Analyzing Features for Automatic Age Estimation on Cross-Sectional Data

Werner Spiegl[1], *Georg Stemmer*[2], *Eva Lasarcyk*[3], *Varada Kolhatkar*[4],
Andrew Cassidy[5], *Blaise Potard*[6], *Stephen Shum*[7], *Young Chol Song*[8], *Puyang Xu*[5],
Peter Beyerlein[9], *James Harnsberger*[10], *Elmar Nöth*[1]

[1]Chair of Pattern Recognition (LME), University Erlangen-Nuremberg, Germany
[2]SVOX Deutschland GmbH, Munich, Germany
[3]Dep. of Computational Linguistics and Phonetics, Saarland University, Germany
[4]Dep. of Computer Science, University of Minnesota Duluth, USA
[5]The Center for Language and Speech Processing, Johns Hopkins University, Baltimore, USA
[6]CRIN, Nancy, France
[7]International Computer Science Institute, University of California at Berkeley, USA
[8]Dep. of Computer Science, Stony Brook University, USA
[9]Dep. Bioinformatics, University of Applied Sciences Wildau, Berlin, Germany
[10]Speech Perception Laboratory, University of Florida, USA

spiegl@i5.informatik.uni-erlangen.de

Abstract

We develop an acoustic feature set for the estimation of a person's age from a recorded speech signal. The baseline features are Mel-frequency cepstral coefficients (MFCCs) which are extended by various prosodic features, pitch and formant frequencies. From experiments on the University of Florida Vocal Aging Database we can draw different conclusions. On the one hand, adding prosodic, pitch and formant features to the MFCC baseline leads to relative reductions of the mean absolute error between 4-20%. Improvements are even larger when perceptual age labels are taken as a reference. On the other hand, reasonable results with a mean absolute error in age estimation of about 12 years are already achieved using a simple gender-independent setup and MFCCs only. Future experiments will evaluate the robustness of the prosodic features against channel variability on other databases and investigate the differences between perceptual and chronological age labels.

Index Terms: Age regression, age estimation, vocal aging, prosodic features, support vector regression (SVR)

1. Introduction

This paper investigates the problem of automatic age estimation of adult speakers' voices. The goal is to develop a suitable acoustic feature set for this task. Linville [2] has described a number of acoustic properties that listeners often consider to be characteristic for aged speakers, like coalescence in the pitch of adult male and female voices, increased harshness, strain, vocal tremor and breathiness. Other features include reduced loudness, slower speaking rate and longer pause duration. Linville's findings suggest that virtually all aspects of the speech signal should be included in the search for an appropriate feature set for age estimation. This is supported by results in the literature on physiological changes (refer to [3] for a recent overview): the speech generation process is affected by age in many different ways, for instance the vocal tract length can be increased by a lowered glottis position, the pulmonary function may be reduced and the vocal folds may become stiffer.

Therefore our study strives to cover a very broad range of different feature types, including not only short-term cepstral features but also long-term prosodic features. In order to evaluate to what extent the features contain useful information we measure the mean absolute error of a support vector machine for the estimation of a speaker's age. For the experiments the *University of Florida Vocal Aging Database (UF-VAD)* is utilized. UF-VAD is a collection of read speech by male and female adult speakers representing equally young, middle aged, and older speakers (see [4] for a full description of the database). For each speaker in the database, not only the actual age but also perceived age judgements are available. This allows us to benchmark recognition rates of the automatic classifier against human age perception.

The problem of automatic age classification from the speech signal has already been investigated by others. Minematsu et al. showed in [8] a high correlation between linear discriminant analysis (LDA) scores and five perceptual age classes. Gaussian mixture models (GMM) for the different age classes are combined with two prosodic features for speaking rate and local power perturbation. Metze et al. compare four different systems for age and gender classification in [10]. Four different age groups (children, young speakers, adults, and seniors) are distinguished. The best performing system is based on Mel-frequency cepstrum coefficients (MFCCs) as features and phone recognizers as acoustic models for each age group. A system based on a combination of GMM and support vector machines, similar to many state-of-the-art speaker recognizers, has been described by Bocklet et al. in [1]. Müller and Burkhardt compare different methods to combine a long-term pitch feature with the short-term MFCC-based feature vectors in [5]. The focus of the above mentioned papers is on developing a system for the disambiguation of few age classes of practical relevance, sometimes by extending the conventional cepstral feature vector with one or two selected prosodic features. This is in contrast to our work which concentrates on the development of a large feature vector that allows to estimate an adult speaker's age as precisely as possible. Therefore this paper is much more

