Impact of Pathologies on Automatic Age Estimation

Leo Schwinn[1], Tino Haderlein[1], Elmar Nöth[1], Andreas Maier[1]

[1]*Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Lehrstuhl für Informatik 5 (Mustererkennung), 91058, Erlangen, E-Mail: leo.schwinn@fau.de*

Introduction

Automatic Age estimation using speech is a challenging problem. Among other things, the many influences that lead to a change in the voice make it difficult to estimate the exact age. Examples are the microphone used, the distance to this microphone or the sex of the speaker. Stable results can be achieved by extracting Mel-Frequency Cepstral Coefficients (MFCCs) from the speech signal and processing them into i-vectors, classification and regression tools such as Support Vector Regression (SVR) can then be used for the age estimation from these features. An additional factor influencing the age estimation are speech or voice disorders of a speaker. Rarely is this impact assessed, resulting in systems that are not tailored to the needs of pathological speakers like Siri or Amazon echo. Another example are companies which use automatic age estimation to forward calls to persons of the same age as the caller. These systems are also adapted to the characteristics of healthy speakers. This paper examines the impacts of such pathologies on age estimation and assess the possibility of reducing their influence by using the Word Accuracy (WA) of the speakers. This measure gives information about the intelligibility of a speaker and should help to reduce the variance of the extracted features. The features with low variance should provide more stable results in the age estimation.

To achieve this, we compare the results of an age estimation for four different groups of speakers. All speakers at once, only pathological speakers, only healthy speakers and training the SVR with healthy speakers while testing with pathological speakers. Each of these groups are further separated by the WA of the speakers.

Materials and Methods

Test Data

All used speech data are based on screen-read recordings of the German version of "The North Wind and the Sun". The recordings were made using the PEAKS software [1]. No personal data of the speakers are available, each speaker can only be identified by a random number to guarantee anonymity.

The PEAKS software can be used as an online tool for medical studies. Since 2009, a total of 4987 voice recordings have been made in 28 different studies. Speakers were removed from the data unless at least 9 other speakers of the same age were present. The remaining speakers are between 11 and 50 years old. Of a total of 2672 speakers, 959 are female and 1713 male. 909 have a speech or voice disorder which was diagnosed by the doctor doing the respective study. The pathologies were grouped as shown in Figure 1.

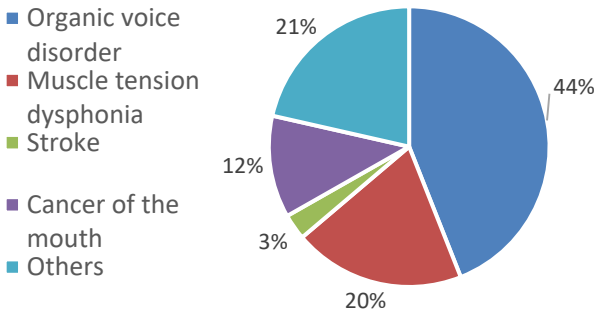

Figure 1: Pathology distribution of the speakers

The distribution of the chronological age of all speakers is shown in Figure 2.

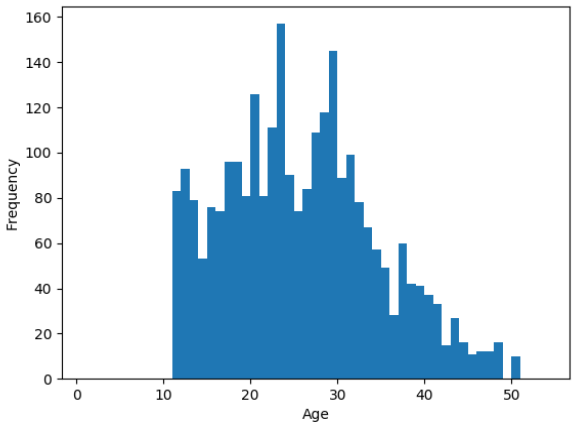

Figure 2: Age distribution of the speakers

The recorded text consists of 108 words, the vocabulary size is 71. However, the age estimation is done using a text independent framework. The records were made with a sampling frequency of 16 kHz [1].

Feature extraction

For the step to extract characteristics, MFCCs are calculated for each utterance. They are created using Hamming windows with a width of 25 ms and a step width of 10 ms. These contain the information relevant for the age estimation and guarantee a compact size of the feature vector. This feature vector consists of 36 coefficients: 12 static coefficients, each supplemented by the first and second derivative to better model the dynamic properties of the signal. The total amount of MFCCs varies for each utterance and is later normalized for the i-vector extraction.

