AGE CLASSIFICATION: COMPARISON OF HUMAN VS MACHINE PERFORMANCE IN PROMPTED AND SPONTANEOUS SPEECH

Felix Burkhardt[1,2]*, Markus Brückl*[1] *and Björn W. Schuller*[2,3,4]
[1]*Technische Universität Berlin,* [2]*audEERING GmbH,* [3] *Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany,* [4]*GLAM – Group on Language, Audio, & Music, Imperial College London, UK*
f.burkhardt@tu-berlin.de

Abstract: We report on the results of an investigation to classify speaker age in vocal utterances with state-of-the-art machine learning algorithms on a small data set. We compare results of manual measurement, i. e., supervised automated extraction of phonetically interpretable measures and observation (by hypothesis tests via procedures like regression analysis) with the outcomes of experiments based on recent machine learning. On isolated vowels the machine outperformed the human estimates.

1 Introduction

Age classification is an crucial part of automatic speaker traits assessment. In contrast to subjective phenomena such as emotional arousal, the (chronological) age of a person may be objectively determined, like for example body size, by an exact measurement. But just like emotional arousal, age is only one of many factors that influence the acoustic speech signal. And the (human) acoustic speech signal is produced and determined by the biological subsystem "vocal apparatus" that may age differently than other human subsystems and moreover its parts (lungs, brains, vocal folds) again may age differently—as compared to the physically exact passing of time that constitutes 'chronological' aging. So, especially the difference between chronological and biological age(s), the latter mainly being determined by how a person lived, makes it nearly impossible to estimate chronological age exactly from the voice.

We discussed the automatic classification with respect to specific age groups in [1, 2, 3], but in these studies the acoustic material consisted of a large number of samples collected over telephone lines in low acoustic quality. In contrast, the current study investigates a comparatively small manually recorded database in studio conditions collected [4]. We investigated with two hypotheses in mind:

- A machine classifier trained on parts of the data achieves accuracy comparable to human age estimation.
- With respect to relevant manually measured acoustic features, they correspond with the most important features used by the machine classifier.

Age classification based on machine learning as such has been investigated numerous times in the past decades. During the 2010 Interspeech Compare Challenge [5], age classification was one of the topics. [6] report on the Agender database [7] by fusing the results of ensemble classifiers trained on subgroups of a larger feature set and get 42.47 % UAR on four age groups which is worse than the baseline with 46.22 % UAR. [8] use a similar configuration with respect to classifiers and feature sets as in this paper to fuse acoustic and metadata for child speech detection. This paper is structured as follows. Section 2 introduces the database. Section 3 discusses the experiments we conducted with respect to machine learning. Finally, Section 4 summarises results, and Section 5 concludes this paper with an outlook.

2 The database

The audio data were collected within the DFG-project "Young and old voices"[1] (cf. [4]).

2.1 The speakers

Although the population to which sample results should be generalisable can be named as "all (female) speakers of German", it was (with a reasonable effort) not possible to determine all subjects of this population and thus, a real random drawing of a sample could not be achieved. So, we drew an ad-hoc sample of 88 speakers of German that comprises females that were within reach and willing to participate in the survey. Their chronological ages range from 20 to 87 years (AM = 50.42 years; SD = 17.64 years). Only one biological sex was chosen in order to reduce variance in the audio data that would be introduced by different biological ageing in speaking organs of different sexes. Only adults were considered, since the process of biological upgrowth should not be confounded with the process of degeneration, which usually is referred to as aging.

2.2 The utterances

The audio data was recorded at the speakers' homes or comparable (non-studio) environments via a head-mounted condenser microphone (AKG C-410) with a DAT-Recorder (TASCAM DA-P1) at 48 kHz sampling rate and 16 Bit resolution.

Each speaker produced five different utterance types, namely read and spontaneous speech and the three cardinal vowels /a/, /i/ and /u/ in sustained phonation. It is assumed that these utterance types provide different amounts of information on the speakers' age and maybe also different ways of encoding this information in the acoustic signal.

