

Flexibility and efficiency through personalisation?

Experiments with a conversational Program Guide Information System

**Péter Pál Boda, Suresh Chande,
Elviira Hartikainen, Nidhi Gupta**

Nokia Research Center
Helsinki, Finland
peter.boda@nokia.com

Sirpa Autere

Elisa Communications Corporation
Research Center
Helsinki Finland
sirpa.autere@elisa.fi

Abstract

Results of a subjective evaluation with a personalised conversational interface to a Program Guide Information System are discussed in the paper. The goal of the work presented here is to examine whether personalisation of a speech-enabled system offers noticeable benefits to the users.

1 Introduction

Personalisation has several meanings. In general terms, service personalisation means that a service is adapted to fit someone's preferences or needs. Its ultimate goal is to offer a more preferred interaction with the system, to retrieve the content faster and/or to present the results in a preferred way.

Personalisation is especially important for mobile devices with limited input and output capabilities (Ruuska-Kalliokulju et al., 2001). It is also clear that personalisation also serves as a product differentiator: by letting users to shape their own information and communication environment in terms of devices, service interaction, etc. has huge potentials. Furthermore, personalisation can provide a richer and even intimate experience. This aspect is especially important when the desire of self-expressions and individual preferences gain more and more importance in today's world.

In our interest is the combination of rendering the content and adapting presentation according to the user's preferences. There are three important aspects to be considered here: how to collect the user preferences, how to render the retrieved results and how to assess users' satisfaction regard-

ing the personalised interface. We decided to collect user preferences in an explicit manner: instead of context and/or user behaviour pattern detection we implemented a web-based interface that enables users to set preferences, *aliases*, for an application. Rendering the retrieved results was implemented as a ranking algorithm. Finally, assessment of users' acceptance was measured via a set of subjective evaluation tests with a personalised Program Guide Information System, accessible through a conversational interface.

The main goal of the work carried out is to provide insights to the acceptance of the personalised interface. Our hypotheses were that users will notice the difference between the personalised and non-personalised versions, and that they will appreciate the personalised version by its advantages (e.g. faster delivery of tailored responses).

The paper is organised as follows. Section 2 explains briefly the Program Guide Information System, the presentation and access techniques. The personalisation method is covered by Section 3. The aspects of the subjective evaluation and the implemented methodology are discussed in Section 4, the achieved results and major findings are summarised in Section 5. Finally, conclusions are drawn and future directions are outlined.

2 Technical Implementation

2.1 General overview of PGIS

The Program Guide Information System (PGIS), developed as the Helsinki demonstrator in the CATCH-2004 (www.catch2004.org, 2002) IST project under the European Union's 5th Framework Research Programme, is a telephony-based conver-

sational demo application which enables users to query information about TV programs. The PGIS application was developed on IBM's ViaVoice™ technology to provide access both in English and Finnish languages. Two monolingual versions, Finnish and English, were implemented, as well as a multilingual solution where implicit language identification was utilised (Harrikari et al., 2002). For the personalisation tests only the monolingual English version was used.

The NLU (Natural Language Understanding) component of the system was based on mixed-initiative dialogue technique and it provides means for verification, focus shift, disambiguation, etc. The basic functions, i.e. the semantic universe of the domain, are listed in Table I.

- | |
|--|
| <ul style="list-style-type: none"> • Program name • Program type (music, movie, news, etc.) • Channel (BBC World, CNN, TCM, Eurosport) • Performer • Date • Time • Language • Other: starting and ending time, duration, country of origin, episode title, short description, possible re-runs, etc. |
|--|

Table I. The semantic aspects of the PGIS domain.

Appendix A shows an example of a conversation conducted with the English monolingual version of the system. The underlined expressions in the transcription stand for semantic units understood and interpreted properly by the system.

2.2 The PGIS architecture

The Program Guide Information System (PGIS) architecture is depicted in Figure 1. The complete PGIS system consists of three layers, namely the User Agents, Server Side System and the Backend System.

The User Agents layer interfaces to various input/output devices, including (from left to right on the top of Figure 1) smart handheld devices and kiosks with multimodal capabilities, web interfaces with GUI-only (Graphical User Interface) and telephones with voice-only input/output modalities.

