Design and Usability of Interactive Voice Response Systems for Emerging Markets♦

Abhimanyu Nohwar[1], Beena Prabhu[1], Kalika Bali, Girish Prabhu
HP Laboratories India
HPL-2005-49
March 17, 2005*

interactive voice response, emerging markets, dialog design, wizard of oz test

Interactive Voice Response Models (IVR) models, which are common in the West, have a fundamental relevance in the context of emerging markets and developing economies like India because of low literacy levels, multiple languages and low levels of English use. India, specifically, has a very 'verbal' culture, where, in addition to English and Hindi, 14 other officially recognized languages are spoken. HP labs is conducting research in enhancing the reach of Information Technology to the masses by creating applications in local languages in verticals like transportation, and is developing a generic system that will cater to any service like healthcare, business, banking etc. Currently, a proof of concept is being built to handle railway reservation and enquiry for the Indian railways, and supports both Indian English and Hindi speech input and output. Since the system is targeted towards the general Indian population which is expected to have a percentage of people that are not technology savvy, development of a user-friendly and easy-to-use system is imperative. Accordingly, this project deals extensively with speech interaction design and usability testing through an iterative, user centered design process.

The system has been through multiple usability tests and iterative design cycles. The system is being modified based on the input from these usability tests and will be taken into a field study in a live environment. In this paper we will discuss the methodology and results of the user-centred design process used to implement the IVR system, and present our experiences and learning from conducting these tests.

* Internal Accession Date Only
[1]Human Factors International, 4th Floor, Chemtex House, Hiranandani Gardens, Mumbai – 400 076, India
♦HCI International 2005, 22-27 July 2005, Las Vegas, Nevada, USA
Approved for External Publication

Design and Usability of Interactive Voice Response Systems For Emerging Markets

Abhimanyu Nohwar[1], Beena Prabhu[1]

Human Factors International
4th Floor, Chemtex House
Hiranandani Gardens
Mumbai – 400 076
{abhimanyu.nohwar, beena.prabhu}@hp.com

Kalika Bali and Girish Prabhu

HP labs
24, Salarpuria Arena,
Hosur Main Road, Adugodi,
Bangalore, India
{kalika, girish.prabhu}@hp.com

[1] *Currently at HP labs*

Abstract

Interactive Voice Response Models (IVR) models, which are common in the West, have a fundamental relevance in the context of emerging markets and developing economies like India because of low literacy levels, multiple languages and low levels of English use. India, specifically, has a very 'verbal' culture, where, in addition to English and Hindi, 14 other officially recognized languages are spoken. HP labs is conducting research in enhancing the reach of Information Technology to the masses by creating applications in local languages in verticals like transportation, and is developing a generic system that will cater to any service like healthcare, business, banking etc. Currently, a proof of concept is being built to handle railway reservation and enquiry for the Indian railways, and supports both Indian English and Hindi speech input and output. Since the system is targeted towards the general Indian population which is expected to have a percentage of people that are not technology savvy, development of a user-friendly and easy-to-use system is imperative. Accordingly, this project deals extensively with speech interaction design and usability testing through an iterative, user centered design process.

The system has been through multiple usability tests and iterative design cycles. The system is being modified based on the input from these usability tests and will be taken into a field study in a live environment. In this paper we will discuss the methodology and results of the user-centred design process used to implement the IVR system, and present our experiences and learning from conducting these tests.

1 Keywords

Interactive Voice Response, Emerging Markets, Dialog Design, Wizard of Oz Test

2 Need for a Speech-Based Technology Interface in Emerging Markets

Due to non-availability of input methods in local languages, computer usage in India is largely restricted to the English-speaking classes, which are a minority in India. This limits access to services based on computer and internet networks to a group of 3.6 million users [see Table 1], out of a national population of 1.08 billion. For example, English newspapers enjoyed a market share of only 15% in the year 2002 (Dionne, 2002), compared to a vast majority of readership in regional languages.