2.3 The perceptual rating of the speakers' age

For the auditory rating of the speakers' age the vowels were cut into three segments of 2.2 s duration each, the onset, the quasi-stationary middle part and the offset, resulting in 11 utterances of each speaker to be rated separately. In total 115 listeners rated the auditorily perceivable speaker age of these 88×11 speech samples (in 4 different perception experiments that were found to yield equally good estimates, cf. [4, pp. 143 ff.]). On average every utterance was rated by 28.98 listeners. In order to test the reliability and, if reliable, to summarize these ratings (in sets with missing data) to a group measure a two-way agreement intra-class correlation (ICC) for unbalanced datasets was used, cf. [9]. The listeners' estimates on all utterance types were highly (significant) reliable with spontaneous speech achieving the most reliable ICC(A,k) = 0.992 and the /u/-offset-parts the least with ICC(A,k) = 0.871. The best estimates per utterance derived form this ICC are here used as measure for the human level performance (HLP, cf. Table 1).

2.4 Best features from manual measurement

The manual search for acoustic indicators of speaker age was driven by findings of physiological and cognitive age-related changes and herefrom derived hypothetical acoustic changes in humans. These hypotheses were tested with correlation and multiple linear regression procedures and identified the following parameter groups as best candidates to convey information on (increasing) age: (lowered) fundamental frequency, (raised) magnitude of vocal tremor, (raised) amplitude

[1]The Project "Akustische und perzeptive Korrelate von Stimme und Sprechweise junger und alter Sprecher und Sprecherinnen" funded by the German Research Foundation

perturbation, (more) noisy energy in higher (3-7 kHz) frequency bands, and slowed / less precise articulation when reading. Age estimates that are predicted by these regression models are used to calculate the "manual" multiple linear regression performance (MLRP, cf. Table 1).

3 Experiments with machine classification

3.1 Data preparation

As the data set (in-domain data) is comparably small, we carried out experiments to add other data sets to the training, some of them stating the speaker age only in decades and not in years (Mozilla common voice). Following, we provide an overview on the two additional training databases.

Mozilla common voice The common voice corpus [10] consists of over-the-web donated speech samples. We selected for this experiment only German female speakers within the target age span (20-90 years old). The age in this data set is given not in years but in decades, which was the main reason for binning into decades for the following experiments. The overwhelming number of speaker is aged between 20 and 60 years, so we did not obtain many samples for elderly speakers. We selected randomly at most 2 000 samples per decade.

DTAG Agender The DTAG Agender corpus [7] was collected over the telephone be Deutsche Telekom AG and has been made available originally for the Interspeech Paralinguistic Challenge 2010 [5]. This database contains a-law coded speech and has been converted to 8 kHz. As the age groups are not balanced, we randomly selected at most 1 000 samples from female speakers per decade.

To be able to compare all results with each other, and because classification results are easier to interpret for humans than regression measures, we binned the age into groups. If possible, we still trained on a regression problem by binning the groups after training a regression model with the classifiers.

The age in years was binned into two groups:

- a seven classes group representing the decades from twenties to eighties (corresponding to the Mozilla format).
- a three classes age group: as the the number of samples became to small, when text-type was taken into account, we binned the age additionally into only three groups: young (from zero to 40 years), middle aged (from 40 to 60 years) and elderly (above 60 years). This resulted in an almost even distribution, with the young group slightly falling behind.

All data were divided randomly into a speaker disjoint train, development, and test set, using 50 % of the speakers for the training, and each 25 % for test and development.

As the data is highly imbalanced, we performed oversampling with it, meaning that we added samples to the underrepresented classes. This was done with the SMOTE (synthetic minority over-sampling technique) algorithm [11] which adds samples by synthesising them on a feature level based on distance to central class representatives. We used this with all three data sets. Originally, we executed some of the experiments that did not require meta parameter tuning with half of the in-domain data as training, and half as the test set, but as this resulted into too sparse data sets for the SMOTE algorithm to work, we generally trained for the final results (after parameter tuning) on 3/4 of the data (speaker disjunct). We tried the LOSO (Leave one

speaker out) technique but because the sample number per speaker is very small we did not get meaningful results. We did not use x-fold cross validation because to run the ANN experiments was already with fixed training and development sets very time consuming.

3.2 Different classifiers

We compare three kind of machine classifiers in this work, which showed good performance in related experiments. For the experiments we use the implementations from python packages sklearn (SVM), xgboost (XGB), and pytorch (MLP) python packages, respectively.

SVM A support vector machine is a statistical classifier that constructs hyper planes based on kernel functions to distinguish samples. It is especially well suited if only a few samples are available, but with a high distinction with respect to the target class membership. To find the best meta parameters for the given data, we perform a grid search on the development set, varying the parameters *C: 0.1, 0.01, 0.001, 0.0001* and *max_iter: 2500, 5000, 7500*.

