AN OVERVIEW OF THE EURESCOM MIVA PROJECT

Denis Johnston

BT Adastral Park, Ipswich IP5 3RE ,England,UK

e-mail: denis.johnston@bt.co.uk

Phone: +44 1473 64 2128

ABSTRACT

The goal of the MIVA project was to answer a number of fundamental questions concerned with the exploitation of speech technology enabled systems. The experimental service chosen was designed to help foreign people travelling in the country to find emergency and embassy numbers, country and area codes, useful numbers (directory service, country direct, etc.) and how to use Telecom and credit cards for placing calls.

Services were implemented in each of the countries taking part and two stages of experimentation were undertaken. The first of these was a mono-lingual experiment carried out in each country to optimise performance for each country /language combination. The second was a fully multi-lingual service in which each of these optimised services was re-implemented in all languages. All systems were evaluated over combinations of local and international environments. Correlations derived from a subset of the subjective and objective results were used to provide a predictive model of users opinions and the remaining subset of data used to test these predictions.

1. INTRODUCTION

Many services based upon advanced speech technology such as ASR have been implemented in recent years but in the main these services have been limited to one language. In this project we undertook basic research into how such systems might be implemented, applied and evaluated in a multi-lingual environment. With so many interacting factors likely to impact upon users' perceptions selecting the best combination is far from trivial. Previous experience, recogniser accuracy, recogniser speed, recogniser threshold settings, vocabulary choice, characteristics of spoken prompts, line types, phone types, dialogue characteristics are all important and interact strongly. Simultaneously optimising them can be exceedingly difficult. Dealing with is complexity was expected to be a major challenge. This multi-dimensional problem has generally resulted in two broad approaches to evaluation. One has been to arbitrarily choose one of the more obvious and directly measurable factors such as recogniser threshold settings, attempt to hold all others constant and then measure some other quantifiable effect such as total transaction time. The second has been to set up the service, invite people to use it and then use questionnaires or interviews to collect opinions about the quality or usability of the service.

However neither of these is really satisfactory. The problem with the first is that there are few factors that are directly measurable and meaningful. More often it is the unmeasurable and intangible features of the service such voice quality and dialogue styles that tend to dominate user perceptions. On the other hand in the field trial approach, reliable data collection (especially of users perceptions) is difficult and expensive. And subsequent analysis and interpretation of the results in what is an almost totally uncontrolled environment is often impossible.

In the MIVA project we substantially circumvented these problems by adopting a methodology based upon multi-factor experimental designs. This approach has been widely applied in other disciplines and is one of the standard methodologies applied in the life sciences[1]. We show how this approach overcome all of the disadvantages of the methods described above and allowed decisions concerning the best 'mix' of characteristics to be determined.

2. ASKING THE RIGHT QUESTIONS

2.1 High level questions

The starting point for such a process is a list of high level questions. This was done in consultation with the technical and marketing departments of our organisations. This ensured a high level customer led drive for the project and simplified the agreement on what issues were of importance to all concerned. Once the questions were established they were prioritised.

The key questions identified were:

- Do users prefer speech recognition to DTMF input for Interactive Voice based services?
- Is the performance of speech recognisers significantly impaired when services are accessed over GSM networks?
- Are there certain preferred dialogue structures ?
- What speech recogniser parameters (e.g. rejection settings, cut- through strategies) are preferred?
- What is the best way to prompt users at the start of a dialogue so that their language can be identified?
- Do recognisers perform differently when accessed over international links?

In examining these questions it became apparent that they could best be answered using two separate series of experiments. The first series would be undertaken independently in each country and would address those questions (numbers 1-3) that did not have a direct multi-lingual dimension. This first phase would also be used to optimise platforms and dialogues and identify the appropriate prompts and words to be recognised. The second phase would embrace those tests that demanded multi-linguality.

2.2.Supplementary questions

During the course of project it became apparent that we could and should, address other questions such as:

How should the problem of 'unanswerable' queries be resolved? For example if a person accesses an information system expecting to obtain information about telephone fault repair how do you deal with the problem that this information is not in the database? What, if any, are the important differences, due to networks, country size, languages, cultures etc. which must be taken into account when providing multilingual services?

