Automatic prediction of speaker age using CART

Susanne Schötz
Dept. of Linguistics and Phonetics, Lund University
susanne.schotz@ling.lu.se

ABSTRACT

This paper describes a small attempt to automatically estimate speaker age aimed at increasing the phonetic knowledge of age. Acoustic features were extracted from the four phonemes of the Swedish word /rɑ:sa/ (collapse) produced by 428 adult Swedish speakers, and then used to build CARTs (Classification and Regression Trees) for prediction of age, age group and gender. Results showed that the CARTs used different strategies to estimate different phonemes, and that age predictors for /ɑ:/ and /s/ performed best. The best CARTs made about 91% correct judgements for gender, about 72% for age group, while the correlation between biological and predicted age was about 0.45. When comparing this results to those of an earlier study of human age perception, it was found that although humans and CARTs used similar cues, the human listeners were somewhat better at estimating age. More studies with larger and more varied speech material are needed in further pursuit of a good automatic age predictor.

1 Introduction

Verbal human-computer communication distinguishes itself from human-to-human communication in many ways. One difference is that most systems fail to identify the speaker-specific or paralinguistic information present in every voice. Human listeners almost instantly recognize the gender, emotional state, attitude and state of health of a speaker. Even age is fairly well judged by listeners. If human-computer interfaces were able to capture some of these properties, man-machine communication would become more natural. Spoken dialog systems would be able to adapt to the gender, age and personality of the user, which could lead to increasing performance. This paper describes a small attempt to automatically predict one speaker-specific quality: age, using one of the most important techniques in pattern recognition: CART, and then comparing the results to age judgements of human listeners.

1.1 Background

While it is generally believed that human listeners are able to judge speaker age to within ±10 years, few computers have had a go at this task. One reason for this may be that it is far from easy. There are acoustic correlates to age in every phonetic dimension, and their relative importance to age perception has still not been fully explored (Hollien, 1987; Jacques & Rastatter, 1990; Linville, 1987; Ptacek & Sander, 1966; Schötz, 2003).

Earlier attempts to automatically predict age include Minematsu et al. (2003), who carried out age estimation tests with 30 listeners for some 400 male speakers, and then used two methods to model the speakers with GMMs (Gaussian Mixture Models). The first method modelled one speaker for each perceived age, and the second was based on the normal distributions of the age estimations. Tests of the models resulted in a correlation of about 0.9 between the automatic prediction and the judgements of human listeners.

A study of human perception of speaker age with resynthesized stimuli led to the conclusion that spectral features and segment duration seem more important than F_0 to age perception (Schötz, 2004). In the same study, 30 listeners judged the exact age (in years) of 24 speakers from a single word. Significant correlations between biological and perceived age were found for the older speakers (0.825 for female, 0.944 for male speakers), but not for the younger ones (0.097 for female, 0.522 for male speakers). Reasons for this result may include the short word durations, misjudgements of atypical speakers (speakers, who sound older or younger than their biological age (Schötz, 2003)) and the fact that the range of biological age was wider in the older group. The results found by Schötz (2004) will be used in the comparisons of human and automatic age estimations in the present study.

One of the most powerful methods in pattern recognition, besides HMMs (Hidden Markow Models) is CARTs (Classification And Regression Trees). CART is a technique that uses both statistical learning and expert knowledge to construct binary decision trees, formulated as a set of ordered yes-no questions about the features in the data. The best predictions based on the training data are stored in the leaf nodes of the CART. Its advantages over other pattern recognition methods include human-readable rules, compact storage, handling of incomplete and non-standard data structures, robustness to outliers and mislabelled data samples, and efficient prediction of categorical (classification) as well as continuous (regression) feature data. (Huang et al., 2001)

The CART method has been used to predict a number of phonetic qualities, including rules for allophones and prosodic features. For Swedish, Frid (2003) automatically modelled rules for segmental as well as prosodic qualities. His LTS (letter-to-sound) conversion rules for 78125 words resulted in 96.87% correct predictions for all letters. Frid also used CART learning to predict prosody both by letter and by whole-word patterns. The result for the prosody prediction was 88.6% correct predictions. Frid also had some success in predicting Swedish word accent and dialect.

