

### **3 Automatic User Classification for Speech Dialog Systems**

**Caroline Clemens**

Center of Human-Machine-Systems, Berlin Institute of Technology,  
Germany

**Thomas Hempel**

Siemens AG, Corporate Technology, User Interface Design, Munich,  
Germany, now with: Siemens Audiologische Technik GmbH, Erlangen,  
Deutschland

#### **3.1 Introduction**

Although the usability lacks of early speech dialog systems had been discussed in expert circles for quite some years, it took until the 2004 ‘study on acceptance and usability of speech applications’ to make this evident to the German scientific and industrial community. In this study, Peissner et al. (2004) had come up with a systematic and at the same time pragmatic approach to measuring acceptance and usability of existing speech dialog systems. Among others, results show:

- The frequency of use of such systems is very low in German speaking countries.
- Many users are not satisfied by the overall quality of the systems and thus are reluctant to use speech dialog systems at all.
- The success of a dialog system is strongly dependent on the usability of the voice user interface.
- So, from a usability perspective focusing on word recognition rates alone is by far not sufficient to achieve desired user acceptance.

Fortunately, today these issues do not sound too ground-breaking any more and most were known to experts even before. But 2004 these findings had huge impact on the voice business in the German speaking world. It led to re-designs of existing systems and influenced the community in a way to incorporate usability issues at much earlier stages of the system development process. In 2006 a second study states that the usage of speech

dialog systems has continued to increase in Germany but also emphasizes the need for further efforts for increasing the usability of such systems (Peissner et al. 2006).

So, whereas in the 1990s technical functionality was the focus, nowadays development processes increasingly concentrate on usability aspects. To guarantee an elaborated dialog design it is necessary to target a prospective system development process from the early development phases on.

Put in technical terms human beings are perfect speech dialog systems of which an important characteristic is that they adapt to their conversation partner. Even in conversations without face to face contact – like telephone calls – people consciously or unconsciously optimize their behavior to the respective dialog partner. In the near future, computers will probably never reach this ability in a comparable performance. This would demand a machine that passes the Turing test. Science fiction computers<sup>1</sup> impress with perfect speech interfaces that allow spoken language conversation like talking to a real person. Some of those have a distinctive personality and are humorous and charming.<sup>2</sup> It might be considered a hypothesis whether they raise popular expectations and make current speech dialog systems seem disappointing but it is a fact that today's systems are not that powerful and cannot copy human communication skills. Concerning this gap between expectation and performance, we detected an interesting effect in our usability tests: If the persona design of a speech dialog system is very realistic (i. e. human like), some users behave in a way that decreases the dialog success. It is known that users who do not respect the limitations of the systems decrease the performance. In our studies, some users overestimated the speech dialog system and expected it to behave like a real person. The recognition problems then increased and the users became disappointed. If this assumption is evidenced one could conclude that it is a disadvantage if a speech dialog system is designed to be too human-like unless it performs correspondingly. The aim should be designing successful dialog systems in terms of user satisfaction, dialog efficiency and matching the user's expectations. Obviously, this cannot be achieved by just humanizing the system.

In our opinion the work of voice interface designers in commercial projects still is underestimated and misunderstood. Reducing their working field to the surface of a dialog system with its audio style and prompt writing is on the same level like saying that graphical user interface (GUI) designers choose nice colors and shapes of buttons. Designing speech dialog

---

<sup>1</sup> The computer of *Starship Enterprise* or *HAL*, the computer of *2001: A Space Odyssey*, for instance.

<sup>2</sup> The car *KITT* in *Knight Rider* or *Eddie* on board *Heart of Gold* in Adams' *Hitchhiker's Guide to the Galaxy* series, for instance.

systems concerns the complete design process on all levels. It begins with an analysis of the aims of the dialog, its functions and requirements. When the target group was defined, decisions are made about the target style. Subsequent implementation steps will be based on this, and the architecture, menu and navigation are going to be determined. After that, keywords, audio design, texts, recordings and grammars can be set. For all these steps today designers still rely on their experience and an intuitive estimation of the users. Of course there are books that provide general guidelines (e. g., Morgan and Ballentine 1999; Pitt and Edwards 2002) but it is fruitless to do a literature review in order to find requirements that are characteristic for different types of users. Designers often try to put themselves in the position of a typical user. But who is this typical user? What are her or his attitudes, needs, wishes and preferences? And even more difficult, what about different types of users? Experience shows that there are definitely big differences between users. But how are they characterized? Can this be measured? Novice vs. power user is an established distinction, but again: how are they characterized? Questionnaires (Bearden and Netemeyer 1998; Brunner and Hensel 1994) can be used to observe the user's behavior, but, of course, it is impossible for every user to fill in a questionnaire before using a dialog system.

