxCeleb	SRE08/10	-/-	9.47/0.61	9.26/0.62	7.70/0.65	10.06/0.56	**7.56/0.76**	8.02/0.64	7.87/0.66

Table 1. Performance of age estimation system (MAE/Pearson coefficient) using various losses. B1 and B2 cannot be directly comparable with our results because of the difference in the data partition between training and testing.

Another loss function used in distribution learning is the Mean-Variance loss proposed in [23]. This loss is a linear combination of mean and variance losses. The mean loss tries to minimize the difference in the means of the output softmax distribution and the label distribution, defined as

$$L_m = \frac{1}{2N}\sum_{i=1}^{N}(\sum_{j=1}^{K} j * p_{i,j} - y_i)^2 \tag{3}$$

The variance loss tries to minimize the spread of the output softmax distribution, defined as

$$L_v = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} p_{i,j} * (j - \sum_{k=1}^{K} k * p_{i,k})^2 \tag{4}$$

The advantage of the mean-variance loss is that it can estimate the age distribution without explicitly requiring to know the distribution of the groundtruth.

5. EXPERIMENTAL RESULTS

In our experiments, we trained one model for every chosen loss function for both the NIST SRE08/10 and AgeVoxCeleb datasets. Then we evaluated their performance for both match and mismatch conditions. With SRE08/10, we had 15 folds, so we compute the average metrics over the 15 folds in the results. In the cross-dataset case, the age ranges predicted by the model do not match that of the ground truth. Hence, we force the age range to be the shorter of the two by making all the age values beyond the range to be equal to the extremity on that side. Table 1 details the performance of each system for each of the test conditions in terms of MAE and Pearson coefficient. Each column corresponds to one system. In the first two columns, we provide baseline systems for both datasets for comparison. Please note that we created 15 random folds for the SRE08/10 protocol similar to [26] (denoted B1 in the table), but it is impossible to reproduce the exact split of data, and the results are not directly comparable with B2 [11] that uses a different split of the data. The hyper-parameters (number of layers and number of units per layer) of the DNN models were selected after an exhaustive empirical study. Additionally, σ (used in KLD and GJM) was empirically found to provide similar results for a range of [15,25]. Therefore, in the remainder of this section, we report the results for σ equal 20. Moreover, we assess the impact of the seed by running 10 different experiments with varying seed. We noticed that MAE varies slightly in the range of [7.91,8.12] with a mean of 8.03 and a standard deviation of 0.07 for KLD (reported MAE=8.00). In the case of matching datasets, the distribution algorithms take three out of the top four ranks. In the case of mismatch in datasets, we observe that the distribution algorithms take three out of top four ranks when trained on AgeVoxCeleb and even surpass the baseline B2, but take the lowest ranks when trained on SRE. However, if we take the weighted average of the metrics using the counts given in Section 2, we see that overall, for match conditions, the ranking of the algorithms is GJM, MVL, KLD, MSE, MAE and CCE. Similarly, for the weighted averages over the mismatch conditions, they rank in the order MAE, KLD, MVL, GJM, MSE and CCE. It is observed therefore, that overall, the distribution learning algorithms are succesful in improving the performance of an age estimation system.

When compared to B1 and B2, it is worth noting that our system does not take advantage of end-to-end training like B1 [26], or finetuning the last layer of the front end speaker embedding for age estimation and using a combination of losses like B2 [11].

6. CONCLUSIONS

Ranking-based distribution learning methods open up a rich potential for automatic age estimation from speech. The three loss functions we chose out of a myriad of distribution similarity measures were those showing promise in computer-vision. While their application is more constrained than facial age estimation, we show that those methods are very competitive and in most cases outperform traditional regression and classification algorithms for both matched and mismatched conditions. Future work will focus on improving those distribution learning methods and extending them to end-to-end architecture for age estimation. Contrary to the baselines B1 and B2, our approach uses off-the-shelf x-vectors since the focus was on the comparative study between the losses; however we believe that, for example, using stats pooling like in B2 could provide better performance. Contrary to B2, our approach does not use a combination of losses and we believe that combining losses will bring further improvement. Currently, we assume a fixed σ that was empirically set, which could be sub-optimal. Future work will investigate ideas of learning σ for every speech utterance.

7. REFERENCES

[1] Edward D Mysak, “Pitch and duration characteristics of older males,” *Journal of Speech and Hearing Research*, vol. 2, no. 1, pp. 46–54, 1959.

[2] Björn Schuller et al., “The INTERSPEECH 2010 paralinguistic challenge,” in *INTERSPEECH*, 2010.

