A Comparison of Human and Machine Estimation of Speaker Age

Mark Huckvale, Aimee Webb

Speech, Hearing and Phonetic Sciences
University College London, London, U.K.

m.huckvale@ucl.ac.uk, aimee.webb.11@ucl.ac.uk

Abstract. The estimation of the age of a speaker from his or her voice has both forensic and commercial applications. Previous studies have shown that human listeners are able to estimate the age of a speaker to within 10 years on average, while recent machine age estimation systems seem to show superior performance with average errors as low as 6 years. However the machine studies have used highly non-uniform test sets, for which knowledge of the age distribution offers considerable advantage to the system. In this study we compare human and machine performance on the same test data chosen to be uniformly distributed in age. We show that in this case human and machine accuracy is more similar with average errors of 9.8 and 8.6 years respectively, although if panels of listeners are consulted, human accuracy can be improved to a value closer to 7.5 years. Both human and machines have difficulty in accurately predicting the ages of older speakers.

Keywords: speaker profiling, speaker age prediction, computational paralinguistics.

1 Introduction

The estimation of the age of a speaker from an analysis of his or her voice has forensic applications – for example the profiling of perpetrators of crimes [1], commercial applications – for example targeted advertising, and technological applications – for example adaptation of a speech recognition system to a speaker [2].

Many previous studies have looked at the performance of both human listeners and machine-learning systems for the estimation of age from speech. Unfortunately, variations in the data set, task and performance metric make these studies hard to compare. In our work we take the view that the natural task should be numerical estimation of the age of the speaker, and the natural performance metric should be the mean absolute error (MAE) of estimation. The MAE answers the question "how close is the average estimate to the actual age?"

A recent review of previous studies on human listener judgments of speaker age may be found in [3]. Of the studies reported which used numerical age estimation and

MAE, most seem to suggest human performance has an MAE of about 10 years. Table 1 provides a summary.

Table 1. Previous studies on human listener age estimation

Study	MAE (yr)	Notes
Braun et al, 1999 [4]	10.5	German speakers & listeners
Braun et al, 1999 [4]	8.5	Italian speakers & listeners
Krauss et al, 2002 [5]	7.1	Limited age-range
Amilon et al, 2009 [6]	9.7	
Moyse et al, 2014 [7]	10.8	

There have also been many studies in the machine prediction of speaker age from speech, see [8] for a review of the state of the art. The studies also vary greatly in terms of the data set, audio quality, audio duration, audio feature set, recognition task and machine learning approach taken. Machine learning methods have included support vector machines, Gaussian mixture models (GMM), GMM supervectors, i-vectors and phoneme recognisers. [9] provides system design and performance figures for a range of contemporary approaches together with a fusion of systems for age estimation of very short speech excerpts (<2s). The different approaches only varied by a few percentage points (43.1-47.5% age categories correctly identified) suggesting that the choice of machine learning algorithm is not a critical factor.

The results of some machine studies that addressed the problem of numerical age estimation evaluated with MAE are shown in Table 2.

Table 2. Previous studies on machine age estimation

Study	MAE (yr)	Notes
Bocklet et al, 2008 [10]	0.8	Children 7-10 yrs
Feld et al, 2009 [11]	7.2-12.8	Same & cross-language
Doby et al, 2011 [12]	9.29-10.00	Depending on gender
Bahari et al, 2011 [13]	7.48	Null model = 8.88
Bahari et al, 2012 [14]	7.9	
Bahari et al, 2014 [8]	6.08	Null model = 10.3

The best performing system on adult speech described in [8] used the i-vector approach followed by support vector regression and demonstrated an MAE of 6.1 years. While at first glance this looks considerably better than the MAE figures quoted for human performance, it is important to note that the test data used in this study had a non-uniform age distribution, with significantly more speakers in the 20-29 age band than in other bands. This uneven distribution means that even a null model which always predicted the mean age of the training speakers would show an MAE of 10.6 years for female speakers and 10.1 years for male speakers. The superiority of the machine system might therefore have arisen from the unfair knowledge it had of the prior age distribution. Since all machine systems in Table 2 may have exploited a

prior on the test speaker age, this makes it impossible to compare any of them fairly with the human listeners, who were not given that information.

