Proceedings of the 18th Conference on Computer Science and Intelligence Systems pp. 1087–1091
DOI: 10.15439/2023F2483
ISSN 2300-5963 ACSIS, Vol. 35

Automatic speaker's age classification in the Common Voice database

Adam Nowakowski, Włodzimierz Kasprzak
0000-0000-0000-0000, 0000-0002-4840-8860
Warsaw University of Technology, Institute of Control and Computation Eng.
ul. Nowowiejska 15/19, 00-665 Warszawa, Poland
Email: wlodzimierz.kasprzak@pw.edu.pl

***Abstract*—An approach to speaker's age classification using deep neural networks is described. Preliminary signal features are extracted, based on mel-frequency cepstral coefficients (MFCC). For gender classification an MLP network appears to be a satisfactory lightweight solution. For the age modelling and classification problem, two network types, ResNet34 and x-vectors, were tested and compared. The impact of signal processing parameters and gender information (both theoretic perfect realistic imperfect) onto the age classification performance was experimentally studied. The neural networks were trained and verified on the large "Common Voice" dataset of English speech recordings.**

I. Introduction

ALTHOUGH work on speaker age recognition dates back to the 1950s [1], this problem is still difficult to solve in practice. Several reasons for this state of affairs can be listed. The speaker's perceived age and his/her chronological age can differ significantly. To train the age classifier well, a very large database of recordings labeled with the age of the speakers will be required. By its nature, the sound of the same speaker's speech depends on many factors independent of age, such as gender, weight, temperament, mood, ethnicity. The first systems with relatively good efficiency in estimating the age of the speaker were developed some 20 years ago [2], [3]. Currently developed systems for estimating the age of the speaker are based on acoustic modeling used in the speaker recognition (speaker identification and verification) systems [4]. The basic classic machine learning methodologies used for this problem are UBM-GMM (Universal Background Model - Gaussian Mixture Model) and "i-vectors" [5], [6], [7]. An early solution based on deep neural networks is the "x-vector" [8], [9], [10] network. Other deep network architectures, such as LSTM [11] or ResNet [12], were also proposed for this purpose.

In this paper, we propose a gender-informed approach to speaker's age classification using the large "Common Voice" database [13] for neural network training and testing. In section 2, this database is introduced in more details. The implemented system SAR (speaker's age recognition) is presented in section 3. Here, we already present results of initial experiments, aimed to find optimal settings of two signal processing parameters. The main experiments and a summary of age classification performance, follows in section 4. At the end, in section 5, we conclude the work with a summary of results.

This work was supported by "Narodowe Centrum Badań i Rozwoju", Warszawa, Poland, grant No. CYBERSECIDENT/455132/III/NCBR/2020. The publication was funded by Warsaw University of Technology.

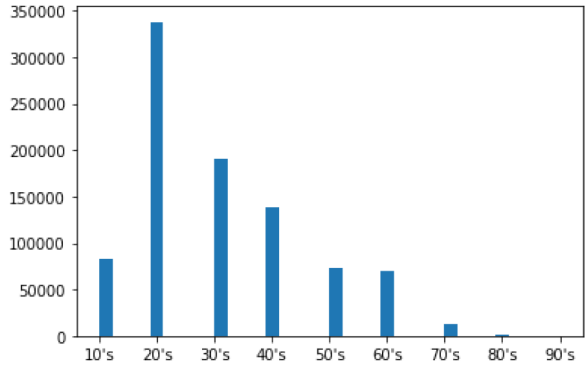

Fig. 1. Statistics of the English language subset of "Common Voice" recordings according to age groups

II. Datasets

A. "Common Voice"

Two large databases of tagged recordings were analyzed for the speaker's age recognition (SAR) system. The first one is "Age-VOX-Celeb" [14], which contains age tags of celebrity recordings, downloaded from "YouTube". The second base is "Common Voice" [13], a Mozilla project, dedicated to record the speech of ordinary Internet user. Everyone can register and record his voice. Other users can listen to the recordings and evaluate their correctness. The tags included speaker's accent, age and gender. The content of the database is growing every day – it contains over 37 million audio files with speech samples in many languages. The "Common Voice" database was chosen here, due to its easy accessibility and a large containment of almost 900,000 recordings in English from over 18,000 people (Fig. 1).