March 2003	March 2004		Growth Rate
1. PC based Internet Subscriber Base (in millions)	3.6	4.2	15%
2. Telephone Subscriber Base (in millions)	54.5	76.2	40%
3. Ratio of Telephone/Internet Penetration	15 times	18 times	2.7 times

Table 1: Telephone and Internet Subscription in India

As can be seen from above table, telephone subscription figures in India have grown 40% between 2003 and 2004 (Chakrabarti et. al., 2004). As emerging markets like India exhibit shared model of usage, it is clear that the access to telephones has significantly increased. Hence, telephone could act as a medium to provide access to services like travel enquiry and booking, healthcare, banking and stock trading, to name a few. These types of telephone based services are beneficial to the masses that would otherwise need to travel to the service provider's nearest office, which could be in a different city altogether, or depend on a mediator who spoke the language of interaction.

An interaction model that connects computer databases to telephones, in the local language, has the potential to reach a much larger audience than that of internet users only. It would accommodate uneducated users and provide easy access to information, in their language. To make computerized services available to the masses, it is necessary to redesign the interaction model of electronic databases by which users can access information and services in their native language. Speech is a natural medium for communication, particularly for the uneducated majorities in emerging economies. A system that can utilise speech as an interface is expected to effectively allow the majority of the populace to utilise its benefits.

3 Nilgiri IVR – User Centred Design

To demonstrate the power of local language services, HP Labs has been developing a generic voice interactive system with adaptability for a wide range of verticals like business, travel, healthcare, city administration etc. To achieve this degree of adaptability, the system has been built using three types of components (Anjaneyulu, et. al., 2005):

- *Application-specific*: the business logic of the subject for which the system is being built, e.g. railway reservation.
- *Dialog-specific*: the components that deal with the Voice User Interface (VUI) of the application, and
- *Neutral components*: components that take inputs from the configuration files and collaborate with other components as dictated by the configuration files.

The project code named "Nilgiri" is a reservation and enquiry system built for the Indian Railways. The scope of the project has been limited to make it possible to concentrate on application development rather than handling logistics related the Indian Railways which is one of the largest transportation systems in the world. It supports trains running between the four metropolitan cities in India - Delhi, Mumbai, Chennai and Kolkata. The system is speaker independent and supports input and output in Indian English and Hindi. The live information is provided by the Indian railways website that is integrated with the Nilgiri application via a web-service.

The Nilgiri Project used an user centred iterative design approach, where based on initial understanding of user needs a working proof of concept system was built using technology and integration of open-source tools for voice-based services. This proof-of-concept was then taken through extensive business logic design, system task flows, dialog designs and few rounds of usability testing. The improvised systems is currently being implemented which will taken through a field study.

3.1 Nilgiri Version 1

3.1.1 System Design

The system logic for Nilgiri Version 1 consisted of a total of 18 states. The user had to navigate through thirteen data entry states to make an enquiry, and an additional five states to purchase a ticket. Every activity required for either making a telephone enquiry or booking a ticket had been assigned its own state, For example, date entry, month entry, credit card number, and credit card expiry date had their own states. Further, the flow for each state was designed independently. This resulted in the development of 18 separate flowcharts for the system (one for each state). A typical flowchart is presented in Figure 1.

