
Voice Portal for Public City Transportation

Łukasz Brocki

PJWSTK
Warsaw, Poland
lucas@pjwstk.edu.pl

Danijel Koržinek

PJWSTK
Warsaw, Poland
danijel@pjwstk.edu.pl

Krzysztof Marasek

PJWSTK
Warsaw, Poland
kmarasek@pjwstk.edu.pl

Abstract

This paper discusses the main aspects of the user-center design for developing a voice portal and the impact it had while creating such a system in the Warsaw city transportation call center. The system was shown to be effective and support its users in their needs: about 30% of users completed their requests through the automated system without talking to human operators.

Keywords

voice portal, user-centered design, speech recognition, call center automation

Introduction

Until the late 80s the call center was an information service provided by the telephone company and staffed by human agents. However, after the information revolution the agent became simply an intermediary between the caller and the Internet. Thanks to the advances in automated voice technologies the process was facilitated by leaving only the challenging problems to the human agents.

Automated voice portals use a combination of technologies, like automated speech recognition, naturally sounding speech synthesis, Internet technologies, application servers and database technologies to give the caller the possibility to use

interactive voice services of various kinds, ranging from local and travel information up to stock and bank trading. While they are still treated as mostly experimental, they have been used for many years with great success. The main issue being user experience and satisfaction means special care has to be used while developing such systems.

User-centered design and its application to speech-based services

The preparation of services for naive users of computer technology has been a research topic for many years, not only resulting in rules or guidelines for proper human-computer interaction (HCI) but also methodologies of user interface design. User-centered design (UCD, ISO 13407: Human-Centered Design Process) places the end-user in the middle of the application preparation process for visually based HCI. Speech as a mode of interaction calls for different aspects of user behavior than visually-based technologies, however several authors agree on the use of UCD as the main methodology for the preparation of voice portals [15]. Usability is a term describing the ease and clarity of human computer interaction for a given computer program or web page. For voice portals usability is not easy to define, even if there is common agreement what the term "usability" means in this context. The classical notion of this term in case of HCI takes into account four main dimensions [14]:

- efficiency: the extent to which a system supports user performance, can the task be accomplished;
- effectiveness: if it takes less time to accomplish a particular task;

- learn ability: easier to learn;
- satisfaction: more satisfying to use.

Usability, because of its often subjective character, is not easy to measure, even if several methods of usability evaluation exist [6] and design methods (in case of visual interfaces) are well known. For a better understanding of the application of HCI design methods to the voice user interface (VUI), first some differences between graphical user interfaces (GUI) and spoken interfaces have to be pointed out:

1. *Time*. Visual information is usually static and lets the reader take as much time as necessary to read it, understand it, or read it again and again. Application dialogs may let the caller repeat an utterance or ask the application to repeat, but at least a part of dialog has to be memorized by a user. Thus, the dialog and prompts (menus) need to have an appropriately simple structure, e.g. in one dialog step the number of options presented to the user has to be limited. It is typically suggested to limit the number of options to 5-7 depending on the existence of barge-in feature. This causes the users to focus on a much narrower context.
2. *Sequential access*. GUI are able to present a lot of information in parallel. A user can scan hundreds of items to quickly get to the desired information. Speech is not so effective: information is presented in sequence, users must carefully listen to various lists, dialog flow cues, and help prompts before they can proceed with an appropriate action. It is well known that most people can only remember between five and nine numbers for around twenty seconds after hearing them. Consequently, listening to long lists of choices is

unreasonable, and purely hierarchical, menu driven applications are exhausting.

3. *User Control.* In the case of GUI, users can work at their own pace - in the case of VUI, users need to wait for instructions and then react. In a GUI, the available options are visible on the screen all the time which is not the case for the VUI.

4. *Mental Model.* The schemes employed in most web-based applications are familiar and consistent, which has given users a very clear mental model of how any GUI is likely to work. In contrast, a first time user of a VUI cannot predict the next dialog step without prior experience with the particular dialog system.