The database contains recordings of voices ranging from teenagers to people in their nineties, but the distribution of age groups is significantly uneven. There are only few participants over the age of seventy. Therefore, for the first series of experiments we decided to limit the age classification to the

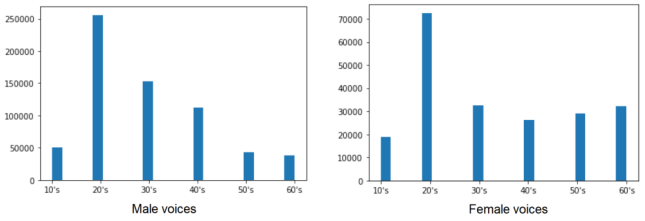

Fig. 2. Distribution of the English subset of "Common Voice" recordings according to speaker's gender and age

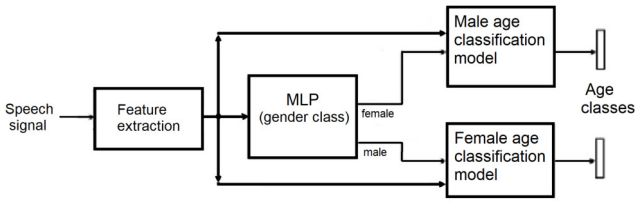

Fig. 3. Structure of SAR

first six age groups. The number of recordings from females is around 200,000 and are much lower than the number of recordings from males, which are over 650,000 (Fig. 2). This leads to our separate treating of age classes for women and men.

This large number of speakers supports extensive speaker classification studies but it also requires extensive computational resources for training and testing neural network models. As a tradeoff solution, we decided to select a training subset of 25,000 recordings of men from each age group and 15,000 recordings of women from each age group. The test set consisted of thousand recordings for each age group, regardless of gender. In total, 12 classes were distinguished, which resulted from taking into account the 6 age groups and 2 gender of the speaker.

III. Solution SAR

The overall structure of the speaker's age recognition (SAR) system is shown in Figure 3. There are three processing stages: feature extraction, gender classification and age group classification.

A. Feature extraction

Standard speech features were chosen based on "mel frequency cepstral coefficents" (MFCC), delivered by popular audio processing library - the LibRosa library [15]. A feature vector is generated for every signal frame. It consists of 70 coefficients, i.e. 3×23 MFCC-based (i.e., MFCCs, delta MFCCs and delta-delta MFCCs) plus 1 energy coefficient.

Nevertheless the standard MFCC-based signal parametrization, some signal segmentation parameters still need to be set optimally. We experimented with different settings of two parameters: the window length (n_fft) and the delay between consecutive windows (hop_ length).

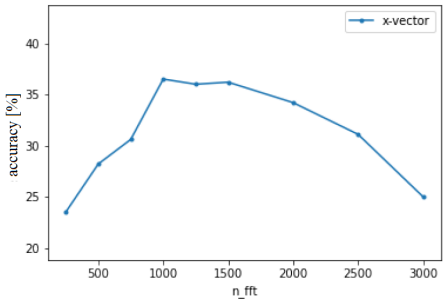

Fig. 4. Experiments with different window lengths

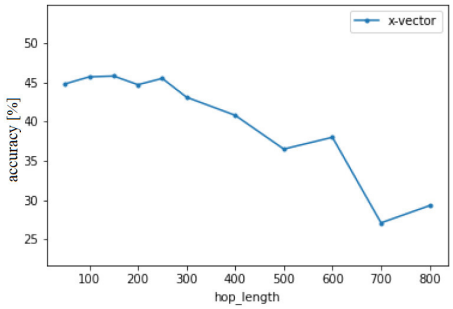

Fig. 5. Experiments with different hop lengths

In the first type of experiments various window lengths, have been set, the x-vector network was trained and its test performance was evaluated. In this test series, the hop_length was set to half of n_fft. The classification performance as a function of window length is shown in Figure 4. The highest quality of results is observed for n_fft in the range between 1000 and 1500 signal samples. (i.e., 45 – 68 ms). Both for smaller and larger windows, the performance is deteriorated. Hence, we selected a window size of 1000 samples, using a FFT of order 1024.

