End-to-End Deep Neural Network Age Estimation

Pegah Ghahremani[1], Phani Sankar Nidadavolu[1], Nanxin Chen[1], Jesús Villalba[1], Daniel Povey[1,2],
Sanjeev Khudanpur[1,2], Najim Dehak[1]

[1]Center for Language and Speech Processing
[2]Human Language Technology Center Of Excellence,
Johns Hopkins University, Baltimore, MD

{pghahre1,snidada1,bobchennan,jvilla17, ndehak3,khudanpur}@jhu.edu, dpovey@gmail.com

Abstract

In this paper, we apply the recently proposed x-vector neural network architecture for the task of age estimation. This architecture maps a variable length utterance into a fixed dimensional embedding which retains the relevant sequence level information. This is achieved by a temporal pooling layer. From the embedding, a series of layers is applied to make predictions. The full network is trained end-to-end in a discriminative fashion. This kind of network is starting to outperform the state-of-the-art i-vector embeddings in tasks like speaker and language recognition. Motivated by this, we investigated the optimum way to train x-vectors for the age estimation task. Despite that a regression objective is typical for this task, we found that optimizing a mixture of classification and regression losses provides better results. We trained our models on the NIST SRE08 dataset and evaluated on SRE10. The proposed approach improved mean absolute error (MAE) by 12% w.r.t the i-vector baseline.

Index Terms: Age identification, x-vector, i-vector

Table 1: *x-vector network architecture*

Layer	Layer type	Size	Context
1	TDNN	400	0
2	TDNN	400	[-2,0,2]
3	TDNN	400	[-3,0,3]
4	TDNN	400	0
5	Stats-pooling (Mean+StdDev)	1500+1500	[0:s:T)[1]
6	Dense (embedding) + ReLU	400	sequence
7	Dense + ReLU	400	sequence
8	Dense softmax	Age bins	sequence
8[2]	Affine	1	sequence

1: T frame-level outputs of layer 4 are Aggregated every s frames.
2: Separate linear transform layer is added for regression objective.

1. Introduction

Speech is a common physiological signal for face-to-face communication and human-computer interaction. Nowadays, mobile phones, smart homes, as well as other assistant devices have accelerated the development of various speech applications. In addition to the dominant linguistic information, the speech also carries paralinguistic information such as speaker identity, emotional estate, age, and ethnicity. Recently, research on automatic extraction of such information has increased, since it can lead to applications, like age-dependent advertisements, caller-agent pairing and other customized service[1].

In this paper, we focus on the age estimation problem. Several approaches have been proposed in the literature, either to classify the age range (young, youth, adult and senior) [1, 2, 3], or predict the actual age [4, 5]. As early as the 1950s, Mysak proposed to use long-term and short-term features to predict age [6]. In [2], Minematsu used MFCC features with delta coefficients with Gaussian mixture models (GMM) for binary age classification. Schötz [7] studied the complex correlation between speech rate, sound pressure, fundamental frequency and speaker's age. Another work [3] proposed to combine support vector machines (SVM) with GMM to combine short-term cepstral features and long-term features, and improve performance. In recent years, Fedorova [4] started to use i-vectors [8] combined with a separate neural network back-end for regression. IBM researchers [5] proposed to apply support vector regression (SVR) on DNN i-vectors extracted using fMLLR features leading to state-of-the-art performance. In [9], the authors studied how the length of the speech segments impacts the age estimator performance. They propose using LSTM networks to improve performance with short segments. All previous work either depends on handcrafted features, or uses combination of multiple components.

Recently the x-vector neural network architecture has been proposed [10] attaining great performance for speaker verification [11], speaker diarization [12] and language recognition [13]. The x-vector architecture converts a variable length feature sequence into a fixed-dimension embedding which contains the relevant information of the utterance. The x-vector embedding is extracted via a temporal pooling layer which summarize information along the time axis. After getting this embedding, utterance level labels, like speaker identity, age, and gender, can be used for discriminative network training. Thus, end-to-end training becomes possible, jointly optimizing both feature extraction and prediction. In this work, we applied the x-vector approach trained with a weighted sum of classification and regression objectives to estimate age given the speech features. Combining two different objective functions led to Mean Absolute Error (MAE) improvement w.r.t. using just the traditional regression objective. With respect to the i-vector baseline, this system achieved 12% improvement.

The rest of the paper is organized as follows. Section 2 introduces the proposed x-vector age estimation system with classification and regression objectives. Section 3 describes the i-vector baseline system. Section 4, explains our experimental setup using NIST SRE08-10; and the system hyper-parameters. Section 5 presents our results. Finally, Section 6 presents the conclusions.