In the present paper, to separate the CART method from the actual trees, the term 'CART' will denote a single decision tree, while 'CARTs' will be used about more than on tree, and when referring to the method, the term will be used only in phrases, i.e. 'the CART method', or 'prediction using CART'.

1.2 Purpose and aim

The purpose of this study was to gain more phonetic knowledge about correlates to speaker age found in different types of phonemes, and to take a first step towards building an automatic predictor of age. Attempting to predict exact *age* (in years), *age group* (old or young) and *gender* (to be used as an input feature to age predictors) by means of a very tentative strategy, the aim was not to construct a state-of-the-art predictor, but rather to answer two questions and to test two hypotheses:

Questions:

1. Which features would an automatic predictor of adult speaker age need, which features seem to be the most important, and how do they correlate with the cues used by human listeners?
2. Could an automatic predictor of adult speaker age, constructed with an easily understandable method using limited features and speech data, actually perform reasonably well, and if so - how would it compare to human perception of age described in an earlier study (Schötz, 2004)?

Hypotheses:

1. Automatic predictors would use separate strategies (i.e. features) for different segments, as many phoneme types (e.g. vowels, fricatives) contain different kinds of phonetic information
2. Gender would be a good input feature for automatic prediction of adult speaker age, as men and women age differently (Schötz, 2004).

2 Material

In order to be able to compare the results of this experiment with the study of human age perception (Schötz, 2004), which was based on 24 elicitations of the single Swedish word *rasa* [ˈʁɑːsa] (collapse) produced by 24 speakers from two villages in southern Sweden, and taken from the SWEDIA 2000 speech database (Bruce et al., 1999), the same type of material was used here. It consisted of 2048 elicitations of *rasa* produced semi-spontaneously in isolation by 428 adult equally many female and male speakers aged 17 to 84 years from 36 villages in southern Sweden (Götaland). Each speaker had contributed 3 to 14 elicitations of the word, and all were included to provide some within-speaker variation in the experiment. The words were normalized for intensity, just as in the human study.

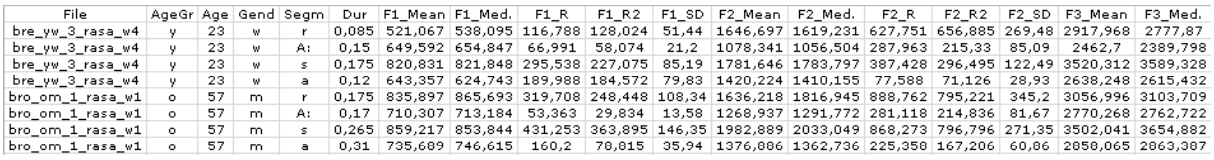
Using a number of scripts (developed by Johan Frid, Dept. of Linguistics & Phonetics, Lund University) for the speech analysis tool Praat (www.praat.org), some of which were further adjusted to suit the purpose of this study, the material was prepared for the CART experiments. First, the words were semi-automatically segmented and transcribed to the SAMPA (Speech Assessment Methods Phonetic Alphabet) alphabet as `rA:sa`. Provided with the orthographic transcription, the script used resynthesis of the word to segment and transcribe the words with fairly good accuracy. Automatic segmentation was preferred over manual in order to save time. Another script extracted 51 acoustic features from each segment, including measurements of fundamental and formant frequencies (F_0 and F_1-F_5) as well as relative intensity (mean, median, range, SD), segment duration, HNR (Harmonics-to-Noise Ratio), spectral emphasis, spectral tilt and several measurements of jitter and shimmer. There were a number of reasons why the features were extracted for each segment instead of e.g. once every

10 ms, which would have given more precise measurements. As the phonetic information contained in separate phonemes varies, the CART is likely to use different features to predict the various segments in order to generate better trees. Another reason was to keep the data size at a reasonable "pilot study level".