5. *Errors.* During interaction with graphical WIMP-style interface (Windows, Icons, Menus, Pointers) user errors may occur, but the correct interpretation of user interaction is generally assured. Despite the progress of Automatic Speech Recognition (ASR) technology, a proper hypothesis cannot be guaranteed: an appropriate acknowledge dialog is necessary.

6. *Latency.* If broadband a Internet connection is used, the latency of Web based systems (at least to the first reactions) can be very limited. In a spoken dialog long pauses are very unnatural and may occur if the dialog system is not fast enough, breaking the naturalness of the voice communication.

Despite the problems mentioned above, user centered design methodology can be adopted to the preparation of voice portals. In this iterative design process (Fig.1) four main design phases are applied.

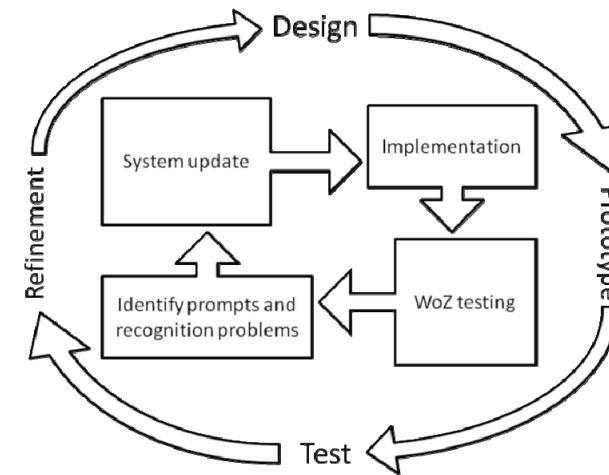


Figure 1. Iterative voice portal design process (after [14]).

In the first phase of voice portal preparation, design, the call flow is defined, as are the initial scripts for the dialog between the application and the user. Here the data collected from the actual operators of services one wants to automate is especially useful. Additionally, interviews with potential users, analyses of their needs and profiles and usage scenarios are applied. The second phase, prototyping, usually uses the Wizard of Oz (WoZ) testing technique. It is used to simulate the behavior of an automated system, by having a human agent play the role of the computer system. A call from a user is directed by a WoZ to the application based on a scenario (script) written according to the design phase. WoZ testing helps in understanding user behavior and gives a better view of potential user requests and expected application reactions. The information collected in the design and prototype phases allows us to build the first version of the voice

portal and initiate the testing phase. Here the main effort lies in tracking and analysis of the test calls to the automated system. During the testing phase, attention is concentrated on errors of the ASR, quality of grammars and dictionaries, Text-To-Speech (TTS) text generation, database interface, etc. One may also conduct surveys to get the callers' general feeling about the application during all preparation phases or just after the first deployment phase.

Experiment using UCD methodology

We prepared the voice portal according to the methodology described earlier in the paper. The system works at present at the Warsaw Transport Authority (Zarząd Transportu Miejskiego m.st. Warszawy - ZTM) which is the biggest city public transportation institution in Poland. It manages a call center that is accessible around the clock under the number +48-22-1-9484 and employs 10-20 operators working in different shifts. The call center provides information about departure times of city public transportation (buses, trams, metro and local railway), giving advice in choosing the best transport to reach a certain destination and other information pertaining to public transportation in Warsaw. The call center receives almost 30.000 calls per month with an average call duration of ca. 1 minute. Special attention was paid to the analysis of user requests and WoZ experiments.

Design and prototype

Within the framework of the LUNA project (EC 6 FR IST 033549) a huge effort of speech data collection of real telephone dialogs has been completed: a corpus of human-to-human dialogs was collected at the Warsaw Transport Authority telephone information hotline and this data allow us to understand how the callers request

information and how human operator complete the users' requests. During data collection, we observed on-site call center operator activities to obtain a better insight into the center's work-flow, group dynamics and interactions between operators and callers. We found, that much of the required data is accessible via the ZTM web page, but some of it remains in an unstructured form used only by operators (text memos, detailed city maps with transportation lines).