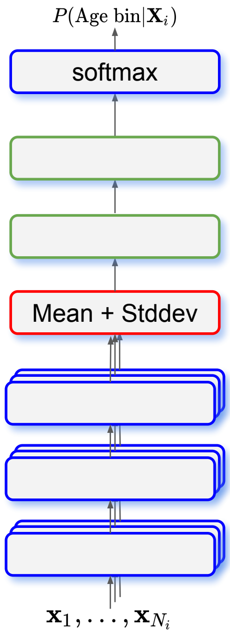
$P(\text{Age bin}|\mathbf{X}_i)$
softmax
Mean + Stddev
$\mathbf{x}_1, \dots, \mathbf{x}_{N_i}$

Figure 1: *x-Vector neural network.*

2. x-Vector age estimation

2.1. Network architecture

Recent works [14, 10] introduced a successful neural network architecture to map sequences into speaker discriminant fixed-length vectors. Authors denominated these embeddings as x-vectors. Figure 1 depicts a generic x-vector neural network. The network receives a sequence of feature frames, which are processed by several layers. The result is summarized by a pooling layer that computes mean and standard deviation over time. We compute these statistics for every 3 frames over all possible outputs of layer 4. Mean and standard deviation are concatenated together and propagated to the output through a series of feed-forward layers. The output is a dense layer with softmax activation predicting the class posteriors–discrete age bins in our application. Before the pooling layer, we used a time delay neural network (TDNN) (a.k.a. 1D convolutions). The sequence embedding is extracted from the first affine transform after the pooling layer (before applying the non-linear activation). Table 1 summarizes the network architecture.

The results in [11] indicate that x-vector can outperform i-vectors and be robust across datasets. However, x-vectors is a data greedy approach and only is able to beat i-vectors when we have a large amount of training data. For this reason, we need to resort to data augmentation schemes–speed perturbation, noise, reverberation– to make x-vectors work optimally. We augment the training data with additive noise and reverberation, where in reverberation , the audio convolves with room impulse responses. We use simulated RIRs described in [15] for reverberation and MUSAN dataset is used for additive noise augmentation, which contains 900 noises, 42 hours from various genres and 60 hours of speech from twelve languages [16]. The detail for 3-fold augmentation is described in [11]. x-Vector embeddings have also been effective for language recognition, being the most successful approach in NIST LRE17 evaluation [17, 13].

2.2. Regression vs. classification

Age estimation is typically considered as a regression problem. However, as ages are discrete values, we decided to explore the idea of treating it as a classification problem with cross-entropy objective. We use softmax layer at the end of network to give the network freedom to model any distribution for separate age classes. Each age class was modeled by an output node in the network softmax layer. We limited the number of age classes, where a new age class was added if the amount of frames for that class is larger than $\delta \times \frac{N}{M}$ with $\delta = 0.01$ and N is the total number of frames and M is the number of ages in training data. The ages with a low number of samples were mapped to the previous age class. The age values belonging to the same age class is mapped to smallest age value in the group during test time.

The classification objective has the drawback that all errors are penalized the same, no matter the distance between the true and predicted ages, which can degrades MAE as a performance measure. For this reason, we also experimented with a combined classification-regression that minimizes cross-entropy and mean square error,

$$L = -\sum_{i=1}^{N} \log[P(y_i = t_i|\mathbf{X}_i)] + \lambda(z_i - t_i)^2 \qquad (1)$$

As shown in table 1, two separate regression and classification layers are added after final layer with output dimension of number of age bins and 1 respectively. The linear affine classification layer is followed by softmax layer. t_i is true ages and y_i and z_i are the output of softmax and regression layer respectively.

3. i-Vector baseline

3.1. i-Vector extraction

The i-vector paradigm [8] is the state-of-the-art method to convert a variable length feature sequence into a single fixed-dimensional vector. This vector termed as i-vector, becomes a new feature for pattern classification algorithms like SVM [18] and PLDA [19, 20].

i-Vectors is an extension of the GMM-UBM approach [21], where each speech segment is modeled by a Gaussian mixture model (GMM). The super-vector mean $\mathbf{M}$ of the segment GMM is assumed to be

$$\mathbf{M} = \mathbf{m} + \mathbf{T}\mathbf{w} \qquad (2)$$

where $\mathbf{m}$ is the UBM means super-vector, $\mathbf{T}$ is a low-rank matrix and $\mathbf{w}$ is a standard normal distributed vector. $\mathbf{M}$ defines the total variability space, i.e. the directions in which we can move the UBM to adapt it to a specific segment.