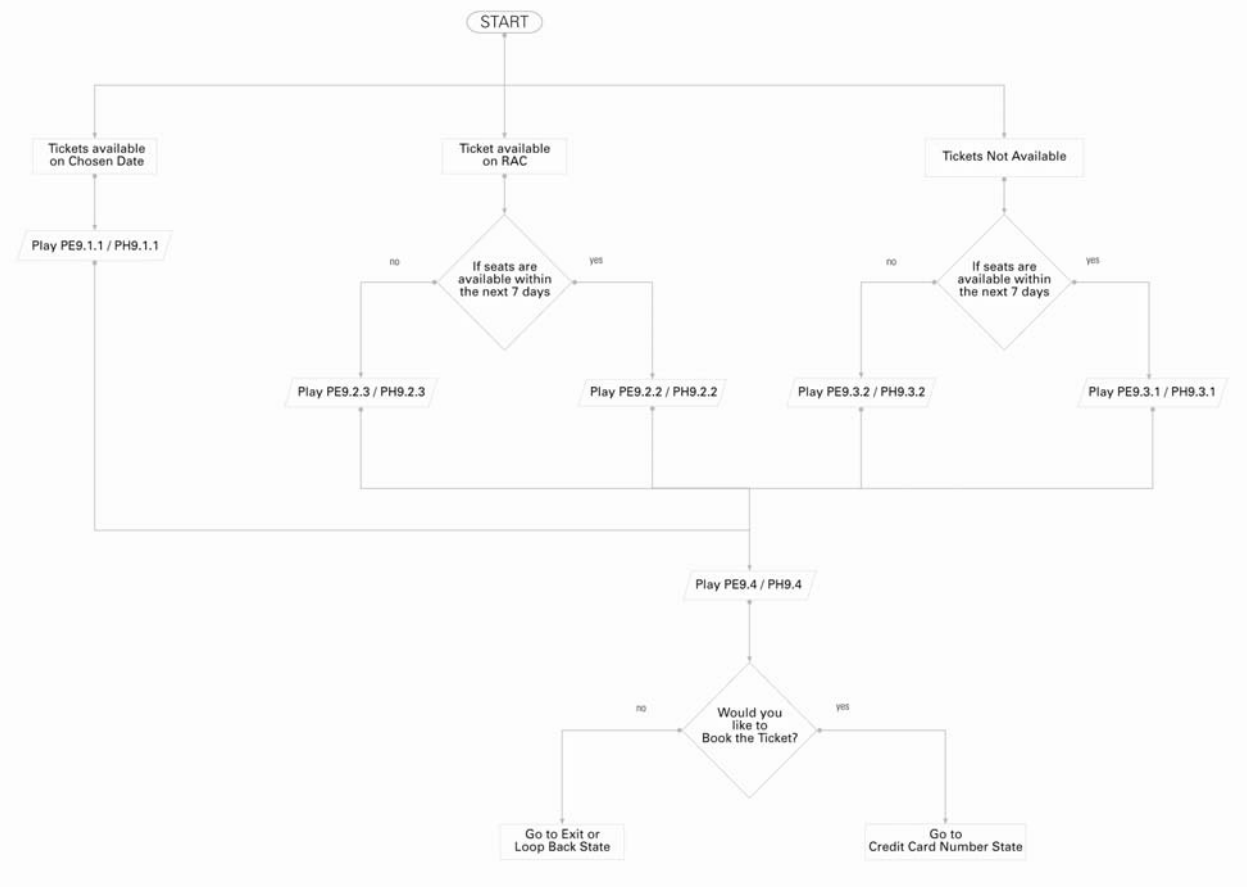

Figure 1: One of the eighteen State Flow diagrams for Nilgiri Version 1

3.1.2 Remote Usability Test

A remote usability test of the Nilgiri application was conducted using 190 participants. Although the primary objective at this stage in the project was to study the technical performance of the recognizer engine and to record recognition-accuracy related data, it was decided to take this opportunity to assess the system's usability issues as well.

3.1.2.1 Methodology:

All the 190 participants were selected from HP India offices. All the participants were given a telephone number to call and check for ticket availability by interacting with the automated IVR system. A sample task was as shown here: "Check the availability of a ticket on Rajdhani Express, for a journey on the 22nd of March, from New Delhi to Mumbai, travelling by 1st class." At the end of each task, the users were to fill in the questionnaire relating to that task. Each question was to be answered on a five point Likert Scale as well as with an open ended comment. The users filled in the questionnaire and returned them via e-mail. All calls were recorded to be analysed later for recognition accuracy. The questionnaire addressed issues such as clarity of the dialog, timeliness and adequacy of feedback messages/prompts, user's confidence in the task flow, error recognition and recovery-for both, user and system errors, overall reaction to the system, and general comments and suggestions for improvements.

3.1.2.2 Results:

The test results indicated that most users were happy with the performance of the system, (917 out of 1104 answers were favourable responses). Also, each answer was accompanied by a subjective comment, which explained the reasoning behind the answer. The total of 1104 responses to subjective comments was categorized into 3 categories,

namely, technical issues, dialog-related issues and users' feelings/reactions. Further, similar responses were clubbed together and a simple frequency distribution was made to highlight the most common comments. Insights resulting from each response were listed alongside the comment (see Table 2).