[3] Sara Skoog Waller, Mårten Eriksson, and Patrik Sörqvist, “Can you hear my age? Influences of speech rate and speech spontaneity on estimation of speaker age,” *Frontiers in psychology*, vol. 6, pp. 978, 2015.

[4] Pegah Ghahremani et al., “End-to-end Deep Neural Network Age Estimation,” in *INTERSPEECH*, 2018.

[5] Douglas A Reynolds, Thomas F Quatieri, and Robert B Dunn, “Speaker verification using adapted Gaussian mixture models,” *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[6] Tobias Bocklet et al., “Age and gender recognition for telephone applications based on GMM supervectors and support vector machines,” in *ICASSP*. IEEE, 2008, pp. 1605–1608.

[7] Najim Dehak et al., “Front-End Factor Analysis for Speaker Verification,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, pp. 788–798, 2011.

[8] Mohamad Hasan Bahari, Mitchell McLaren, Hugo Van hamme, and David Van Leeuwen, “Age Estimation from Telephone Speech using i-vectors,” in *INTERSPEECH*, 2012.

[9] David Snyder et al., “X-vectors: Robust DNN embeddings for speaker recognition,” in *ICASSP*. IEEE, 2018, pp. 5329–5333.

[10] Jesús Villalba 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, pp. 101026, 2020.

[11] Naohiro Tawara, Atsunori Ogawa, Yuki Kitagishi, and Hosana Kamiyama, “Age-VOX-Celeb: Multi-Modal Corpus for Facial and Speech Estimation,” in *ICASSP*. IEEE, 2021, pp. 6963–6967.

[12] Raphael Angulu, Jules R Tapamo, and Aderemi O Adewumi, “Age estimation via face images: a survey,” *EURASIP Journal on Image and Video Processing*, vol. 2018, no. 1, pp. 1–35, 2018.

[13] Bin-Bin Gao et al., “Deep label distribution learning with label ambiguity,” *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 2825–2838, 2017.

[14] Ali Akbari, Muhammad Awais, Manijeh Bashar, and Josef Kittler, “How Does Loss Function Affect Generalization Performance of Deep Learning? Application to Human Age Estimation,” in *ICML*. PMLR, 2021.

[15] Ivan Huerta et al., “A deep analysis on age estimation,” *Pattern Recognition Letters*, vol. 68, pp. 239–249, 2015.

[16] Rasmus Rothe, Radu Timofte, and Luc Van Gool, “Deep expectation of real and apparent age from a single image without facial landmarks,” *International Journal of Computer Vision*, vol. 126, no. 2, pp. 144–157, 2018.

[17] Guodong Guo, Yun Fu, Charles R Dyer, and Thomas S Huang, “Image-based human age estimation by manifold learning and locally adjusted robust regression,” *IEEE Transactions on Image Processing*, vol. 17, no. 7, pp. 1178–1188, 2008.

[18] Dong Yi, Zhen Lei, and Stan Z Li, “Age estimation by multi-scale convolutional network,” in *Asian conference on computer vision*. Springer, 2014, pp. 144–158.

[19] Vladimir N. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag New York, Inc., 1995.

[20] Zhiguang Yang and Haizhou Ai, “Demographic classification with local binary patterns,” in *International Conference on Biometrics*. Springer, 2007, pp. 464–473.

[21] Shixing Chen et al., “Using ranking-CNN for age estimation,” in *IEEE CVPR*, 2017.

[22] Bin-Bin Gao, Hong-Yu Zhou, Jianxin Wu, and Xin Geng, “Age Estimation Using Expectation of Label Distribution Learning,” in *IJCAI*, 2018, pp. 712–718.

[23] Hongyu Pan, Hu Han, Shiguang Shan, and Xilin Chen, “Mean-variance loss for deep age estimation from a face,” in *IEEE CVPR*, 2018.

[24] Alvin F Martin and Craig S Greenberg, “NIST 2008 speaker recognition evaluation: Performance across telephone and room microphone channels,” in *INTERSPEECH*, 2009.

[25] Alvin F Martin and Craig S Greenberg, “The NIST 2010 speaker recognition evaluation,” in *INTERSPEECH*, 2010.

[26] Rube Zazo et al., “Age estimation in short speech utterances based on LSTM recurrent neural networks,” *IEEE Access*, vol. 6, pp. 22524–22530, 2018.

[27] Hao Wang et al., “Cosface: Large margin cosine loss for deep face recognition,” in *IEEE CVPR*, 2018.

[28] Xin Geng et al., “Learning from facial aging patterns for automatic age estimation,” in *ACM international conference on Multimedia*, 2006, pp. 307–316.