Copyright © 2009 ISCA

comparable to Schötz's publication [6]. Schötz evaluates several different prosodic and spectral features like fundamental frequency, formants, energy, jitter, shimmer, and duration for the task of age estimation using CARTs. Our work differs in the sense that we are less interested in comparing selected features but more in the development of a feature vector integrating many different cepstral, spectral and prosodic parameters to get a low error rate.

The rest of the paper is structured as follows: In section 2 we describe the database we used in our experiments. Then we present in section 3 the analyzed features and the applied feature selection algorithm. In the subsequent section the setup of the experiments is described. After discussing the results the paper ends with the conclusions and a short outlook on future work.

2. Data

The data used for our experiments is taken from the *University of Florida Vocal Aging Database (UF-VAD)* [4], a corpus of American English recorded between 2003 and 2007. The database itself features 150 different speakers and 1350 utterances of read speech originating from known material such as the *Rainbow Passage*, the *Grandfather Passage*, and *SPIN sentences*. Each subject reads approximately 2 minutes of the same text into the same microphone and recording conditions, aggregating a total corpus length of about 5 hours. Moreover, the contributing speakers are evenly distributed in terms of gender and three general age groups. That is, we have 25 male and 25 female speakers from each of young (18-29), middle-aged (40-55), and old (62-92) categories. This gives 75 representatives for each gender, as well as 50 representatives for each age group. The mean ages for each age group are 21, 48, and 79, respectively.

For our purposes, the consistency of the UF-VAD helps us normalize the significant variabilities caused by factors other than age. The use of the same microphone, recording equipment and environment reduces the likelihood of channel dependence in our resulting age-classifier, while the use of the same read text across all speakers reduces its dependence on varying linguistic content.

Finally, the corpus also provides a separate set of data on the *perceived ages* of the contributing speakers, 147 listeners estimated the respective ages of the corpus' contributing speakers from the sentence-level material in the database (16 sentences total). The set of results were included in the UF-VAD, which allow for the possibility of comparing our age-classification system with the abilities of human perception.

3. Features

We extract several different types of features to cover a broad range of phonetic dimensions. As baseline we use the well-known Mel-frequency cepstrum coefficients (MFCCs). Three additional groups of features are computed from voiced segments of the speech signal: pitch $f0$, the first four formants $F1 - F4$ and prosodic features. In combination we get 220 features. In order to reduce the total number of features we apply a selection process called $\mathrm{MAX\ R}$ which is described at the end of the section.

3.1. MFCCs

The MFCCs are the standard features in speech processing. Here the MFCC vector has a dimension of 24, consisting of the log-energy, the first 11 static MFCCs and 12 dynamic features, which are calculated using a regression-line over the 5 surrounding frames. The window size of each frame is 16 ms and the frame shift is 10 ms.

3.2. Pitch and formants

The pitch $f0$ is computed using the normalized cross correlation function and dynamic programming. The formant trajectories are estimated by Linear Prediction Coding (LPC) and optimized with dynamic programming as well. For each frame the first four formant frequencies $F1 - F4$ are extracted together with the corresponding bandwidths $B1 - B4$. Pitch and formants are computed with a frame shift of 10 ms each. As implementation of both extraction algorithms we used the Snack Toolkit [15].

3.3. Prosodic features

Once the so-called basic features pitch and energy and the voiced-unvoiced decision have been computed for each frame of the signal, a high-dimensional prosodic feature vector modeling structured prosodic features is generated for each voiced speech segment. These prosodic features are derived from the basic features, the duration of the segments, the speech pauses and the speech quality (i.e. jitter and shimmer). Various attributes model the prosodic properties of each voiced segment. As an example Fig. 1 shows the attributes for the basic feature pitch $f0$. Each of the attributes results in a prosodic feature. In addition a context of five voiced segments is taken into account yielding an even larger context-dependent prosodic feature vector. For an in-depth description of the prosodic feature set see [12, 13]. All in all we get a vector of 187 different prosodic features, each of them belongs to one of the following five feature groups: *Pitch*, *Energy*, *Duration*, *Pause* and *Quality*.