Response Category	Insights	User Comments	Frequency
Dialog-related issues	1. Most users preferred barge-in.	The process is a bit slow. Confirming all answers may not be necessary, maybe only at the end.	16
		It is simpler and faster on the web; also, one has more time to think on the internet.	7
	2. Users preferred using common speech rather than technically accurate language.	Once I answered wrongly, I couldn't go back to the previous or main menu.	3
		By mistake I selected the wrong destination and later I was not able to cancel the selection. I got disconnected.	2
		I could not get the system to repeat the message.	1

Table 2: A part of the User Test results analysis chart

The study resulted in a number of insights that led to major changes in the design of the dialog and system flow. Some of the changes to the system that resulted from this user test were:

1. Barge-in was incorporated and the users were informed at the beginning of the session that barge-in was acceptable.
2. User-initiated error-recovery was changed to a system-initiated one.
3. Error recovery was modified such that, in the event of incorrect user input, a corresponding error message was added and the user was guided on how to recover, rather than disconnecting him after a timeout or after three consecutive errors (Balentine and Morgan, 1999).
4. Vocabulary of each prompt was increased to accept a certain degree of colloquial input (e.g. "ya" for "yes").
5. Explicit confirmation was reduced after each user entry.

3.2 Nilgiri Version 2

The Version 1 user test provided a platform to refine the technical performance and system architecture of the system. Issues of recognition accuracy and system integration were addressed. The next phase of the project saw this focus shift to usability and dialog design.

3.2.1 Design Changes

3.2.1.1 Changes to the System Architecture

After the first user test with Version 1, we found that it was difficult for users to correct errors and they could not recall the help and error recovery keywords. Following the recommendations from the usability test, the architecture was redesigned and simplified [see Figure 2]. Where initially each state had its own flow, now all components of the user's task flow were handled by one flow structure. The error recovery system was changed from user-initiated to machine-initiated, in order to lessen the cognitive load of having to remember particular correction keywords and numbers on the keypad (Balentine and Morgan, 1999).