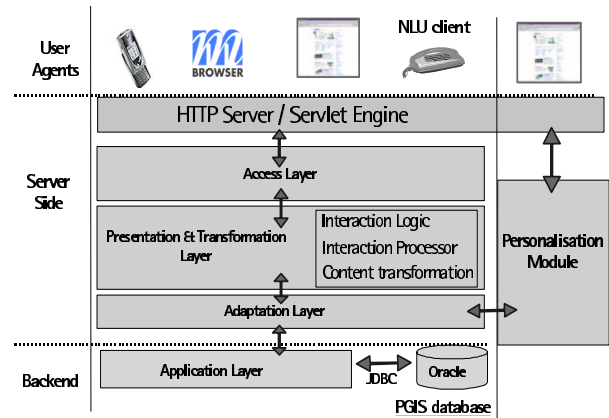


Figure 1. The entire PGIS architecture.

The Web Browser is a normal web-based client (WAP Browser or HTML browser), used mainly in the development phase to test database responses for simulated user queries. Other utilisation might be a kiosk implementation without voice capabilities.

The Multimodal Browser is capable of rendering multiple modalities such as voice and visual modalities utilising markup authoring languages such as VoiceXML, WML and XHTML. The browser framework was developed as part of the CATCH-2004 project (www.catch2004.org, 2002). User interface actions on a particular single-modal browser are captured, converted to events and distributed to the other browsers participating in the framework. A Synchronization Protocol was defined which assists in distributing the events generated by each of the modality specific browser (Kleindienst et al., 2002a & 2002b). This enabled to bring the value of each of the modalities in the most economic way in order to offer the user a habitable, easy-to-use interaction style. As an earlier study on the Multimodal Browser by Halonen et al. (2002) reveals, the speech modality is especially appreciated for searches, since random access with speech allows the user to speak the item directly instead of locating it, after a potentially lengthy browsing, on a long list.

The conversational interface, NLU client, is implemented using IBM's ViaVoice™ and Natural Language Understanding technology. The NLU client of PGIS can be accessed via any telephone (fixed or mobile) using voice-only connection.

The Server Side System provides web-based (HTTP) access to the Backend System. It accepts a user request, interpreted by one of the User Agents, and in-turn it interacts with the PGIS database to deliver the relevant results in the required format. Once the required program details are received, the Server Side adapts the response to the user's preferences. The Server Side embeds the Multimodal Application Framework and is deployed onto the Web Container of a J2EE Application Server (JBoss).

The Backend System is deployed into the EJB Container of the J2EE Application Server. The Backend System supports the interpretation of the user's request: it queries the database and delivers the list of resulting program(s) to the Server Side System. Physically, in our demonstration implementation the PGIS database resides in the premises of Elisa Communications and is connected to the rest of the architecture over the Internet.

The Personalisation Module (PM) is defined in a generic way so that it can be used for content personalisation by a set of applications. This module is based on Java for creating, storing and maintaining user preferences. The Personalisation Module architecturally consists of several layers as depicted in Figure 2.

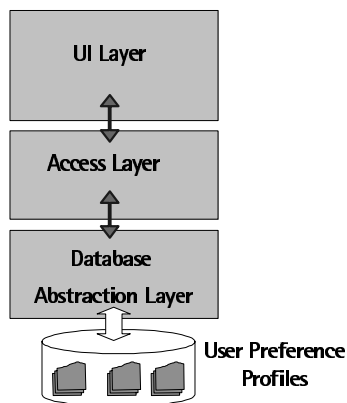


Figure 2. The Personalisation Module.

The Access Layer is used for registration, authentication, query and update of user profiles. The Database Abstraction layer provides a unified way of accessing the backend. The User Preference Profile is broadly divided into three parts – personal data, general preferences (which remain the same across all applications, e.g. the user's name, mother tongue, etc.), and service preferences that are spe-

cific to a particular application. There is a web-based User Interface which uses the Access layer to facilitate registration, authentication, viewing and updating of user preferences (see Appendix B).

The communication between the PGIS Server Side System and the Personalisation Module is shown in Figure 3.