A description file containing all the feature names was created, and the extracted features were stored as vectors in two data files together with the following features:

- segment label (as different phonemes contain different acoustic information)
- biological age (in exact years, defined as a continuous feature, as not every age was included)
- age group (a binary feature, where 'old' was stipulated as 42 years or older, 42 being the youngest age defined as 'old' in the SWEDIA database, and 'young' as younger than 42)
- gender (a binary feature, which might influence age prediction).

One file was used only as a test set for comparison with the human listener study. It contained only the same 24 speakers and words (24 words * 4 segments = 96 vectors) that had been used in the human perception study. The other file comprised the other 404 speakers (1924 words * 4 segments = 7696 vectors), and was further split into a training set and a test set with 90% of the data for training and 10% for testing the CARTs. An example excerpt of the extracted data vectors is shown in Figure 1.



File	AgeGr	Age	Gend	Segm	Dur	F1_Mean	F1_Med.	F1_R	F1_R2	F1_SD	F2_Mean	F2_Med.	F2_R	F2_R2	F2_SD	F3_Mean	F3_Med.
bre_yw_3_rasa_w4	y	23	w	r	0,085	521,067	538,095	116,788	128,024	51,44	1646,697	1619,231	627,751	656,885	269,48	2917,968	2777,87
bre_yw_3_rasa_w4	y	23	w	A:	0,15	649,592	654,847	66,991	58,074	21,2	1078,341	1056,504	287,963	215,33	85,09	2462,7	2389,798
bre_yw_3_rasa_w4	y	23	w	s	0,175	820,831	821,848	295,538	227,075	85,19	1781,646	1783,797	387,428	296,495	122,49	3520,312	3589,328
bre_yw_3_rasa_w4	y	23	w	a	0,12	643,357	624,743	189,988	184,572	79,83	1420,224	1410,155	77,588	71,126	28,93	2638,248	2615,432
bro_om_1_rasa_w1	o	57	m	r	0,175	835,897	865,693	319,708	248,448	108,34	1636,218	1816,945	888,762	795,221	345,2	3056,996	3103,709
bro_om_1_rasa_w1	o	57	m	A:	0,17	710,307	713,184	53,363	29,834	13,58	1268,937	1291,772	281,118	214,836	81,67	2770,268	2762,722
bro_om_1_rasa_w1	o	57	m	s	0,265	859,217	853,844	431,253	363,895	146,35	1982,889	2033,049	868,273	796,796	271,35	3502,041	3654,882
bro_om_1_rasa_w1	o	57	m	a	0,31	735,689	746,615	160,2	78,815	35,94	1376,886	1362,736	225,358	167,206	60,86	2858,065	2863,387

Figure 1. An example of the acoustic features extracted for further storage in data- and description files.

3 Method

The preferred method for this study would be straightforward and easy to use. Combining statistical learning with expert (human) knowledge, the CART technique could use features that quite easily compare to the cues used by the human listeners in Schötz (2004). In addition, the existence of a ready-to-use application successfully used in previous phonetic studies (Frid, 2003) and the fact that the CART technique produces fairly human-readable trees, made the choice of method an easy one. The procedure for this limited time pilot study was tentative and unorthodox. Several problems were solved with similar methods to the ones used by Frid (2003) in his CART experiments.