XGBoost XGBoost (eXtreme Gradient Boosting) is a very successful alternative to SVM based on selecting ensembles of random forests. We ran a grid search for optimal classifier meta parameters, and varied in *subsample: .5, .7*, *n_estimators: 50, 80, 200*, and *max_depth: 1, 6*.

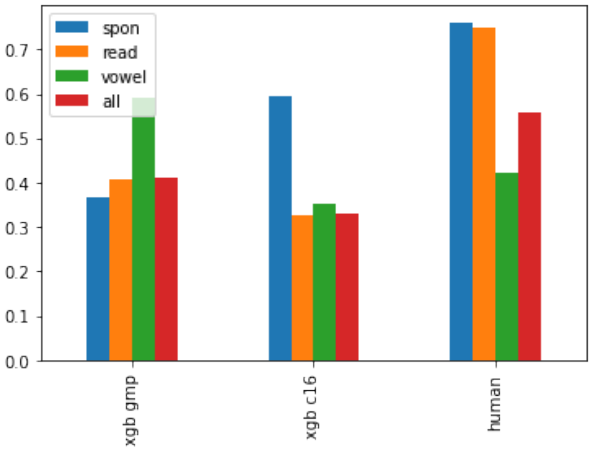
Multi Layer Perceptron The MLP classifier is an extension of the perceptron algorithm to multiple layers, some of them "hidden". It is the most basic of the artificial neural network architectures, and we feature it, as it can work with fixed size numerical vector inputs as given in our features sets (see below). We investigated the performance of two loss functions, namely MSE Loss (Mean Squared Error) as a regression loss on the winner class, and the Cross Entropy Loss over the distribution of the seven classes. In the experiment section, these loss functions are denoted by *reg*, *class*, and *mix* for a 50 % mixture of them. In order to prevent overfitting, we use a relatively small configuration with two hidden layers with 128 and 16 neurons and add a drop out probability of 30 %. As optimiser serves SGD (stochastic gradient descent). We ran each training for 100 epochs if not noted otherwise. Because Artificial neural nets are initialized by random numbers we ran each experiment 10 times and report the average result.

3.3 Feature sets

One of the most interesting challenges is to find a good feature set that represents the attribute of interest (in this case: age) well. A modern approach would be to learn these features by ANNs (artificial neural nets) and represent the acoustic input as some form of spectrogram, a more traditional one is to use expert features that can be of a high number when multiplying low level features based on frame analysis with functionals that aggregate them over a chunk such as an utterance.

GeMAPS The GeMAPS (Geneva Minimalistic Acoustic Parameter Set) feature set has been manually selected to be used as a basic standard acoustic parameter set for various areas of automatic voice analysis, such as paralinguistics or clinical speech analysis [12]. It contains only 88 features and usually performs well on detecting affective speech. openSMILE [13] is an open source framework to extract acoustic features from audio and provides GeMAPS features.

Figure 1 – Results as UAR for humans and classifiers and text types for three age groups



Compare 2016 feature set (all) This feature set represents a "brute force" approach and has been introduced at the Interspeech Paralinguistic Challenge 2016 [14]. It contains 6373 OpenSMILE features, resulting from the combination of low level features and functionals.

Compare 2016 reduced feature set (top) Because the MLP approach (see Section 3.2) does not scale well with respect to the size of the input features (all inputs get multiplied by number of neurons in the second layer) we learnt a reduced feature set from the ComParE 16 set by selecting the 512 best performing features according to the XGBoost classifier. The top ten features all deal with the openSMILE parameter *audspec_l1norm_sma* and statistical functionals thereof, namely *minpos, maxpos, range, quartile 1, 2 and 3, inter quartile range 1-2, 2-3 and 1-3, and percentile 1_0. audspec* is a shortcut for auditory spectrum, *l1norm* the sum of its absolute values, and *sma* a moving average filter based on this. So all of the most important features deal with the distribution of energy in spectral bands changing over the utterance, basically features that correspond to loudness in spectral bands.

As can be seen in the figure, this feature indeed varies for the different age groups, but counter-intuitively not in a linear fashion.

