Language-Independent Age Estimation from Speech Using Phonological and Phonemic Features

Tino Haderlein[1], Catherine Middag[2], Florian Hönig[1], Jean-Pierre Martens[2], Michael Döllinger[3], Anne Schützenberger[3], and Elmar Nöth[1]

[1] Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Lehrstuhl für Mustererkennung (Informatik 5), Martensstraße 3, 91058 Erlangen, Germany
http://www5.cs.fau.de
Tino.Haderlein@cs.fau.de

[2] Universiteit Gent, Vakgroep voor Elektronica en Informatiesystemen (ELIS), Sint-Pietersnieuwstraat 41, 9000 Gent, Belgium

[3] Klinikum der Universität Erlangen-Nürnberg, Phoniatrische und pädaudiologische Abteilung in der HNO-Klinik, Bohlenplatz 21, 91054 Erlangen, Germany

Abstract. Language-independent and alignment-free phonological and phonemic features were applied for automatic age estimation based on voice and speech properties. 110 persons (average: 75.7 years) read the German version of the text “The North Wind and the Sun”. For comparison with the automatic approach, five listeners estimated the speakers’ age perceptually. Support Vector Regression and feature selection were used to compute the best model of aging. This model was found to use the following features: (a) the percentage of voiced frames, (b) eight phonological features, representing vowel height, nasality in consonants, turbulence, and position of the lips, and finally, (c) seven phonemic features. The latter features might be relevant due to altered articulation because of dentures. The mean absolute error between computed and chronological age was 5.2 years (RMSE: 7.0). It was 7.7 years (RMSE: 9.6) for an optimistic trivial estimator and 10.5 years (RMSE: 11.9) for the average listener.

Keywords: age estimation, phonological features, phonemic features, SVR

1 Introduction

In speech science, increasing attention is given to age-dependent characteristics of speech, as life expectancy and the percentage of elderly population are growing fast, especially in Europe and North America. Better understanding of aging effects on speech performance will provide better insight into models of the anatomical, physiological, and linguistic consequences of aging. The accuracy of a model for vocal aging can be tested by classifying a speaker’s age automatically. A person with a large discrepancy between chronological and perceived or computed age should be examined more in detail in order to reveal possible symptoms of beginning diseases. The analysis of healthy speech may provide key contributions to the early diagnosis of neurodegenerative disorders, such as shown for Parkinson’s disease [1]. In other scenarios, age-specific recognition systems can be applied, when the user’s age has been estimated automatically, for instance by choosing specific speed, volume, or music for system prompts.

The focus of many papers in that field is on the disambiguation of only a few age classes of practical relevance. Our work concentrates on the development of a large feature vector that allows to estimate an adult speaker's age as precisely as possible.

Phonological and phonemic features capture many voice and also speech properties. They were successfully used for language-independent detection of voice quality and speech intelligibility and can even be used to visualize these aspects [2–4]. Hence, we regarded them also suitable for automatic estimation of a person's age from speech.

This paper is organized as follows: Section 2 introduces the speech data used for the experiments, Sect. 3 describes the features computed from the data and the Support Vector Regression for creating the aging model. The results will be discussed in Sect. 4.

2 Test Data and Subjective Evaluation

110 German persons (31 men, 79 women) without voice or speech problems and between 50 and 94 years of age participated in this study. The average age was 75.7 years with a standard deviation of 9.6 years (Fig. 1, Table 2). They were recruited from senior community centers, senior meetings, and assisted living facilities. Persons receiving voice-related medical treatment, in need of skilled nursing care and/or with relevant cognitive limitations (e.g. dementia) were excluded from the study [5]. Each person read the phonetically rich text "Der Nordwind und die Sonne" ("The North Wind and the Sun", [6]), which is frequently used in medical speech evaluation in German-speaking countries. It contains 108 words (71 distinct) with 172 syllables. The data were recorded with a sampling frequency of 16 kHz and 16 bit amplitude resolution. The study respected the World Medical Association (WMA) Declaration of Helsinki on ethical principles for medical research involving human subjects and has been approved by the ethics committee of the University Erlangen-Nürnberg (FAU).