The goals of this study are to make a fair comparison between human and machine speaker age estimation. This will be done by: (i) comparing human and machines on same test data, (ii) comparing them on the on same task – numerical age estimation, (iii) evaluating both using the same performance metric – MAE, and (iv) removing any advantage of knowing a prior on the test set by using a uniform test age distribution.

We describe the data set used for the task, results of a human listening task and results of a machine age estimation system constructed to be similar to the best performing systems in Table 2.

2 Speech Corpus

The work described here uses the Accents of the British Isles corpus (version 2) available from The Speech Ark [15]. The ABI-2 corpus consists of recordings of 262 speakers covering 13 accent areas of the British Isles. Each speaker is recorded reading a range of English language materials; although for this work we used only the first part of the "accent diagnostic" passage which has a median duration of 39.2s. The recordings are supplied as wide bandwidth audio of good quality, recorded using a close-talking microphone at 22050 samples/sec.

The corpus was divided into a test set containing 52 speakers, and a training set of the remaining 210 speakers. The test set was chosen to have equal representation of men and women for all 5-year age bands between 15 and 80. Figure 1 shows the age distribution by gender for the test and training sets.

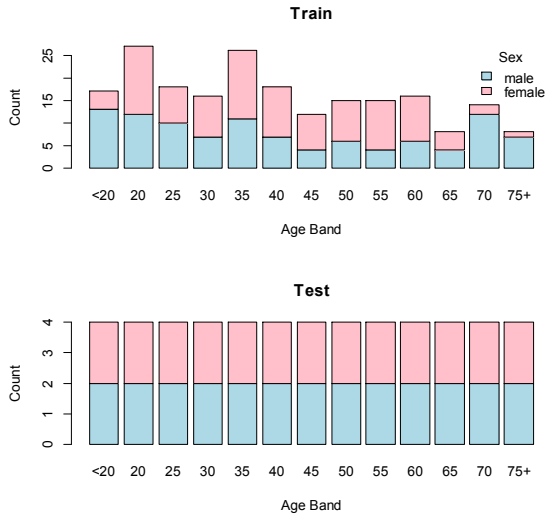

Fig. 1. Age and gender distribution for the train and test sets.

The mean age of the training set was 42.6 years. Used as a null prediction for the test set this value would score a mean absolute error of 16.7 years.

3 Human Prediction Performance

To obtain human age prediction performance a web-based data collection protocol was used. Listeners were able to listen to each test recording then make an age estimation using a sliding scale between 15 and 80. Estimates were recorded as whole numbers of years. Recordings were presented in a random order different for each listener. Listeners could make their age estimate at any time while the recording was playing, or could listen to the audio multiple times. Listeners conducted the test in their own homes, but were asked to listen over headphones. The web interface may be seen in Figure 2.

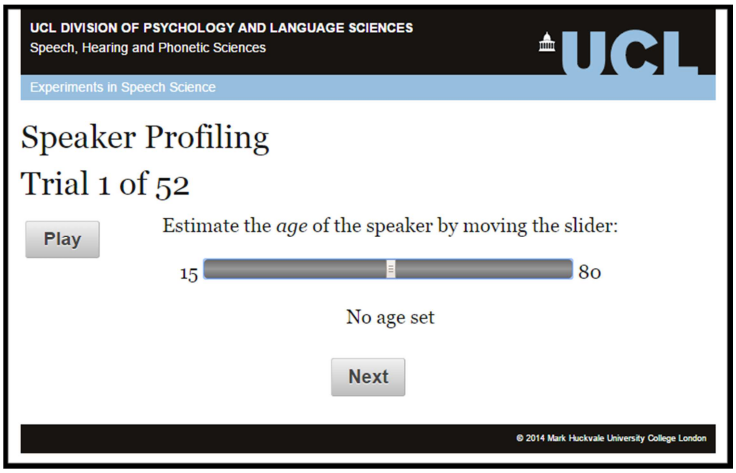

Fig. 2. Web experiment interface.

