Exploiting Speech for Recognizing Elderly Users to Respond to their Special Needs

Christian Müller, Frank Wittig, Jörg Baus

Department of Computer Science
Saarland University, Germany
{cmueller, wittig, baus}@cs.uni-sb.de

Abstract

In this paper we show how to exploit raw speech data to gain higher level information about the user in a mobile context. In particular we introduce an approach for the estimation of age and gender using well known machine learning techniques. On the basis of this information, systems like for example a mobile pedestrian navigation system, can be made adaptive to the special needs of a specific user group (here the elderly). First we provide a motivation why we consider such an adaptation as necessary, then we outline some adaptation strategies that are adequate for mobile assistants. The major part of the paper is about (a) identifying and extracting features of speech that are relevant for age and gender estimation and (b) classifying a particular speaker, treating uncertainty, and updating the user model over time. Finally we provide a short outlook on current work.

1. Introduction

With systems getting more ubiquitous and mobile, designers are faced with the challenge of universal usability. This involves the need to accommodate for context and user diversity. The notion of context diversity covers areas such as different environments (indoor/outdoor) and different machines (desktop/pocket PC). User diversity refers to the problem that interfaces have to be designed to be usable by people with a wide range of needs and capabilities. One example of this diversity is the comparison between average aged adults and the elderly. Elderly people are one of the last groups to benefit from access to computers. What makes technology difficult for elderly people to use is that elderly people very often suffer from cognitive disabilities like age degenerative processes, motor impairments, short-term memory problems, and reduced visual and auditory capabilities [1]. These disabilities are often magnified by a person's unfamiliarity with the given technology and the different learning curves possessed by individuals.

Making systems easier to use for elderly people raises two questions. First: What kind of adaptation should a system provide, when knowing that the current user belongs to the group of elderly users? And second: How can a system acquire this information?

[2] address the first question by the example of a mobile pedestrian navigation system with a multi-modal dialog component. They suggest among other things that the speech output should be slower/louder and the GUI should be clearer in that the toolbars, buttons, maps and text be displayed in a larger format. Despite these improvements concerning certain features of the information presentation on a mobile assistant, there are more possibilities to adapt to different users and situations, especially in a navigation task. Navigational instructions could be given verbally, graphically or in a combination of these presentation modes. They could range from complete route directions to incremental instructions, which are given step by step. A user in a hurry probably prefers short instructions because in this case complicated presentations that must be decoded under time pressure often lead to vicious circles, where the sense of not understanding the presentations results in stress reducing the ability to decode the presentation and so on. In contrast to the aforementioned situation imagine a sightseeing tour where more elaborate utterances or highlighting interesting sites a long the way seem to be more adequate. Information about known or preferred routes enables a navigation system to optimize routes, e.g., minimizing the amount of turning points, thereby reducing the number of possibilities to take wrong turns. And last but not least, the user might be familiar with some types of presentation methods used in navigation tasks, e.g., ordinary street maps use certain coding conventions to convey their information and they rely on the ability of user to follow these conventions and to read maps. With regard to graphical presentations we have to adapt the level of detail in graphical presentations that is sufficient to successfully depict a route.

Whereas the before-mentioned general adaptation options for mobile assistants are investigated comprehensively (cf. [3]), specific heuristics to support elderly people are not. Results of several experiments suggest that acquiring and maintaining spatial orientation is more problematic for elderly than for younger people (cf. [4, 5]). [6] point out that age-related declines in spatial orientation may at least partly due specific injuries and disease processes that are statistically associated with aging, rather than being a concomitant of the aging process itself. There has also been a lot of discussion on male-female differences in path finding and navigation tasks, as in the case of aging. Women tend to use strategies appropriate to tracking and piloting, while men use strategies appropriate for navigation. It seems that men and women are attending to different cues during a traverse. They will pick up different information about the environment resulting in different navigation strategies (cf. [6]).

