Speech as a Source for Ubiquitous User Modeling

Christian Müller and **Frank Wittig**
Department of Computer Science
Saarland University
{cmueller, wittig}@cs.uni-sb.de

Abstract

In this paper, we present an approach on how to use speech as a source for user modeling in a mobile and ubiquitous context. In particular, we exploit different abstraction levels of speech features to estimate the user's age and gender. To solve the classification task, we compared several well known machine learning techniques such as artificial neural networks and support vector machines. The results of our study imply that one can indeed successfully extract higher level information from the raw speech data. We show how this approach is integrated into a generic resource adaptive system architecture. One particular instance of this system is an implementation of a mobile pedestrian navigation system.

1 Introduction

In ubiquitous computing, speech plays an important role as an interaction modality. Application scenarios like mobile navigation systems, shopping, tourist or museum guides, imply hands-free eyes-free situations, where the users can interact with the system by speech only. Therefore, we consider speech as an important and rich source for ubiquitous user modeling. Speech contains information about the speaker: 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.

The question of how this information can be made available for user modeling is currently addressed in the project M3I[1]. M3I is part of the BMB+F[2] funded project COLLATE[3] at the Saarland University and the DFKI[4]. The goal of the M3I project is to develop a framework for resource adaptive multi-modal dialog with mobile devices.

[1]A Mobile, Multi-modal, and Modular Interface

[2]Bundesministerium für Bildung und Forschung (Federal Ministry of Education and Research)

[3]Computational Linguistics and Language Technology for Real Life Applications

[4]Deutsches Forschungszentrum für Künstliche Intelligenz (German Research Center for Artifical Intelligence)

It consists of a central server component with which mobile devices can communicate via wireless network connections. While the mobile devices function on a stand-alone basis, the availability of the server improves the coverage and quality of the services. For example, the speech recognition that is implemented on the mobile device is limited due to lack of computing power and working memory. When connected to the server, the speech can be processed in parallel on the server much faster and with a larger vocabulary. Besides this, the server provides additional services like topic detection. That means that the server possesses a speech recognition module with a general language model that recognizes the domain to which the utterance of the user belongs. The integrated speech recognition module uses this information to load specific language models for this domain. The M3I framework is designed to implemented new components on the server side first. In this manner, the approach can be tested and improved easier before a slim embedded version is implemented. An example of such a component is the speech-based user modeling module that is described in this paper.

2 Features of Speech That Are Relevant for User Modeling

Regarding speech as a source for ubiquitous user modeling, the relevant features can be divided into three different levels of abstraction (see figure 1).

On the lowest level, there are *acoustic* features that are related to the signal's power and frequency and their changes over time. An example of such a feature is the jitter value that describes frequency variations between voiced periods of speech. Because they are based on physical properties, acoustic features are relatively easy to extract from the signal and independent from the language. They can be extracted before the actual speech recognition process is done. On the other hand, those features are sensitive to changes in the acoustic environment and the recording quality.

On the next level, there are *prosodic* features. Prosody refers to all aspects of sound above the level of segmental sounds, like intonation, stress and rhythm. Speech rate and pauses can also be assigned to this group. In most cases, prosodic features cannot be immediately derived from the physical properties. The extraction is therefore more expensive. In the case of speech rate and pauses they also have to

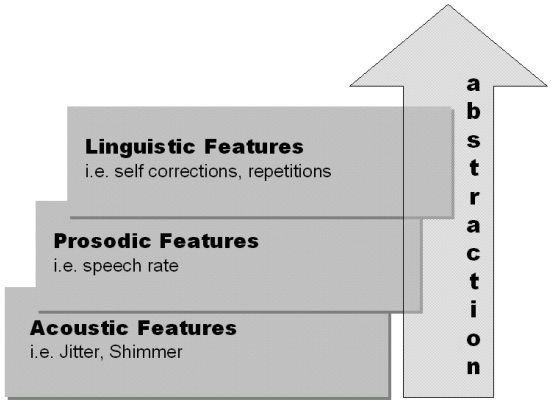

Figure 1: Three levels of abstraction of speech features

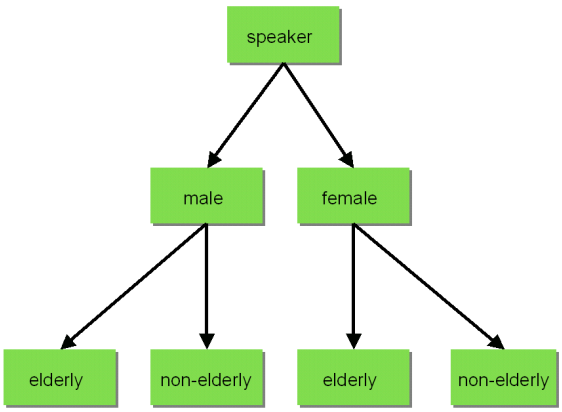

Figure 2: Classification hierarchy

be compared to a baseline, either of the individual speaker or of a group of speakers. Still, the extraction can be performed without understanding the content of the utterance.