The next step is to apply the Cepstral Mean and Variance Normalization (CMVN) to the resulting MFCCs. This is done to generate a new set of characteristics where the mean is zero and the variance is one [2]. Finally, based on these normalized characteristics, we can train our i-vector extractor to extract i-vectors with a dimensionality of 400. The dimension of the i-vectors was chosen due to the results in [2]. These features were selected because they proved to be suitable for age estimation [3, 4]. The fundamental frequency was initially included in the i-vectors as feature but was later excluded due to an overall bad influence on the performance of the age estimation. All features were computed with the Kaldi toolkit [5].

Age estimation using support vector regression

The age of the speakers is estimated by an SVR in the open source machine learning toolkit WEKA [6]. For the underlying Support Vector Machine (SVM), a "NormalizedPolyKernel" is used because it could obtain the best results compared to other kernels. The complexity constant C was set to 1 after evaluating the results of changing C by powers of 10, like the method proposed in [7]. The test is performed with a 10-fold cross-validation. The results of the regression are evaluated with the Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE) of the predictions compared to the chronological age. For comparison, the standard deviation of the age is also provided (Std). In addition, the Pearson correlation r(CA-PA) between the chronological age and the predicted age of the speakers is given. The best result is compared with that of an optimistic trivial estimator. The optimistic trivial estimator always estimates the mean value of the age distribution as the age. Only one utterance of each speaker was used for the age estimation.

Age estimation including word accuracy

The regression step is done multiple times with different subsets of the original data. These are generated by thresholding the speakers with their WA. The WA serves as a measure of how many words of an utterance could be recognized and thus gives us information about the intelligibility of a speaker. It was calculated automatically as part of the PEAKS software [1]. The correlation r-(WA-CA) of the chronological of the speakers and the WA is also part of the evaluation.

Feature and data verification

The quality of the extracted features was verified with a speaker identification. For this purpose, a total of 530 utterances were used. These are from 500 different speakers of which 30 each have 2 utterances. These additional 2nd utterances are now compared with the remaining 500. The aim is to find the other utterance of the same speaker for each of the 30 utterances. A nearest neighbor classifier is used to assign the utterances. The cosine similarity is used as a measure of similarity. A total of 29 of the 30 utterances could be assigned to the correct speaker. This indicates that the features contain relevant speaker information.

Results

The results are divided into 4 tests on different groups of speakers and one gender independency verification of the age estimation.

All Speakers

Table 1 shows how the data are distributed for different WA thresholds. Speakers, female, male represent the number of total, female and male speakers. Pat. and age range represent the percentage of pathological speakers and the age range of the speakers. It should be noted that the distribution of speakers regarding their gender, as well as the total amount of speakers, varies with the threshold value of the WA.

Table 1: Data distribution

Min. WA	speakers	Female [%]	Male [%]	pat. [%]	age range
None	2672	36	64	34	11–50
50	1756	45	55	32	11–50
70	561	47	53	28	11–47

Figure 3 shows the regression results based on the SVR described above without excluding speakers based on their WA.

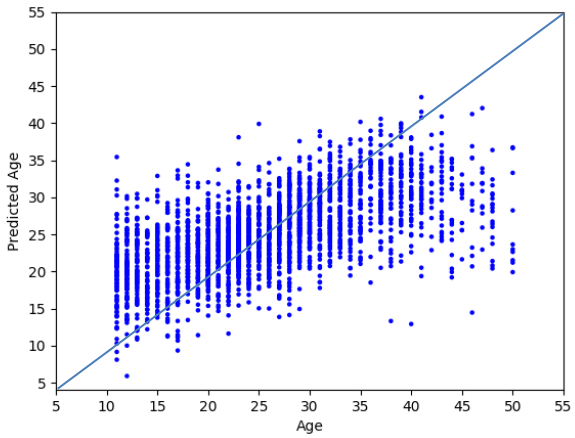

Figure 3: Result of SVR, without regarding the WA threshold

The next two figures show the results of the regression, excluding speakers based on the WA threshold.