One female and four male raters evaluated the audio data perceptually by assigning the age to each speaker after listening to the respective audio sample. One male rater was speaking German as a second language, the others were native German speakers. They did not know about the distribution and the range of age in the data in advance.

3 Features Computed from the Speech Data

Since it is expected that speech of elderly persons is not only affected by voice aging but also by changes in articulation, we use phonological and phonemic features to capture these effects. They were designed for Flemish, but in recent studies [2, 4] they have been also successfully used for German. The pre-processing stage returns 12 Mel-frequency cepstral coefficients (MFCCs) and an energy value for each 25 ms speech frame (frame shift: 10 ms). From this spectro-temporal representation of the acoustic signal, speaker features are extracted which constitute a compact characterization of the speech of the tested person. Based on the stream of MFCCs, two text-independent feature extraction methods, focusing on phonological and phonemic aspects, have been explored.

Alignment-free phonological features (ALF-PLFs): First described in [7], these features follow from a tracking of the temporal evolutions of the individual outputs of an

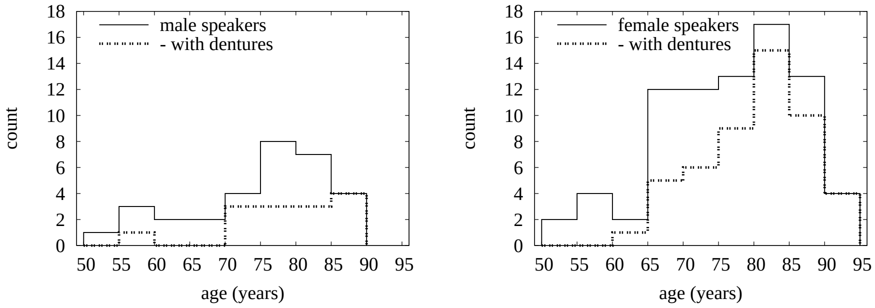

Fig. 1. Distribution of age and dentures among the 31 male and 79 female speakers; for one woman, no information about the dental status was available.

artificial neural network that was trained (see [3, 7] for more details) to generate 14 phonological properties per frame. These properties describe:

- vocal source: voicing
- manner of articulation: silence, consonant-nasality, vowel-nasality, turbulence (referring to fricative and plosive sounds)
- place of consonant articulation: labial, labio-dental, alveolar, velar, glottal, palatal
- vowel features: vowel height, vowel place, vowel rounding

Every phonological property is analyzed by two sub-networks. One of them determines whether the property is relevant at a given time (e.g. it is not relevant to investigate vowel place during utterance of a consonant); the other one determines whether the characteristic (e.g. “labial”) is actually present or not. The hypothesis is that temporal fluctuations in the network outputs can reveal articulatory deficiencies, regardless of the exact phonetic content of the text that was read, at least as long as this text is sufficiently rich in phonetic content. The temporal analysis of each network output generates a set of parameters, such as the mean and standard deviation, the percentage of the time the output is high (above 0.66), intermediate or low (below 0.33), respectively, the mean height of the peaks (maxima), and the mean time it takes to make a transition from low to high. The overall number of output features is 504, and it is acknowledged that several of them may carry similar information. These speaker features are computed without knowledge of the text that was read. Hence, we expect them to be text-independent.

Alignment-free phonemic features (ALF-PMFs): The features, introduced in [3], were originally based on the hypothesis that intelligibility degradation is correlated with problems in realizing a certain *combination* of phonological classes that is needed for the production of a certain phone. Therefore, the ALF-PMFs follow from a plain analysis of posterior phone probabilities. Considering all frames for which the maximal probability is assigned to a particular phone, one computes the mean and standard deviation of that probability, and the mean of the peaks (maxima) and the valleys (minima) found in its temporal evolution. In addition, the percentage of the time a frame is assigned to the phone, and the mean probability of this phone over all frames are computed. Clearly,