An attempt was made to recruit listeners over a range of ages and genders, although the balance was not perfect. In all, 36 native English listeners completed the test; Table 3 shows their distribution by age and gender.

Table 3. Distribution of listeners by age and gender

Number	20-29	30-39	40-49	50-59	60-69
Male	4	4	3	3	2
Female	5	4	3	6	2

The raw age predictions are plotted against the true speaker ages in Figure 3. The line of best fit has a slope of 0.68 and an intercept of 12.7years. The correlation coef-

ficient is 0.759 and the mean absolute error (MAE) of prediction is 9.79 years (male speakers only 10.1, female speakers only 9.51).

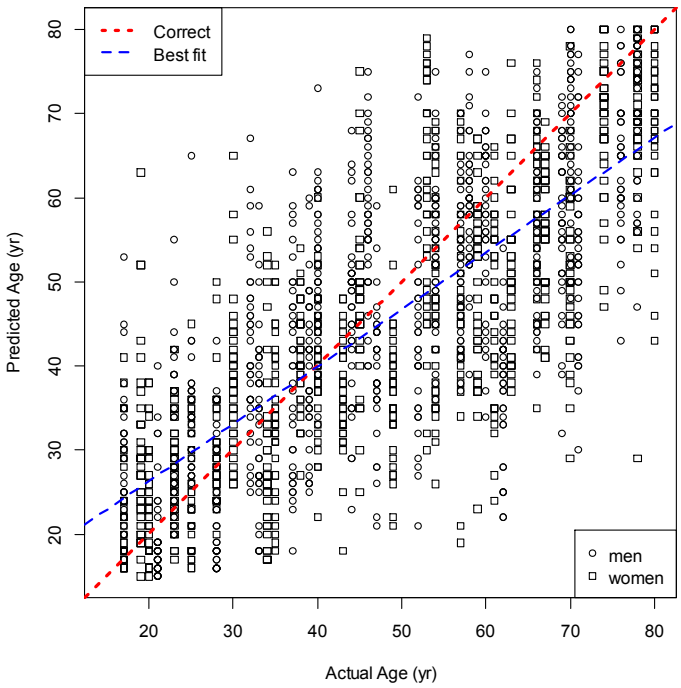

Fig. 3. Age predictions of 52 test speakers by 36 listeners.

The MAE as a function of the age and sex of the speaker is shown in Table 4, and MAE as a function of the age and sex of the listener is shown in Table 5.

Table 4. Mean Absolute Error of prediction as a function of age and sex of the speaker

MAE(yr)	**20-29**	**30-39**	**40-49**	**50-59**	**60-69**	**70-79**
Male	7.42	9.32	10.52	7.99	13.85	11.14
Female	5.63	8.00	8.71	12.07	12.10	11.30

Table 5. Mean Absolute Error of prediction as a function of age and sex of the listener

MAE (yr)	**20-29**	**30-39**	**40-49**	**50-59**	**60-69**
Male	8.34	8.22	11.36	10.01	13.24
Female	9.95	9.39	10.57	9.38	10.10

Generalised linear mixed-effects models of the predictions were estimated using Markov chain Monte Carlo techniques with the MCMCglmm package [16]. The

models were used to determine if the absolute error in age prediction was affected by the sex or age of the speaker, or the sex or age of the listener.

The speaker model was trained with the identity of the listener as a random factor. The sex of the speaker was found not to have significant effect. The age-band of the speaker did have significant effect, with the ages of speakers in the 20-29 age band being significantly better estimated than the other bands.

The listener model was trained with the identity of the speakers as a random factor. The sex of the listener was found not to have a significant effect. The age-band of the listener did have significant effect with listeners in age-bands 40-49 and 60-69 giving significantly worse predictions than listeners in the 20-29 age band.

The fact that the line of best fit of the estimates does not have a gradient of one might be due to the limited range of the age slider in the web task creating floor and ceiling effects. Listeners were unable to estimate ages lower than 15 years or greater than 80 years even if these would in fact have been in error.

A distribution of the age prediction errors is shown in Figure 4. It may be seen that errors are approximately symmetric about zero. This suggests that an averaging of age predictions over listeners would provide a better age estimate.