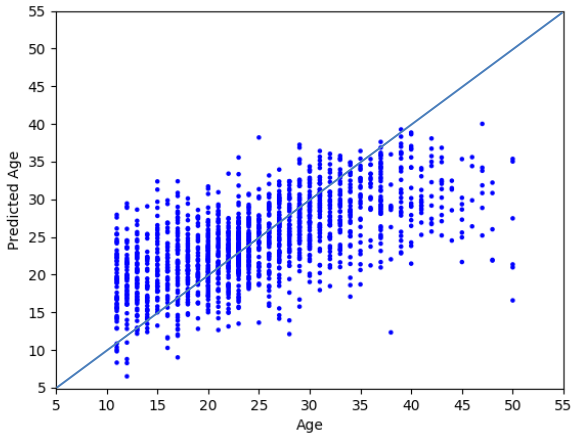

Figure 4: Results of SVR, WA ≥ 50

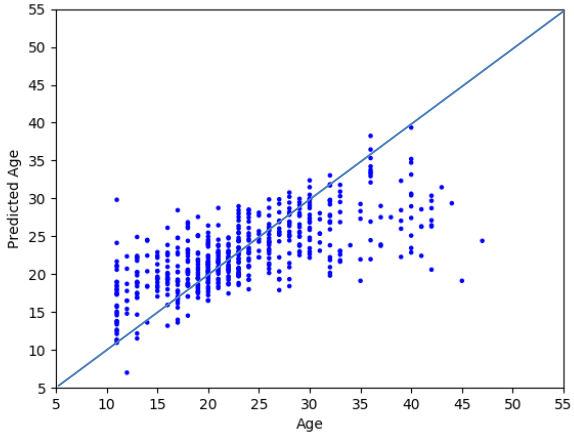

Figure 5: Results of SVR, WA ≥ 70

Erreur ! Source du renvoi introuvable. gives an overview of the error quantities and the correlation depending on the WA.

Table 2: Error quantities and correlation depending on the WA threshold including all speakers

Min. WA	RMSE	MAE	Std	r(WA-CA)	r(CA-PA)
None	7.0	5.2	8.8	-0.10	0.61
50	6.6	4.9	8.6	-0.15	0.64
70	5.5	4.0	7.6	-0.13	0.70

Healthy Speakers

Looking at the age estimation using only the feature vectors generated from speakers without speech or voice disorders shows similar results. The quality of the age estimation still improves with the WA threshold. For a larger group size, the results of the age estimation without pathological speakers still provides better results than the age estimation with pathological speakers with a smaller group size and similar or higher Std. The results are shown in Table 3.

Table 3: Data distribution of non-pathological speakers

Min. WA	speakers	Female [%]	Male [%]	age range
None	1763	36	64	11–50
50	1096	44	56	11–50
70	403	49	51	11–47

Pathological Speakers

The results of using only speakers with pathologies indicate the same by showing the worst correlations r(WA-CA) for all WA thresholds in comparison to the other groups.

Table 4: Error quantities and correlation depending on the WA threshold of only pathological speakers

Min. WA	RMSE	MAE	Std	r(WA-CA)	r(CA-PA)
None	7.7	5.8	8.8	-0.06	0.50
50	6.8	5.1	8.6	-0.17	0.56
70	5.5	3.9	6.2	-0.04	0.56

Table 5: Data distribution of only pathological speakers

Min. WA	speakers	Female [%]	Male [%]	age range
None	909	36	64	11–50
50	560	41	59	11–50
70	158	42	58	11–47

Split training and testing into pathological and healthy speakers

Furthermore, testing the SVR with only pathological speakers while training with only non-pathological speakers shows even worse results. Those point to large differences between the extracted features of pathological and non-pathological speakers. Still, increasing the WA threshold also increased the correlation between computed and chronological age. No tendency could be observed that the age of pathological speakers was consistently estimated to be older or younger. The exact results are shown in Table 6:

Table 6: Error quantities and correlation depending on the WA of a regression trained with only non-pathological speakers and tested with exclusively pathological ones.

Min. WA	RMSE	MAE	Std	r(WA-CA)	r(CA-PA)
None	8.4	6.4	8.8	-0.06	0.38
50	7.7	5.8	8.6	-0.08	0.41
70	6.1	4.8	6.2	-0.1	0.44

Gender independency verification

Looking at the different sexes individually, the regression shows no bias in the age estimation as can be seen in **Erreur ! Source du renvoi introuvable.**. The regression lines of the female and male speakers show the same slope of 0.29 and the same offset of 2.61 years, having the same mean age of 24 years for both groups. Thus, the age estimation is gender independent.