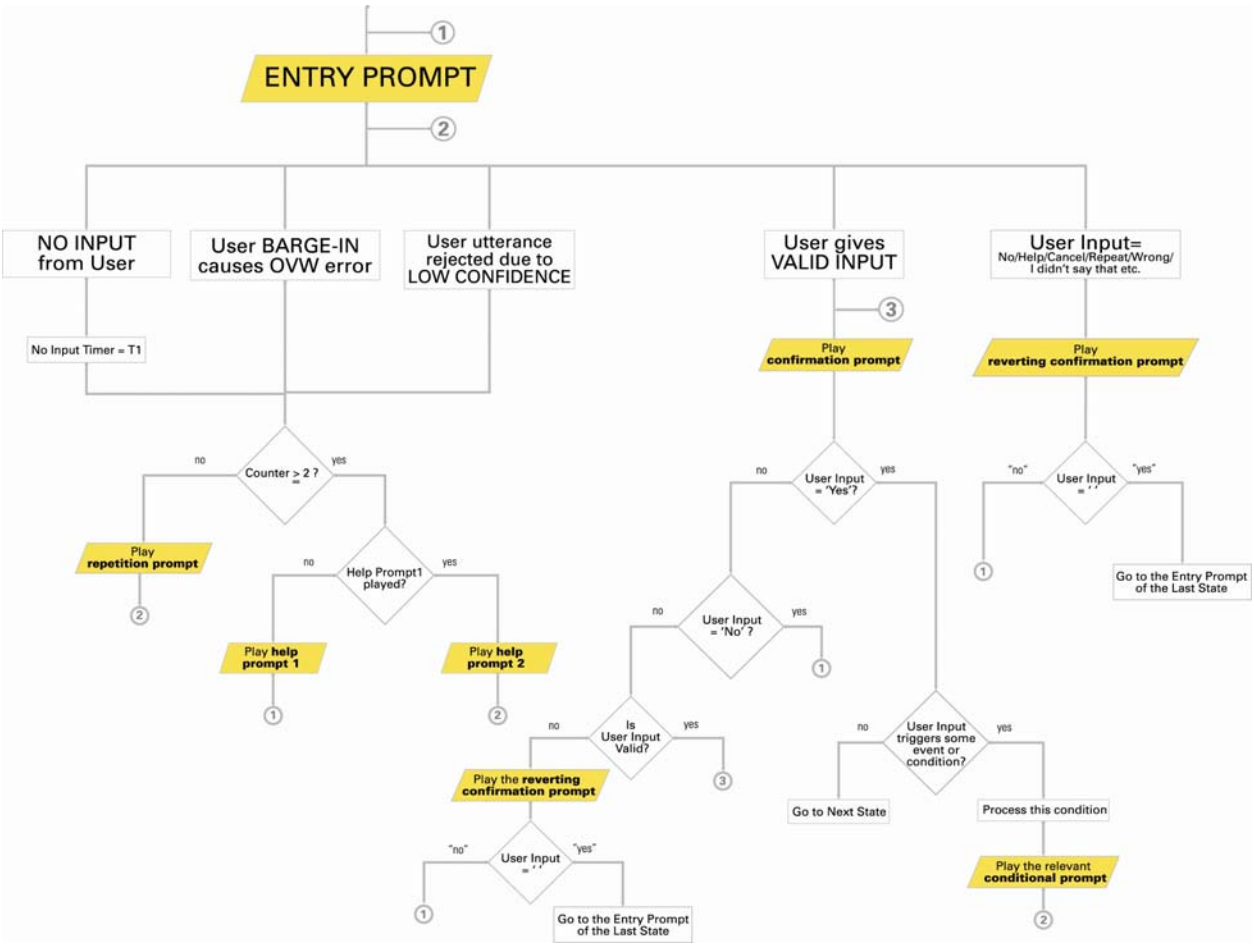

Figure 2: Integrated system flow of Nilgiri Version 2

3.2.1.2 Implementing Hindi Dialog

Designing the Hindi dialog for Nilgiri was a challenging exercise partly due to the structure of the language and partly due to nature of multilingual society. The syntactic structure of Hindi follows the order Subject-Object-Verb as opposed to Subject-Verb-Object in English. Further differences between other syntactic components of the language structure make a one-to-one translation between the two languages fairly difficult. For example, in English almost all nouns can occur as verbs, while in Hindi, the verbalization of nouns results in complex lexical items of the form nominal + simple verb. Thus, the verb "(to) start" in English will be roughly translated into Hindi as "aarambh karnaa" [start (noun) + (to) do (verb)] (see Chakrabarti et. al., 2004). It was crucial to get the right variety and style of Hindi for the dialogue. As a diglossic language, Hindi has at least two varieties where one is recognised as a "high" variety that is used for more formal domains and the other is a relatively "low" variety that is used more "informally" (Schiffman, H., 1997). In designing the Hindi dialog the challenge lay in choosing a "medium path", a variety that was informal while still being polite. Several of the terms used in the English dialogue, like "ticket", "seat", "station" etc. had either no equivalents or very formal literary equivalents in Hindi. Most of these English words have been borrowed into Hindi with alteration in their pronunciation inline with Hindi sound-system. Users sometime spoke in English even though they had selected the Hindi mode. Constant code-switching between English and Hindi is fairly common place in a multilingual society like India where even those who may not be able to speak English have a passive knowledge of some English vocabulary, especially that of common use like numbers, days of the week etc. (Mohan, 2005).

3.2.1.3 Changes to the Dialog Structure

The dialog structure of Version 1 had three tiers of prompts: Level 1 had prompts for novice users, which were elaborate; Level 2 had regular prompts for intermediate users and Level 3 prompts were shortened to keywords for advanced users. This was reduced to one tier of prompts for all profiles of users. This change resulted in reduced development time and increased system robustness as the system flow did not require adapting for different levels. It also reduced the error rate since the language of prompts remained consistent as against the previous model where the user would hear different prompts according to his performance—elaborated prompts for novices and shortened ones if the system classified the user as an advanced user.