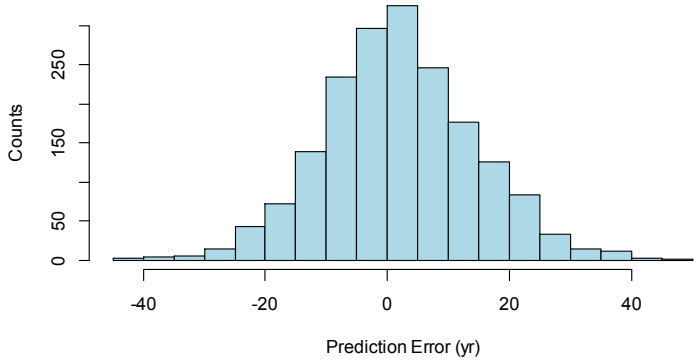

Fig. 4. Distribution of age prediction errors by human listeners.

To estimate the benefit of averaging across listeners, panels of size 2 to 12 were built post hoc from random selections of listeners. The average MAE calculated over 50 random panels of each size is plotted in Figure 5. It is seen that considerable advantage may be had by consulting a listener panel, with a panel of 10 listeners for example having an MAE of 7.41 years.

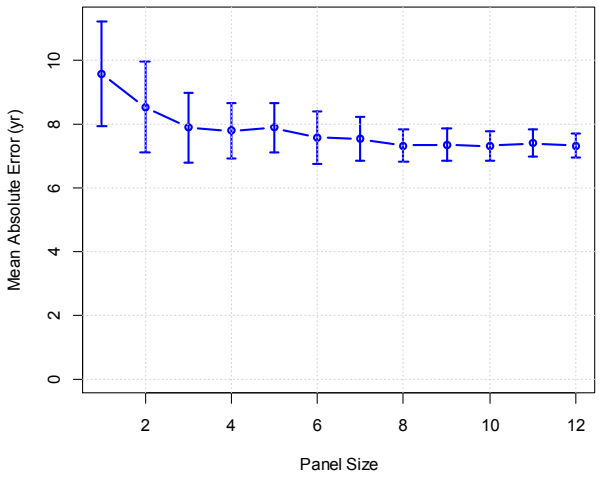

Fig. 5. Distribution of mean absolute error of age prediction by listener panel size. Average over 50 random panels. Bars show 1 s.d.

4 Machine Prediction Performance

4.1 Feature analysis

Following on from the acoustic feature analysis used in the Interspeech Computational Paralinguistics challenges, we have used the OpenSMILE toolkit [17] to generate a large feature vector for each audio recording. The specific set of parameters was those used for the 2014 challenge [18]. This feature set comprises 65 low-level descriptors which are extracted from short-term windows on the signal. These describe speech signal properties such as energy, spectral envelope, pitch and voice quality. The descriptors are then summarized over each file using a large number of statistical measures such as means, medians, quantiles, differences, and so on. The output is a vector of 6373 features for each file.

4.2 Machine Learning

The method chosen for learning the prediction model was Support Vector Regression (SVR) [19] as used by previous authors [10,11,12,13]. The "e1071" package for the "R" statistics library was the chosen implementation [20]. In support vector regression a subset of the training vectors are chosen to represent the optimal regression hyperplane.

To reduce the training complexity, a feature selection process was implemented. Only features which had an absolute value of correlation greater than an arbitrary threshold of 0.1 with the age of the speaker were passed to SVR. This selection was

made on the training set only and left 2538 features. Performance was not strongly affected by the choice of this threshold providing enough features were included.

At the front end of the SVR, a radial-basis function kernel is applied – this provides an additional tunable non-linearity applied to the feature values. Also the SVR algorithm applies a feature normalization step to ensure all features have a similar dynamic range.

Optimal control parameters were found using a cross-validation procedure on subsets of the training data only. The optimal parameters were: C=8, gamma=0.25/number-of-features, epsilon=0.1.

Separate SVR systems were trained for male and female speakers as in [8].