Trill feature set (trill) There are several ways to derive feature sets for paralinguistic speech classification: an alternative for hand-crafted features like the openSMILE features is to use so called "embeddings", such as the weights of the pen-ultimate layer of a deep neural net. The idea is that the net learnt a level of abstraction on the given task that can be used as a kind of transfer learning.

In [15], the authors describe a new set of acoustic features based on a deep neural net trained with a triplet loss to distinguish near-by from far away acoustic snippets based on multitude of data sets collected for speaker, language, emotion or health classification, all in all about .5 million samples and provided by Google[2].

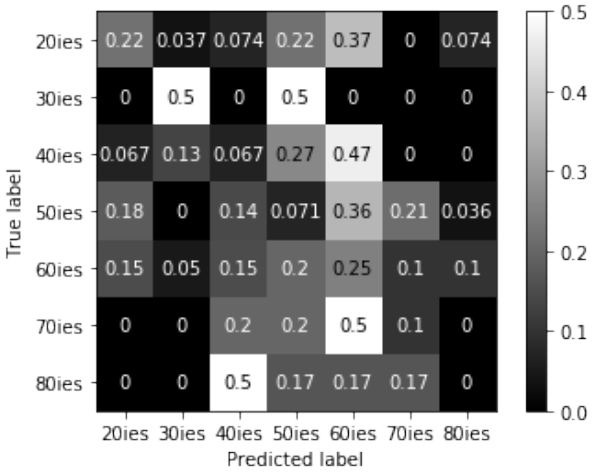
4 Results and discussion

4.1 Investigating the effect of text material

As mentioned in Section 3.1, we had to bin the age into three groups (young, middle, old) when distinguishing text type as the data would have been too sparse for smaller groups.

[2]https://ai.googleblog.com/2020/06/improving-speech-representations-and.html

Figure 2 – Confusion matrix for the spontaneous speech samples



We trained and tested the classifiers SVM and XGBoost (see section 3.2) on the GeMAPS and ComParE 16 feature sets only using material from the three text types spontaneous speech (*spon*), read speech (*read*) and isolated vowels (*vowel*). We compare the outcomes in Figure 1. The SVM classifier failed to converge despite the grid search for an optimal parameterization so we left the results out. The data set was too small to compute MLP models reasonably.

As can be seen, the human estimators performed clearly better when listening to read or spontaneous speech, but do have problems to judge the age from isolated vowels. With respect to the classifiers, there is surprisingly high performance for the XGBoost classifier using GeMAPS features on isolated vowels, given that this is a reduced feature set not primarily targeting speaker age but rather emotional expressions. Such findings will have to be verified with more data.

4.2 Comparing classifiers and features

The results of the main experiment are summarised in Table 1. These results are based on the seven age group classifications and all text types combined. Although we added databases to the training for some of the experiments (D1: Agender, D2: Mozilla common voice, see Section 3.1), the human performance is about twice the level of the machine performance. This contradicts earlier findings where super human performance was achieved [1]. It is probably based on the sparse data situation and the diversity of the cross databases.

Looking at the most important features selected by the XGBoost approach and some results that are barely above chance level, we are a bit sceptical that in all cases generally useful age representations have been learnt. Nonetheless, we are confident that at least in some cases this has been successful. For example Figure 2 shows an exemplary confusion plot, in this case based on the MLP classifier with C16-top features. Irrespective of the low UAR, a trend for high values near the diagonal can be observed.

The best performing experiment is based on a training with all three databases and the feature set learnt from even larger databases (Trill features). We tested the significance of the differences with paired t-tests.

5 Conclusion and Outlook

We investigated the machine classification of speaker age on a small database. With respect to our hypotheses, we could support only one of them: the machine performance is comparable to the human one, but the most important features of the manual investigation do not correspond with those of the machine classifier. The lack of super performance is explainable by little data

Table 1 – Overview of results: all values denote the Unweighted Average Recall (UAR)

feature set		top	all	trill
stat. classifier	SVM	.219	.210	.113
	XGB	.142	.222	.156
art. neural net	MLP mix	.148	-	.165
	MLP reg	.169	-	.173
	MLP class	.158	-	.172
	MLP+D1	.177	-	.225
	MLP+D2	.152	-	.171
	MLP+D1+D2	.161	-	**.237**
	MLP D1	.161	-	.194
	MLP D1	.200	-	.137
	MLP D1 and D2	.217	-	.217
manual regression	MLRP	.191		
human group	HLP	**.299**		

from similar domains and one should revisit this experiment with a more general age model as a background. On isolated vowels the machine outperformed the human estimates.