This paper leaves the question about appropriate adaptation strategies unanswered and focuses on the problem on how to acquire the necessary information, i.e. whether a specific user belongs to the group of elderly users or not. We consider speech as an important and rich source for gaining information about the speaker (user): Hearing someone's voice, we can in most cases recognize the gender of the speaker, estimate the age, and maybe even get an idea of what mood the speaker is in with regard to stress or emotions.

2. Extracting Relevant Features from Speech

Considering speech as a source to gain information about the user (speaker), the set of features can be divided into three groups: acoustic, prosodic, and linguistic features. Attending the latter group raises the problem, that the utterances have to be interpreted beforehand. Acoustic and prosodic features however can be extracted relatively easy before the actual speech recognition process. On this account, we focus on acoustic and prosodic features, and, by reviewing the literature, identified the acoustic features *jitter* and *shimmer* as appropriate to determine the age (and also the gender) of the speaker [7, 8, 9]. We implemented feature extractors for jitter and shimmer using the open source phonetic analyzing tool PRAAT.[1] PRAAT provides several algorithms for jitter and shimmer measurements yielding eight different values in sum (five jitter values and three shimmer values). The prosodic feature speech rate also emerged as a candidate for age estimation, but has not yet been taken into consideration.

2.1. Jitter

Jitter is defined as the maximum pertubation of fundamental frequency (F_0). Jitter values are expressed as a percentage of the duration of the pitch period. Large values for jitter variation are known to be encountered in pathological (and old) voices. Jitter in normal voices is generally less than one percent of the pitch period. We used five different jitter algorithms that are provided by PRAAT. Among them are: *Jitter Ratio* (JR) , *Period Variability Index* (PVI), and *Relative Average Pertubation* (RAP), that are well known from the literature ([10]), as well as the standard PRAAT jitter algorithm that is similar to RAP. The major differences are the following: JR determines cycle-to-cycle variability whereas PVI calculates a value that is akin to the standard derivation of a period. RAP compares the average of three cycles to a given period. In this vein, the effects of long term F_0 changes, such as slowly rising or falling pitch, are reduced.

2.2. Shimmer

Shimmer represents the maximum variation in peak amplitudes of successive pitch periods. Large values for shimmer variation are known to be encountered in pathological (and old) voices. Shimmer in normal voices is generally less than about 0.7db. Again, several algorithms (three) were used to retrieve multiple shimmer values. The differences are similar to the differences of the jitter algorithms. The *Amplitude Perturbation Quotient* for example attempts to desensitize long-term amplitude changes like RAP does for frequency variations. APQ uses eleven point averaging (average of eleven cycles). For a detailed description of jitter and shimmer algorithms, we refer to [10].

3. Learning Classifiers

In this section, we present an empirical study that we conducted in order to find out whether it is possible at all to classify users according to their gender and age on the basis of their speech. We used the features based on shimmer and jitter that have been discussed in the previous section. The study consisted of a comparison of the most commonly used machine learning (ML) approaches for classification tasks.

[1] www.praat.org

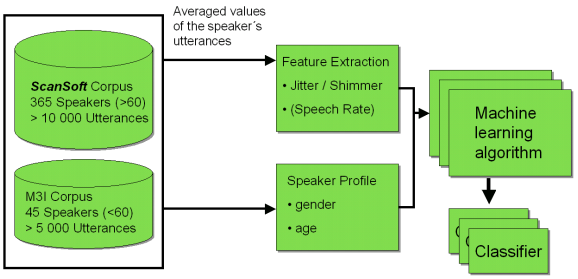

Figure 1: Age estimation procedure

Figure 1 depicts our approach. We analyzed a corpus with speech from elderly people that was provided by SCANSOFT[2] for this purpose. This corpus contained more than ten thousand utterances from 347 different speakers with an age of over 60 years. A second corpus that was collected within the M3I project contained about five thousand utterances from 46 speakers under 60 years. Both corpuses together consist of 162 female and 231 male speakers.

For our study, the average values of the 8 jitter and shimmer features per person were used. For a test whether an automatic classification is possible at all, the present data suffices, although we have to keep the uneven distribution regarding elderly and non-elderly speakers in mind when discussing the results.