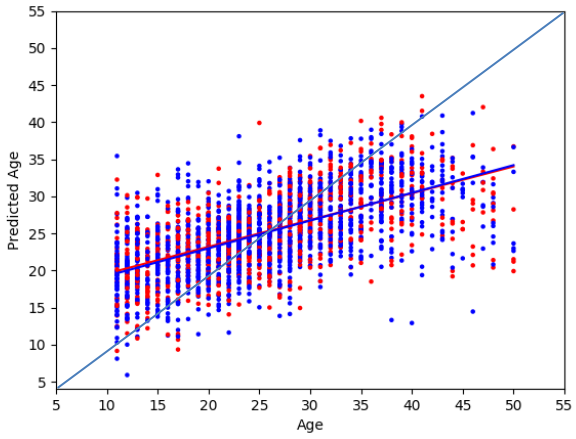

Figure 6: Result of the SVR without regarding the WA. Red dots represent female speakers, blue dots male speakers.

Discussion

From Table 2, 3, 5 and 7 it can be seen that the limitation of WA has an influence on the quality of the regression, which is linked to the degree of speech intelligibility in our data set [8].

Increasing the WA threshold improves the result of our age estimator which can be seen by the lower MAE and RMSE values compared to the Std. The correlation between predicted and chronological age also increased with the WA threshold. The r(WA-CA) shows no correlation between chronological age and WA. The better results from high WA speakers doesn't result from a correlation between WA and age but rather from the better features extracted from high WA recordings.

Conclusion and Outlook

Summarizing the results, training a regressor with no information about pathological speakers makes it unable to correctly predict the age of such speakers. This is shown with the error quantities in Table . The results can be improved by including pathological speakers in the training data and improving the intelligibility of the utterances by using only high WA recordings. Applying such measures can lead to a stable age estimation for pathological and healthy speakers.

To further improve the age estimation for pathological and healthy speakers other prosodic features can be considered. This could help to capture also long-term temporal differences in the age signal as MFCCs capture only short-term temporal differences. This additional information might help to find other influences of the pathologies on the age estimation. Such features could be fundamental frequency, average jitter in voiced frames, percentage of voiced frames and others.

Another approach would be to use a Convolutional Neural Network for feature extraction and a Fully Connected Layer with a linear activation function for the regression. Different filter sizes can be used to extract long-term and short-term temporal differences in the speech signal. This could lead to an end to end age estimation pipeline without handcrafted features. For this approach the signal can be represented by a spectrogram or scalogram. Similar approaches have been made for other application using speech and could possible also applied here.

References

[1] Maier A., Haderlein T., Eysholdt U., Rosanowski F., Batliner A., Schuster M. and Nöth E.: PEAKS – A system for the automatic evaluation of voice and speech disorders, Speech Communication 51 (2009), 425-437

[2] Silnova A. (2015). Exploring i-Vector Based Speaker Age Estimation, Master's thesis, University of Eastern Finland, Joensuu, Finland

[3] Bahari M. H., M. McLaren, Hamme H. V., and van Leeuwen D., Age estimation from telephone speech using i-vectors, in Proceedings of Interspeech (2012), 506-509

[4] Bahari M. H., McLaren M., Hamme H. V., and van Leeuwen D., Speaker age estimation using i-vectors, Engineering Applications of Artificial Intelligence (2014), vol. 34, 99–108

[5] Povey D., Ghoshal A., Boulianne G., Burget L., Glembek O., Goel N., Hannemann M., Motlicek P., Qian Y., Schwarz P., Silovsky J., Stemmer G., and Vesely K.: The Kaldi Speech Recognition Toolkit, Technical Report, Idiap-RR-04-2012, Idiap Research Institute, Martigny, Switzerland, 2012

[6] Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P. and I. H. Witten: The WEKA data mining software: an update, in: ACM SIGKDD Explorations Newsletter 11 (2009), 10-18

[7] Haderlein T., Middag C., Hönig F., Martens J.P., Döllinger M., Schützenberger A. and Nöth E.: Language-Independent Age Estimation from Speech using Phonological and Phonemic Features, Proc. Text, Speech and Dialogue (TSD 2015), vol. 9302 of LNAI, Springer International Publishing Switzerland, Cham, 165-173

[8] Haderlein T., Schützenberger A., Döllinger M. and Nöth E.: Subtext Word Accuracy and Prosodic Features for Automatic Intelligibility Assessment, Proc, Text, Speech and Dialogue; 21st International Conference (TSD 2018) vol. 11107 LNAI of Springer Nature Switzerland, 473-481