4.3 Raw Prediction Performance

The raw prediction performance of SVR is shown in Figure 6. The line of best fit has a slope of 0.53 and an intercept of 18.9 years. The correlation coefficient is 0.82, and the mean absolute error is 9.13 years (male speakers only 7.98, female speakers only 10.29). A gender independent model gave a correlation of 0.81 and an MAE of 9.18 years. As mentioned previously, a null model has an MAE of 16.7 years.

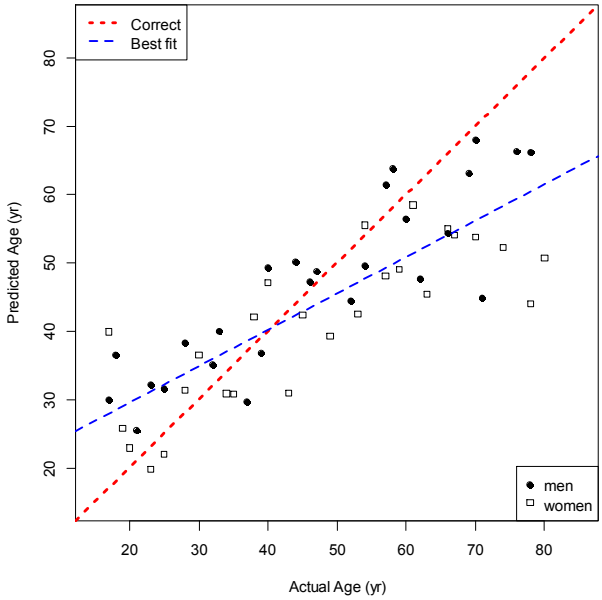

Fig. 6. Machine prediction of age using SVR.

Like the human listener predictions, the machine predictions also overestimate the ages of younger speakers and underestimate the ages of older speakers. Table 6 shows the MAE as a function of the age and sex of the speaker. It is noticeable that the greatest estimation errors are with the speakers older than 50. In the next section we

try to rebalance the training set to investigate whether this bias is just a reflection of the uneven age distribution in the training data.

Table 6. Mean Absolute Error of prediction as a function of age and sex of the speaker

MAE(yr)	**20-29**	**30-39**	**40-49**	**50-59**	**60-69**	**70-79**
Male	7.64	4.89	4.68	5.54	8.87	12.40
Female	3.12	4.46	7.86	7.72	11.00	23.97

4.4 Effect of balancing the training set

Since our original motivation was to make a fair comparison with human listeners on a balanced test set, it may be that we have now disadvantaged the machine system by only providing an unbalanced training set. The machine predictions are worse for the older speakers (Table 6) who are under-represented in the training data (Figure 1). The training of predictive models under circumstances of imbalanced data is an ongoing area of research both for classification and regression tasks [21]. To explore the effect of imbalanced training data in this task, we explore the synthetic creation of training data samples using a variation of the SMOTE algorithm [22] designed for regression [23].

Here we present results in which we artificially generate additional training samples from linear interpolations between existing vectors. We even out the number of samples for male and female speakers and boost the number of training samples for speakers of ages >50 years. Each new sample is generated from two randomly-chosen instances of the same sex and age band by choosing a random point along the interpolation joining the two vectors. The new age value is interpolated from the ages of the two samples at the same fraction. In total a further 271 vectors were added.

Figure 7 shows the distribution of training samples by decade before and after balancing.

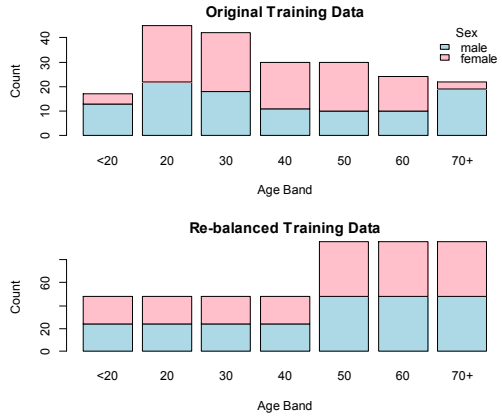

Fig. 7. Results of boosting the frequency of the older speakers in the training data.