Using this model, we can compute the posterior distribution of $\mathbf{w}$ given the utterance features. The mean of the Gaussian posterior is the i-vector embedding.

3.2. Back-end

The role of the back-end is to predict the age label of the utterance given the input i-vector. The results in [4] showed that choice of the back-end between a DNN and an SVR has little effect on the performance of age estimation from i-vectors. In this work, we used a two hidden layer DNN back-end for both i-vector, and fusion of i-vector and x-vector. DNN was configured for regression with a linear output layer and was optimized to minimize the mean square error between the true and predicted ages.

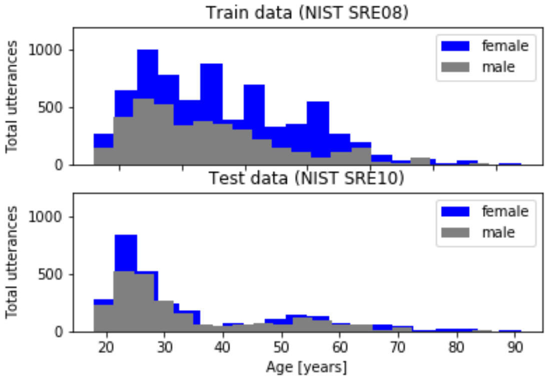

Figure 2: *Histogram showing number of utterances for train and test for both male and female.*

4. Experimental setup

4.1. NIST SRE08-10 dataset

We experimented on 2008-2010 NIST speaker recognition evaluation (SRE08-10) datasets with a configuration similar to the one in [5]. The x-vector and back-end neural networks were trained on telephone speech from SRE08, which consists of 11205 utterances from 1227 (769 female and 458 male) speakers. We used conversations in English and non-English languages. The performance was evaluated on NIST SRE10 telephone condition, which consists of 5331 utterances corresponding to 492 (256 female and 236 male). No speaker overlap exists between the train and test sets. Figure 2 shows the histograms of number of utterances w.r.t. age bins for training and test sets.

4.2. x-Vector system configuration

The x-Vector setup is described in detail in this section. The features were 23-dim MFCC short-time mean normalized over sliding the window of 3 seconds. An energy-based SAD (speech activity detection) was used to remove non-speech frames. The DNN configuration is outlined in table 1. The time-delay deep network layer (TDNN) with the rectified linear unit (ReLU) non-linearity was used. Batch normalization was also used after the non-linearity. The first 4 layers had small temporal context centered at the current frame t, e.g. layer with $[-3, 0, +3]$ concatenates frames from $t-3$, t to $t+3$ and builds the layer on them. The statistics pooling layer aggregates information across time over all T frame-level outputs at layer 4. The computed mean and standard deviation are concatenated and propagated through next layers and the output softmax layer. The training examples consisted of mini-batches of speech utterances with the corresponding age label. To make the network robust to variable-length test utterances, we trained the DNN model using variable length chunks randomly sampled from the full-length utterances. The effect of fixed versus variable length training chunk length is investigated in details in Section 5.2. Since x-vector is data greedy approach, we also used data augmentation, consisting in adding noise and reverberation to increase the amount of training data, which improved the mean absolute error from 5.9 to 4.9.

Table 2: *MAE, regression vs. classification objectives for x-vector training on SRE08 and testing on SRE10*

Classification weight	Regression weight	MAE
1	0	5.1
0	1	11.9
1	0.001	**4.9**

4.3. i-Vector system configuration

The i-vector system also used MFCC features with short-time centering. The UBM and i-vector extractors were trained on NIST SRE04-06 English telephone speech containing 1936 female speakers and 679 male speakers. We used a 2048 component GMM-UBM model with full covariance matrices. Total variability subspace dimension was set to 400. It is worth mentioning that there is no speaker overlap between the data used to train the i-vector extractor and data used to train and test the age estimation backend/x-Vector system.

The input dimension for the DNN back-end was 400 for i-vectors and 800 for the fusion experiments–concatenation i-vector and x-vector. We also experimented applying LDA to the input embeddings, the LDA dimension was set to 50. The LDA transform was learned with respect to the age labels.

We split the SRE08 data into train and validation sets without speaker overlap to training the back-end. All the speakers with more than 6 utterances were used to train the back-end. All the other speakers were used for validation. To be consistent with the x-vector experiments, all ages with few training samples were mapped to the previous age class.

