Implicit Feedback for User-Adaptive Systems by Analyzing the Users' Speech

Frank Wittig and **Christian Müller**

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

Abstract

This paper describes an approach to recognize the gender and age of a user on the basis of her/his speech. Within a two-level framework, machine learning techniques are applied to learn several appropriate classifiers whose results are combined on the top level to improve the overall classification of the user. As a part of the user model, the information about the user's age and gender can be an important basis for appropriate adaptation decisions.

1 Introduction

The development of more and more mobile and ubiquitous systems faces the systems designers with challenging problems regarding the interface design. Since in many mobile scenarios hands-free interaction is preferred, speech as a communication and interaction channel becomes more and more important. Therefore, the impact of speech as a source for user modeling increases and every piece of information about the user that can be extracted of her/his speech in an implicit way, i.e. without asking the user explicitly, can give valuable hints how to adapt the system's behavior in an adequate manner.

In the present paper, we will outline an approach to estimate a user's gender and age on the basis of her/his speech. Especially the age of the user may contribute to derive adaptation decisions that can be important to make it easier for the elderly to access modern computing facilities in their everyday lives, e.g. concerning museum assistant systems or talking to computers at call-centers on the phone. In the latter case for example, a reduction of the number of menu items per level in the menu hierarchy may make it easier for the elderly user to navigate to the desired functionality (although it may took a longer time).

The described method is a general framework that we plan to apply to other tasks in future, such as the estimation of a user's cognitive load and her/his affective state while (s)he is interacting with the system. Both can in some contexts be important aspects of the user model that is used by the user-adaptive systems to determine appropriate adaptation decisions. The age and gender recognition can be seen as a particular instance of this general framework.

2 Speech as a Source for User Modeling

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 (see Figure 1).
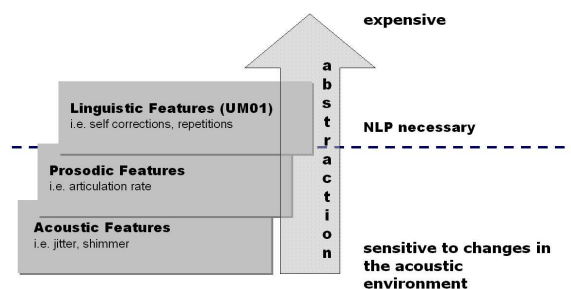

Figure 1: Three levels of abstraction of speech features

On the lowest level, there are *acoustic* features that are related to the signal's power and frequency and their changes over time. 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 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, speech recognition has to be done first.

Attending the latter group of speech features 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.

Müller *et al*. [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 were: 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] and Wittig [2003] showed that this network can be successfully used for this particular classification task.

Here, 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 [Linville, 2001; Schötz, 2001; Minematsu *et al*., 2002]. 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 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. We used five different jitter algorithms that are provided by PRAAT. Among them are: Jitter Ratio (JR), Period Variability Index (PVI), and Relative Average Perturbation (RAP), that are well known from the literature [Baken and 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.

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 Baken and Orlikoff [2000].

---

[1] http://www.praat.org

3 Recognizing a User's Age and Gender

We developed a two-level machine learning approach to recognize a system user's age and gender. In the following subsections we will briefly summarize the techniques and results related to the lower level since they have been described in more detail before, see Müller *et al*. [2003]. We will concentrate on the top level that addresses the meta reasoning process to combine the separate results of the low-level classifiers. We begin by presenting an outline of the overall approach.

3.1 Two-Level Machine Learning Approach

The basic idea is to apply traditional classifiers on the lowest level that are able to cope with raw sensor data, i.e. the shimmer and jitter values. These classifiers are learned using standard machine learning methods. We identified artificial neural networks (ANNs) to be the most appropriate classifier for our task, see Müller *et al*. [2003]. For both partial classification problems—age and gender recognition—ANNs (using backpropagation with sigmoid nodes) outperform other alternatives like support vector machines, decision trees and naive Bayes.