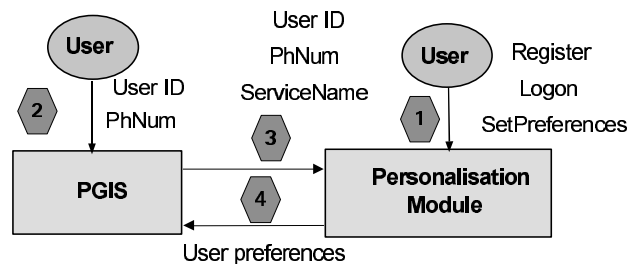


Figure 3. Personalisation Module operation.

The steps involved in the personalisation of the PGIS application are listed below:

Step 1: Initially, the user needs to register, to login and then to set the preferences for the PGIS application (explained below);

Step 2: When the user interacts with the PGIS application over the phone, the user must be identified first. This can be done with the telephone number or by explicit request of a User ID;

Step 3: The PGIS application requests the PM to deliver the user's preferences by passing over the User ID and the service identifier. The service identifier tells the PM about the actual service requesting the information;

Step 4: The Personalisation Module checks the access rights for the actual service and in case it is allowed it delivers the user's preference for the actual service, now to the PGIS application.

3 Personalisation

The Personalisation Module is responsible for storing the users' application specific preferences. The Personalisation Module itself does not contain separate preferences for different User Agents, rather the application renders the results for different User Agents according to the common preferences.

In our paper we consider only the voice-only, i.e. the NLU access case.

3.1 Aliases: setting preferences

Once the user logged into the Personalisation Module via a web interface (see Appendix B), several aliases can be set. An *alias* is a shortcut that gathers several program preferences under one name. For the sake of simplicity we pre-defined the alias names in advance as *MySports*, *MyNews*, *MyMovies*, *MyMusic* and *MyTravel*.

The length of the dialogue can be considerably reduced with the aliases since they capture pre-defined channels, program types and sub-types, as well as the preferred viewing time. Furthermore, having an easily memorable alias name helps the user in the interaction with the NLU client.

3.2 Ranking: presenting results

Typically, after the user made a query the PGIS application interacts with the Backend System. It is normal that for a quite open question like “*What’s going on after 8pm on BBC World?*” the system will provide a rather long list of programs. From this list the user has to find the relevant program she or he is looking for. Personalisation can help in re-ordering this long list so, that the programs matching the user’s preferences will come first. On this way the interaction time can be reduced significantly, since no follow-up questions are necessary. The flow chart below describes the PGIS Ranking mechanism.

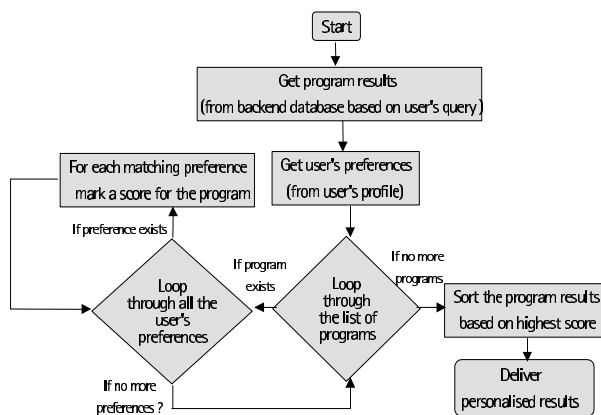


Figure 4. PGIS Ranking.

In case the user has no preferences set, for a question like “*Tell me about MySports*” the system will fetch the data from the backend by sending a (*Type=sports*) request. However, if the user set the

preferences e.g. like it is indicated in Appendix B, the system will issue the following request:

(Channel=Eurosport | BBC World) & (Type=news | sports) & (Subtype=golf | hockey | football) & (Time=Evening 18:00 – 24:00)

The results returned in both cases are ranked according to the algorithm depicted above. The ranking is interpreted by the NLU module and the results presented accordingly by the text-to-speech synthesis engine.

3.3 Modifications in the NLU client

The personalisation feature required only two modifications in the NLU client.