Table 1. The AMPEX features (for details, see [10])

feature	description
PVF	percentage of all frames in the recording that were labeled voiced
PVS	percentage of speech frames that were labeled voiced
AVE	average voicing evidence in voiced frames
PVFU	percentage of voiced frames with an unreliable F_0
Jit	average F_0-jitter in voiced frames
Jc	average F_0-jitter in voiced frames with a reliable F_0
VL90	90th percentile (in seconds) of the voiced fragment durations
Tmax	duration (in seconds) of the longest speech fragment (not interrupted by a pause)

these features are computed without any knowledge of the text that was read and can therefore be expected to be text-independent. There are 495 different ALF-PMFs.

All the neural networks for the computation of ALF-PLFs and ALF-PMFs had been trained with Flemish speech data and were now used with German test data. Their general independence of the language had been shown before [2, 4].

Prosodic features (AMPEX): They originate from a holistic analysis of the frame-level volume, fundamental frequency, and voicing evidences. This analysis can be conducted on arbitrary speech, irrespective of the language that is spoken. The frame-level prosodic features are converted into 8 AMPEX features [8]. The voicing evidence and the signal loudness (see [9, 10]) are used to label the frames as voiced/unvoiced and as speech/silence, and to locate pauses, defined as intervals of more than 200 ms long. Based on these classifications, the AMPEX feature extractor computes the features listed in Table 1. They can be grouped into voicing-related parameters (e.g. the percentage of speech frames classified as voiced) and F_0-related features (e.g. average jitter of the fundamental frequency F_0 in voiced frames). They were computed for the whole length of each speech sample. In earlier studies, supplementing phonological features with these F_0-and-voicing related speaker characteristics enhanced intelligibility prediction [3]. We assumed that they may also support automatic age estimation.

Support Vector Regression (SVR): In order to determine the best subset of all phonological, phonemic, and prosodic features to model the chronological age, Support Vector Regression (SVR, [11]) was used. The underlying SVM used a linear kernel. The complexity constant C for the SVR was set to 0.01 after a short series of experiments with heuristic changes to C by powers of 10. Each training example for the regression consisted of a set of features (the inputs) and a chronological age (the target output). The sequential minimal optimization algorithm (SMO, [11]) of the Weka toolbox [12] was applied in a 10-fold cross-validation manner.

For the selection of the attributes, the Greedy Stepwise algorithm was applied. The standard settings were not changed. All input features were standardized (mean value: μ=0, standard deviation: σ=1) for the analysis. For the final regression, the most relevant features were used, precisely those who had been selected between 7 and 10 times during the 10 folds of the process.

4 Results and Discussion

For the final set of the aging model (Table 3), one AMPEX feature (PVF) was selected. Eight phonological features were in the best feature set. The most relevant (selected 9 and 10 times) are:

- *highlow_presence_meanmin:* the mean minimum probability of the vowel height throughout all vowels in the text
- *highlow_presence_tneg:* the mean duration (in number of frames) of a segment in which a low vowel is present
- *consonantnasality_presence_meanneg:* the mean probability of nasality in a consonant that sounds non-nasal
- *consonantnasality_relevance_negdelta1:* the time needed to do a transition from consonant to vowel

Concerning the vowel height, or the vowel trapezium in general, there is no proof that the measured effects are caused by anatomical changes due to aging. Instead of the voice, altered articulation may be the reason. Earlier studies reported that the pronunciation of phones can change during time due to changes in the language, or in the way one speaker uses a language [13, 14]. Hence, also the vowel space, i.e. the area enclosed by the vowel trapezium, can change. Older speakers of English were reported to undergo a shift in the speaker space roughly along a diagonal in the phonetic height × backness plane [15]. It is not sure so far that these findings can be generalized to other languages, however. Nevertheless, we regard our results important for age estimation from speech – as opposed to age estimation from voice which uses less information.

Transitions, as represented by *consonantnasality_relevance_negdelta1*, might in older people be slower. The negative sign in the regression weight (Table 3) supports this assumption. The weights are very low, however. Different weights for men and women in the formulae, especially for *highlow_presence_meanmin*, might be related to the different F_0 or to a more rapidly falling F_0 in women as result of aging processes.