3.1 Tools

In this study, *Wagon*, a CART implementation from the Edinburgh Speech Tools package, was used (Taylor et al., 1999). It consists of two separate applications: *wagon* for building the trees, and *wagon_test* for testing the trained trees with new data. *Wagon* supports discrete well as continuous features in both input and output. It also contains a large number of options for controlling the tree-building processes, of which only the three options controlled in the present study will be briefly explained here. A more detailed description of the *Wagon* tree building algorithm and its control options is given in Taylor et al. (1999). The *stop* value was used for fine-tuning the tree to the training set; the lower the value, the more fine-tuned and the larger the risk of an overtrained tree. If a low stop value is used, the overtrained tree can be pruned using the *held_out* option, where a subset is removed from the training set and then used for pruning to build smaller CARTs. All trees in this study were built with the *stepwise* option switched on, which instead of considering all features, looked for and incrementally used the individual best features in order to build smaller and more general trees, but at a larger computational cost.

3.2 Procedure

A number of test runs were carried out in search for the best decision trees for each feature. *Age* and *age group* were predicted both with and without gender as an input feature. *Gender* was then predicted using neither age nor age group as input features.

To reduce computation time, a subset of the data (489 words * 4 segments = 1956 vectors) was used in an initial search for the option values that would generate the best trees. The *stop* value was in turn set to 2, 3, 4, 5, 10, 20, 50 or 100, and the *held_out* value for pruning was varied with 0%, 10% or

20% of the data. These tests suggested that *stop* values of 3, 5 and 10 in combination with all three *held_out* values would generate the best prediction trees. In the remaining tests the options were restricted to these values.

Baselines were not easy to estimate, especially for *age*, as not every age was represented, and as the ages included in the training set were not equally distributed. As there were 54 ages in the data, a rough baseline for age might perhaps be calculated as 1/54 (≈ 1.85%), but this value is neither comparable to the correlation between predicted and biological age nor does it account for predictions of speakers with ages not included in the set. Both *age group* and *gender* were binary features. Female speakers were found in 3928 out of the 7696 vectors, so while one possible baseline for *gender* would be 51.04% (3928/7696), another would be 50%, given an expected equal distribution in the population to be predicted. For *age group*, a rough baseline might be 50%, since there were equally many (3848) vectors for older as for younger speakers. However, since the range of biological age was 42 (distributed as 36 different ages) for the old group, but only 18 (every age from 17 to 35) for the young group, this is not really a representative value. Thus, the baselines suggested in the result tables below should only be regarded as rough estimates of the performance of a baseline predictor.

In the first actual test runs, the whole data set containing all segments was used. Then, additional tests using only the vectors of one segment at the time were run in order to get some idea of which of the phonemes contained the best information for age and gender prediction, i.e. generated the best trees, but also to find out if the CARTs used different features from different segments for prediction.

Finally, tests of the same words used in the study with human listeners were run using the best CARTs for each segment and the results compared to the human results. The first (=best) features of the trees were compared to the cues used by the human listeners. The method and results of the study with human listeners is described in more detail in Schötz (2004).

4 Results

4.1 Tests with the whole data set

Control options and results (represented by *Wagon_test* as the correlation coefficient (*r*) between input and predicted feature for age, and by the percentage of correct predictions for age group and gender) for the best CARTs found with the whole data set (with all of the segments) are shown in Table 1.

Table 1. The results from the best CARTs using the whole data set for the features age, age group and gender.

continuous feature	*prediction…*	*stop*	*held_out*	*correlation*	*baseline*
age	…without gender	10	10	**0.344**	0.0185?
	…with gender	10	10	**0.385**	0.0185?
discrete feature	*prediction…*	*stop*	*held_out*	*correct (%)*	*baseline (%)*
age group	…without gender	10	0	**65.37**	50?
	…with gender	10	0	**66.80**	50?
gender	-	10	20	**83.63**	51.04?

The best predictions were achieved for *gender* (83.63% correct). For *age* and *age group*, the trees built with the input feature gender were only slightly better than the ones build without gender information.

4.2 Tests with one segment at the time

Table 2 shows the best prediction results for each segment. The best results for all features were obtained for the stressed vowel `A:`. Including gender as an input feature only marginally influenced the results of the trees. For `A:`, the best correlation between predicted and biological age was about 0.45, the best tree for age group predicted 72.14% correctly, and for gender this value was 90.62%.

