Age Recognition for Spoken Dialogue Systems: Do We Need It?

Maria Wolters, Ravichander Vipperla, Steve Renals

Centre for Speech Technology Research
School of Informatics, University of Edinburgh, Edinburgh, Scotland
maria.wolters@ed.ac.uk, r.c.vipperla@sms.ed.ac.uk, s.renals@ed.ac.uk

Abstract

When deciding whether to adapt relevant aspects of the system to the particular needs of older users, spoken dialogue systems often rely on automatic detection of chronological age. In this paper, we show that vocal ageing as measured by acoustic features is an unreliable indicator of the need for adaptation. Simple lexical features greatly improve the prediction of both relevant aspects of cognition and interactions style. Lexical features also boost age group prediction. We suggest that adaptation should be based on observed behaviour, not on chronological age, unless it is not feasible to build classifiers for relevant adaptation decisions.

Index Terms: age recognition, pitch, keyword spotting, cognitive ageing

1. Introduction

Spoken dialogue systems (SDS) provide an invaluable source of information about user characteristics such as age, gender, and emotion: the user's speech. Detecting these characteristics from speech input allows systems to automatically adapt to user needs [1].

In this paper, we look at one particular instance of this paradigm: detection of older voices in order to adapt systems to the particular needs of older people. Older people potentially require the system to accommodate age-related changes in voice, speech, and cognition [1, 2]. Other aspects that may be adapted to older people may include the products and services offered initially, the formality of system messages, and the level of politeness [1]. To complicate matters, older users are also more likely to be dissatisfied with badly designed systems or systems that cannot accommodate their particular interaction style [3, 4, 5].

Recent research has focussed on inferring chronological age from utterance-level and frame-based acoustic features [6, 7, 8]. But is the predicted age group sufficient for automatically adapting SDS to the needs of older users? If there is a lowest common denominator of the literature on ageing, it is diversity. This makes older people very difficult to design for [9]. Hence, it might be more useful to predict the need for specific adaptations directly, without a detour via age recognition.

In this study, we predict two relevant user characteristics, information processing speed and interaction style using both predicted age group and specially constructed regression models. In our models, we use only features that are easy to extract from the speech waveform: pitch-related features, vocal tract length warping factor, speaking rate, mel-frequency cepstral coefficients (MFCCs) and the frequency of word groups that can be determined using simple keyword spotters. We find that age is not a useful hidden variable for determining relevant user characteristics. Specialised models almost always outperform models that only use predicted age. Lexical features in particular boost classification.

2. Background

We can estimate speaker age from vocal cues because age-related changes in anatomy and physiology affect the vocal folds and the vocal tract [10]. In particular, F0-related measures such as jitter, shimmer, and overall F0 statistics have been shown to correlate with ageing [11]. Long-term average spectra also change [12]. State-of-the-art algorithms for age recognition exploit MFCCs and other features commonly used in speaker recognition [7].

Once age has been estimated, we can accommodate the specific requirements of older users. For example, the system can switch to different acoustic and language models [13, 14], change the number of options presented to accommodate lower working memory spans [15], or adapt dialogue management strategies [5].

However, this approach is not without problems. The age that can be estimated from vocal cues ("perceived age") is not necessarily a good indicator of chronological age [10]. Therefore, we need to assess whether estimated age can be used to predict relevant characteristics of older users. In particular, we will compare age-based classifiers with models derived from easily extracted acoustic and lexical features.

3. Material

3.1. Participants

Our data set consists of 448 appointment scheduling dialogues between 50 participants and 9 spoken dialogue systems. 26 participants were older, with an average age of 66 (SD = 9.1, range: 52–84) and 24 participants were younger, with an average age of 22 (SD = 2.7, range: 18–29). All participants underwent a comprehensive battery of cognitive tests [4]. There were significant differences between younger and older participants on all tests in the expected direction [4]. Older users had a shorter working memory span, slower information processing speed, lower fluid intelligence, and higher crystallised intelligence.

3.2. Data Collection

The dialogue data was collected using a Wizard-of-Oz paradigm. Users were asked to schedule one appointment

Copyright © 2009 ISCA

each with nine system-initiative simulated SDS. The systems differed in the number of options presented at each stage and the confirmation strategies used. For a comprehensive description of the experimental setup, recording, transcription, and annotation, see [4, 16]. Task success was measured using task completion, number of tasks completed correctly, and appointment recall, while user satisfaction was measured using a questionnaire.