In experiments with different hop lengths, the window size was fixed. The delay parameter was studied in the range from 50 to 800 (Figure 5). As expected, the results show a tendency of increased classification accuracy with decreased hop length. The best performance was stable obtained for delays between 50 and 250 samples (i.e., ca 2 – 11 ms). Among them, we have chosen 250 samples for reasons of computational efficiency. The accuracy of the x-vectors network was 45.5%, only slightly worse than 45.7% and 45.8% for 100 and 150 samples, respectively. Please note, that already with this optimal setting of the hop length parameter, the multi-class classification accuracy of the x-vectors network has been increased by 9% (from 36.5% to 45.5%).

B. Gender classification

For the purpose of gender classification, the feature vectors of all the frames of given recording, are combined into a single vector. This constitutes the input of a multi-layer perceptron (MLP) with two hidden layers. The network is trained on samples annotated with gender information. We evaluated the

Layer	Layer context	Total context	Input x output
Conv1D	[t-2,t+2]	5	210x512
Conv1D	{t-2,t,t+2}	9	1536x512
Conv1D	{t-3,t,t+3}	15	1536x512
Conv1D	{t}	15	512x512
Conv1D	{t}	15	512x1500
stat pooling	[0,T]	T	1500Tx3000
linear	{0}	T	3000x512
linear	{0}	T	512x512
softmax	{0}	T	512xN

Fig. 6. The x-vectors network [8]

trained model on a small database, containing recordings of 24 speakers (the RAVDESS database), and its accuracy was 98.7%.

C. x-vectors

X-vectors is a very popular speech recognition network, recently proposed in [8] and already ighly cited in the literature. The architecture of the x-vector network is summarized in Figure 6. The discussed architecture uses a one-dimensional convolution layer - the filter kernel operates along the time domain, while each feature is treated separately. The first five layers operate in this way. Statistical data (mean and deviation) are extracted from the last convolutional layer for each output feature (each channel). This operation is to ensure a fixed length of the output vector, which will later be processed by two dense layers. Finally, there is a softmax layer that maps their outputs to 12 age classes.

D. ResNet34

The second, much more complicated neural network used in our work, is ResNet34 (Figure 7). This architecture also enjoys popularity [16], [17]. It is based on convolutional networks and residual connections. It starts with a convolutional layer with 64 output channels and a 3×3 reception area. Then, there is a ResNet block of 3 layers, each one composed of 2 convolutional layers, with a 3×3 area and 64 outputs. Next, three more ResNet blocks follow, with same kernel size but growing number of outputs. An average pooling layer, a dense layer and softmax complete the network.

We conclude, that the main difference between the two considered architectures is the use of different convolutional layers. X-vectors is using a lightweight 1-D convolution along time axis, whereas ResNet34 applies a true 2-D convolution along time and feature indices.

IV. Results

The technological stack of the system implementation consists of: the Python 3.9.6 language, Jupyter Notebook 8 interactive code editor Librosa 9 library in Python for audio signal processing. PyTorch 10 library for neural network tools and utility libraries, like NumPy and Pandas.

Layer	Layer context
Conv2D	3x3, 64
ResNetBlock1	$\begin{bmatrix} 3x3, 64 \\ 3x3, 64 \end{bmatrix}$ x3
ResNetBlock2	$\begin{bmatrix} 3x3, 128 \\ 3x3, 128 \end{bmatrix}$ x4
ResNetBlock3	$\begin{bmatrix} 3x3, 256 \\ 3x3, 256 \end{bmatrix}$ x6
ResNetBlock4	$\begin{bmatrix} 3x3, 512 \\ 3x3, 512 \end{bmatrix}$ x3
stat pooling	avarage pooling
linear	512x12
softmax	-

Fig. 7. The ResNet34 network [16]

	x-vectors	ResNet34
6 classes - no gender information	25,5%	32,7%
12 classes (2 x 6) - single model - gender information	53,8%	67,1%
6 classes x two separate models - gender information	59,6%	70,8%
3 classes x two separate models - gender information	77,3%	86,1%

Fig. 8. Summary of age classification results obtained on the "Common Voice" subset

A. Age classification

In Figure 8, we give results of several training and test series of the two considered network architectures for the class-balanced subset of the „Common Voice" English dataset. Both x-vectors and ResNet34 were applied in the same way - they were trained in 15 epochs with the feature set. The second model is more complex than x-vectors (one needs ca. 10 times more parameters to train) but it shows a better performance than the first one in all experiments.