Table 2. Results for the best CART predictions of age, age group and gender for each segment (best values in boldface, *stop*/*held_out* values within parentheses).

segment	*age (without gender)*	*age (with gender)*	*age group (without gender)*	*age group (with gender)*	*gender*
`r`	0.299 (10/10)	0.299 (5/20)	65.10% (5/20)	65.10% (5/20)	77.34% (10/20)
`A:`	**0.446** (5/0)	**0.454** (10/0)	**72.14%** (10/20)	**72.14%** (10/20)	**90.62%** (10/0)
`s`	0.406 (5/20)	0.393 (10/0)	64.06% (3/10)	64.84% (10/0)	80.99% (3/10)
`a`	0.273 (10/20)	0.286 (10/0)	63.28% (3/20)	63.28% (3/20)	87.50% (10/20)
baseline	0.0185?	0.0185?	50%?	50%?	51.04%?

The features used in the first yes-no questions in the best CARTs for each segment are shown in Table 3. For *age*, questions about the formant frequencies dominated, but F_0, relative intensity (Int.), HNR and shimmer were also used. Important cues for the *age group* CARTs included F_1-F_5, HNR, spectral emphasis (Sp.Emph.), relative intensity and shimmer, but F_0 was not included in the top three features of the trees. Not the same features were used in the first questions when gender was included in the input features as when it was excluded, and the feature gender was never used in any of the first three questions. The trees for *gender* prediction depended on first questions about F_0 values, but also on questions about F_1, F_2, F_5, relative intensity and spectral emphasis.

Table 3a-e. Top three features used by the best CARTs for each segment to predict age, age group and gender.

a) age (without gender)

a	r	A:	s	a
1st	F_3 (median)	F_4 (mean)	F_1 (range)	F_1 (range)
2nd	F_1 (range)	F_5 (range)	F_2 (range&mean)	F_4 (median)
3rd	Int. (mean)	F_0 (range)	F_4 (mean)	F_5 (median)

b) age (with gender)

b	r	A:	s	a
1st	F_3 (median)	F_4 (median)	F_1 (range)	F_2 (median)
2nd	F_2 (SD)	HNR	Int. (range&mean)	Shimmer
3rd	Int. (mean)	Int. (range)	F_2 (range)	F_0 (mean)

c) age group (without gender)

c	r	A:	s	a
1st	F_3 (mean)	HNR	F_1 (range)	F_1 (range)
2nd	F_4 (range)	F_1 (median)	F_3 (mean)	F_2 (mean)
3rd	F_2 (median)	Sp.Emph.	F_5 (mean)	Shimmer

d) age group (with gender)

d	r	A:	s	a
1st	F_3 (mean)	HNR	F_1 (SD&range)	F_1 (range)
2nd	F_4 (range)	F_1 (median)	Int. (mean)	F_2 (mean)
3rd	F_2 (median)	Sp.Emph.	F_2 (range)	Shimmer

e) gender.

e	r	A:	s	a
1st	F_0 (median&mean)	F_0 (median)	F_5 (median)	F_0 (median)
2nd	Int (range)	F_1 (mean)	F_2 (median)	F_1 (median)
3rd	Sp.Emph	F_5 (mean)	F_0 (mean)	F_0 (range)

4.3 Comparisons of results by the CARTs and the human listeners

In Table 4 the mean estimated ages for the 24 speakers by the 30 human listeners in the study by Schötz (2004) were compared to the predictions of the best CART. Human estimations were better for 13 speakers, while the CART more accurately predicted 9 of the speakers. Two speakers were estimated equally well by both humans and the CART. Neither the human listeners not the automatic predictor was considerably better than the other at judging the age of female or male speakers.