### 3.2 Automatic Information Retrieval and Adaptation

In the majority of cases, speech dialog systems are made for a large number of users. In many systems the users are anonymous and do not log in, so there is no advanced information about the user. In these cases a detailed analysis of each individual user by human experts is impossible. Therefore, information about the user and their behavior has to be gathered automatically. Three fields of benefit can be seen for automatic user classification – adaptivity of systems, clues for the designers, and marketing:

#### *Adaptivity of systems*

First of all it is needed to clarify the concept of adaptivity, because there are similar concepts like personalization, individualization, adjustment, adaptation and adaptivity. *Personalized* means customized to a specific person. Personalization or *individualization* is only possible if the user is known and identified by the system. This could be done using a password for the login. An example of *adjustment* is a car seat that can be adjusted to suit the user. If a system can adapt itself automatically to modified conditions it is called *adaptive*. *Adaptivity* is the possibility to adjust a system according to the requirements. User oriented systems adapt the system behavior to the user behavior. As there are big differences between

users (heterogeneous audience) it is clear that adaptive dialogs are a great step forward. For this, an automatic detection of user features is necessary. In speech dialog systems, several levels can be adapted to gain a user-friendly dialog, such as content, presentation and interaction style. It has been shown that effects of an automatic adaptation can be higher system performance, more efficiently resolved dialog anomalies, thus higher overall dialog quality (Chu-Carroll and Nickerson 2000; Litman and Pan 1999).

### *Clues for the designers*

Today's research has to focus on the possibilities of building dialogs that fit as well as possible to the actual user. Only a few studies offer helpful design rules for speech dialog systems. Of course there are standards (ISO 9241-110 2006; ISO 9241-11 1998) which contain dialog design guidelines. But such standards are very general and do not help the designer concerning a detailed design decision. If automatic user classification was in place it could automatically derive patterns of use and user interaction and provide it to the designers.

### *Marketing*

Marketing needs to know as much as possible about the target group. Automatic user classification could deliver information in a very economical way. Extracting user features automatically would be of high value for complementing user profiles.

After analyzing existing adaptive dialogs we inspected the technological feasibility of user classification. Although we used telephone-based speech dialog systems, most non-specific characteristics are also relevant for non-telephone-based systems. User interface designers came together in expert workshops to discuss scenarios of automatic adaptation and evaluated estimated benefits. An extract of the results is presented in the following. Sources of information for automatic user classification and adaptation are mainly audio signals, log files and – in case of telephone based dialog systems – the telephone number:

- The audio signal contains the speech signal of the user as well as background noise. Room characteristics and the transmission influence on the acoustic signal. In the speech signal there is information about individual voice, speech and linguistic characteristics like age, gender or accent.
- In a running computer program, log files automatically document when certain program steps take place. They offer the opportunity to register events in a speech dialog system with a precise timestamp. Many events of a dialog are initiated by the user. It is recorded which menu point the

user has chosen at a particular time. There is also documentation of which input is recognized by the speech recognition system or if there is a no-match or no-input. The reaction time of the user can be calculated from the timestamps.

- Where the telephone number is transmitted it represents an information source. The number consists of a prefix and the extension. If the call is coming from a foreign country the country prefix is transmitted, too. The prefix is either an area code or mobile network number. So you know whether the user has called from a landline or a mobile phone. The area code tells you which region the user is calling from or the mobile network number tells you which mobile network is being used. It is technically possible to detect how often and at what times a certain telephone number is calling the speech dialog system.

The following examples show how the dialog system can adapt to automatically retrieved information. All adaptations suggested here are technically feasible and some of them have already been realized, such as a classification of gender and age (Stegmann et al. 2006, Metze et al. 2007) and adaptations to the user's observed level of expertise (Jokinen 2006).