Information Processing Speed(DSST): Although the dialogue strategy used did not affect task success [4], we found that users with lower information processing speed had more problems recalling the appointment correctly. None of the other cognitive variables correlated with task success. Since information processing speed correlates with task success, we chose it as a target variable for prediction. We measured information processing speed using the digit/symbol substitution subtest of the Wechsler Adult Intelligence Scale [17]. The resulting variable is called DSST. Our older participants had an average DSST score of 51 (SD = 11, range: 21–70), while younger participants had an average DSST score of 75 (SD = 8.6, range: 63–93).

Interaction Style (IS): We also found significant differences in interaction style between older and younger users [5]. "Social" users produced a large number of `social` speech acts, showed a high level of user initiative, and tended to use synonyms for simple answers such as "yes" and "hello" as well as words used in interpersonal communication such as "hello" and "please". "Social" users were significantly less happy with the simulated SDS than "factual" users. They also did not adapt their interaction style to the dialogue systems over the course of nine dialogues. 62% of all older users (n=16) and 4.2% of all younger users (n=1) used a Social interaction style.[1]

3.3. Features

All features were rescaled to have a mean of 0 and a standard deviation of 1.

Acoustic Features: Mel Frequency Cepstral Coefficients (MFCC) were computed from the speech utterances using a window size of 25 ms and a frame shift of 10 ms. 14 features were used, one per MFCC plus energy. The first 10 seconds of each speaker's utterances were used as input to the classifier. Speaking rate in phonemes per second was computed based on forced alignment of the waveforms with the transcription. The vocal tract length normalisation warping factor VTLN was computed using the HTK toolkit [18]. For all vowels, we obtained five shimmer values, five jitter values, the mean noise-to-harmonics (NTH) ratio, the mean harmonics-to-noise (HTN) ratio in dB, the fraction of unvoiced frames (% UV), and minimum, maximum, mean, and median F0 in Hz using the Praat voice profile [19]. Vowels with F0 values greater than the 95^{th} quantile of all values or smaller than the 2.5^{th} quantile for a given variable (mean, median, min, max) were excluded from analysis, since these values were likely to have been affected by pitch detection errors.

Table 1: *Definition of word groups selected for final regression models. (etc. = and variants)*

Cat.	Description	Cat.	Description
`yes`	"yes" etc.	`no`	"no" etc.
`pos`	positive feedback, not `yes`	`neg`	negative feedback, not `no`
`please`	"please" etc.	`sorry`	"sorry" etc.
`hello`	"hello" etc.	`bye`	"good-bye" etc.
`hes`	hesitations	`trunc`	truncated words

Lexical Features: In addition to the acoustic features, we used word-class frequencies as defined in [5]. Table 1 summarises all word classes used as the input for feature selection. These frequencies can be easily computed online using a keyword spotter. We also included a count of hesitations and truncated words because these might indicate disfluencies, a potential sign of high cognitive load.

4. Method

The classifiers constructed for this study use linear regression for the normally distributed target variable DSST (Shapiro test, $p < 0.85$) and logistic regression for the binary variables IS and Agel. For the MFCC-based classifier, we used L2 regularised logistic regression and support vector machines [20], because this is better suited to dealing with large amounts of training data; for all others, we used the R functions `lm` and `glm` [21]. We use regression because the resulting models are easy to interpret and allow us to establish and quantify the strength of links between predictors and target variables.

Since our data set is comparatively small with 50 data points, it is particularly vulnerable to overfitting when using large feature vectors. Therefore, we use feature selection to construct models that generalise well. We first selected promising jitter, shimmer, and pitch features using stepwise feature selection (`stepAIC`, [22]). Starting from a constant baseline model, the method selects the predictor from the variable pool that explains the largest amount of residual deviance until the Akaike Information Criterion (AIC, [23]) no longer decreases. We then combined those features with the remaining variables (mean NTH, mean HTN, % unvoiced, rate, VTLN) for final selection of all voice-related features (`Voice` model). Next, we used the same feature selection procedure to find a set of appropriate lexical features (`Lex` model). Finally, we applied the feature selection procedure to the combined set of voice features and lexical features to yield a model that incorporates both voice and lexical features (`Voice+Lex` model).

[1] Each user's interaction style was described by a cluster of three feature sets: overall dialogue statistics, speech act group frequency, and word group frequency. We clustered user behaviours both based on each feature set in isolation and on a combination of all three feature sets. For all four input vectors (dialogue stats, speech act groups, word groups, complete feature sets), the best solutions consisted of two clusters, which overlapped to a large degree [5]. For our regression experiments, we predict interaction style derived from speech act frequencies in order to avoid circularity. The two outliers excluded from the original analysis are assigned to the "Social" cluster, since they represent extreme examples of "social" users.