Table 4. Biological age and age estimations by human listeners and the CART for A: for the 24 speakers (closest estimations in boldface, speaker ID = village (a, s) + age group (o, y) + gender (m, w) + number (1-3)).

spkr ID:	*syw1*	*sym2*	*sym1*	*syw3*	*aym1*	*aym2*	*ayw2*	*ayw3*	*aym3*	*sym3*	*syw2*	*ayw1*	*aom3*	*aow3*	*aow1*	*som1*	*aom1*	*som3*	*sow3*	*aom2*	*sow2*	*sow1*	*som2*	*aow2*
bio. age	18	20	22	24	27	27	28	28	29	29	30	31	60	60	61	62	66	70	70	71	72	73	76	82
human	36	49	39	27	**43**	**28**	**30**	**24**	41	**34**	45	**35**	46	47	**61**	**51**	**60**	**68**	57	**62**	**66**	**75**	**70**	**75**
CART	**24**	**48**	**25**	**24**	67	**26**	34	57	**28**	53	**32**	44	**72**	**70**	55	**73**	45	51	**65**	26	64	65	48	64

A comparison of the misjudgements (in years) made by the humans and the best CART is shown in Figure 2. The largest errors were made by the CART trying to predict the age of one young (aym1) and one old (aom2) male speaker. The mean absolute error for the CART predictions was 14.45 years, while the same figure for the human listeners was 8.89 years.

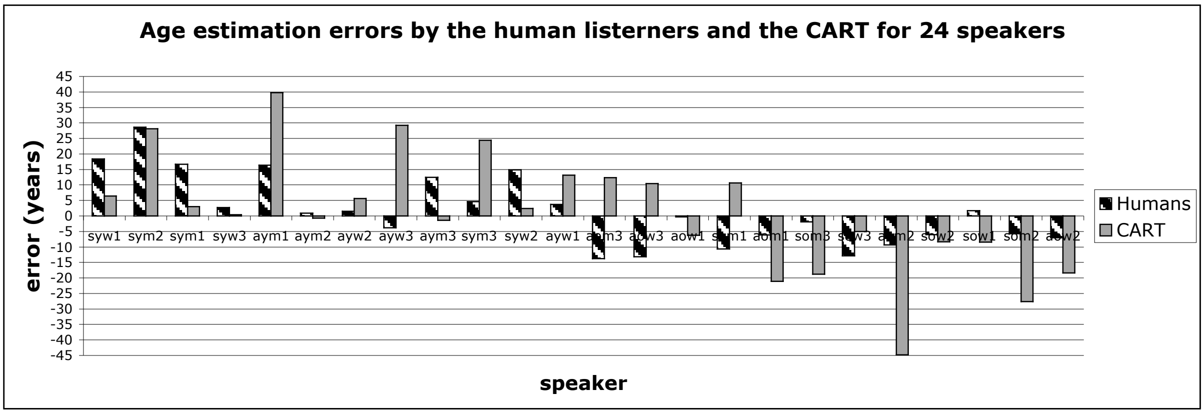


Figure 2. The deviation of the age estimations from biological age for human listeners (mean value) and the predictions made by the CART for the best segment of the tests (A:) for each speaker.

When comparing the features used by the CARTs to predict age with the acoustic correlates to the cues used in human listener study, several similarities were found. Spectral cues (e.g. formant frequencies) were dominant to F_0 for both humans and CARTs. However, the human study also found duration to be an important cue to age, while the CARTs did not use duration in their first questions.

5 Discussion and future research

The present study was a first attempt to build an automatic age predictor with the CART method to gain more phonetic knowledge about age. Although the CARTs did not predict age as well as humans, they still provided some interesting results, which point towards a number of problems yet to be solved in pursuit of a state-of-the-art predictor of speaker age. Some of these questions are discussed here along with a several other reflections and suggestions for future studies.

5.1 Reflections on the speech material and the method

Questions and suggestions related to the speech material include the choice of speakers, language and dialect, types of speech as well as the preparation of the material for the CART tests.