We performed a ten-fold cross-validation procedure for each learning/classification method that we considered (for the informed reader we list the key parameters in parentheses, if any): C4.5 decision tree induction (DT), artificial neural networks (ANN, learning rate 0.15, momentum 0.2, 500 iterations), k-nearest neighbors (kNN, k=5, simple distance weighting), naive Bayes (NB) and support vector machines (SVM, C=20, polynomial kernel with degree 4). Particularly, we used the implementations of the WEKA collection of machine learning tools [11].

Table 1 shows the results with regard to the predictive accuracy. As a baseline (BL), we included the results for a simple classifier that always predicts the more frequently occurring class, i.e. elderly (88%) and male (59%) samples, respectively. This enables us to interpret the results in a more adequate manner instead of simply looking at the raw percentages that may lead to wrong conclusions—due to the uneven distribution of elderly vs. non-elderly and female vs. male speakers, respectively.

	C4.5	**ANN**	**kNN**	**NB**	**SVM**	**BL**
gender	67.79	81.09	75.62	67.34	70.51	58.78
age	92.68	96.57	95.71	91.15	96.52	88.30

Table 1: Results: prediction accuracy (percentages)

Overall, the different results show that it is indeed possible to create classifiers that are able to successfully predict age and gender on the basis of low-level acoustic features of the user's speech. Each method performs significantly better than the baselines 58.78 and 88.3 (two-tailed t-test, $p<0.01$). Artificial neural networks perform best in our study. This is a reasonable result since this method is known to be successfully

[2] www.scansoft.com

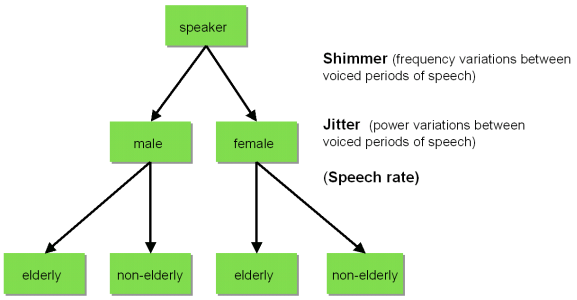

Figure 2: Classification hierarchy

applied frequently in such situations where raw sensor data has to be exploited. Note, that naive Bayes is in both cases the alternative that performs worst. This is most likely due to the fact that our data violates the basic assumption underlying the naive Bayes classifier: the independence of the feature values given the class value. Those 8 features used in our experimental setup are obviously not independent of each other. There are subsets that reflect mainly the same acoustic features of speech, i.e. variations of jitter and shimmer.

To get a better understanding of the performances of the classifiers with regard to our unbalanced data set, we present the true positive rates in Table 2, i.e., the rates of correct predictions for the two separate classes. Particularly, we present the results for the artifical neural network.

non-elderly	elderly	female	male
85.6	98.3	74.0	86.1

Table 2: Results: true positive rates

These results show that although our data is way from being evenly distributed, the classifiers are able to predict each class correctly with a rate higher than 74%.

Nevertheless, it is of minor interest which particular instance of the different learning/classification algorithms is able to outperform the others, the main result of our exploratory study is that we can indeed learn successful classifiers and that it is therefore worth to follow this line of research more intensively.

4. Taking Gender-Specific Aging Into Account

An important step towards the development of a successful recognition procedure is based on the observation that voices of men and women age differently [7]. In the following, we describe a two-level approach to incorporate this observation into the classification procedure.

Basically, we use gender-specific age classifiers that are trained using only speech data of men or women, respectively. If the user's gender is known, we could apply the corresponding classifier to decide whether she is an elderly or non-elderly person.

A straightforward solution would be to first classify the new user according to his/her gender followed by a gender-specific classification with regard to his/her age as outline in Figure 2. The main weakness of such a classification hierarchy is that if the user's gender is wrongly determined this error has an immediate negative influence on the ability to classify the user's age correctly (by choosing the wrong age classifier). A promising approach is to integrate the two separate parts of the gender/age classification procedure on the basic level within a probabilistic meta-reasoning framework such as Bayesian networks (BNs) [12].