2.3 From questions to hypotheses

Once the questions had been established we were able to move to the experimental design stage. The first step in that direction was to convert the questions into the 'null hypothesis' format necessary to allow statistical tests to be performed at the subsequent analysis stage.

Formulations of null hypotheses for the above questions are:

- Users show no preference or behavioural patterns between speech recognition or DTMF
- The performance of speech recognisers is not impaired when services are accessed over GSM networks?
- All dialogue structures are equally efficient and effective.
- Recogniser parameters such as rejection settings, cut- through strategies make no difference to performance or user behaviours.

The value of recasting the questions into a null hypothesis format comes from the fact that the onus shifts from having to prove something true to proving it untrue.

2.4 Selecting an appropriate service

Having established the basic 'scientific' hypotheses the next stage was to establish the framework of the service. Other issues, such as data exchange agreements and protocols were also established at this point. Our choice of service was determined by a number of factors amongst which were

- The potential for using Automatic Speech Recognition
- Usefulness in a multilingual environment
- Relevance to our parent Telecommunications companies
- Feasibility within the time frame of the project.

One service that met these all of these constraints was originally conceived as a pan-European multilingual help-line. This would provide help on how to use the telephone network in foreign countries. For example an Italian speaker visiting France would be able to obtain guidance, in Italian, on how to use the major facilities of the France Telecom network. Examples of the types of information to be provided were emergency and embassy numbers, country and area codes, national and international directory service numbers, country direct numbers, tariffs and how to use Credit and Telephone Cards.

Clearly such a service could be implemented with various degrees of sophistication ranging from a simple DTMF service to one using a full natural language interface. It was also apparent that each implementation would have to be 'tuned' to the services actually available in each host country.

By this stage the architectural structure of Fig 1 was beginning to emerge. However just how well this would work in all countries was not clear, so in parallel with addressing the first set of questions, the first experimental phase was designed to optimise the dialogue structure for the service environment of each.

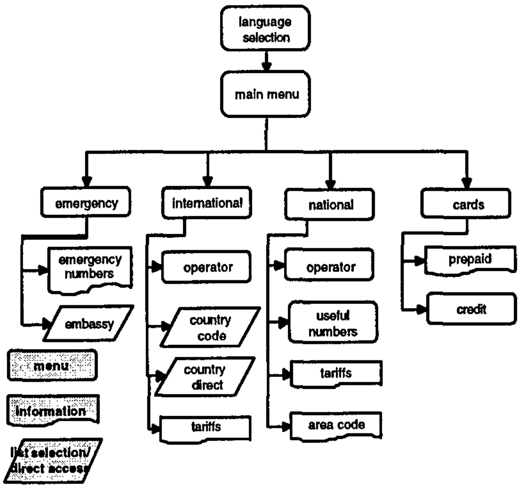

Fig 1 – Outline of the basic service structure

3. THE MONOLINGUAL EXPERIMENTS

3.1 General

The first series of experiments were mono-lingual experiments. In these all partners independently implemented and tested their platforms but adhered to the common structure and methodology.

The aim of these was to undertake experiments concerned with:

fixed/cellphone comparisons
prompt optimisation
recognition/Touch-tone comparisons
recognition threshold settings which are preferred

The following factors were used in each case
A total of 16 test conditions were devised by combining three recogniser settings + 1 'Touch-tone' with four dialogue structures. At least two different types of phone (GSM and fixed network) were then used in the test process.

3.2 Tasks

For each 'conversation' with the system subjects were given a task to complete. These were divided into two categories - those, which had answerable questions, and those which the system could not answer. The two types were chosen because in real systems many users request information which the system does not contain. We wanted to examine how different dialogue strategies coped with this situation and how users reacted to them . Examples of the answerable questions were:

What is the number for the French Embassy?
What number do you call for an operator?
What number do you dial for chargecard information?

The unanswerable questions were very similar – but the database did not support them e.g.:

What is the number of the Russian Embassy?
What number to you call to install a new phone line?
What is minimum charge for calls made using credit cards?

3.3 Test Procedures

In principle, with 16 test conditions/treatments it should be sufficient to undertake a fully balanced experiment with 16 subjects and 16 tasks. However the complicating factor of the answerable/unanswerable questions and the slight imbalance created by having 3 recogniser based system and one DTMF system meant that a modified design involving 16 subjects and 20 tasks were used.