This is no longer the case for *linguistic* features. Those features refer to the syntactical structure of the utterance, the number and category of the words, or even to their semantic content. To extract these features, natural language processing has to be done first.

Müller, Großmann-Hutter, Jameson, Rummer, and Wittig (2001) describe a study where prosodic and linguistic features were used to recognize the user's cognitive load and time pressure. The features were called "symptoms of cognitive load and time pressure" and were extracted manually from the speech by fully transliterating the utterances and rating the quality of the content. Some of the prosodic features that were found to be relevant for this task are: articulation rate (the number of syllables articulated per second of speaking time), silent pauses, and filled pauses (e.g., "Uhh"). Besides this, the following linguistic features were considered: (a) disfluencies (the logical disjunction of several binary variables, each of which indexes one feature of speech that involves its formal quality: self-corrections involving either syntax or content; false starts; or interrupting speech in the middle of a sentence or a word) and (b) content quality (the average quality assigned to the utterance).

The results were used for learning a Bayesian network (Pearl, 1988) that reflects the causal dependencies between the symptoms and the cognitive load and time pressure of the user. Müller et al. (2001) showed that this network can be successfully used for this particular classification task.

In the remainder of this paper we present an approach on how to extract acoustic and prosodic features of the speech for the purpose of recognizing the gender and the age of a user. In section 3 we provide a motivation, why the age of a user should be estimated by a system. In section 4, we present a case study on how to use machine learning techniques to induce classifiers for age and gender, and describe, how this approach is integrated into the above mentioned M3I architecture. In section 5 we briefly outline how the different abstract speech features introduced in figure 1 can be exploited within a single framework. We conclude in section 7 by directly referring to the relevant workshop questions.

3 The Elderly as a User Group

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 they very often suffer from cognitive disabilities like age degenerative processes, motor impairments, short-term memory problems, and reduced visual and auditory capabilities (Jorge, 2001). 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? Müller and Wasinger (2002) 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 and the GUI should be clearer in that the toolbars, buttons, maps and text be displayed in a larger format. In this paper, we focus on the second question: How an appropriate user model can be obtained automatically on the basis of speech.

4 Case Study: Using Machine Learning to Induce Classifiers for Age and Gender on the Basis of Acoustic Features

In the following, we will present an initial exploratory study regarding the classification of users on the basis of low-level acoustic features according to their gender and age. Within a comparison of the most commonly used machine learning (ML) approaches, we aimed to find out whether it is possible at all to identify a user's age and/or gender with ML methods.

The voices of men and women age differently (Linville, 2001). Therefore it is reasonable to first try to determine the gender of the user, before the age is estimated. Figure 2 depicts the corresponding classification hierarchy.

By reviewing the literature, we identified *jitter* and *shimmer* as appropriate feature to determine the gender and the

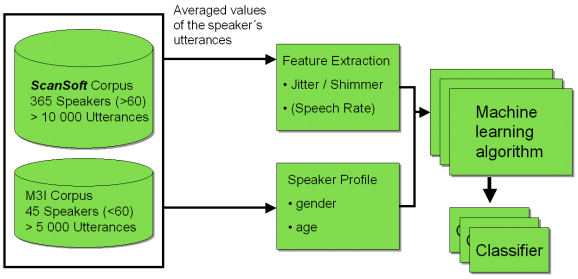

Figure 3: Age estimation procedure

age of the user (Linville, 2001; Schötz, 2001; Minematsu, Sekiguchi, & Hirose, 2002). Both features belong to the group of acoustic features according to the classification that was introduced in section 1. Besides this, the prosodic feature speech rate is also a candidate for age estimation, but has not yet been taken into consideration.