A BN consist of two parts: (a) a directed acyclic graph G that encodes the causal relationships between the considered random variables and (b) a set of conditional probability tables (CPTs) that quantify the uncertain relationships represented by the links of G. Figure 3 shows the structure of the BN that we used to solve our classification task. There are direct causal dependencies between the user's actual gender/age and the results produced by the classifiers for a particular speech sample.

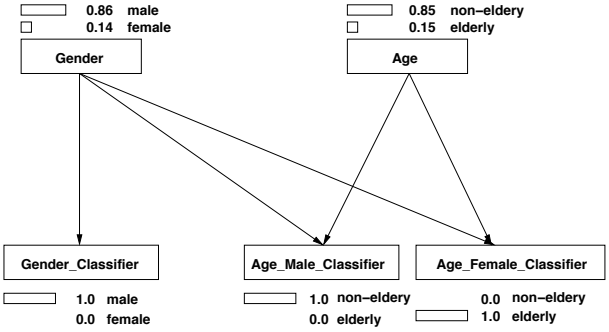

Figure 3: Bayesian network used to integrate the classification results.

The second part of the BN, the conditional probabilities can be computed as the true positive/negative rates of the learned classifiers when applied to corresponding samples. For example, the probability that the age_male_classifier indeed classifies an elderly male as elderly $P(\text{age\_male\_classifier} = \text{'elderly'}|\text{age} = \text{'elderly'}, \text{gender} = \text{'male'})$ can be estimated by testing the classifier's prediction accuracy only using data of elderly males. To estimate the corresponding negative rate, in our example $P(\text{age\_male\_classifier} = \text{'elderly'}|\text{age} = \text{'}\textit{non-elderly}\text{'}, \text{gender} = \text{'male'})$ a male dataset that has not been used for training consisting of samples of non-elderly males, has to be used as a test set. To exploit a limited dataset optimally, well-known cross-validation techniques could be used. For our dataset, we received values of 0.95 and 0.11 for the two probabilities, respectively. The remaining probabilities can be estimated analogously.

Figure 3 includes an example how the BN is used to estimate the user's gender and age. The results of the three (artificial neural network) classifiers are used as evidences for the variables at the figure's bottom that are interpreted by applying the BN reasoning mechanisms. This procedure yields posterior probabilities (conditioned on the basic classification results) for the states of both variables of interest on the figure's top line—gender (female/male) and age (elderly/non-elderly). The example situation represented in Figure 3 (male, non-elderly, elderly) yields a 85.08% probability that the user is non-elderly and a 86.37% probability that she/he is male by combining the partial classification results.

Using BNs as described for the integration of the different classification results has several benefits: (a) it yields explicit confidence values for the results by providing probabilities for the hypotheses under consideration, (b) it improves the quality of the results compared to a classification hierarchy with sequential binary decisions (first gender, then age) by taking into

account the uncertainty regarding the two possible outcomes female/male within the Bayesian reasoning procedure, and (c) potentially available confidence values of a classification result as provided by some classifier can be incorporated in a natural manner on the BN meta-reasoning level. Additionally, as we will briefly outline in the next section, the presented framework can be easily extended by modeling further aspects of the context, e.g. the presence of background noise.

5. Continuously Updating the Age and Gender Estimates

A question that remains to be answered is on how many speech samples the classification should be based. On the one side, to receive highly accurate estimates, a larger number is preferred to minimize the effects of random fluctuations in the user's speech or contextual influences such as noise. On the other side, the systems needs almost right from the start at least an initial estimation to base its adaptation decisions on. To address this issue, we embedded the classification procedure described in the previous section in a framework that is able to update its estimates over time based on dynamic Bayesian networks (DBNs, [13]).

Essentially, each time a new speech sample becomes available, a new instance of the BN of Figure 3 is connected to the previous ones, the current results of the classifiers are taken into account and the standard BN reasoning algorithms are applied as described before. Figure 4 shows an example where three utterances have been recorded and analyzed by the system. In this way, the estimates for age and gender are updated—converging to the real values as more and more speech samples are processed.