This highlights the sequence of conditions that a typical subject experienced. Each subject was first given 3 'practice' sessions to ensure they were comfortable with the procedures. The above design is balanced in that every subject experiences all the conditions and all the tasks. However no subject experiences exactly the same combination of tasks and conditions as any other. The advantage of using such a balanced approach becomes evident when the analysis stage is reached for it becomes possible to partition the data in many different ways. For example exactly half of all calls will have been to fixed networks and exactly half to GSM networks. The balancing process guarantees that each of these groups will contain exactly the same set of tasks and exactly the same distributions of talkers.

How this helps in the analysis is shown below, but before that we examine the responses.

3.4 Responses

There are a number of well known subjective response scales available for this type of subjective procedure[2]. However for our purposes none were immediately suitable. The added complication of task completion - which could be successful or unsuccessful had to be taken into account as there is an important distinction to be observed between satisfaction with the result obtained and satisfaction with the system used to obtain it.

The three dimensions identified were:

- Did the service deliver what it was supposed to?
- Was it easy to use?
- Was it pleasant to use?

To deal with this complication the following three simple subjective responses were collected after each transaction.

Please think first about the **quality of the result** you have just obtained, and mark one of the following to show your opinion.

- Fully satisfactory
- Satisfactory in the main, but left something to be desired.
- Unsatisfactory or misleading
- Irrelevant or positively wrong
- No result obtained at all

Now please think about your **satisfaction or dissatisfaction with the system** that you have just used to obtain this result.

How easy or difficult was this system to understand and use?
Allocate a figure of merit in the range 1 to 10 where 1 represents "Very difficult to understand or use" and 10 represents "Very easy to understand and use". [ ]

How pleasant or unpleasant was this system to use?

Allocate a figure of merit in the range 1 to 10 where 1 represents "Very unpleasant, slow or tedious to use" and 10 represents "Very pleasant and interesting to use" [ ].
Besides these subjective responses to each transaction, the following measurements were made:

Time taken per transaction.
Number of error-correcting dialogues entered.
Numbers of substitution or insertion errors.
Correctness or incorrectness of each result obtained.

Each task and result was presented separately to each subject who then had to respond on paper with each sheet collected after each call.

3.5 Analysis of results

The factorial design allows the results to be analysed in several ways. It also allows statistical tests to be undertaken to determine the significance of the components.

To illustrate the richness of the output data , some examples of these types of results are shown in Figures 3 and 4. These show the various objective (speed, error-rate and length) responses and the three subjective responses (Result, effort and Pleasant) for each case with the data partitioned between fixed and GSM results.