A new SVR model was trained on the re-balanced data. 3378 features were selected using the same correlation threshold. Cross validation on the training data suggested the best control parameters were now C=32, gamma=0.125/number of features, epsilon=0.001.

Figure 8 shows the age predictions for the test set after training with the re-balanced training data. The line of best fit had a slope of 0.554 and an intercept of 17.8 years. The correlation was 0.852 and the MAE 8.64 years (male speakers only 7.87, female speakers only 9.42). A gender independent model gave a correlation of 0.81 and an MAE of 9.49 years. Table 7 shows the mean absolute error of prediction as function of the age and sex of the speaker.

While some performance improvements are seen in comparison to results with the original training set, overall the improvement is small. It may be that in this task, the SVR model does not gain any useful information from the synthetic samples.

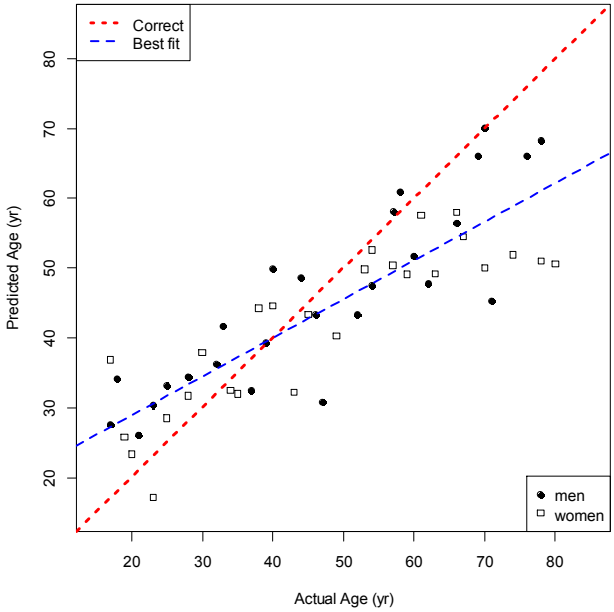

Fig. 8. Machine prediction of age using SVR trained on rebalanced data.

Table 7. Mean Absolute Error of prediction as a function of age and sex of the speaker using SVR trained on rebalanced data

MAE(yr)	**20-29**	**30-39**	**40-49**	**50-59**	**60-69**	**70-79**
Male	6.81	4.41	8.32	4.79	8.74	11.38
Female	4.10	4.68	6.45	5.27	9.45	23.04

5 Discussion

In this study we have made direct comparison between human listeners and machine learning on the problem of speaker age estimation. By nullifying any advantage a machine system may have by knowing about the prior distribution of test speakers, we have shown that humans and machines are more similar in estimation performance compared to results published in previous studies.

Nevertheless the machine system showed a slight advantage. The best machine performance had an MAE of 8.64 years, while the human listeners had an MAE of 9.79 years. The machine system was able to outperform two-thirds (25/36) of the human listeners. However even a panel of 2 listeners had superior average performance than the machine system in this particular experiment.

Interestingly, both human and machine had problems with the extremes of the age range, both showing lines of best fit with slopes significantly less than unity. We showed that boosting the number of older speakers in the training set had very little effect, perhaps because the SVR model did not extract any more information from the interpolated samples than it could extract from the original samples. The difficulty of predicting the ages of older speakers may be due to some inherent characteristics of the data – perhaps the voice characteristics of older speakers are more variable for a given age. This would fit with other research [24] that has shown how cognitive abilities become increasingly heterogeneous with advancing age. Further research into this issue, and improved machine performance, is likely to come from data sets with a larger number of speakers and a larger range of ages.