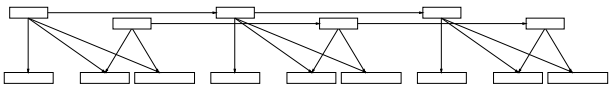

Figure 4: Example DBN.

The described approach using DBNs can be straightforwardly extended to incorporate changing aspects of the context, e.g. the quality of the microphones and the level of potentially present noise. Additional variables to model these contextual factors can be included in the time slices. For example, by connecting them to the classification results variables and specifying the conditional probabilities accordingly, we can represent the negative influence, e.g. a 10% accuracy reduction, of a higher level of noise present in the environment (which can be recognized quite easily) on the reliability of the results produced by the classifiers.

6. Summary and Current Work

In this paper we showed how to exploit raw speech data to gain higher level information about the user in a mobile context. In particular we introduced an approach for the estimation of age and gender using well known machine learning techniques.

Application scenarios include a mobile pedestrian navigation system with a multi-modal interface. Such an application benefits from the advanced user modeling by (a) the facility adapting the interface with regard to the special needs of a particular user group (the elderly) and (b) the improved speech recognition quality using specific acoustic models.

Currently, one line of work is collecting more data to balance the corpus and the implementation of extractors for prosodic features such as speech rate. Another line consists of the developing concrete strategies to adapt the systems behavior according to the special needs of elderly people.

7. Acknowlegdement

The authors want to thank Hewlett Packard for partially supporting the research presented in this paper under their *HP Voice Web Initiative Program*.

8. References

[1] J. A. Jorge, "Adaptive tools for the elderly: new devices to cope with age-induced cognitive disabilities," in *Proceedings of the 2001 EC/NSF workshop on Universal accessibility of ubiquitous computing*. ACM Press, 2001, pp. 66–70.

[2] C. Müller and R. Wasinger, "Adapting Multimodal Dialog for the Elderly," in *Proceedings of the ABIS-Workshop 2002 on Personalization for the Mobile World*, 2002.

[3] J. Baus, A. Krüger, and W. Wahlster, "A Resource-Adaptive Mobile Navigation System," in *IUI2002: International Conference on Intelligent User Interfaces*. New York: ACM Press, 2002, pp. 15–22.

[4] K. C. Kirasic, "Age differences in adults' spatial abilities, learning environmental layout, and wayfinding behavior," *Spatial Cognition and Computation*, vol. 2, pp. 117–134, 2000.

[5] U. Lindenberger and J. Kray, "Kognitive Entwicklung," in *Entwicklungspsychologie des mittleren und höheren Erwachsenenalters*, S.-H. Filipp and U. M. Staudinger, Eds. Hofgrefe Verlag, 2002 (In Druck).

[6] E. Hunt and D. Waller, "Orientation and wayfinding: A Review," Office of Naval Research, Arlington, Tech. Rep. N00014-96-0380, 1999.

[7] S. E. Linville, *Vocal Aging*. San Diego, Ca: Singular, 2001.

[8] S. Schötz, "A perceptual study of speaker age," in *Proceedings of Fonetik 2001*, A. Karsson and J. Van de Weijer, Eds. Lund Working Papers, 2001, pp. 136–139.

[9] N. Minematsu, M. Sekiguchi, and K. Hirose, "A perceptual study of speaker age," in *Proceedings of the International Conference of Acoustics Speech and Siganl Processing*, 2002, pp. 123–140.

[10] R. Baken and R. Orlikoff, *Clinical measurement of speech and voice (2 nd edition)*. San Diego: Singular publishing Group, 2000.

[11] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques With Java Implementations*. Morgan Kaufmann Publishers, 1999.

[12] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann, 1988.

[13] P. Dagum, A. Galper, and E. Horvitz, "Dynamic network models for forecasting," in *Uncertainty in Artificial Intelligence: Proceedings of the Eight Conference*. San Francisco: Morgan Kaufman, 1992, pp. 41–48.