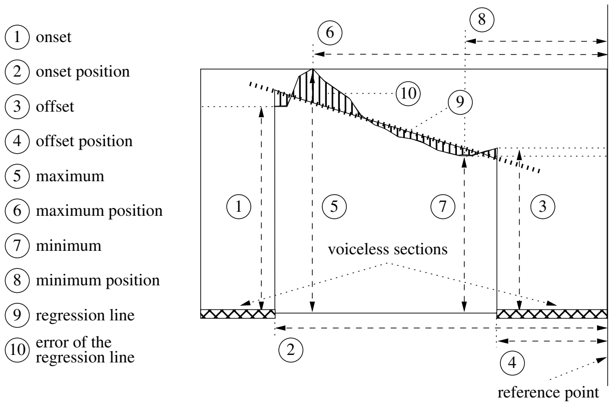

Figure 1: Modeling the pitch contour in a voiced segment (after [12]).

3.4. Feature selection with $\mathrm{MAX\ R}$

Feature selection is performed using multiple regression/correlation analysis as described in [11]. This procedure is a sped up alternative to the $\mathrm{MAX\ R}$ algorithm presented in [14]. The basic idea of the algorithm is to select the best feature subset by maximizing the explained variance R^2 with R as correlation. In this respect the approach iteratively determines the best subset consisting of one feature, two features and so on. For this $\mathrm{MAX\ R}$ calculates all possible exchanges of additional features (after [14, p. 86]) until there is no better subset. If no additional feature improves the results, the algorithm stops. For each subset the weighted correlations of the features are computed by solving the least square error to the system of linear equations $y = X\beta$ where X is the matrix of regressor effects (feature vectors in the rows) and β is the vector with the regression parameters (feature weights). y is the vector of response values, i.e. in our case the ages of the speakers.

4. Experiments

To get an objective measurement of how the different features reflect the changes of the aging voice, a regression system is constructed that estimates the age of each speaker in the database.

4.1. Setup

For each speaker in the database a single high-dimensional feature vector is generated which contains all information that is available: the *meta feature vector*. The age estimation system is trained and tested on the basis of this meta feature vector.

4.1.1. Feature modeling

The age estimation experiments require a constant dimension of the meta feature vector that is independent of the length of the utterances and of the underlying features which can be extracted on the frame or segment level. Therefore the meta feature vector consists only of the speaker-wise means μ and standard deviations σ of the different features: a Gaussian model. So, the dimension of the meta feature vector is twice the sum of the dimensions of the underlying base features. With the 220 different features we obtain for each speaker a meta feature vector of dimension 440, containing the μ and σ of the pitch, the formants, MFCCs and the prosodic features.

4.1.2. Regression

For the age estimation a *Support Vector Regression (SVR)* with a linear kernel is applied. The method of *SVR* adopts the principle of support vectors known from classification with *Support Vector Machines (SVM)* for the area of regression [7]. The system is evaluated with a n-fold cross validation (leave one speaker out) where n equals the number of instances in the database. Afterwards of all folds the mean absolute error (MAE) in years is calculated.

4.2. Results

The age estimation experiments are realized under various conditions. On the one hand they are arranged under the aspect of speaker's gender: the *SVR* system estimates the chronological and the perceived age of males and females. Additionally males and females are combined to a gender independent set. On the other hand we look at the features and process them separately: First we choose the different feature sets by hand: pitch, formants, MFCCs and prosodic features. Second we used $\mathrm{MAX\ R}$ to analytically select features. The results of our age estimation experiments are summarized in Tab. 1. In the last column one can find the results on all features.