This study used only 428 speakers of southern Swedish dialects. Due to the aim of the SWEDIA project to document only a younger and an older generation of adult speakers, not every biological age could be represented in the speech material. Although gender was evenly distributed, with 214 female and 214 male speakers, no speakers were under 17, over 84 or between 36 and 42 years old. Most younger speakers were between 20 and 33 years, and most older speakers between 55 and 77 years. This must have affected the CARTs. There was, however, a considerable dialectal variation present in the data, including variations of the Swedish grave word accent, as well as allophonic variation of the phonemes /r/ and unstressed /a/, with pronunciations from the standard Swedish [ˈɹɑːsa] to [ˈʁɑːsa], [ˈʁɑːsə] and even [ˈwɑːsə]. In future studies, the purpose of the predictor would determine how much and what kind of speech data is needed to build general enough trees, as more speakers, dialects and languages provide more between-speaker variability, and more types of speech from each speaker implies more within-speaker variation,

The right choice and combination of acoustic features are likely to build better CARTs. More and improved methods to automatically extract acoustic features, like better inverse-filtering techniques for laryngeal features, ways to extract reliable values for LTAS (Long Time Average Spectra), formant bandwidths (B_1-B_5) and levels (L_1-L_5) may also improve the trees. Other possible methods include building segment-independent predictors of age by extracting features at regular time intervals, e.g. every 10 ms.

Features were extracted automatically in this study. Though timesaving when compared to manual feature extraction, one should always double check automatic methods to reduce the influence of outliers and artefacts. This was done only to some extent in this study.

Due to the small data size, one cannot be certain that the features used by the CARTs in this study actually mirror important age cues. More research with larger material is needed to determine this.

5.2 Comparing the tests with whole data set to the ones for each segment

The trees based on the whole data set did not perform as well as the ones that used only the segments `A:` or `s`. Most speech researchers agree that stressed vowels contain the most phonetic information, and the fact that the CARTs for `s` performed relatively well is in line with Schötz (2003), where it was found that the typical energy platform for [s] begins at higher frequencies for younger-sounding speakers. The segment `r` displayed a large allophonic variation among the speakers, which may explain the poor results of the CARTs for `r`. Segment durations may be another reason why the predictors for `A:` and `s` outperformed the ones for `r` and `a`. However, although `r` indeed was the shortest segment, the durations for `a` resembled those for `A:` and `s`, and none of the trees actually contained any early questions about duration. Future automatic predictors of age might use a technique to identify and extract only the longest segments containing the most acoustic information (e.g. stressed vowels and voiceless fricatives) from longer sequences of (spontaneous) speech and to base their predictions on them.

5.3 Comments on comparisons of CARTs with human age perception

It can be argued that the humans were better at predicting age than the CARTs, since the mean absolute error for the CART predictions was 14.45 years, but only 8.89 years for the human listeners. Such figures are hard to interpret for several reasons. How much did the outliers in the CART predictions influence the results? Is a machine that misjudges the age of speakers by approximately ±14 years a good or a bad predictor, compared to human listeners, and compared to chance? These questions are not easily answered, especially not when the results are based on such a limited material. The goal when building an automatic age predictor would probably not be to get absolutely correct predictions, but rather to be able to place a speaker in “her early twenties” or “his mid-seventies”.

Although age cues for human listeners displayed similarities with the features used by the CARTs, this does not mean that humans and automatic predictors use the similar strategies when estimating age. The features used by the CARTs may, however, give some indication on where to look for acoustic correlates to the cues of human age perception.

5.4 Additional comments and reflections

Is there really any practical use for an automatic predictor of age? Why can’t the system just ask the users about their age? There are at least two situations where this is difficult. One may occur in forensic situations, where objective age estimations of unknown potential suspects leaving a message on an answering machine may be of help. The other reason is more of a psychological or social nature. A number of users might be offended when asked how old they are. Not even computers should ask a lady about her age.