B. Age classification without gender information

Consider first a 6-class problem, when there is one model for both male and female speakers and no gender information controls the classification process. Both networks perform poorly (25,5% x-vectors, 32,7% ResNet34). If perfect gender information is available to the system, one can expect better results.

C. Age classification with perfect gender information

Consider first a single network model created for a 12-class problem (2 gender $\times 6$ age groups), with the additional information about gender, that allows to select the most likely output from the proper 6-class subset. The ResNet34 shows a better total accuracy of 67,1% in this classification case, compared to an accuracy of 53,8% of the x-vectors.

Now, remember our proposed architecture, shown in Figure 3, with two separate models, trained separately on the two gender samples. Each of the two models is classifying a

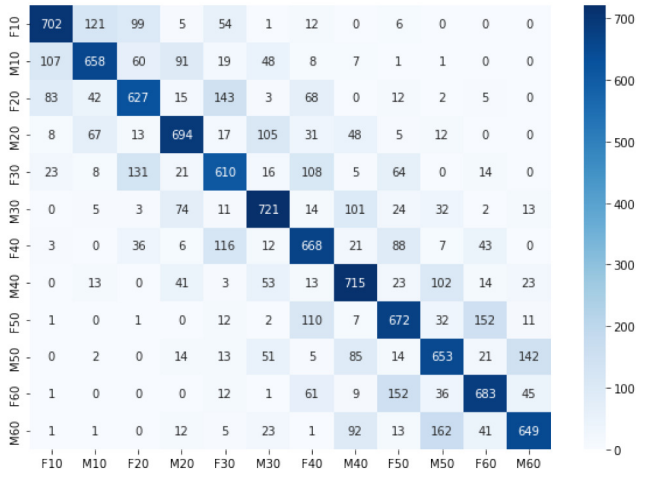

Fig. 9. Confusion matrix of age classification into 12 classes under known gender information, obtained on the "Common Voice" dataset using the ResNet34 network

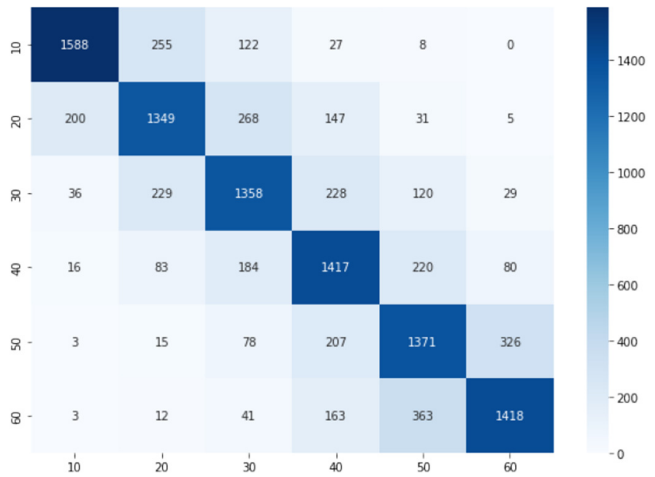

Fig. 10. Confusion matrix of age classification for separate networks, each created for 6 classes, under known gender information (presenting combined results of the two networks), obtained on the "Common Voice" dataset using the ResNet34 network

speaker into one of 6 age classes. In this solution, the performance of both model types is again increased - ResNet34 achieves 70,8% and x-vectors achieves 59,6%.