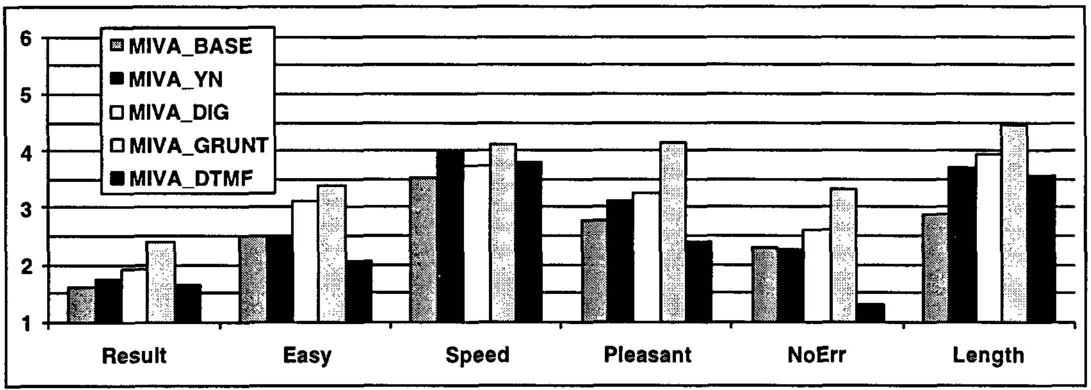

Figure 3 - Fixed phone responses

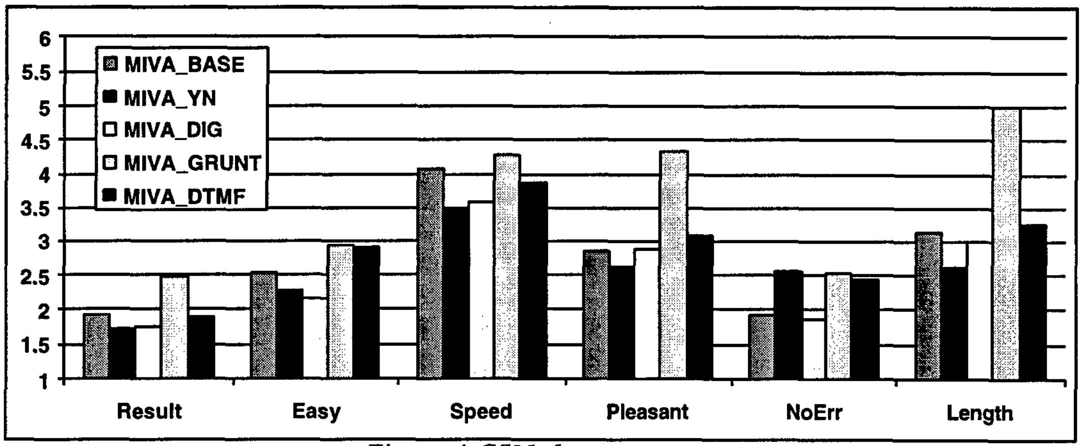

Figure 4 GSM phone responses

Direct observation shows that there are few substantial differences between the two sets of results and that the differences between, for example, dialogue types are much greater. Statistical tests were undertaken to establish any significant differences. The main conclusions from the analyses at this point were:

- All partners found that for menu driven systems of the type explored here, accessed over the fixed network, DTMF was easier to use than speech recognition.
- When the same services were accessed from mobile phones (mainly GSM) speech recognition was marginally preferred by all partners with the exception Deutsche Telekom where DTMF was still slightly preferred.
- 'Simple' recognition strategies e.g. using 'grunt' detection or numbered items are preferred less than proper word based recognition for all cases of fixed and mobile phone access.
- The absolute accuracy of recognition over mobile networks (as compared to the performance of the over the fixed network) was found to be slightly worse than that for fixed network in Deutsche Telekom and Italia Telecom. However no significant difference was found in the BT and Portugal Telecom systems.

4.MULTILINGUAL EXPERIMENTS

4.1 Introduction

The mono-lingual experiment had addressed the bulk of the technical questions but had not addressed any of the multi-lingual aspects. It had also tested out the individual platforms and allowed us to define the country specific dialogues.

The multi-lingual experiments built upon these results and answer the remaining questions by exploring

- Effects of cut-through
- What is the best way to prompt users to say a language
- Comparison of national versus International access
- Language/cultural differences.
- Predictive modelling of objective/subjective results.

As each partner had by now implemented and optimised their own local service in their own language the basic structures of the services were now well established. The next stage was to migrate to the totally multi-lingual environment. This meant that every suite of dialogue prompts in each language had to be translated, sent to the 'mother tongue' country for recording and then sent back. At the same time a data collection exercise to collect the necessary words to train the speech recognisers in every language had to be undertaken. There were, on average, about 70 words to be recorded per dialogue. Appropriate vocabularies had to be collected from a representative set of over 800 people in each country, labelled and stored in a form that would allow the specific service to be implemented in every location

Although the logistics of this seemed quite complex, in reality the use of digitised speech recordings stored as files on CD-ROMS proved so reliable that the recording and distribution processes ran extremely smoothly. To check quality at various stages the recognition components were tested in the laboratory with a common test set -country experiment.

With the five platforms each supporting the five language variants the experimental phase could start.

4.2 Multi-lingual Experimental design

Balanced experimental designs similar to those for the mono-lingual experiments were used as the basis for collecting objective performance data and subjective preferences.

This time the variables of interest were factored by the 5 languages and 2 dialogue types to give a total of 10 treatments. The dialogue types were either "Short and concise" or "Descriptive/verbose" and were further subdivided into technologies (by countries) to the extent that two classes of service could be identified. The first had essential information first (e.g. country direct number), followed by details on how to use it. This structure was implemented on the FT, BT and PT platforms.

The second had descriptive information first, followed by information content successively. This structure was typically implemented on the IT and DT platforms.

Also two countries (France and Italy) deployed 'cut-through' in their systems.