One feature in the set is related to the position of the lips in vowels (mean of negative relevance of *roundedspread*). Three features are related to turbulence in the voice.

The most relevant phonemic features (Table 3) refer to minimal, maximal, and mean probabilities of some phones. The mean and maximum probabilities of /s/ over all frames where it was recognized, and the percentage of positive presence values for /Z/ (as in French ‘journal’; SAMPA notation) may indicate altered articulation due to dentures. This is an important aspect to consider when estimating age from speech. The same reason may hold for the occurrence of /n/, /v/, and /l/ (*l_meanpos* is the mean positive value for the presence of /l/). /Y/ denotes the short u-umlaut. Its role in the set is currently unclear.

The absolute error between the automatically estimated and the chronological age was 5.2 years for all speakers together (root mean square error RMSE: 7.0), for men and women separately, it was slightly higher (Table 3). A trivial estimator optimizing with respect to the mean absolute error would have had an error of 7.7 years (RMSE: 9.6). Hence, our results show an error which is lower by more than two years. The average human rater estimated with an absolute error of 10.5 years (RMSE: 11.9). The greatest mismatch occurred for a woman who was estimated 42 years younger by rater 3.

Table 2. Chronological age, perceived age for single raters and their average, and automatically determined age (SVR); $\mu(|e|)$ denotes mean absolute error, min/max e denotes minimal/maximal error; perceptual ratings were given in integer numbers (without decimal places)

	all speakers				women				men			
	μ(age)	$\mu(\lvert e\rvert)$	min e	max e	μ(age)	$\mu(\lvert e\rvert)$	min e	max e	μ(age)	$\mu(\lvert e\rvert)$	min e	max e
chron.	75.7	—	—	—	76.3	—	—	—	74.0	—	—	—
rater 1	72.9	7.2	–34	27	73.0	7.1	–34	18	72.8	7.6	–14	27
rater 2	74.5	6.6	–22	22	75.6	6.6	–22	20	71.9	6.7	–15	22
rater 3	57.3	19.1	–42	11	57.9	18.9	–42	11	55.9	19.7	–39	11
rater 4	63.5	13.3	–34	25	63.0	13.9	–34	9	64.7	12.0	–24	25
rater 5	63.4	13.9	–31	24	63.7	14.1	–31	15	62.6	13.5	–25	24
rater avg.	66.3	10.5	–24.0	19.2	66.6	10.4	–24.0	7.4	65.6	10.7	–19.2	18.4
SVR	76.0	5.2	–24.8	14.8	76.2	5.4	–19.0	14.0	76.5	5.8	–24.4	7.2

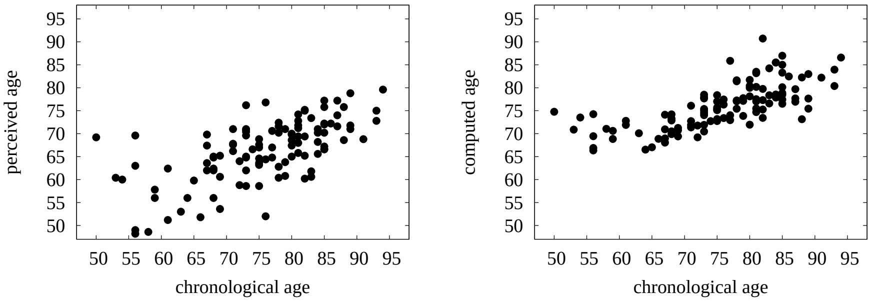

Fig. 2. Average chronological vs. perceived *(left graphics)* and computed age *(right graphics)*

The correlation of the automatically estimated and the chronological age was r=0.72 for all speakers, r=0.73 for men, and r=0.69 for women only. The average human rating showed a correlation of r=0.65 to the 'real' age of the whole speaker group. No single rater performed as good as the machine (rater 1: r=0.41; 2: r=0.66; 3: r=0.59; 4: r=0.43; 5: r=0.30). The inter-rater agreement on all the data, i.e. the correlation of one rater against the average of the others, was r=0.69 for three of the raters, r=0.62 for rater 1, and r=0.48 for rater 5. The chronological, perceived, and computed age are also shown in Fig. 2. The smaller range of the computed values is caused by the error minimization during the SVR training. Without the somewhat optimistic feature selection, the human-machine correlation was r=0.52 (mean error: 6.5 years, RMSE: 8.6) for all speakers.