D. Grouping of age classes

An obvious way further to improve the results is to decrease the number of classes, by grouping difficult-to-distinguish classes. In many applications, the speaker's age classification problem can be reduced to three age classes: teenagers (class "10"), adults (classes "20", "30", "40") and senior adults (classes "50", "60"). In this case, the x-vector-based solution has shown increased accuracy from 59.6% (for 6 classes) to 77.3% (for 3 classes), and the ResNet34 - from 70.8% (6 classes) to 86,1% (3 classes).

E. Confusion matrices

The above results can be justified, when confusion matrices are studied. Such a matrix for results on 12 classes, obtained with ResNet34, is presented in Figure 9. Already a general view leads to the conclusion, that main errors happen between neighbour age groups, as the "errors" (represented by big numbers outside the diagonal) concentrate in the direct neighborhood of the diagonal axis. For example, a misclassification of a teenager as a 60+ senior is practically excluded. Similar observation comes from an error matrix created for a 6-class problem, with gender information, were the results for the same ages classes of man and women are combined (Figure 10).

F. Real gender and age classification

In the second series of experiments, we simulated the realistic case of imperfect gender information. We divided 25000 recordings into three sets: a) for training the age model (now for 8 age classes), b) for training the gender model, c) for coupling age and gender into 16 age/gender classes, i.e., class 10_male, 20_male, ... , 80_male, 10_female, 20_female, ... , 80_female. The results on the test subsets, obtained for models created after 30 learning epochs, are presented by two confusion matrices, given in Figures 11 and 12.

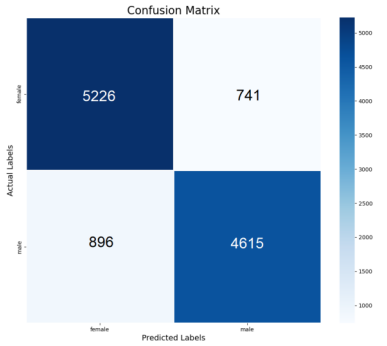

Fig. 11. Confusion matrix of gender classification in the gender-and-age experiment

The gender success rate (Recall) for females is 87.58% and for males – 83.74%. Thus, the weighted gender-average success rate is 85.74%. The overall age success rate (average Recall) of the combined gender–and–age classifier is 41%, which applies to a 16–classes problem. Interestingly, an increasing age is correlated with growing success rate. The oldest age groups of people 70+ and 80+ are recognized very well.

V. Conclusions

The proposed three-stage approach to speaker's age classification has been trained and tested on a large dataset containing recordings of a high number of speakers. Presented results have shown the positive impact of perfect gender information onto age classification performance. We have also received realistic performance scores under non-perfect gender information. The "heavy-weight" ResNet34 network models

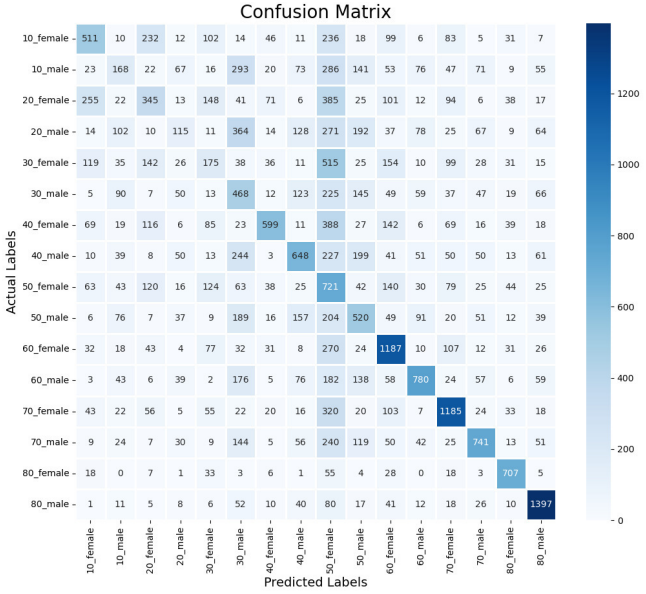

Fig. 12. Confusion matrix of age classification into 16 classes (8 age groups × 2 gender) with imperfect, realistic gender information

has clearly outperformed the "x-vectors" model, which is a popular DNN approach to speaker recognition.