Jitter is defined as the maximum perturbation 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. 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 (see (Baken & Orlikoff, 2000));

Figure 3 depicts our approach. We analyzed a corpus with speech from elderly people that was provided by SCANSOFT [5] 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. Table 1 summarizes both corpuses including the number of female vs. male speakers. We implemented feature extractors for jitter and shimmer using the open source phonetic analyzing tool PRAAT.[6]

We used five different jitter and three different shimmer algorithms that are provided by PRAAT. The main jitter algorithms are: *Jitter Ratio* (JR) , *Period Variability Index* (PVI), and *Relative Average Perturbation* (RAP), that are well known from the literature ((Baken & Orlikoff, 2000)), 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. The differences between the shimmer algorithms are similar to the differences between 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 (Baken & Orlikoff, 2000).

All together, we received 8 features that can be used for classification. For the following initial study, the average values per person were used. In a future phase of the project, we plan to apply methods that continuously update the system's estimate of age and gender as more and more speech samples of the current user become observed. We are currently collecting more speech samples in order to work with a more balanced set, i.e. containing a larger number of non-elderly persons. Nevertheless, for an initial test whether the automatic classification is possible at all, the present data suffices, although we have to keep the uneven distribution of our samples in mind when discussing the results.

non-elderly	**elderly**
46	347

female	**male**
162	231

Table 1: Number of (non-)elderly and female vs. male persons in the data set

We performed for each learning/classification method that we applied a ten-fold cross-validation procedure. In our study, we considered the following methods (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 (Witten & Frank, 1999).

Table 2 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.

	C4.5	**ANN**	**kNN**	**NB**	**SVM**	**BL**
gender	69.10	81.73	76.41	67.26	70.43	58.78
age	92.41	96.75	95.76	91.25	96.45	88.30

Table 2: 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 applied frequently in such situations where raw sensor data has to be exploited. Note, that naive Bayes is in both cases—the prediction of age as well as gender—the alternative that

[5] www.scansoft.com

[6] www.praat.org

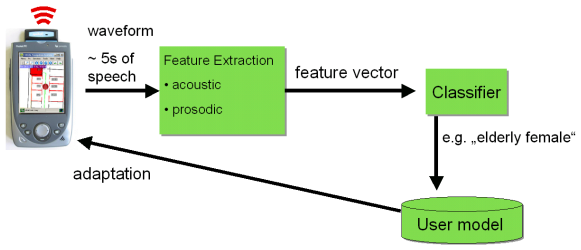

Figure 4: Integration into the M3I architecture

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 negative and positive rates in Table 3, respectively, 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
82.6	98.3	69.8	88.7

Table 3: 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 70%.

Nevertheless, as already mentioned, 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.

Figure 4 shows in a simplified way, how the age estimation component was integrated into the above mentioned M3I architecture. It is currently implemented as a server side service. The speech of the user is recorded on the mobile device (PDA) and then streamed to the server over a wireless network connection. On the server side, the relevant features are extracted and the corresponding values are used to classify the speaker according to age (elderly / non-elderly) and gender (female / male). The information is written into the user model and serves as a basis for adaptation.

5 Towards an Integrative Approach for Exploiting a Variety of Features of the User's Speech

As discussed in the introduction of this paper, speech in general is an important and rich source of information for ubiquitous user modeling. Each one of the three levels of abstraction of speech features as shown in Figure 1 may make its

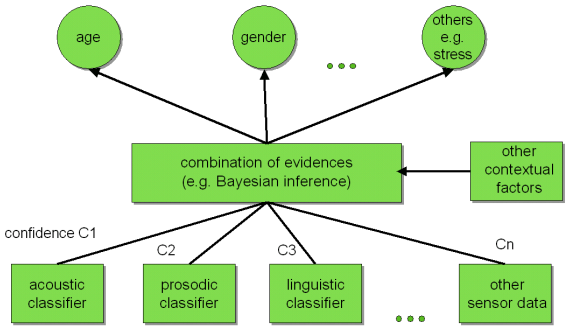

Figure 5: Integrative framework for exploiting a variety of features of the user's speech

own contribution in the context and user modeling process. Müller et al. (2001) have shown a system that is able to estimate a user's level of cognitive load and time pressure he/she is suffering from by interpreting high-level speech symptoms, e.g. self corrections, sentence breaks and so on. In this paper, we have described an initial study that strongly indicates that it is possible to successfully recognize some information about the user such as age and gender on the basis of low-level acoustic features of his/her speech. Related literature suggests that these features could also be used to reason about cognitive load and time pressure (Minematsu et al., 2002). In order to exploit a huge variety of available speech-related information for user modeling on the three different levels of abstraction, we briefly present an outline of an integrative approach along these lines.