In order to offer a more pleasant presentation for a personalised answer, and to indicate to the user that the response is according to his/her preferences, the system prompt was modified as follows:

```

if personalisation == true
  system_prompt = "I have found the following
                  preferred programs ..."
else
  system_prompt = "I have found the following programs"
  
```

Table II. Modification of NLU prompting.

The other change is related to the user identification. In our particular test environment analogue telephone cards were used thus we introduced an extra step in the beginning of the session to request the user to key in a four digit User ID. Although this is not the most advanced and seamless way of identifying the user, this served properly our test purposes.

4 Subjective evaluation

Subjective evaluation was organised within Nokia Research Center in July 2002. The experiments were conducted with the English NLU client.

4.1 Preliminary hypotheses

Before starting the tests, the initial expectations were identified:

1. Subjects will notice the difference in the operation of the personalised and non-personalised systems;

2. Test users will appreciate the advantages offered by personalisation;
3. The aspects that will be appreciated most will be easiness, accelerated interaction and delivering the correct and desired answer;
4. Users who experiment first the non-personalised system and only after that the personalised one will appreciate more the advantages of the personalised system, compared to those users who experiment in the reversed order;
5. Users are ready to use a personalised PGIS and a web-based preference setting is acceptable.

4.2 Methodology

The tests were organised with two groups of users, in each group having 6 persons with technical background. Four out of the 12 persons tried out an earlier version of the PGIS application, for the rest it was entirely unknown. None of the tests users had English as a mother tongue.

The tests were executed in two steps for each group. Group A first called the personalised and then the non-personalised version, while Group B executed the tests in the reversed order. The two rounds for each group were organised on different days. After each test round a web-based questionnaire was filled up by the user. Once both tests were conducted, the users received by e-mail a final questionnaire for comparing the two systems.

4.3 Test scenarios

Before starting the tests, each user was asked to pick up one favourite alias and they were instructed to fill up the web-based personalisation settings. The scenario, the sequence of queries to conduct with the system, is described in Appendix C. The very same scenario was given to all users in both rounds of the test.

5 Results

5.1 Objective evaluation of the NLU client

Conversational interfaces, utilising Natural Language Understanding and dialogue technology, are gradually becoming mature enough to serve in lar-

ger deployments. Our attempt to use such technology for the Program Guide Information System is justified by the achieved objective results as shown below.

Number of calls (2 groups x 2 runs x 6 persons)	24
Number of user inputs altogether	165
Average number of queries over all sessions	5.69
personalised system	4.81
non-personalised system	6.77
Average length of a session	2 min 54 s
Average word error rate	24.25%
Average task (~concept) accuracy	91.0%
personalised system	93.9%
non-personalised system	88.2%

Table III. Objective system performance figures for both systems.

Looking into these results one must take into considerations the following issues:

1. The call success rate was 100%, each test person could conduct the session successfully;
2. There were more than one task per call, since often test users asked additional information (e.g. about director, duration of program, etc.). In our case task and concept understanding accuracy are almost identical since due to the given scenario a task consisted of only one semantic unit;
3. Although the word error rate is relatively high, one must remember that this is a conversational system with a middle size vocabulary including ca. 7000 words. Errors in the recognition often occur for words irrelevant from the understanding point of view;
4. The word error rate in the April 2002 monolingual tests were 28.62%. Compared to those result the word error rate is lower now (24.25%) due to shorter sessions with a rather limited test scenario (the acoustic models were not updated after April 2002);
5. The task success rate was higher for the personalised system, which partly can be explained by the fact that the sessions with this version were shorter by almost 30% (4.81 queries vs. 6.77).

5.2 After-call evaluation

After each call the users immediately filled up a questionnaire. There we asked how they liked the just tested system, how they felt about learnability, how fast/easily they got the requested answer and how they perceived the understanding capabilities of the system. The results for the two groups are presented below:

	Aspect	1 st syst.	2 nd syst.
Group A	Acceptance	3.5	3.17
	Learnability	4.17	4.17
	Task completion	3.67	3.0
	Flexibility	3.33	3.17
	Understanding	3.17	4.17
	<i>Average</i>	3.57	3.54
Group B	Acceptance	3.5	3.5
	Learnability	3.67	4.17
	Task completion	3.0	3.5
	Flexibility	3.0	3.33
	Understanding	3.5	3.83
	<i>Average</i>	3.33	3.67

Table IV. Questionnaire results for the two groups: means for each group for different usability aspects. NB.: for Group A the 1st system was the personalised one, for Group B the 2nd one (the shaded areas).