Every user input had explicit confirmation at two points in the dialog. The intention was to give the user a second chance to correct any errors. However, this proved to be too much confirmation for the users' liking. Hence, this was educed to one confirmation message. While it is a common rule to keep confirmation prompts implicit in the dialog (Balentine and Morgan, 1999), the user studies at showed that the participants were comfortable with a degree of explicit confirmation. Observations at the railway reservation counter showed that people asked for confirmation of ticket details several times. While this has not been subject to organised study as yet, it will be tested in the field study.

3.2.2 Wizard of Oz test

The new design was tested using a Wizard of Oz test, a low-fidelity usability test that enables the design team to test the usability of a system without having to actually build a working prototype. One member of the testing team - the 'wizard' - played the role of the IVR system and spoke over a phone with the participant, who sat in an adjacent room. The aim is to simulate an actual IVR system and to get the participants to behave as they would in that situation.

3.2.2.1 Objectives of the Test

The primary objective of the Wiz of Oz test was to assess the usability of the systems task flow, ease of error recovery, and clarity of the system prompts:

1. Systems Task flow - After the remote usability test, the system flow diagram was redesigned and condensed into one flow that accommodated all the system states and error-recovery conditions. It became necessary to test how well this logic flow would work and if it could still handle all possible scenarios of use.
2. Ease of error recovery - Design of error recovery logic such that the system would attempt to automatically identify erroneous input and prompt the user to make the correction. This needed to be tested to see if the system logic would be comfortable for the user to deal with, when it gave instructions on how to recover from an error.
3. Clarity of system prompts - To validate issues of clarity in the language of the prompt, and comprehensibility of the Text-To-Speech synthesis of the prompts.

The aim was also to investigate if technology savvy of participants had any influence on the overall performance using the IVR system.

3.2.2.2 Methodology:

The Wizard of Oz test was conducted with 16 participants recruited from HP office as shown in Table 3. The participants were recruited via e-mail and screened for age and technology savvies. However, the nature of the test was not revealed and they did not have any idea that they were going to interact with a person and not a fully functioning IVR system. They were asked to fill screener questionnaires to collect their personal profiles. We managed to recruit eight Hindi and English speaking participants each with a 50-50 gender mix and a good age group distribution. We looked for participants who were primarily from a non-technical background and had varied prior experience with IVR systems – naïve (6), average (6) experienced (4).

Tech savvy	ENGLISH SPEAKERS (8)			HINDI SPEAKERS (8)		
	Age Group			Age Group		
	20-30	**30-40**	**Abv40**	**20-30**	**30-40**	**Abv40**
Naive	2		1	1	2	
Average	1	1	1	1	2	
Experienced		1	1	2		

Table 3: Distribution of participants

The test was scheduled over four days and it took from 1 to 1.5 hours per participant. The general ask was to book a ticket on a given date and train, in a particular class, from different start and destination stations. The participants were given four variations of the above task and errors were introduced into three of these to study their reactions as follows:

- Task1- no errors.
- Task2- one recognition error of the train name.
- Task3- no tickets available–asking the user to go back in the task flow and make changes to his enquiry preferences.
- Task4-Multiple errors: one substitution error with confusable date pairs, where the recognizer substitutes a like-sounding word for another (e.g. 13/30, 7/11). One recognition error with credit card date, rejecting user's voice input and forcing him to use DTMF after playing the DTMF Help prompt.

At the end of each task, the users were asked questions about their experience. The facilitator probed to get insights into unusual behaviour and user reactions. Users' feedback and observations from the study were incorporated into the architecture and the dialog of version 3.