All regression models were evaluated using leave-one-out cross-validation. For MFCC features, the model was trained on all frames from n-1 speakers and then evaluated on the frames of the nth speaker. Age group was assigned based on the majority vote. The leave-one-out results of the `Voice`, `Lex`, and `Voice+Lex` age models were used as "predicted age" for predicting IS and DSST. The output vector of the MFCC-based classifier was not tested separately because performance is very similar to the `Voice` model.

5. Results

The resulting regression models for age, IS, and DSST are given in Table 2. Table 3 summarises accuracy, precision, and recall for detecting whether a user is older. All models represent a clear improvement over the baseline classifier, which only predicts the most frequent age group (i.e. "older"; accuracy and precision: 52%, recall: 100%). Voice features and MFCCs have similar accuracy. Using lexical features allows a more precise age detection than using voice features or MFCCs alone; adding lexical features to voice features mainly improves precision.

Table 2: *Predictors used in Regression Models*

Variable	Model	Predictors
Age	Voice	min F0, med F0, mean F0, shimmer AB
	Lex	`please`, `pos`, `thanks`, `hello`
	Voice+Lex	`please`, `pos`, min F0, med F0
IS	Voice	Rate, med F0, VTLN
	Lex	`please`, `sorry`, `hes`, `yes`
	Voice+Lex	`sorry`, VTLN, `hes`
DSST	Voice	med F0, shimmer AB, mean F0, mean NTH, % UV, mean HTN
	Lex	`pos`, `sorry`, `neg`
	Voice+Lex	`pos`, mean NTH, mean F0, med F0, `sorry`

Since most of our participants use a "factual" interaction style, accuracy is acceptable at 66% if we default to this class as our baseline (cf. Table 4). However, we are interested in finding participants who are likely to have a "social" interaction style. None of the age-based logistic regression models improved on the baseline; they always predicted the most frequent interaction style. When we estimate IS using age-specific defaults (i.e. older users are "social", younger users are "factual") and real age groups, we obtain near-perfect recall (94.12%) and respectable precision (61.54%). Although the voice-based model fails to improve on this performance, using lexical features gives us nearly perfect scores. Most importantly, precision is greatly improved, which means that the more sophisticated model can accurately discriminate between "social" older users and "factual" older users.

The baseline RMS error of a classifier that predicts the same DSST score for all users is 15.50. Using just the user's real age group, we can decrease RMS to 10.23. Table 5 lists RMS values for the three models based on predicted age and the three models constructed from scratch. The classifier trained on both acoustic and lexical features is slightly, but not significantly better than the model using real age. The classifiers based on voice features alone are significantly worse than the classifier based on real age (one-sided Mann-Whitney test, p $<$ 0.0026 using age predicted from voice, p $<$ 0.049 using voice features).

Table 3: *Results for Age Prediction (Target: Older)*

Model	Acc	Prec	Recall
Voice	70.00	84.62	66.67
Lex	82.00	76.92	86.96
Voice+Lex	90.00	92.31	88.89
MFCC	70.00	73.90	65.40

Table 4: *Results for Interaction Style (Target: Social)*

Model	Acc	Prec	Recall
Voice	76.00	41.18	77.78
Lex	96.00	100.00	89.47
Voice+Lex	100.00	100.00	100.00

6. Discussion

We constructed classifiers based on voice and lexical features for two user characteristics that predict task success and user satisfaction of older users in an appointment scheduling task: information processing speed (DSST) and interaction style (IS). For both target variables, classifiers based on predicted age were outperformed by models that were trained to predict the variables of interest directly. Lexical features in particular improved performance. This result is independent of the performance of our age prediction algorithms. A dedicated IS model greatly outperforms the best possible classifier based on actual age, while a specific DSST model approximates the performance of a classifier based on actual age. This suggests that systems may benefit from predicting relevant user characteristics directly whenever feasible, not infer them from predicted age. We acknowledge that often, this is not possible or desirable (e.g. for age-specific product suggestions).

The best performance is seen when lexical features are combined with voice features, while voice features on their own perform relatively poorly. The same lexical features used for dedicated classifiers also boost age prediction. Since the features we used are not task specific, they should transfer reasonably well to other tasks and classifiers. We plan to investigate this issue in future work.

The greatest disadvantage of our approach is that compiling appropriate lexical frequency statistics takes time, whereas systems such as the one discussed by Metze et al. [1] use voice features determined from the first response to adapt almost all aspects of the subsequent interaction, including products presented to the user. Although some of the word groups (`please`, `hello`) used are likely to occur in the first utterance or two, word groups

Table 5: *Results for DSST. RMS-A: Models based on Predicted Age from Feature Set, RMS-M: Models directly derived from Feature Set*

Feature Set	RMS-A	RMS-M
Voice	14.36	12.29
Lex	13.29	12.20
Voice+Lex	11.34	9.77

such as `pos`, `sorry`, or `hes` can typically only be detected during the dialogue. One solution would be to compile relevant lexical frequency statistics as part of the dialogue history, and use them to change system behaviour when problems arise.