6 Acknowledgements

We thank the SEMULIN research project by the German BMWi (FK: 19A20012B) for partial funding of this work.

References

[1] METZE, F., J. AJMERA, R. ENGLERT, U. BUB, F. BURKHARDT, J. STEGMANN, C. MÜLLER, R. HUBER, B. ANDRASSY, J. BAUER, and B. LITTEL: *Comparison of four approaches to age and gender recognition for telephone applications*. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 4. 2007. doi:10.1109/ICASSP.2007.367263.

[2] BOCKLET, T., A. MAIER, J. BAUER, F. BURKHARDT, and E. NÖTH: *Age and gender recognition for telephone applications based on GMM supervectors and support vector machines*. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing – Proceedings*. 2008. doi:10.1109/ICASSP.2008.4517932.

[3] FELD, M., F. BURKHARDT, and C. MÜLLER: *Automatic speaker age and gender recognition in the car for tailoring dialog and mobile services*. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010*. 2010.

[4] BRÜCKL, M.: *Altersbedingte Veränderungen der Stimme und Sprechweise von Frauen*, vol. 7 of *Mündliche Kommunikation*. Logos Verlag, Berlin, 2011.

[5] SCHULLER, B., S. STEIDL, A. BATLINER, F. BURKHARDT, L. DEVILLERS, C. MÜLLER, and S. NARAYANAN: *The INTERSPEECH 2010 paralinguistic challenge*. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010*. 2010.

[6] LINGENFELSER, F., J. WAGNER, T. VOGT, J. KIM, and E. ANDRÉ: *Age and gender classification from speech using decision level fusion and ensemble based techniques*. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010*. 2010.

[7] BURKHARDT, F., M. ECKERT, W. JOHANNSEN, and J. STEGMANN: *A database of age and gender annotated telephone speech*. In *Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010*. 2010.

[8] KATERENCHUK, D.: *Age group classification with speech and metadata multimodality fusion*. In *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference*. 2017. doi:10.18653/v1/e17-2030. 1803.00721.

[9] BRÜCKL, M. and F. HEUER: *irrNA: Coefficients of Interrater Reliability – Generalized for Randomly Incomplete Datasets*, 2018. URL https://CRAN.R-project.org/package=irrNA. R package version 0.1.4.

[10] ARDILA, R., M. BRANSON, K. DAVIS, M. HENRETTY, M. KOHLER, J. MEYER, R. MORAIS, L. SAUNDERS, F. M. TYERS, and G. WEBER: *Common voice: A massively-multilingual speech corpus*. In *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*. 2020. 1912.06670.

[11] CHAWLA, N. V., K. W. BOWYER, L. O. HALL, and W. P. KEGELMEYER: *SMOTE: Synthetic minority over-sampling technique*. *Journal of Artificial Intelligence Research*, 2002. doi:10.1613/jair.953. 1106.1813.

[12] EYBEN, F., K. R. SCHERER, B. W. SCHULLER, J. SUNDBERG, E. ANDRE, C. BUSSO, L. Y. DEVILLERS, J. EPPS, P. LAUKKA, S. S. NARAYANAN, and K. P. TRUONG: *The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing*. *IEEE Transactions on Affective Computing*, 2016.

[13] EYBEN, F., M. WÖLLMER, and B. SCHULLER: *openSMILE — the Munich versatile and fast open-source audio feature extractor*. In *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1459–1462. 2010.

[14] SCHULLER, B., S. STEIDL, A. BATLINER, J. HIRSCHBERG, J.AND BURGOON, A. BAIRD, A. ELKINS, Y. ZHANG1, E. COUTINHO, and K. EVANINI: *The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity & Native Language*. In *Proceedings of the 17th Annual Conference of the International Speech Communication Association, INTERSPEECH 2016*. 2016.

[15] SHOR, J., A. JANSEN, R. MAOR, O. LANG, O. TUVAL, F. DE CHAUMONT QUITRY, M. TAGLIASACCHI, I. SHAVITT, D. EMANUEL, and Y. HAVIV: *Towards Learning a Universal Non-Semantic Representation of Speech*. In *Proc. Interspeech 2020*, pp. 140–144. 2020. doi:10.21437/Interspeech.2020-1242. URL http://dx.doi.org/10.21437/Interspeech.2020-1242.