What can we see here? An intra-group comparison reveals that there is no significant difference between the 1st and 2nd system results in both test groups. However, it can be noticed that in Group A the average scores for the 2nd system (non-personalised one) are remained the same or lower than for the 1st personalised system. An opposite trend is noticeable for Group B, where the presentation of the systems were in reversed order and the scores are remained the same or increased. The only aspect which do not fit to this trend is Understanding for Group A. The average score increased from 3.17 to 4.17 and it suggests some kind of learning effect, i.e. for the second run the users accepted and learned the capabilities of the system, they gained some practice, talked more clearly, managed to perform a better interaction and thus detected a better understanding by the system.

Group B preferred the latter tested personalised system over the non-personalised one more than Group A preferred the first tested system, thus there is a relevance in which order the users executed the tasks. Under this basis we can say that our 4th hypothesis holds and the appreciation of the personalised system, according to the 1st hypothesis, is also valid (although not in significant terms).

The increasing and decreasing trends for both groups hints that the users set their absolute level of judgment by the first round of tests. It is reasonable to think that the novelty of the systems was important in the first test round and users scored somehow on an absolute level. In the second round their scoring was already relatively scaled to the experience of the first round.

5.3 Comparative evaluation

After the second test round the users immediately filled up the after-call evaluation questionnaire. Later, the same day, they received by e-mail a questionnaire to compare the two systems. The result of the comparative analysis revealed the following major findings:

1. The speed of the personalised system is appreciated, especially by Group B which first evaluated the non-personalised version. Clearly, the non-personalised system, i.e. the order of test, had an effect on them. 58.3% of all users considered the personalised version faster. 83.3% of Group B answered that the personalised system was faster, while only 33.3% of Group A users judged the personalised system faster;
2. Delivering the correct and desired answer by the personalised system is appreciated in only 41.7% of all answers;
3. Difference in preference is also noticeable. 66.7% of users in Group A preferred in general the non-personalised system while 66.7% of Group B users preferred in general the personalised system. This is most probably due to the order how the systems were presented to the test users. Another reason might be a sort of learning effect, since both groups preferred the systems that they tested in the second round;
4. A similar trend is present for the aspect of answer quality: while 2/3 in Group A considered the non-personalised system higher quality the

same amount stated similarly in Group B about the personalised system;

5. Every fourth user considered the non-personalised system friendlier than the personalised one. One explanation can be that the personalised system was faster and it provided shorter program lists in the answers. Probably this “to-the-point” strategy was considered as a somehow unfriendly approach.

6 Summary

The five hypotheses set before the tests were mostly justified, however, some can hold only with some modifications. Below there is a summary of the achieved results grouped according to the hypotheses.

1. *“Subjects will notice the difference in the operation of the personalised and non-personalised systems.”* - Yes, the users noticed the differences, especially in Group A (users who tested first the personalised system) 83.3% of the users reported noticeable differences (e.g. speed). 66.7% reported larger amount of programs presented by one of the systems (the non-personalised one, but it was not revealed to them). In overall, 66.7% of all test persons (both Group A and B) reported noticeable differences.
2. *“Test users will appreciate the advantages offered by personalisation.”* – Two third of the users in Group A preferred in general the non-personalised system while a similar portion of Group B test users preferred the personalised system. A similar trend was detected when the quality of the systems was questioned. We can thus conclude that test users appreciate the advantages offered by the personalised version if they are already familiar with the non-personalised version. Thus our hypothesis should be modified as: “Personalisation will be appreciated if the user is already familiar with a similar system”.
3. *“The aspects that will be appreciated most will be easiness, accelerated interaction and delivering the correct and desired answer.”* – This hypothesis holds only partly. The users did not rate easiness similarly within the two groups: both groups found the second system easier to use (personalised one for Group B). When the

delivery of the correct answer was asked, only 41.7% of all users’ appreciated the responses of the personalised system – this is not what we expected. However, regarding speed both groups favoured the personalised system. In overall, 58.3% of all users considered the personalised system faster.