On a sample of more than 500 information seeking dialogs five proximate topic classes were identified:

- information requests on the itinerary between given points in the city;
- timetable for a given stop and given line and the travel time from given stop to destination;
- information on line routes, type of bus, tram (e.g. is wheelchair access available or not);
- information on stops (the nearest from a given point in the city, stops for a given line, transfer stops, etc.);
- information on fare reductions and fare-free transportation for specified groups of citizens (children, youth, seniors, disabled persons, etc.).

For these categories a WoZ experiment has been prepared. In fig. 2 the architecture of the system is given [9].

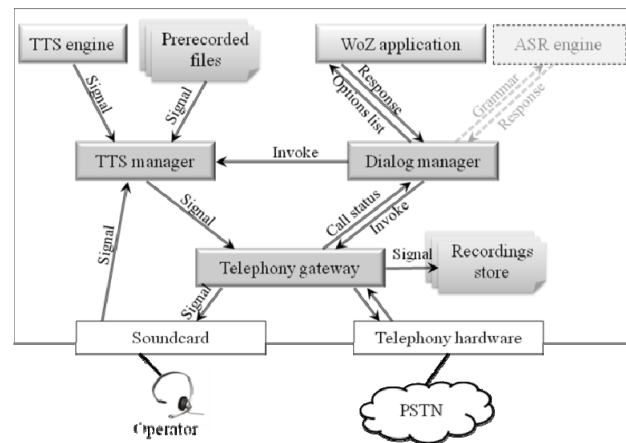


Figure 2. Wizard-of-Oz system [9].

The main parts of the system are the dialog manager and the TTS manager which coordinate the signal transmission through the telephony gateway. The WoZ operator chooses a response from a short list to simulate machine dialog steps. The WoZ simulation was done in its entirety for the domain of city transport fares reductions. For other domains, after a few steps of preliminary information collection from the user by the WoZ system, the operator would switch it off and continue the dialog with the caller. The WoZ experiments gave us a better view into user requests. From 844 recorded calls, only 459 of them were retained and classified as belonging to one of the mentioned topics. Quite often, users (155) wanted to speak directly with the operator, ignoring the system or waiting silently for the operator to respond. This showed, that the acceptance of automated service is quite low [9]. Moreover, many users were surprised that the system is automatic, so dialogs started with difficulty, contained long pauses, hesitations and WoZ

prompts repetitions. It was a clear indication that the system voice should be as natural as possible. In our preliminary experiments we found that latency introduced by our TTS engine is unacceptable to the users of the service. Thus, we decided to switch to a commercial product offered by Loquendo, characterized by high speed and high quality of speech synthesis. Moreover, the most frequent and static prompts have been replaced by prerecorded speech. Careful analysis of the WoZ data also allowed to build better, more flexible grammars for the ASR module and to prepare more informative prompts spoken to the user. The analysis of those dialogs shows that it is almost impossible to automate all classes of dialogs, mostly because of the complicated structure of the Warsaw public transport system, stops name conventions and imprecise users requests (see [13] for details). Thus we decided to choose the following dialog topics for automated services:

ZTM voice portal features

After dialing in, the user connects automatically with the voice portal. It is possible to connect with the human operator at any moment by pressing any key on the phone keypad. If all operators are busy, the user is put into a queue and they are informed about their place in the queue. The user can always decide to leave the queue and to use the automatic features of the voice portal instead. These include: departure times, complaints, ticket prices, fare reductions and news. There are two more options: connect with operator and city routes, both of which connect directly to the human operator. Automated functions are described below.