#### *Age:*

Existing automatic age classifiers can differentiate between broad age groups. In languages that use different forms of address like German ("Du"/"Sie") or French ("Tu"/"Vous") the form used by the system could be fitted to the age group of the user. Speaking and language style as well as target group focused advertising could be chosen.

#### *Gender:*

There are classifiers that can detect the gender of a speaker with a high success rate. According to the detected gender the system could be adapted adequately. For instance, if it is known that men and women prefer different speakers then a suitable speaker could be chosen.

#### *Car:*

An acoustic analysis can detect if the speaker is in a car. While driving, the user might get distracted by the driving task and not react immediately. Then the timer could allow longer reaction times for a no-input time out.

#### *Emotions:*

Emotions such as anger or frustration can be detected automatically (Ang et al. 2002). The system could react by transferring to an operator and thus avoid a hang up.

*Native language:*

There are techniques to find it out which language a user is speaking in and the dialog system could switch to this.

*Number of barge-ins:*

Many barge-ins that do not produce no-matches can indicate an experienced power user. The system could react with a faster navigation, shorter prompts and fewer explanations.

*Timestamp of barge-in:*

Early barge-ins show that the power user knows the next input before the prompt is played completely. Shortcuts could be offered to allow a shorter dialog for the next call.

*Number of no-matches or no-inputs:*

If many no-matches are produced the user may need more explanations. It might be helpful to reduce the input possibilities to keywords in order to improve the recognition rate. If recognition is still problematic, the user could be transferred to an operator.

*Path and pattern of navigation:*

If a user wanders around in the navigation, the system could help with declarations and a more directed dialog.

*Number of help requests:*

A large number of help requests shows that the user needs more declarations and a more directed dialog style.

*Telephone extension number:*

Storing telephone numbers in a database allows checking whether a number has called before, how often, and at what time of day. Extensions of some companies or public authorities all begin with the same numbers. When it comes to saving data, privacy has to be respected. Of course the same telephone number does not guarantee the same user which makes it necessary to handle sensitive data responsibly.

*Telephone prefix:*

If the prefix is an area code, local services could be offered for the location the user is calling from. The prompt texts could be chosen in accordance with local conventions.

To supplement the results of the expert examination two focus groups were organized with five members each. After a short introduction to telephone based speech dialog systems, there was time for a moderated discussion about possible adaptations of such a system. The results of the two focus groups confirmed those of the expert examination. The focus groups delivered some additional ideas but showed up the same critical points and no really new aspects. After both the expert and the focus group studies were examined we decided to focus on log files in further studies, because as a source of information they offer possibilities that are easy to realize and powerful at the same time. Other sources of information are either too complex and therefore too expensive to develop or their automatic classifiers are still not mature enough for state of the art performance.

### 3.3 User Features and Log Files

The following list shows a clustered collection that describes the wide range of personal, cultural and situational factors that influence the user's behavior:

- situation, location, attendant persons, time pressure, other activities and tasks beside using the system itself (e. g. driving as major task);
- demographic factors like age, gender, native language;
- experience, knowledge, skill, expertise, intelligence, cognitive workload, memory decline, attention, capability, expectation, preferences, habits, interests;
- personality, traits, type, character, attitude;
- mood, health, emotions, stress, condition, state;
- the person's task and aim of use;
- the speech dialog system itself.

Determining all these factors is impossible. It is important to know how the user behaves and how the dialog can adapt to that behavior in the best way. The features can be grouped in four dimensions: *User features* concern the user and the user's interaction features; *Task features* picture the task of the concrete use case; *Technical features* are set by devices and infrastructure used for solving the task; *Surrounding features* like situation, setting, local and temporal context.

Given our focus, user features have to be extracted from parameters that are retrieved from log files. Depending on the architecture of the system there is more than one module that creates a log file like the speech recognition system or the dialog management server. If several log files have to be combined it is important to assure synchronism. Typically, log file

entries will have a form like [timestamp|event|detail1|detail2|...]. Events that are logged are basically: Start of the dialog, switches to sub-dialogs or menu points, events of the recognition system including recognition rate and recognized words, and end of dialog. From the logged dialog events, parameters have to be extracted that can describe the user's interaction:

- Dialog duration;
- User turn duration;
- User response delay;
- Number of user turns;
- Words per user turn;
- Number of help requests;
- Number of time-out prompts;
- Number of ASR rejections;
- Number of error messages;
- Number of barge-ins;
- Number of cancel attempts.