7. Acknowledgements

This research was funded by the Strategic Research and Development Grant MATCH (Scottish Funding Council grant no. HR04016). This work has made use of the resources provided by the Edinburgh Compute and Data Facility which is partially supported by the eDIKT initiative (http://www.edikt.org.uk).

8. References

[1] F. Metze, R. Englert, U. Bub, F. Burkhardt, and J. Stegmann, "Getting closer – tailored human-computer speech dialog," *Universal Access in the Information Society*, 2009.

[2] M. Zajicek, "A methodology for interface design for older adults," in *Enterprise Information Systems Vi*, 2006, pp. 285–292.

[3] E. C. Stephens, C. M. Carswell, and M. M. Schumacher, "Evidence for an Elders' Advantage in the Naive Product Usability Judgments of Older and Younger Adults," *Human Factors*, vol. 48, pp. 422–433, 2006.

[4] M. Wolters, K. Georgila, R. Logie, S. MacPherson, J. Moore, and M. Watson, "Reducing working memory load in spoken dialogues," *Interacting with Computers*, vol. in press, 2009.

[5] M. Wolters, K. Georgila, S. MacPherson, and J. Moore, "Being old doesn't mean acting old: Older users' interaction with spoken dialogue systems," *ACM Transactions on Accessible Computing*, in press.

[6] J. Ajmera and F. Burkhardt, "Age and gender classification using modulation cepstrum," *Proc. Speaker Odyssey*, 2008.

[7] F. Metze, 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*, vol. 4, 2007, pp. 1089–1092.

[8] C. Müller and F. Burkhardt, "Combining short-term cepstral and long-term prosodic features for automatic recognition of speaker age," *In Proceedings of Interspeech*, 2007.

[9] P. Gregor, A. Newell, and M. Zajicek, "Designing for Dynamic Diversity - interfaces for older people," in *Proc. ASSETS*, 2002, pp. 151–156.

[10] L. O. Ramig, S. Gray, K. Baker, K. Corbin-Lewis, E. Buder, E. Luschei, H. Coon, and M. Smith, "The aging voice: a review, treatment data and familial and genetic perspectives," *Folia Phoniatrica and Logopaedica*, vol. 53, pp. 252–265, 2001.

[11] S. E. Linville, "The sound of senescence," *Journal of Voice*, vol. 10, pp. 190–200, 1996.

[12] ——, "Source characteristics of aged voice assessed from long-term average spectra," *Journal of Voice*, vol. 16, pp. 472–479, 2002.

[13] S. Anderson, N. Liberman, E. Bernstein, S. Foster, E. Cate, B. Levin, and R. Hudson, "Recognition of elderly speech and voice-driven document retrieval," in *ICASSP*, vol. 1, 1999, pp. 145–148.

[14] R. Vipperla, M. Wolters, K. Georgila, and S. Renals, "Speech Input from Older Users in Smart Environments: Challenges and Perspectives," in *Proceedings of HCI International*, San Diego, CA, 2009.

[15] P. M. Commarford, J. R. Lewis, J. Al-Awar Smither, and M. D. Gentzler, "A Comparison of Broad Versus Deep Auditory Menu Structures," *Human Factors*, vol. 50, no. 1, pp. 77–89, 2008.

[16] K. Georgila, M. Wolters, V. Karaiskos, M. Kronenthal, R. Logie, N. Mayo, J. Moore, and M. Watson, "A fully annotated corpus for studying the effect of cognitive ageing on users' interactions with spoken dialogue systems," in *Proc. of LREC*, 2008.

[17] D. Wechsler, *Manual for the Wechsler Adult Intelligence Scale-Revised*, The Psychological Corporation, New York, 1981.

[18] S. Young, G. Evermann, *et al.*, *The HTK book version 3.4*, Engineering Department, University of Cambridge, 2006. [Online]. Available: http://htk.eng.cam.ac.uk

[19] P. Boersma and D. Weenink, *Praat: Doing phonetics by computer*, 2008. [Online]. Available: http://www.praat.org

[20] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[21] R Development Core Team, *R: A Language and Environment for Statistical Computing (Version 2.6.2)*, R Foundation for Statistical Computing, Vienna, Austria, 2008. [Online]. Available: http://www.r-project.org

[22] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S-PLUS*, 2nd ed. New York, NY: Springer, 1997.

[23] H. Akaike, "A new look at statistical model identification," *IEEE Trans. Automatic Control*, vol. 19, 1974.