Studies on age estimation have been presented before. For instance, Schötz [16] used prosodic and spectral features, such as F_0, formants, energy, jitter, shimmer, and duration, for age estimation by Classification And Regression Trees (CARTs). The average error in the classified age was above 15 years on 24 speakers from two age classes (18 to 31 and 60 to 82 years). F_0, the formants F_1 to F_4, and prosodic features were also

Table 3. Regression weights for the best feature set when applied to all speakers, to women and men separately, respectively. The human-machine correlation and the respective errors between computed and chronological age are given in the lower part of the table

feature	type	chosen	all	women	men
PVF	AMPEX	7	–0.097	–0.121	–0.009
Y_min	ALF-PMF	7	–0.097	–0.052	–0.019
l_meanpos	ALF-PMF	7	–0.175	–0.118	–0.090
n_min	ALF-PMF	8	–0.169	–0.163	–0.054
s_mean	ALF-PMF	8	–0.066	–0.060	–0.071
s_max	ALF-PMF	9	–0.131	–0.127	–0.069
v_min	ALF-PMF	7	–0.055	–0.074	–0.050
Z_posperc	ALF-PMF	9	0.138	0.147	0.088
consonantnasality_relevance_negdelta1	ALF-PLF	10	–0.048	–0.034	–0.023
consonantnasality_presence_meanneg	ALF-PLF	10	0.177	0.154	0.066
highlow_presence_tneg	ALF-PLF	9	–0.114	–0.106	–0.027
highlow_presence_meanmin	ALF-PLF	10	0.173	0.183	0.016
roundedspread_relevance_meanneg	ALF-PLF	7	0.129	0.149	0.050
turbulence_relevance_meanmax	ALF-PLF	7	–0.076	–0.033	–0.088
turbulence_presence_mean	ALF-PLF	7	–0.024	–0.044	–0.051
turbulence_presence_tneg	ALF-PLF	8	0.099	0.101	0.061
correlation r to chronological age	—	—	0.72	0.73	0.69
mean abs. error to chronological age	—	—	5.2	5.4	5.8
RMSE to chronological age	—	—	7.0	7.0	8.2

used together with MFCCs on the University of Florida Vocal Aging Database [17]. The mean absolute error of listeners was 6.4 years, and the error of the machine was 10.0 years for gender-independent classification. Different years of age were not represented continuously in the data, however, but in three separate age groups with gaps between them. This may also be the reason why the humans were better in their evaluation than the machine. In our experiments, there was a monomodal distribution of age.

Minematsu et al. [18] reported correlations between perceived and computed age of up to r=0.88, but on audio data showing a clear trimodal age distribution. Their approach was based on Gaussian Mixture Models (GMMs). In a study of Bocklet et al. with children in preschool and primary school age (average: 8.3 years), a system based on GMMs and SVR showed a mean absolute error of 0.8 years and a maximal error of 3 years [19]. The ratio of error and average age was smaller in our system; however.

This study showed the potential of the presented features for language- and gender-independent estimation of age from speech data. The method may be helpful for clinical screening tests and for applications based on automatic speech recognition in general.

Acknowledgments: This work was partially funded by the Else Kröner-Fresenius-Stiftung (Bad Homburg v. d. H., Germany; grant 2011_A167) and by Hessen Agentur/ Hessen ModellProjekte (Wiesbaden, Germany; project 463/15-05). Further supporters were “Kom op tegen Kanker”, the campaign of the Vlaamse Liga tegen Kanker VZW, and The Netherlands Cancer Institute/Antoni van Leeuwenhoek Hospital (Amsterdam).