Interaction features extracted from log files have already been used to predict the quality of a dialog system (Möller 2004). To make the interaction features more comprehensible they can be aggregated:

Speed:

- Reaction time of user in prompts;
- Mean duration of inputs;
- Call duration;
- Difference between the length of the user's path in the dialog and a defined optimal path for solving the task.

Competence:

- Use of barge-ins;
- Number of no-inputs;
- Number of no-matches;
- Number of help requests;
- Goal-oriented navigation.

Cooperativeness:

- Use of proposed keywords;
- Number of words per utterance.



Learning effect:

- Decreasing number of no-inputs;
- Decreasing number of matches.

Politeness, conformance with conversation rules of human-human dialogs:

- Number of barge-ins;
- End of call: leave-taking/good-bye.

### 3.4 Testing and Findings

A preceding usability test was conducted in cooperation with T-Systems Enterprise Service GmbH, Berlin, Germany. We examined differences between users of different gender and age, resulting in the user groups: adult male, adult female, senior male, senior female, and children. After a questionnaire about demographic data, the test person made calls with a Wizard of Oz speech dialog system. This was a fictitious timetable information system for public transport that included an automatic age and gender classifier. After the test calls, the user's opinions and experience were observed in guided interviews.

The test showed a higher drop-out rate for senior test persons. The problematic or unsuccessful calls showed noticeable characteristics like a high number of no-matches, long inputs with many words in long phrases and few keywords only. No noticeable learning effect could be observed and the interaction of the test person conformed strictly to human-human interaction conventions. Additionally, those test persons appeared nervous, aroused or frustrated, some reported they felt misunderstood by the system.

All user groups in the test preferred a more directive dialog style in contrast to mixed-initiative dialogs. The more calls the test persons made the more target oriented they performed. This learning effect is reflected in lower error rates.

The automatic classification of gender and age to the user groups was correct in 56% of the cases. All wrong classifications regarded the age of the test person whereas all gender judgments were right.

In our main study, data was collected from test persons in three age groups (children 9–14 years, adults 20–45 years, seniors 63–72 years). Additionally, personality related data were gathered. We retrieved log files of around 12 calls per person with different tasks and scenarios as well as additional material such as video and audio recordings and notes of the test leader.

The telephone-based speech dialog system we used included automatic speech recognition for spoken language input and played recorded audio

files as acoustic output. Common global navigation keywords were active and barge-ins were accepted. The current state<sup>3</sup> of analysis points out the following issues:

- The collected data indicates that there is a relation between users' execution speed and their overall technical experience.
- Increasing experience with the dialog system under test leads to changes in some users' behavior (learning effect).<sup>3</sup> The common terms novice and power user are in fact useful.
- There is a relationship between age and learning effect.
- There is a bigger variance in the results of the group of seniors than in the group of younger adults.
- There are more elderly with low tempo and little technical experience.
- There is a greater variance in performance of users with little or no experience compared to the variance among experienced users.

### **3.5 Summary and Outlook**

The usability of a speech dialog system depends crucially on the quality of the dialog design. For obtaining high user acceptance for the system, the user centered dialog design should already be focused on during the early phases of the system development process.

From the perspective of our studies the usage of adaptive speech dialogs is a big technological and creative step since this enables both new user-centered speech dialog designs and new marketing opportunities.

Furthermore, automatic classification and analysis data is about to lead to yet undiscovered relations within measurable parameter sets in users' interaction behavior. Therefore, it is important to collect the named interaction features during the dialog – and based on such results research is about to develop more advanced classifiers.

Also, it is important to determine details like the time needed by a classifier to reach a reliable classification result during a user's call. But even if it takes too long technically to be able to use the result in the current user dialog, the result could be stored and be available for the next call from the same telephone number.

There might be more detectable patterns in the navigation behavior of users. Recurring navigation patterns like loops, jumps or iterations in the dialog flow can correlate with known user features. But even in some years

---

<sup>3</sup> Parts of the analysis are still pending. For further findings and a detailed description of procedure and results see first author's doctoral thesis (to be published 2008).

when automatic classification is expected to work even more reliable, voice dialog designers will face their true challenge of how to adapt the dialog to the satisfaction of the respective user group. Results of user tests of the different user groups have to be carefully analyzed. Only then the results could help to formulate adequate style guidelines for the design of better – user focused – dialog systems.