Figure 5 represents the basic architecture. On the top, we have the variables of primary interest in our user/context model. These are connected to different classifiers that are used to interpret the different data streams on the different speech abstraction levels. To combine their results, each (single result of the) classifier comes along with a confidence value that measures the average success of the classifier. These confidence values could be estimated or computed on the basis of empirical studies as described in this paper or by Müller et al. (2001). If we interpret this architecture as the structure of a Bayesian network, these confidence values play the role of the conditional probabilities annotated at the links between the top row variables and the classifier variables, i.e., the probability that a particular classifier is able to correctly classify the speech symptoms in the situation under consideration. The implementation of this "meta-reasoning" scheme about the results of the different classifiers using BNs allows a flexible integration of other contextual factors or aspects of the domain that are relevant for the user model, such as input from bio sensors. The BN provides at any time up-to-date estimations in the form of probability distributions (conditioned on the available different pieces of information).

6 Coping with Mobile Speech and a Noisy Environment

When speech is recorded with mobile devices such as handheld PC, the quality of the signal is reduced due to the low

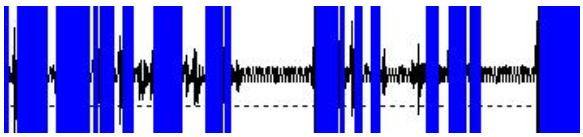

Figure 6: Voiced portions of the speech signal

microphone quality and environmental noise. This raises the question, whether the classifiers that were trained with clean data still perform well, when applied to mobile data.

First tests with a mobile device (HP Jornada) in fact showed, that the classification performance was far from the above mentioned cross validation results. Whereas gender was mostly recognized correct, young voices were often wrongly classified as elderly. However, we ascribe this fact partly to the unbalanced corpus (far more elderly voices) that may lead to a biased classification. In our approach, the effect of noise (environmental and microphone-induced) is reduced by the fact, that jitter and shimmer algorithms are based solely on voiced parts of the signal. As illustrated in figure 6, only a subset of the signal is treated (between 75 Hz and 600 Hz).

Nevertheless, we cannot exclude any impact of signal quality. Currently, we pursue two approaches to investigate this issue: (a) we implement filters that artifically reduce the quality of our training material. Thereby the characteristics of mobile recorded speech are simulated as close as possible. The performance of the resulting classifiers can then be compared with the one that were trained with unmodified data; (b) on a higher level we intend to incorporate a node NOISE into the above mentioned Bayesian network structure to model the impacts of noise explictly (e.g. reduce the confidence values of the classification when noise is likely present). Whereas (a) is under way, (b) requires a better understanding of the impacts of noise and will be focus of further research.

7 Conclusion with Regard to Workshop Questions 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. We classified the relevant speech features into three levels of abstraction each implying their own characteristics with regard to extraction costs and expressiveness.

We introduced the architecture of the M3I project, which copes with the limited resources of the mobile scenario by distributing services between mobiles devices and a server (see section 1). The age and gender estimation component that is described here was integrated into this architecture. A demonstration of the system can be given at the workshop.

Application scenarios within the M3I 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, investigating the impacts of noise as described in section 6, and the implementing extractors for prosodic features such as speech rate. Another line consists of the concrete realization of the above mentioned framework to integrate information of all three levels of speech features.

References

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

Jorge, J. A. (2001). 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* (pp. 66–70). ACM Press.

Linville, S. E. (2001). *Vocal Aging*. San Diego, Ca: Singular.

Minematsu, N., Sekiguchi, M., & Hirose, K. (2002). A perceptual study of speaker age. In *Proceedings of the International Conference of Acoustics Speech and Siganl Processing* (pp. 123–140).

Müller, C., Großmann-Hutter, B., Jameson, A., Rummer, R., & Wittig, F. (2001). Recognizing time pressure and cognitive load on the basis of speech: An experimental study. In M. Bauer, P. Gmytrasiewicz, & J. Vassileva (Eds.), *UM2001, User Modeling: Proceedings of the Eighth International Conference*. Berlin: Springer. (Available from http://dfki.de/∼jameson/abs/MuellerGJ+01.html)

Müller, C., & Wasinger, R. (2002). Adapting Multimodal Dialog for the Elderly. In *Proceedings of the ABIS-Workshop 2002 on Personalization for the Mobile World*.

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

Schötz, S. (2001). A perceptual study of speaker age. In A. Karsson & J. Van de Weijer (Eds.), *Proceedings of Fonetik 2001* (pp. 136–139). Lund Working Papers.

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