6 References

1. Tanner, D.C., Tanner, M.E.: Forensic Aspects of Speech Patterns: Voice Prints, Speaker Profiling, Lie and Intoxication Detection. Tucson: Lawyers & Judges Publishing (2004).
2. Pellegrini, T., Hedayati, V., Trancoso, I., Hämäläinen, A., Dias, M.: Speaker age estimation for elderly speech recognition in European Portuguese. Proceedings InterSpeech 2014, Singapore, 2962-2966 (2014).
3. Moyse, E.: Age Estimation from Faces and Voices: A Review. Psychologica Belgica, 54, 255-265 (2014).
4. Braun, A., Cerrato, L.: Estimating speaker age across languages. Proc. ICPhS 1999, San Francisco, 1369-1372 (1999).
5. Krauss, R., Freyberg, R., Morsella, E.: Inferring speakers' physical attributes from their voices. J. Experimental Social Psychology, 38, 618-625 (2002).
6. Amilon, K., van de Weijer, J., Schötz, S.: The impact of visual and auditory cues in age estimation. In Müller, C. (ed) Speaker Classification II, Lecture notes in artificial intelligence, 10-21. Springer, Berlin (2007).
7. Moyse, E., Beaufort, A., Brédart, S.: Evidence for an own-age bias in age estimation from voices in older persons. European J. Aging, 11, 241-247 (2014).
8. Bahari, M., McLaren, M., van Hamme, H., van Leeuwen, D.: Speaker age estimation using i-vectors. Engineering Applications of Artificial Intelligence, 34, 99-108 (2014).

9. Li, M., Han, K., Narayanan, S.: Automatic speaker age and gender recognition using acoustic and prosodic level information. Computer, Speech and Language, 27, 151-167 (2013).
10. Bocklet, T., Maier, A., Nöth, E.: Age determination of children in preschool and primary school age with GMM based supervectors and support vector machines regression. Proc. Int. Conf. Text, Speech and Dialogue, Brno, 253-260 (2008).
11. Feld, M., Barnard, E., van Heerden, C., Müller, C.: Multilingual spear age recognition: regression analyses on the Lwazi corpus. IEEE workshop on Automatic Speech Recognition and Understanding, 534-539 (2009).
12. Dobry, G., Hecht, R., Avigal, M., Zigel, Y.: Supervector dimension reduction for efficient speaker age estimation based on the acoustic speech signal. IEEE Trans. Audio, Speech and Language Processing 19, 1975-1985 (2011).
13. Bahari, M., van Hamme, H.: Speaker age estimation and gender detection based on supervised non-negative matrix factorization. Proc. IEEE Workshop Biometric Measurements and Systems for Security and Medical Applications, 1-6 (2011).
14. Bahari, M., van Hamme, H.: Speaker age estimation using hidden Markov model weight supervectors. IEEE Int. Conf. Information Science, Signal Processing and their Applications, 517-521 (2012).
15. Speech Ark, Second Accents of the British Isles Corpus. www.thespeechark.com/abi-2-page.html
16. Hadfield, J.: MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package. J. Statistical Software 33, 1-22 (2010).
17. Eyben, F., Weninger, F., Groß, F., Schuller, B.: Recent developments in opensmile, the Munich open-source multimedia feature extractor. Proc. of the 21st ACM International Conference on Multimedia, Barcelna, Spain, 835–838 (2013).
18. Schuller, B., Steidl, S., Batliner, A., Epps, J., Eyben, F., Ringeval, F., Marchi, E., Zhang, Y.: The INTERSPEECH 2014 Computational Paralinguistics Challenge: Cognitive & Physical Load. Interspeech 2014, Singapore (2014).
19. Smola, A., Schölkopf, B.: A tutorial on support vector regression. J. Stat. Computing 14, 199-222 (2004).
20. CRAN Project, E1071 package of functions from Dept. Statistics, TU Wein. cran.r-project.org/web/packages/e1071/index.html
21. Branco, P., Torgo, L., Ribeiro, R.: A Survey of Predictive Modelling under Imbalanced Distributions. CoRR abs/1505.01658 (2015).
22. Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P.: SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 321-357 (2002).
23. Torgo, L., Ribeiro, R. P., Pfahringer, B., & Branco, P.: Smote for regression. Progress in Artificial Intelligence 378-389. Springer Berlin Heidelberg (2013).
24. Ardila, A.: Normal aging increases cognitive heterogeneity: Analysis of dispersion in WAIS-III scores across age. Archives of Clinical Neuropsychology 22, 1003-1011 (2007).