4.3. Discussion

The results shown in Tab. 1 demonstrate that the chronological (actual) age of a speaker can be effectively estimated using a combination of prosodic, spectral and cepstral features: both for males and females the mean absolute error (MAE) is about ten years. When the genders are combined in a single experiment as shown in the last two rows of Tab. 1 there is only a relatively small degradation leading to a MAE of 12 years. In our opinion this indicates that the features we used for age estimation are to a certain extent gender-independent, or, put the other way round, that there are similarities in the vocal aging process of male and female speakers. This is supported by an analysis of the features preferred by the $\mathrm{MAX\ R}$-selection algorithm: Tab. 2 compares the number of features selected from each feature group for both genders. It can easily be seen that both for male and female speakers there is a clear preference of MFCCs, Pitch, and Energy-based features which persists for the combined experiment. For female speakers, formants seem to be more valuable than for male speakers. Our observation of small gender differences in the features is somewhat surprising, because it is contradictory to the results of Schötz [6], who found more prominent differences between the male and female speakers, and to Higgins and Saxman [9] who state that both genders age differently.

Comparing the errors for the different feature groups in Tab. 1 shows that in all cases $f0$ performs worst of all features. For chronological age labels, MFCCs are the best performing feature group. Adding all prosodic features, pitch and formants to the MFCC reduces the MAE by 20% relative for the female speakers and just by 4% relative for the male speakers. For the experiment with both genders combined, no improvement could be found over standard MFCCs by adding other feature groups. Feature selection using the $\mathrm{MAX\ R}$ algorithm leads to a much smaller feature vector dimension together with a small increase in error. Interestingly the relative reduction in MAE by adding prosodic features, pitch and formants to the MFCC baseline is much larger for the perceptual age labels than for the chronological labels: when using $\mathrm{MAX\ R}$ the relative reduction for female speakers is 25% and for male speakers it is 12%. For both genders combined, a relative reduction in MAE of 23% has been achieved.

For all feature types the estimation error is smaller when computed w.r.t. the perceptual age than for the chronological age. This fits to our expectation that it is difficult to estimate the age for a certain subset of the speakers, both for the humans and the computer. By comparing the perceptual age estimations and the chronological age the average human performance can be computed: the MAE for human listeners is 6.4 years. Thus, the human error is about 50% smaller than for the machine.

5. Conclusions and future work

Our age-regression experiments demonstrated that a speaker's age can be effectively estimated using a feature vector of prosodic, spectral and cepstral features. In order to achieve reasonable results it seems that it is not necessary to distinguish male and female speakers beforehand. Feature selection experiments show that MFCCs, pitch and energy are the most important feature groups. The relative error reductions over a standard MFCC baseline feature vector by adding prosodic, spectral and pitch features are between 4-20% relative. For a gender-independent setup, no improvement at all could be measured. Considering the additional effort for extracting these features, we come to the conclusion that MFCC features are sufficient to build a practical system. However, the error of this system is about twice the error of human listeners. Furthermore we observed significant improvements from adding prosodic, spectral and pitch features when using the perceptual age labels as a ref-

Table 1: MAE (Mean Absolute Error) of the experiments with different feature sets. MAX R column shows the results after the MAX R feature selection step described in section 3.4.

MAE (years)		*f0*	*Formants*	*MFCCs*	*Prosodic*	MAX R	*All*
Females	*Perc*	11.3	9.5	9.2	7.3	**6.9**	6.2
	Chrono	18.6	13.5	12.0	14.1	**10.0**	9.5
Males	*Perc*	13.2	11.0	8.6	9.5	**7.6**	7.9
	Chrono	19.1	15.6	**10.5**	11.6	13.3	10.1
Combined	*Perc*	14.5	11.0	9.0	9.2	**6.9**	9.4
	Chrono	21.0	16.6	**11.3**	14.9	12.8	11.5

Table 2: Analysis of the features selected by MAX R. Count of the different feature groups.

	Females		Males		Combined		
	Perc	Chrono	Perc	Chrono	Perc	Chrono	$\sum$
MFCC	5	3	1	3	4	5	21
Formant	3	3	1	1	0	1	9
Pitch	4	3	3	3	2	2	17
Energy	2	4	4	4	3	5	22
Duration	1	2	3	1	2	0	9
Pause	0	0	0	0	0	0	0
Quality	0	2	0	1	1	1	5
Number of features	15	17	12	13	12	14	83

erence. Therefore future investigations will concentrate on the differences between the chronological and the perceptual age labels. Furthermore we plan to measure the robustness of the prosodic features w.r.t. channel variations, an aspect that cannot be evaluated using the UF-VAD database.