Departure times

The most elaborate feature of the voice portal in ZTM is providing departure times for buses and trams. The system recognizes about 4500 different names of bus and tram stops in Warsaw. It also recognizes all the transport line names, dates and times. After asking a few questions, the voice portal retrieves the times from a special database and synthesizes the response. The amount and type of questions the system has to ask depends on the chosen line and stop, and varies between 4-8 questions. The Warsaw transportation system is quite complex and the portal has to deal with a fair amount of disambiguation to be able to provide an accurate response every time. Also, the schedule is subject to slight changes very frequently (every 2-3 days) with not a lot of time in advance (sometimes only one day). This means that the system cannot provide reliable information too far into the future. The dialog about departures starts (the user is warned before that the conversation is recorded) with the question whether the user wants the schedule for current or some other day. The schedules are different for weekends and holidays so this is a very important question. As follows from our research, 90% of users want the current schedule so it is reasonable to start from this question. If the user is interested in future schedules, the system asks for a specific date and accepts many different types of phrases: tomorrow, Monday, January 12th, etc. The next question is about the route number, i.e. the number written on the side of the bus or tram. In Warsaw these are usually a number between 1-999 and sometimes combinations of letters and numbers, e.g. C-6. A list of n-best hypotheses returned by the ASR will sometimes be generated in this step due to acoustic similarity of some sequences in Polish, e.g. E-2 and N-2. In the following question the system asks for the

stop name. This is a difficult step, because the names often consist of several parts and each part has to be recognized individually or in connection with others (e.g. Central Railway Station 05). These names usually originate from streets and sites where the stops are located. The transport system of the city of Warsaw services a large area including the many small suburban localities. Therefore, a common name like "Szkolna" (i.e. School) street may occur several times. In such cases, the portal has to ask an additional question to disambiguate the stop. Finally, several stops in a single trip may have the same name, but are suffixed by an additional number. This infrequently merits an additional question from the portal. Most of the stops have two opposite directions. If a stop has more than one direction for the chosen line, the system asks the user about the direction by giving a list of end destinations for the given line, e.g.: which direction are you interested in: towards the city center, towards Wilanów city district? The final question is an approximate time of departure the user is interested in. Since most of the schedules contain up to 100 different departure times, it would be a waste of time to read them all. Instead the system reads the three closest of the given approximate time (1 earlier and 2 later) and allows the user to navigate the times by giving out commands: next departures, previous departures. The approximate time can also be given in many different ways, e.g.: 12 o'clock, half past eight, 7 in the morning, etc.

Ticket Prices

This feature gives the user information about the price of a ticket. There are up to 20 different ticket types available. The system asks the user what ticket he is interested in by asking a series of questions. There are

a few options per dialog step and after a few steps the user is presented with a price for the chosen ticket.

Fare Reductions

Here, users can find out if they can benefit from a certain fare reduction. The guidelines for ticket reductions are governed by city officials and are quite complicated to understand for the average passenger. Similarly to the ticket prices, this dialog is organized in a tree-like fashion. The depth of the tree is at most 5 steps and each leaf contains comprehensive information about the chosen reduction.

News

This is just a database of synthesized news items updated manually by the ZTM staff. Each item consists of a title and some content. These may contain information about important changes to the infrastructure, lines, ticket prices, etc. The user is simply presented with a list of news topics (titles) and by picking the particular news item (by sequence number) its content is synthesized.

Complaints

The favorite feature of the voice portal among the human operators is the automatic recording of complaints. This function simply asks the user to provide all the information after a tone. This information is recorded and stored for further review. ZTM is an independent, government sponsored organization that governs and audits the different commercial transport carrier companies around the city. One of their main tasks is quality control and communication with the customers. Unfortunately, in times of heavy traffic, accidents or other unpredictable events, lots of harsh complaints from frustrated

passengers pour in, which takes a big toll on the human operators and doesn't serve any purpose. When faced with an automatic complaint recorder, users are less likely to vent their anger and provide useful and constructive information.

Testing and refinement

According to UCD methodology the next step in application preparation is its testing and refinement. 300 calls were analyzed in order to check how many people used automatic information. Around 30% of hotline clients used the voice portal and didn't connect to human operators at all. The most demanded automatic feature is information on departure times. The second most used automatic feature is filing complaints. From the first day of deployment all calls are monitored and usage statistics are collected (Fig.3 and Fig.4). We modified the dialog flow based on the analysis of recorded data, especially ASR error handling. The ASR performs very well, normal procedure is to acknowledge the user input by additional question and simple yes/no confirmation. However, if the system cannot recognize the user input for the third time, the user is prompted to press "0" key on the phone keyboard to redirect the call to the human operator; if the fourth attempt to recognize speech fails, the redirection is automatic. On a side note, in previous versions the system asked the user to press any key if they wanted to connect directly to a human operator. Due to many mistakes we had to change that to "press 0" - callers were often confused to as what "any key" meant.