Each hidden layer had 256 neurons with sigmoid non-linearity. The output layer had one linear neuron to predict the age value. The network was trained to minimize mean-squared-error objective. The mini-batch size was set to 32. We used stochastic gradient descent (SGD) optimizer with an initial learning rate of 0.001 and a momentum of 0.9. We decreased the learning rate by a factor of 2 when the validation loss does not improve for two successive epochs. Minimum learning rate was set to 1e-05. Training is stopped if validation does not improve for three consecutive epochs. The model with best validation loss was used for testing.

4.4. Performance measure

To asses the goodness of our age estimators, we report performance in terms of mean absolute error and Pearson's correlation coefficient. Mean absolute error (MAE) is defined,

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - t_i| \quad (3)$$

where y_i and t_i are the true and predicted age values respectively.

Pearson's correlation coefficient is defined as,

$$\rho = \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{y_i - \mu_y}{\sigma_y}\right)\left(\frac{t_i - \mu_t}{\sigma_t}\right) \quad (4)$$

where μ_y and σ_y are the mean and standard deviation for the predicted ages; and μ_t and σ_t for the true ages. Higher correlation coefficients are better.

Table 3: *Effect of x-vector training/test segment length for age estimation on SRE10.*

MAE	Test segment length(s)			
Train segment length(s)	10	15	20	full
2	13.03	10.92	6.04	6.1
5	13.37	10.97	**4.92**	**4.92**
10	19.46	15.2	4.99	4.97
$4 - 30$	**11.62**	**9.88**	5.16	5.3

Table 4: *i-Vector vs x-vector vs fusion. LDA indicates that we reduce embedding dimension before applying the back-end.*

	Male		Female		Overall	
	MAE	ρ	MAE	ρ	MAE	ρ
i-Vector						
w/o LDA	**6.54**	**0.77**	**5.12**	0.89	**5.77**	**0.84**
with LDA	6.65	0.76	5.15	**0.90**	5.82	**0.84**
x-Vector end-to-end						
Train on 5s chunks	5.78	**0.74**	**4.23**	**0.87**	**4.92**	**0.81**
Train on 10s chunks	**5.55**	**0.74**	4.54	0.85	4.97	0.80
Fusion i-vec+x-vec (x-vec train on 5s)						
w/o LDA	7.78	0.75	5.75	0.90	6.67	0.82
with LDA	6.30	0.80	**4.50**	**0.92**	5.30	0.87
Fusion i-vec+x-vec (x-vec train on 10s)						
w/o LDA	6.01	**0.83**	4.98	**0.92**	5.44	**0.88**
with LDA	**5.84**	**0.83**	4.68	**0.92**	**5.20**	**0.88**

5. Results

5.1. Regression vs. classification

Table 2, shows results for x-vector systems trained with different weights of the regression and classification objectives. While in principle classification is clearly superior to regression, combining both we improved by 4% relative. Note that though the regression weight is much lower than the classification weight, that doesn't mean that regression is less important. The dynamic range of the mean square error objective is larger than the one of the cross-entropy objective, so the low value of the regression weight compensates for that.

5.2. Train/test duration analysis

For the same application of age estimation, we need to be able to estimate the user age as fast as possible. For example, in call centers, we want to pair the customer with an agent that is expert dealing with people in a given age range. Therefore, we experimented with what is the best way to train the x-vector network to work well with short speech durations. When training the network, we randomly select speech chunks from the full-length training recordings. Those chunks can have a fixed duration or we can also sample chunk with random durations. Table 3 shows results for different x-vector training/test durations. For short test durations, variable length chunks performed significantly better–about 10% relative to 10-second tests. Meanwhile, for long durations, fixed length training was around 7% better than variable length. This is different to what is reported on previous x-vector works for speaker verification [11], where variable length training was consistently better.

5.3. System fusion: i-vector + x-vector

Table 4 present results comparing x-vectors end-to-end evaluation with the i-vector baseline. It also presents fusion between x-vector and i-vector, where both embeddings were concatenated a feed into the 2 layer DNN back-end. We consider the case where we don't apply any post-processing to the i-vector or i-vector+x-vector; and the case where we reduce their dimension to 50 using age discriminant LDA. The table also compares x-vectors trained with 5 seconds and 10 seconds chunks. Results are reported in terms of MAE and correlation coefficient (defined above).

The x-vector system significantly outperformed i-vector by 14% relative. Fusion of i-vector and x-vector improves by 9% w.r.t the i-vector system. However, it is still worse than the single end-to-end x-vector. This x-vector result is similar to the best result reported in the literature (to our knowledge) [5], which required a more complicated pipeline including fMLLR features and DNN i-vectors.

LDA dimensionality reduction was not beneficial for the case of using i-vectors as input to the back-end, but it improved when concatenating x-vector and i-vector. This means that the DNN back-end was not able to handle high dimension inputs, probably because of the limited data in the SRE08 dataset.