6. Acknowledgments

The authors appreciate the support of the JHU CLSP 2008 summer workshop and its sponsors for enabling this research. We gratefully acknowledge a number of collaborators that contributed to this project before and during the workshop: Andreas Andreou, Peter Birkholz, Tobias Bocklet, Allen Gorin, Jonathan Harrington, Sanjeev Khudanpur, Andreas Maier and Stefan Steidl.

7. References

[1] Bocklet, T.; Maier A.; Bauer J.; Burkhardt F.; Nöth E., ”Age and Gender Recognition for Telephone Applications Based on GMM Supervectors and Support Vector Machines”, In Proceedings of ICASSP, Las Vegas, 1605-1608, 2008

[2] Linville, S. E., ”The Sound of Senescence”, The Journal of Voice, 10(2), Lippinicott-Raven, 190–200, 1996

[3] Harnsberger, J. D.; Shrivastav, R.; Brown, W. S.; Rothman, H.; Hollien, H., ”Speaking Rate and Fundamental Frequency as Speech Cues to Perceived Age”, The Journal of Voice, 22(1), The Voice Foundation, 58-69, 2008

[4] Harnsberger, J. D.; Brown, W. S.; Shrivastav, R.,; Rothman, H. B., ”Noise and tremor in the perception of vocal aging in males” (accepted), Journal of Voice.

[5] Müller, C.; Burkhardt, F., ”Combining Short-term Cepstral and Long-term Pitch Features for Automatic Recognition of Speaker Age”, In Proceedings of Interspeech 2007, Antwerp, 2277-2280, 2007

[6] Schötz, S., ”Prosodic Cues in Human and Machine Estimation of Female and Male Speaker Age”, Nordic Prosody: Proceedings of the IXth Conference, Lund, 215-223, 2004

[7] Smola, A. J.; Schölkopf, ”A tutorial on support vector regression”, Statistics and Computing 14(3), Hingham, 199-222, 2004

[8] Minematsu, N.; M. Sekiguchi; K. Hirose, ”Automatic Estimation of Perceptual Age Using Speaker Modeling Techniques”, Proceedings of Eurospeech 2003, Geneva, 3005-3008, 2003

[9] Higgins, M. B.; J. H. Saxman, ”A Comparison of Selected Phonatory Behaviours of Healthy Aged and Young Adults”, Journal of Speech and Hearing Research, 13, 1000-1010, 1991

[10] Metze, F.; Ajmera, J.; Englert, R.; Bub, U.; Burkhardt, F.; Stegmann, J.; Müller, C.; Huber, R.; Andrassy, B.; Bauer, J.G.; Littel, B., ”Comparison of Four Approaches to Age and Gender Recognition for Telephone Applications”, In Proceedings of ICASSP, Honolulu, 1089-1092, 2007

[11] Maier, A.; Haderlein, T.; Eysholdt, U.; Rosanowski, F.; Batliner, A.; Schuster, M.; Nöth, E., ”PEAKS A System for the Automatic Evaluation of Voice and Speech Disorders”, Speech Communication, 51, 425-437, 2009

[12] Batliner, A.; Buckow, J.; Niemann, H.; Nöth, E.; Warnke, V., ”The Prosody Module”, Verbmobil: Foundations of Speech-to-Speech Translations, Berlin, 106–121, 2000

[13] Steidl, S., ”Automatic Classification of Emotion-Related User States in Spontaneous Childrenss Speech”, PhD thesis, Chair of Pattern Recognition (LME), University Erlangen-Nuremberg, Germany, 2008.

[14] Clark, V. (Ed.), SAS/STAT® 9.2, User’s Guide, SAS Institute Inc., Cray, NC, USA, 2008

[15] Sjölander, K.; Beskow, J.; Gustafson J.; Lewin, E.; Carlson, R.; Granström, B., ”Web-based Educational Tools for Speech Technology”, Proceedings of ICSLP, 3217–3220, Sydney, 1998.