Also the prompt asking for a date of requested schedule has been modified: we observed that most callers asked for today schedule, so this is now the first option

to choose. Initially the cordial end of dialog was not being recognized (if the caller said "thank you, goodbye" system answered: "sorry, I did not understand"). Now the system reacts properly.

Finally, some users connected via old telephone equipment with rotary dialing instead of the standard touch-tone. This problem was solved by allowing the user to say "connect to operator" at any point in the dialog.

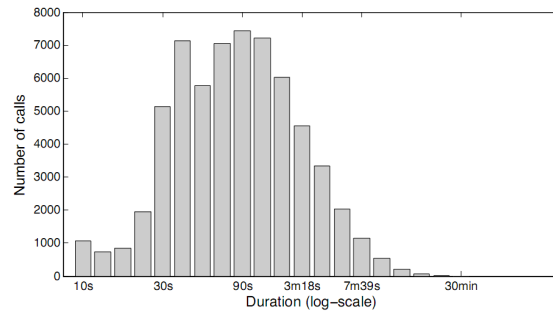


Figure 3. Distribution of call durations, including waiting time (based on ca. 60000 calls).

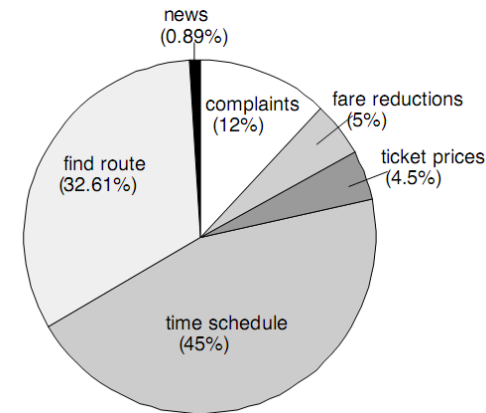


Figure 4. Calls distribution (based on ca. 60000 calls).

Conclusions

In Fig.4 some important statistics are given. Most of the calls took ca. 90 seconds, which is an acceptable call duration (an average dialog with a human operator took on average 2 minutes, but included more complicated requests). The system is effective and supports users in their needs: about 30% of users complete their requests through the automated system. Moreover, because some of the calls can be fully automated, the voice portal successfully shortens the waiting time for connection to the human operator. In the first days after the voice portal was deployed in ZTM hotline not many people were interested in using the automatic service. However, it was observed that with time more and more people started to use the fully automatic features. After three months several hundred people a day used automatic information and did not connect to the operators at all. It seems that the hotline divided customers into two groups. The first group (which consists mostly of pensioners and elderly

people) always wants to speak with the operators. The second group (in which there are many students and younger people) uses automatic dialogs willingly, especially if the queue waiting time is significant. It seems that many hotline clients are satisfied with the automatic service. The main advantage of such a system is that they never have to wait to get information that is available automatically. Most people are interested in departure times of buses and trams. It takes around 45 seconds for the human operator to provide proper information regarding this topic. The automated system needs twice the time. However, when one considers that people have to wait in queue sometimes for five or even ten minutes to get to the operator, the advantage of using the voice portal is obvious. A more natural dialog can be achieved if its flow is not fully controlled by the computer system. We hope to apply the methods elaborated in the LUNA project towards more flexible, mixed- initiative dialog flow.

Acknowledgements

This work is partially supported by the LUNA project (EC 6 FR IST 033549). The voice portal technology described in this paper is the property of Prime-Speech (www.primespeech.pl).