References

1. Rusz, J., Cmejla, R., Ruzickova, H., Ruzicka, E.: Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease. J. Acoust. Soc. Am. **129** (2011) 350–367
2. Middag, C., Bocklet, T., Martens, J.-P., Nöth, E.: Combining phonological and acoustic ASR-free features for pathological speech intelligibility assessment. In: Proc. Interspeech, ISCA (2011) 3005–3008
3. Middag, C.: Automatic Analysis of Pathological Speech. PhD thesis, Ghent University, Ghent, Belgium (2012)
4. Haderlein, T., Middag, C., Maier, A., Martens, J.-P., Döllinger, M., Nöth, E.: Visualization of Intelligibility Measured by Language-Independent Features. In Sojka, P., Horák, A., Kopeček, I., Pala, K., eds.: Proc. Text, Speech and Dialogue. Volume 8655 of LNAI., Cham, Springer (2014) 547–554
5. Schneider, S., Plank, C., Eysholdt, U., Schützenberger, A., Rosanowski, F.: Voice Function and Voice-Related Quality of Life in the Elderly. Gerontology **57** (2011) 109–114
6. International Phonetic Association (IPA): Handbook of the International Phonetic Association. Cambridge University Press, Cambridge (1999)
7. Middag, C., Saeys, Y., Martens, J.-P.: Towards an ASR-free objective analysis of pathological speech. In: Proc. Interspeech, ISCA (2010) 294–297
8. Moerman, M., Pieters, G., Martens, J.-P., van der Borgt, M.-J., Dejonckere, P.: Objective evaluation of the quality of substitution voices. Eur. Arch. Otorhinolaryngol. **261** (2004) 541–547
9. van Immerseel, L., Martens, J.-P.: AMPEX Disordered Voice Analyzer [computer program]. Digital Speech and Signal Processing research group, Ghent University, Ghent, Belgium Available: http://dssp.elis.ugent.be/downloads-software. Last visited May 28, 2015.
10. van Immerseel, L.M., Martens, J.-P.: Pitch and voiced/unvoiced determination with an auditory model. J. Acoust. Soc. Am. **91** (1992) 3511–3526
11. Smola, A.J., Schölkopf, B.: A Tutorial on Support Vector Regression. Statistics and Computing **14** (2004) 199–222
12. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. 2nd edn. Morgan Kaufmann, San Francisco (2005)
13. Harrington, J., Palethorpe, S., Watson, C.I.: Does the Queen speak the Queen's English? Nature **408** (2000) 927–928
14. Watson, P.J., Munson, B.: A Comparison of Vowel Acoustics Between Older and Younger Adults. In: Proc. ICPhS XIV, International Phonetic Association (2007) 561–564
15. Harrington, J., Palethorpe, S., Watson, C.I.: Age-related changes in fundamental frequency and formants: a longitudinal study of four speakers. In: Proc. Interspeech, ISCA (2007) 2753–2756
16. Schötz, S.: Prosodic and Non-Prosodic Cues in Human and Machine Estimation of Female and Male Speaker Age. In Bruce, G., Horne, M., eds.: Nordic Prosody: Proceedings of the IXth Conference, Lund, Sweden (2004) 215–223
17. Spiegl, W., Stemmer, G., Lasarcyk, E., Kolhatkar, V., Cassidy, A., Potard, B., Shum, S., Song, Y.C., Xu, P., Beyerlein, P., Harnsberger, J., Nöth, E.: Analyzing Features for Automatic Age Estimation on Cross-Sectional Data. In: Proc. Interspeech, ISCA (2009) 2923–2926
18. Minematsu, N., Sekiguchi, M., Hirose, K.: Automatic Estimation of Perceptual Age Using Speaker Modeling Techniques. In: Proc. Eurospeech, ISCA (2003) 3005–3008
19. 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. In Sojka, P., Horák, A., Kopeček, I., Pala, K., eds.: Proc. Text, Speech and Dialogue. Volume 5246 of LNAI., Heidelberg, Springer (2008) 253–260