The reliability of these classifiers depends on the context. For example, in our particular scenario the microphone quality and potential noise in the environment influence the classification accuracy. Additionally, there may exists dependencies between the results of different low-level classifiers. In our setting, it is known that the voices of women and men age differently [Linville, 2001]. Therefore, it makes sense to take this into account within the reasoning process. We use Bayesian networks on the top level to combine the results and to represent the causal relations between the classification quality and environmental factors such as noise.

To improve the overall age and gender estimation, we process several subsequent utterances to consolidate our results. This is done by the application of dynamic Bayesian networks the can be used for reasoning over time.

3.2 Classifying Users According to Age and Gender

As already indicated, we performed an initial study to compare several alternative classification techniques. A detailed description can be found in Müller *et al*. [2003]. The overall results showed that ANNs perform best. They are able to yield an average classification accuracy of 0.92 for gender and 0.95 for age, respectively. (These particular values were computed within a replication of the study on the basis of a larger, more balanced dataset compared to that one used for the initial study).

We used the average values of jitter and shimmer of all speech samples available for that person. The classification accuracy has been estimated by a 10-fold cross-validation procedure. Here, it is difficult to integrate additional context information, such as the presence of noise, into the classification procedure. In that case, further data that has been collected under these particular circumstances has to be available. Furthermore, no relationship between the age and gender classification procedure is taken into account. Both aspects of the user model are estimated separately.

3.3 Using Bayesian Networks to Improve the Classification Process

As described in Section 3.1 it is known that the voices of men and women age differently. The question is now: How
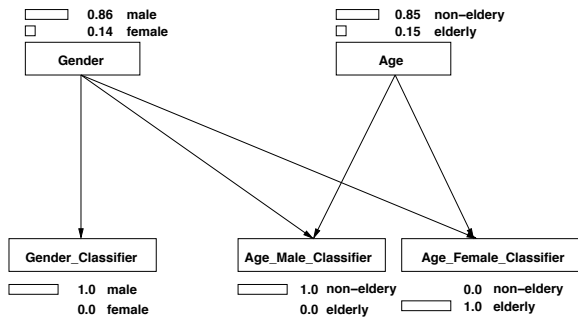
Figure 2: Gender-specific age estimation using a Bayesian network

can this knowledge be exploited within the classification procedure? And furthermore: How can related context information be taken into account? In the following, we will concentrate on the first question, but the latter one can be addressed in a similar way, using the same basic approach.

Our solution is based on *Bayesian networks* (BNs) to model dependencies between the individual classification results. A Bayesian network represents a joint probability distribution over random variables. It consists of two parts: (a) a DAG to represent the conditional (in-)dependencies between the variables, and (b) conditional probability tables that are associated with the links in the DAG. They contain the conditional probabilities of the variables' states conditioned on the combination of their parents' states.

Figure 2 shows the BN, that we used to model the gender-specific aging process of voices: The variables on the bottom line represent the classification results of the low-level classifiers, the ANNs. The first one is an ANN for classifying the user according to her/his gender. Conceptually, this result is causally influenced by the actual gender of the user, thus there is a link from `Gender` to `Gender_Classifier`. The value of the top gender variable given a classification result is estimated by the application of Bayes' theorem.

It remains to describe how we determine the conditional probabilities needed for this computation. These values are estimated by the prediction rates of the low-level classifier, i.e. the percentages with which the ANN is able to classify the users correctly. These rates are determined within a 10-fold cross-validation analysis and account to 0.94 for male and 0.90 for female, respectively.

Regarding the age classification, we can exploit the results of the gender classifier. The basic idea is to learn two "specialized" ANNs that are connected to the actual age variable—in an analogous way as the actual gender variable to the gender classifier variable—, one that is specialized for females, i.e. learned only with data of female speakers, and a second one for the males, i.e. learned only with male speakers. By connecting the gender variable to both of these classification result variables, we can "put more weight" on the result of that classifier that seems to be the more appropriate one, e.g. if the speaker is classified as female, the age classifier that is specialized for females should have more impact in the age classification process.