References

- [1] A. Acero, H. Hsiao-Wuen, H. Xuedon (2001), Spoken Language Processing: A Guide to Theory, Algorithm, and System Development, ISBN 0130226165, Prentice Hall
- [2] A. Black, A. Hunt (1996), Generating F0 contours from ToBI labels using linear regression. Proceedings of ICSLP 96', Philadelphia, USA, 3:1385-1388
- [3] H.A. Bourlard, N. Morgan (1993), Connectionist Speech Recognition: A Hybrid Approach, ISBN:0792393961, Kluwer Academic Publishers
- [4] B. Byrne (2001), Turning GUIs into VUIs: Dialog Design Principles for Making Web Applications Accessible By Telephone, VoiceXML Review, Volume 1, Issue 6
- [5] R. De Mori, F. Béchet, D. Hakkani-Tür, M. McTear, G. Riccardi, G. Tur (2008), Spoken Language Understanding for Conversational Systems, Signal Processing Magazine Special Issue on Spoken Language Technologies, Vol. 25, No. 3, pp. 50-58
- [6] A. Dix, J. Finlay, G. Abowd, R. Beale (2004), Human-Computer Interaction, ISBN:0-13-046109-1 Pearson Prentice Hall
- [7] T. Dutoit (1997), An Introduction to Text-To-Speech Synthesis, ISBN 0-7923-4498, Kluwer Academic Publishers
- [8] IBM International Technical Support Organization (2006), Speech User Interface Guide
- [9] D. Koržinek, Ł. Brocki, R. Gubrynowicz, K. Marasek(2008), Wizard of Oz Experiment for a Telephony Based City Transport Dialog System. In Proceedings of the 16th Int. Conference Intelligent Information Systems, 16-17 June 2008, Zakopane, Poland (in print)
- [10] Ł. Brocki, D. Koržinek, K. Marasek (2006), Recognizing Connected Digit Strings Using Neural

Networks, Text Speech and Dialog 2006, Brno, Czech Republic p.343-350

[11] K. Marasek, R. Gubrynowicz (2005), Multilevel annotation in SpeeCon Polish Speech Database, in: L. Bolc, Z. Michalewicz, T. Nishida (Eds), Intelligent Media Technology for Communicative Intelligence, Second International Workshop, IMTCI 2004, Warsaw, Poland, September 2004, Revised Selected Papers. LNAI 3490, Springer-Verlag, Germany, 58-68.

[12] K. Marasek, R. Gubrynowicz (2008), Design and Data Collection for Spoken Polish Dialogs Database, Language Resources and Evaluation Conference 2008, Marrakech Morocco

[13] A. Mykowiecka, K. Marasek, M. Marciniak, J. Rabięga-Wisniewska, R. Gubrynowicz (2009), Annotation of Polish spoken dialogs in LUNA Project, LTC'07, to appear in Springer LNAI

[14] J. Nielsen (1994), Usability Engineering, ISBN 0-12-518406-9, Morgan Kaufmann Publishers

[15] M. Polkosky (2005), What is speech usability anyway? Speech Technology Magazine

[16] R. A. Redner, H. F. Walker (1984), Mixture densities, maximum likelihood and the EM algorithm, SIAM Review, vol. 26, no. 2, pp. 195-239.

[17] K. Szklanny (2009), Cost function estimation in unit selection speech synthesis, Ph.D. Thesis, PJIIT Warszawa (in preparation, in Polish)

[18] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, K.J. Lang, (1990), Phoneme recognition using time-delay neural networks, in: Readings in speech recognition, ISBN:1-55860-124-4, Morgan Kaufmann

[19] P.J. Werbos (1990), Backpropagation through time: what it does and how to do it, Proc. IEEE, 78(10):1550-1560.

[20] R. Williams, D. Zipser (1989) A learning algorithm for continually running fully recurrent neural networks., Neural Computation, 1(2):270-280.

[21] Y. Jon Rong-Wei, (2003) Corpus-Based Unit Selection for Natural-Sounding Speech Synthesis, Ph.D. Thesis, MIT Dept. of Electrical Engineering and Computer Science