REFERENCES

[1] E. D.Mysak and T. Hanley, "Aging processes in speech: Pitch and duration characteristics", *Journal of Gerontology,* vol. 13, 1958, no. 3, pp. 309–313, https://doi.org/10.1093/GERONJ/13.3.309.

[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", *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP),* vol. 1, 2002, pp. 137–140, https://doi.org/10.1109/ICASSP.2002.5743673.

[3] C. Müller, F. Wittig and J. Baus, "Exploiting speech for recognizing elderly users to respond to their special needs", *Interspeech, Proc. 8th Eur. Conf. Speech Commun. Technol.,* 2003, pp. 1305–1308, https://doi.org/10.21437/Eurospeech.2003-413.

[4] U. Kamath, J. Liu and J. Whitaker, *Deep Learning for NLP and Speech Recognition,* Springer Nature Switzerland AG, Cham, 2019, https://doi.org/10.1007/978-3-030-14596-5.

[5] P. G. Shivakumar, M. Li, V. Dhandhania and S. S. Narayanan, "Simplified and supervised i-vector modeling for speaker age regression", *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2014, pp. 4833–4837, https://doi.org/10.1109/ICASSP.2014.6854520.

[6] M. H. Bahari, M. McLaren, H. Van Hamme and D. A. van Leeuwen, "Speaker age estimation using i-vectors", *Engineering Applications of Artificial Intelligence,* vol. 34, 2014, pp. 99–108, https://doi.org/10.1016/j.engappai.2014.05.003.

[7] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, "Front-End Factor Analysis for Speaker Verification", *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 19, 2011, no. 4, pp. 788–798, https://doi.org/10.1109/TASL.2010.2064307.

[8] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition", *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2018, pp. 5329–5333. https://doi.org/10.1109/ICASSP.2018.8461375.

[9] B. Gu, W. Guo, L. Dai and J. Du, "An Adaptive X-vector Model for Text-independent Speaker Verification", 2020. https://doi.org/10.48550/ARXIV.2002.06049.

[10] L. Zhou, M.Wang, Y. Qian, H. Luo, H. Li and X. Lin, "Text-independent Speaker Recognition Based on X-vector", *2022 7th International Conference on Signal and Image Processing (ICSIP),* 2022, pp. 121–125. https://doi.org/10.1109/ICSIP55141.2022.9887021.

[11] R. Zazo, P. Sankar Nidadavolu, N. Chen, J. Gonzalez-Rodriguez and N. Dehak, "Age Estimation in Short Speech Utterances Based on LSTM Recurrent Neural Networks", *IEEE Access*, vol. 6, pp. 22524–22530, 2018, https://doi.org/10.1109/ACCESS.2018.2816163.

[12] A. I. Mansour and S. S. Abu-Naser, "Classification of Age and Gender Using Resnet - Deep Learning", *International Journal of Academic Engineering Research (IJAER),* vol. 6, 2022, no. 8, 20–29, https://philpapers.org/rec/MANCOA-4/.

[13] R. Ardila, M. Branson, K. Davis et al., "Common Voice: A Massively-Multilingual Speech Corpus", *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020),* 2020, pp. 4218–4222, https://aclanthology.org/2020.lrec-1.520/.

[14] N. Tawara, A. Ogawa, Y. Kitagishi and H. Kamiyama, "Age-VOX-Celeb: Multi-Modal Corpus for Facial and Speech Estimation", *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 2021, pp. 6963–6967, https://doi.org/10.1109/ICASSP39728.2021.9414272.

[15] LibRosa, "Audio and music processing in Python", https://librosa.org/

[16] C. Li, X.Ma, B. Jiang et al., *Deep Speaker: an End-to-End Neural Speaker Embedding System*, May 2017. arXiv:1705.02304 [cs.CL] (or arXiv:1705.02304v1 [cs.CL]) https://doi.org/10.48550/arXiv.1705.02304.

[17] S. Hourri and J. Kharroubi, "A deep learning approach for speaker recognition", *International Journal of Speech Technology*, vol. 23, 2020, pp. 123–131, https://doi.org/10.1007/s10772-019-09665-y.