Again, it remains to describe the computation of the conditional probabilities. In this case, it is somewhat trickier than regarding the gender part. We now have more than one low-level classifier that depends additionally to another variable, i.e. the gender variable (see Figure 2). We have to consider four different situations: For each of

the two gender-specific age classifiers its performance regarding female and male speakers, respectively. For each of these situations the respective prediction rates have to be determined. To accomplish this task, the dataset has to be separated into a female and a male dataset. Both age classifiers have to be evaluated with both gender-specific datasets to receive conditional probabilities in the female and male case, respectively. Then for both genders, each specific classifier has to be tested either using a cross-validation or with test cases. Cross-validation is applied if the "gender" of the dataset equals that one that has been used to train the classifier, e.g. the male-specific ANN is evaluated by 10-fold cross-validation on the male dataset. The conditional probabilities for the opposite situation, e.g. the classification of females with the male-specific ANN are estimated using the female dataset as a test set. These conditional probabilities/prediction rates range from 0.74 to 0.97.

In Figure 2, we describe an example application of the fully specified BN. The classifier that is responsible for the gender determination yields "male". The gender-specific age classifiers do not agree and yield "non-elderly" and "elderly", respectively. By Bayesian reasoning, the system computes a 0.86 chance that the user is indeed male and a 0.85 chance that he is non-elderly. This example shows that in case that the gender-specific age classifiers yield different results, that one that is specialized to the more likely gender "wins" over the other one. This is due to the encoded knowledge about the gender-specific aging process.

It is straightforward to model additional aspects of the context in this BN. For example, the microphone quality and background noise can be represented by additional variables and links in th BN. If the conditional probabilities are correctly chosen or otherwise determined, it is possible to model for example a 10% reduction of the classification accuracy of a particular low-level classifier.

3.4 Combining the Results of Several Utterances

In the previous subsection, we described how to classify a user on the basis of a single utterance. If we take more utterances into account, we should be able to receive a more reliable estimation of a user's age and gender. Often, there are only short statements, e.g. to confirm an action, or the speech signal recorded by the microphone could be of minor quality, e.g. because the user made a head movement. Such samples of speech are not well suited for the classification procedure. Therefore, we developed a method to exploit several utterances that is based on *dynamic* Bayesian networks (DBNs) [Dagum *et al.*, 1992].

A DBN consists of several so-called *timeslices* to represent discrete points in time. In our situation, a timeslice corresponds to a single utterance. Essentially, timeslices are BNs that are connected by links that model the behavior of some variables over time. The BN of Figure 2 serves as a timeslice of our DBN. By keeping in mind that the age and gender of a user do not change over time, we are able to connect the classification results based on subsequent utterances.

Figure 3 shows an example of the DBN used to estimate the age and gender on the basis of a number of utterances. The links between the actual gender and age variables are annotated by conditional probabilities that model the static nature of these variables (not shown in the figure), i.e. all these variables have always the same values and can conceptually be seen as two single (static) variables
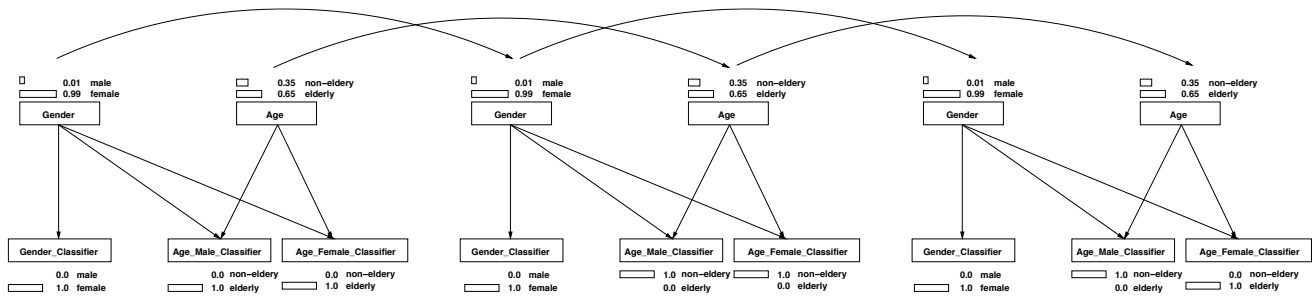
Figure 3: Combining the classifications of three utterances in a dynamic Bayesian network