Comparing x-vector models trained on 5 and 10-second speech chunks, we observe that, though the 5-second model was slightly better for end-to-end x-vector, the 10-second model fused better with the i-vector, with and without LDA.

6. Conclusions

In this paper, we proposed to use x-vector neural network architecture for age estimation from speech. This architecture consists of a series of time delay layers (TDNN) followed by a temporal pooling layer which summarizes the feature sequence into a single fixed dimension embedding. The embedding is fed into a series of feed-forward layers to predict the age value. This is adequate to predict sequence level properties of an utterance, such as speaker identity, language or age. We trained the network with an objective that is the weighted sum of cross-entropy and mean square error, which gave a certain advantage over the typical regression objective used for this task.

We used NIST SRE08 dataset to train our model and NIST SRE10 for evaluation. We obtained mean absolute error of 4.9, which is 14% better than our i-vector baseline. This result is competitive with the best result published on this task.

7. References

[1] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. MüLler, and S. Narayanan, "Paralinguistics in speech and languagestate-of-the-art and the challenge," *Computer Speech & Language*, vol. 27, no. 1, pp. 4–39, 2013.

[2] 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 *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 1. IEEE, 2002, pp. I–137.

[3] C. Müller and F. Burkhardt, "Combining short-term cepstral and long-term pitch features for automatic recognition of speaker age," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[4] A. Fedorova, O. Glembek, T. Kinnunen, and P. Matějka, "Exploring ann back-ends for i-vector based speaker age estimation," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[5] S. O. Sadjadi, S. Ganapathy, and J. W. Pelecanos, “Speaker age estimation on conversational telephone speech using senone posterior based i-vectors,” in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5040–5044.

[6] E. D. Mysak, “Pitch and duration characteristics of older males.” *Journal of Speech & Hearing Research*, 1959.

[7] S. Schötz, “Acoustic analysis of adult speaker age,” in *Speaker Classification I*. Springer, 2007, pp. 88–107.

[8] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, 2011.

[9] 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*, March 2018.

[10] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, “Deep Neural Network Embeddings for Text-Independent Speaker Verification,” in *Proceedings of the 18th Annual Conference of the International Speech Communication Association, INTERSPEECH 2017*. Stockholm, Sweden: ISCA, aug 2017, pp. 999–1003.

[11] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-Vectors : Robust DNN Embeddings for Speaker Recognition,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018*. Alberta, Canada: IEEE, apr 2018.

[12] D. Garcia-Romero, D. Snyder, G. Sell, D. Povey, and A. Mccree, “Speaker Diarization Using Deep Neural Network Embeddings,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017*. New Orleans, LA, USA: IEEE, mar 2017, pp. 4930–4934.

[13] D. Snyder, D. Garcia-Romero, A. Mccree, G. Sell, D. Povey, and S. Khudanpur, “Spoken Language Recognition using X-vectors,” in *submitted to Odyssey 2018*, Les Sables d’Olonne, France, jun 2018.

[14] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur, “Deep neural network-based speaker embeddings for end-to-end speaker verification,” in *Proceedings of the 2016 IEEE Spoken Language Technology Workshop (SLT)*. San Diego, CA, USA: IEEE, dec 2016, pp. 165–170.

[15] 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 *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 5220–5224.

[16] D. Snyder, G. Chen, and D. Povey, “Musan: A music, speech, and noise corpus,” *arXiv preprint arXiv:1510.08484*, 2015.

[17] A. Mccree, D. Snyder, G. Sell, and D. Garcia-Romero, “Language Recognition for Telephone and Video Speech : The JHU HLTCOE Submission for NIST LRE17,” in *submitted to Odyssey 2018*, Les Sables d’Olonne, France, jun 2018.

[18] S. Cumani, O. Glembek, N. Brummer, E. De Villiers, and P. Laface, “Gender independent discriminative speaker recognition in i-vector space,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2012*. Kyoto, Japan: IEEE, mar 2012, pp. 4361–4364.

[19] P. Kenny, “Bayesian Speaker Verification with Heavy-Tailed Priors,” in *Proceedings of Odyssey 2010 - The Speaker and Language Recognition Workshop*. Brno, Czech Republic: ISCA, jul 2010.

[20] J. Villalba and N. Brummer, “Towards Fully Bayesian Speaker Recognition: Integrating Out the Between-Speaker Covariance,” in *Proceedings of the 12th Annual Conference of the International Speech Communication Association, Interspeech 2011*. Florence, Italy: ISCA, aug 2011, pp. 505–508.

[21] 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, jan 2000.