4. *“Users who experiment first the non-personalised system and only after that the personalised system will appreciate more the advantages of the personalised system, compared to those users who experiment in the reversed order.”* – This hypothesis turned out to be entirely valid as it was discussed in the case of the 2nd hypothesis.
5. *“Users are ready to use a personalised PGIS and a web-based preference setting is acceptable.”* – As the feedback analysis revealed, 90,9% of the users who watch TV would be ready to fetch at least some program information from the personalised PGIS system. Although web-based setting is acceptable (score given 4.25), more automation would also be a useful feature, especially Group A indicated so.

7 Conclusions and future work

Answering the question in the title of the paper, we conclude that efficient interaction can be implemented using personalisation. Personalisation brings benefits to the user: faster interaction is clearly the most appreciated feature. However, when the correctness of answers are considered, further work is needed to improve the acceptance of a personalised implementation.

Speech technology is mature enough and its benefits are recognised by the user. Flexible interaction can be offered by Natural Language Understanding based technology.

Future work will utilise the know-how developed in the project and the findings of the here presented personalisation tests.

Acknowledgements

The authors want to thank Heli Harrikari for updating the NLU module for the personalisation tests. The comments and suggestions of the reviewers are highly appreciated.

The first author kindly acknowledges the support of the Nokia Foundation.

References

- Ruuska-Kalliokulju, S. et al.: 2001, "Shaping the Future of Mobile Devices – Results of the Workshop on Future Mobile Device User Interfaces at CHI 2000". SIGCHI Bulletin, January/February 2001.
- CATCH-2004, 5th Framework European Union IST Project (2000-2002), <http://www.catch2004.org>
- Harikari, H., M. Mast, T. Ross & H. Schulz: 2002, "Different Approaches to Build Multilingual Conversational Systems". 5th International Conference on Text, Speech and Dialogue, TSD 2002, Brno, Czech Republic.
- Kleindienst J., L. Seredi, P. Kapanen & J. Bergman: 2002a, "CATCH-2004 Multi-Modal browser: Overview Description with Usability Analysis". IEEE 4th International Conference on Multi-modal Interfaces, Pittsburgh, PA, U.S.A.
- Kleindienst J., L. Seredi, P. Kapanen & J. Bergman: 2002b, "Loosely-coupled approach towards multi-modal browsing", Submitted to Universal Access in Information Society magazine's special issue on Multi-modal User Interfaces.
- Halonen, K, H. Salminen & P. Kapanen: 2002, "Ring-ing Tones for Mobile Phones by Voice - Usability Study". ISCA Tutorial and Research Workshop on Multi-Modal Dialogue in Mobile Environments, Kloster Irsee, Germany.

Appendix C. The test scenario

1. Ask for your preferred programs (e.g. "Tell me about MyMovies.");
2. Act as the system instructs you (e.g. specify day or time, give further criteria);
3. Once a list of programs is given by the system, ask for a description of the first program in the list;
4. When asked whether the system should send you the program details in a short message, say that you do not want it (use natural expressions);
5. End the session and hang up.

Appendix A. An example natural language dialogue (non-personalised PGIS application)

System: Welcome to Program Guide. What would you like to know?

User: Could you please tell me about movies tonight?

S: I found the following movie programs on Monday between 6 pm and 12 pm: It's Always Fair Weather, The Murder Men, Boom Town, and The Hook.

U: What time is the first one?

S: It's Always Fair Weather starts at 5 past 5 pm. How else can I help you?

U: How long does it last?

S: It's Always Fair Weather lasts 1 hours 40 minutes. What information can I offer you next?

U: Who are the actors?

S: The actors are Gene Kelly, Dan Dailey, Michael Kidd, Cyd Charisse. How else can I help you?

U: Can I have a description for it please?

S: I can offer you the following type of description: The unbeatable musical team of Kelly and Donen is behind this happy musical about three serviceman buddies reunited after 10 years..... What else can I do for you?

U: That's all for now. Good-bye.

S: Thank you for using the Program Guide. Good-bye.

Appendix B. The web interface for the personalised settings