representing age and gender.[2] In this example, the system analyzed three utterances of a female elderly user. The low-level classifiers did not yield consistent classifications of this particular user over time. Sometimes she has been misclassified by the age classifier to be non-elderly. But this misclassification on the lower level of our overall classification procedure is—as intended by its development—compensated by the combination of the results of different utterances and different (specialized) classifiers.

Environmental aspects can again be taken into account straightforwardly. If background noise is modeled on the top level, the corresponding variable can be instantiated according to the situation in which the utterance was observed.

4 Discussion of Initial Results in the Application System and Outlook

We integrated the described method into the M3I prototype that is developed as a part of the COLLATE project.[3] Although, we have very good results on our collected dataset, there are still some issues to address to improve the performance in a real application scenario.

The most important one is to cope with the poor quality of microphones that are embedded in portable devices. Potential solutions are (a) collect enough data with these particular devices to learn on that basis adequate low-level classifiers, and (b) preprocess our dataset in order to simulate the minor recording quality (e.g. by down-sampling of the speech signal).

Another line of our work—besides further empirical evaluations—addresses the question on what basis the classifiers should be learned? Should they be learned using average values per person as we have described it in this paper, or should they be learned on an utterance-per-utterance basis? We have to analyze whether the averaging may lead to a loss of information for the search process to learn the classifiers.

As soon as these major issues are sufficiently solved, we plan to generalize the described approach for the recognition of cognitive and affective states of the user using a wide range of biological sensors in the READY project[4] and will evaluate its usefulness for that application scenario.

References

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

[Dagum et al., 1992] P. Dagum, A. Galper, and E. Horvitz. Dynamic network models for forecasting. In Didier Dubois, Michael P. Wellman, Bruce D'Ambrosio, and Phil Smets, editors, *Uncertainty in Artificial Intelligence: Proceedings of the Eighth Conference*, pages 41–48, San Francisco, 1992. Morgan Kaufmann.

[Linville, 2001] Sue Ellen Linville. *Vocal Aging*. Singular, San Diego, Ca, 2001.

[Minematsu et al., 2002] Nobuaki Minematsu, Mariko Sekiguchi, and Keikichi Hirose. A perceptual study of speaker age. In *Proceedings of the International Conference of Acoustics Speech and Signal Processing*, pages 123–140, 2002.

[Müller et al., 2001] Christian Müller, Barbara Großmann-Hutter, Anthony Jameson, Ralf Rummer, and Frank Wittig. Recognizing time pressure and cognitive load on the basis of speech: An experimental study. In Mathias Bauer, Piotr Gmytrasiewicz, and Julita Vassileva, editors, *UM2001, User Modeling: Proceedings of the Eighth International Conference*. Springer, Berlin, 2001.

[Müller et al., 2003] Christian Müller, Frank Wittig, and Jörg Baus. Exploiting speech for recognizing elderly users to respond to their special needs. In *Proceedings of the Eighth European Conference on Speech Communication and Technology (Eurospeech 2003)*, September 2003.

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

[Schötz, 2001] Susanne Schötz. A perceptual study of speaker age. In A. Karsson and J. Van de Weijer, editors, *Proceedings of Fonetik 2001*, pages 136–139. Lund Working Papers, 2001.

[Wittig, 2003] Frank Wittig. *Maschinelles Lernen Bayes'scher Netze für benutzeradaptive Systeme*. DISKI 267. Aka Verlag, Berlin, 2003.

---

[2]We keep the several instances of these variables to make the concept of a timeslice more obvious.

[3]http://www.collate.dfki.de

[4]http://w5.cs.uni-sb.de/~ready