## Acknowledgements

The presented studies were conducted within a research project of Siemens AG, Corporate Technology, User Interface Design in cooperation with the Center of Human-Machine-Systems at Berlin Institute of Technology. They were part of a project on Adaptive Speech Dialogs funded by Siemens COM's Chief Technology Office. The project was part of a co-operation of Siemens AG with Germany's Deutsche Telekom AG. The authors would also like to thank Siemens CT project manager Dr. Ing. Bernt Andrassy (Professional Speech Center, [www.siemens.com/speech](http://www.siemens.com/speech)) and Frank Oberle and his team at T-Systems Enterprise Service GmbH, Berlin, for the fruitful co-operation.

## References

- Ang J, Dhillon R, Krupski A, Shriber, E (2002) Prosody-Based automatic detection of annoyance and frustration in human-computer dialog. In: Proceedings of the 7th ICSLP (International Conference on Spoken Language Processing), Denver
- Bearden WO, Netemeyer RG (1998) Handbook of Marketing Scales: Multi – Item Measures for Marketing and Consumer Behavior Research. 2nd ed. Newbury Park, Sage Publ., Calif
- Borkenau P, Ostendorf F (1993) NEO-Fünf-Faktoren-Inventar (NEO-FFI) nach Costa und McCrae. Hogrefe, Göttingen, pp 5–10, 27–28
- Brunner GC, Hensel PJ (1994) Marketing scales handbook: A compilation of multi-item measures. American Marketing Association, Chicago
- Chu-Carroll J, Nickerson JS (2000) Evaluating automatic dialogue strategy adaptation for a spoken dialogue system. In: Proc. NAACL 1, pp 202–209
- ISO9241-110 (2006), Ergonomics of human-system interaction – Part 110: Dialogue principles
- ISO9241-11 (1998), Ergonomic requirements for office work with visual display terminals (VDTs) – Part 11: Guidance on usability
- Fahrenberg J, Hampel R, Selg H (2001) Freiburger Persönlichkeitsinventar (FPI). Revidierte Fassung (FPI-R) und teilweise geänderte Fassung (FPI-A1). Hogrefe, Göttingen

- Jokinen K, (2006) Adaptation and user expertise modelling in AthosMail. In: Universal Access in the Information Society 4 pp 374–392
- Litman DJ, Pan S (1999) Empirically Evaluating an Adaptable Spoken Dialogue System. In: Proceedings of the 7th International Conference on User Modeling
- Metze F, Ajmera J, Englert R, Bub U, Burkhardt F, Stegmann J, Müller C, Huber R, Andrassy B, Bauer JG, Littel B, Comparison of Four Approaches to Age and Gender Recognition for Telephone Applications, ICASSP 2007. IEEE International Conference on Acoustics, Speech and Signal Processing, 2007, pp IV-1089–IV-1092
- Morgan DP, Balentine B (1999) How to Build a Speech Recognition Application: A Style Guide for Telephony Dialogues, Enterprise Integration Group, San Ramon
- Möller S (2004) Quality of Telephone-Based Spoken Dialogue Systems. Springer-Verlag, Berlin
- Peissner M, Biesterfeldt J, Heidmann F (2004) Akzeptanz und Usability von Sprachapplikationen in Deutschland. Technische Studie, Fraunhofer-Institut für Arbeitswirtschaft und Organisation (IAO), Stuttgart
- Peissner M, Sell D, Steimel B (2006) Akzeptanz von Sprachapplikationen in Deutschland 2006, Fraunhofer-Institut für Arbeitswissenschaft und Organisation (IAO), Stuttgart und Initiative Voice Business, Bad Homburg
- Pitt I, Edwards A (2002) Design of Speech Based-Devices. Springer-Verlag, New York
- Stegmann J, Burkhardt F, Oberle F, Eckert M, Englert R, Müller C (2006) Einsatz der Sprecherklassifizierung in Sprachdialogsystemen. In: Tagungsband der 7. ITG-Fachtagung Sprachkommunikation, Kiel



<http://www.springer.com/978-3-540-78342-8>

Usability of Speech Dialog Systems

Listening to the Target Audience

Hempel, Th. (Ed.)

2008, X, 176 p., Hardcover

ISBN: 978-3-540-78342-8