The experiences made in the present study might serve as a springboard for attempts to automatically predict other paralinguistic features with the CART method, leading to future improvements in speech and speaker recognition applications dealing with issues related to the personality of the user.

Automatic age predictors might also be helpful tools when trying to improve the naturalness of synthetic speech by including speaker-specific features in the synthetic voice. To synthesise speaker age, a CART for age prediction might be traversed from the leaf node of the desired age to the root of the tree, hereby adjusting the acoustic parameters of the synthetic voice.

Age is only one of many speaker-specific or paralinguistic qualities found in speech. In the future a combination of predictors for a number of such qualities, including age, gender, emotions, health, speaking style and even dialect may be of help in many speech and speaker recognition as well in spoken dialog systems. Computers would then be able to interact more naturally with the user, e.g. comfort a sad user, encourage an insecure user and even get angry and refuse to help a rude user. But would we really like a computer to behave like ourselves? In which situations would be acceptable for a spoken dialog system to behave like a human, and which would be completely out of the question? These questions remain to be answered.

6 Conclusions

From the pilot experiments in this study the following tentative conclusions were drawn:

1. Which features to use in state-of-the art automatic age predictors remains unclear. However, important features for the CART method in this study included formant frequencies, HNR and intensity, which is in line with human age perception, where spectral features are likely to dominate over F_0, but with duration as another important cue.
2. The CARTs for prediction of age seemed to use different tree-building strategies (in terms of input features) for different phonemes.
3. It is possible to construct a CART age predictor for one single word based on automatically extracted acoustic feature data with a performance slightly worse than human listeners’.
4. Although gender was predicted with >90% accuracy, information about gender did not seem to considerably influence the age predictions in this study.
5. Studies with methods to extract more acoustic features (laryngeal features, LTAS, B_1-B_5, L_1-L_5) and with larger more varied speech material are needed to further increase the phonetic knowledge about speaker age.

7 References

Bruce, G., Elert, C.-C., Engstrand, O., Eriksson, A. (1999). Phonetics and phonology of the Swedish dialects - a project presentation and a database demonstrator. *Proceedings of the XIVth ICPhS*. San Francisco. 321–324.

Frid, J. (2003) *Lexical and Acoustic Modelling of Swedish Prosody*. Travaux de l'institut de linguistique de Lund 45. Lund, Sweden: Department of Linguistics and Phonetics, Lund University.

Hollien, H. (1987) ”Old voices”: What do we really know about them?, *Journal of Voice*, vol. 1, no 1, pp. 2-13.

Huang, X., Acero, A., Hon, H. (2001) *Spoken Language Processing*. Prentice Hall. Upper Saddle River, New Jersey.

Jacques, R. D. & Rastatter, M. P. (1990) Recognition of speaker age from selected acoustic features as perceived by normal young and older listeners, *Folia Phoniatrica*, vol. 42, pp. 118-124.

Linville, S. E. (1987) Acoustic-perceptual studies of aging voice in women, *Journal of Voice*, vol. 1, no 1, pp. 44-48.

Minematsu, N., Sekiguchi, M., Hirose, K. (2003) Automatic Estimation of Perceptual Age Using Speaker Modeling Techniques. *Proceedings of Eurospeech 2003*. Geneva. Switzerland.

Ptacek, P. H. & Sander, E. K. (1966) Age recognition from voice, *Journal of Speech and Hearing Research*, vol. 9, pp. 273-277.

Schötz, S. (2003) A First Step From Analysis To Synthesis. *Proceedings of the XVth ICPhS.* Barcelona. 2585-2588.

Schötz, S. (2004) The role of F_0 and duration in perception of female and male speaker age. *Proceedings of Speech Prosody 2004*. Nara.

Taylor, P., Caley, R., Black, A., King, S. (1999). Edinburgh Speech Tools Library. System Documentation Edition 1.2. Website: http://festvox.org/docs/speech_tools-1.2.0/book1.htm.