3.2.2.3 Results and Recommendations

The analysis of the user response data indicated that there were no differences on user acceptance as well as usability across gender, age groups or familiarity with IVR systems. The analysis resulted in the following recommendations for changes to the design of the prompts and the system flow:

1. Design of Prompts:
   - Rephrase and simplify confusing terms like 'Reservation Availability' in both languages.
   - Confirmation prompts should respond in the user's format, for example:
     system: In which month do you want to travel?
     user: "eleven"
     system: Did you say eleven…November?
     user: "yes"
     This also applies if the user speaks in English even though he has selected the Hindi mode.
   - Keep the elaborate dates like 'one-three…thirteen' restricted to confusable pairs like 13-30 and 11, 7 etc.
   - For confusable date-pairs, if the user says 'no' to the confirmation prompt and if the confidence level is low, ask about the other in the pair automatically as the next confirmation prompt.
   - If entry was in DTMF, rephrase confirmation prompt to 'you entered' instead of 'you said' (Balentine and Morgan, 1999, Moscovitch, 1997).
   - Add 'okay' and 'thank you' to the vocabulary for confirmation prompts (some users were found to use these phrases)
   - Change Low confidence error from "I'm sorry, I didn't understand" to 'Please repeat that' – guides the user on what to do next (Balentine and Morgan, 1999).
   - Don't use examples (Balentine and Morgan, 1999); just explain the format of entry as users tend to follow the example instead of entering their data.
   - Once the user ID is recognized, the system could play a personalized confirmation/welcome message like "Hello, Arjun.'

One feature of the system that most of the users did not like was too many confirmations. The confirmations after each prompt slowed down the whole process. However, this opinion was taken as representative of the subject profile, which did not cover the entire range of user profiles and which may not be consistent with the entire target audience. The field study will aim to validate this assumption.

2. System Flow:
   - Add a loop to catch false/mistaken confirmation by the user.
   - Add a loop to catch user's attempt at correction.

On the whole, the system architecture worked well and users were able to recover from errors in all cases. There were a few issues with the dialog with respect to comprehensibility of the language, which was then simplified.

4 Next Steps

The new version of Nilgiri based on above recommendation will be tested in an un-moderated field study of the system, with around 200 participants in four distinct geographical regions. All calls will be logged and user feedback recorded for analysis using a graphical data visualisation tool. The focus of the field study is to test the usability and overall acceptability of the system and to identify issues in the voice user interface before the application is launched. Some issues that were not validated in the user tests such as a) acceptance of the system as a function of education and exposure to technology, b) acceptance of local language systems in comparison to just an English-based system, and c) users preference for explicit confirmation of data entry, will be tested in the field study

The field study will be a summative test of the system architecture and dialog design. The main focus of conducting the study, however, is to observe users' reactions to using a telephone-based system, and to get an insight into what aspects of such an interaction were liked or disliked. The main usability objectives will be to evaluate ease of navigation through the system architecture, efficiency of the dialogs, ease of error recovery, and optimization of task completion times.

Following the field study, the Nilgiri project will be used as a benchmark to develop the IVR application further. Future applications will be developed as reusable modules that will be suited for use in various services. We aim to identify which demographics respond well to this system and benefit the most from it, and whether certain sections of people require specific tailoring of content and presentation (like logic flow and dialog). Accordingly, we hope to gain insights into possible further applications of such a system.

5 References

Balentine, B., and Morgan, D. P. (1999), How to Build a Speech Recognition Application. Enterprise Integration Group, 59, 186, 244.

Moscovitch, M. (1997). Designing Voice Menu Applications for Telephones in Helander, M. G., Landauer, T. K., and Prabhu, P. V. (eds.) Handbook of Human-Computer Interaction, Elsevier Science Publishing, B.V. (pp. 1085-1102).

Anjaneyulu, K.S.R., Rao, G.V.D, Sitaram, R.N.V, Reddy, S., and Urs, A. D., (2005). A Reusable Framework for Voice Applications on OCMP. HP Techcon 2005. 1-4.

Dionne B., (2002), The Rise of Print, Frontline. Vol 19-Issue 14, July 06-19.

Chakrabarti, D., Rane, G., and Bhattacharyya, P., (2004). Creation of English and Hindi Verb Hierarchies and their Application to English Hindi MT. International Conference on Global Wordnet (GWC 04), Brno, Czech Republic, January, 2004.

Schiffman, H (1997) Diglossia as a Sociolinguistic Situation. Florian Coulmas (ed.), The Handbook of Sociolinguistics. London: Basil Blackwell, Ltd.

Mohan, P. (2005) Is English the language of India's future?. Seminar 545: India 2004: A symposium on the year that was. January 2005.