The total number of subjects recruited for the test was 100, a panel of 20 subjects per partner. At end of the experiment a total number of 900 calls had been collected and 886 questionnaires completed. Extensive objective data had also been collected.

4.3 Multilingual Analysis

As before, the results could be factored and analysed using analysis of variance techniques. For example comparisons could be drawn between the dialogue types, the use (or not) of cut through and any effects of language.

It was also possible to use the data to see if there were any 'learning' effects taking place. Evidence that there was is illustrated by the following analysis. Learning effects were tested by means of the Chi-square test between opinions for consecutive calls. Subjective measures were then compared once we had identified the number of calls a subject needed to place before he could be considered an expert. Figure 5 which is extracted from [3] illustrates the effect.

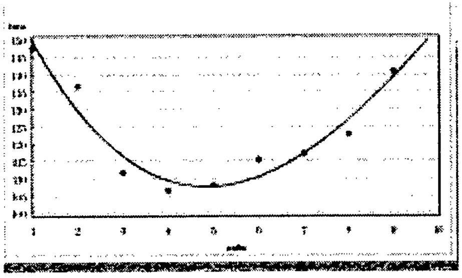

Figure 5. Transaction time as a function of order of call

4.4 Regression analysis.

Regression analysis is a technique used to determine the relationships between variables.

In this case we wanted to see if we could predict subjective responses from the various objectively measured parameters. From an initial examination of the data suggested that the subjective variables with the highest association with satisfaction were transaction time, number of utterances and number of correction turns. Using half of available data a series of regression analyses were performed to find an analytical model that best fitted the data for each parameter.

For the linear case functions and parameters found were:

Ease of Use = 4.440576 - 0.004778* transaction time
Learnability = 3.848315 - 0.080385*numbers of utterances
Pleasantness = 3.881013 - 0.003505*time
Effort = 3.914733 - 0.288068*corrections turn
Correctness = 0.576166 + 0.035409*word recognition rate;
Duration = 3.726034 - 0.005278* transaction time

4.5 Validation of the model

To validate the above the other half of the data was used. Below shows the results when the complementary data was applied to the ease of use data.

	N	Mean time	Observed	Prediction by linear regression
BT	111	96.51	3.89	3.98
DT	104	172.5	3.54	3.62
FT	116	68.0	4.47	4.12
IT	54	177.4	3.61	3.59
PT	101	116.6	3.71	3.88

The differences between observed and predicted values were tested by means of t-test. In both cases the test was not significant: the predicted mean value did not differ from the observed mean demonstrating external validity.

4.6 Observations

The regression methodology provided some further intriguing data concerning different language/nationality behaviours. For example English users strongly associated *satisfaction* with 'transaction time'. Germans on the other hand associated *satisfaction* and *correct recognition feeling* with 'transaction success'. Italian subjects associated *satisfaction, correct recognition* and *effort* ratings with 'correction turns' and 'word recognition'. For the Portuguese subjects *learnability* was the only parameter associated with the objective 'transaction success'. One conclusion may be that the questions were interpreted significantly differently by different groups.

5. CONCLUSIONS

We have described how standard methods of multi-factor experimental design have been adapted to evaluate the relative importance of various aspects of Interactive Voice response systems. The way in which subjective and objective data may be correlated and used as the basis of a predictive model has also been demonstrated. The model itself was then verified using a split data approach.

Although the project covered a great deal, there were some limitations. For example all experimentation was undertaken in a laboratory – as opposed to a market – environment. One exception to this was France Telecom who went one step further and undertook a public trial in the Musee de Lannion.

6. ACKNOWLEDGEMENTS

The authors gratefully acknowledge all the project participants that designed and implemented the MIVA system and collaborated for the realisation of the experiment: Sheyla Militello, Joaquin Azevedo, Nuno Beires, Francis Charpentier, Mark Farrell, Eric Le Flour, Giorgio Micca, Karsten Schroede. We are especially indebted to Juan Siles (EURESCOM supervisor) for his valuable contribution to reviewing all the reports produced during the project.

7. DISCLAIMER

This document may not reflect the technical position of all the EURESCOM Shareholders; its contents and specifications may be subject to further changes without prior notification. This document contains material, which is the copyright of some EURESCOM Project Participants and may not be reproduced or copied without permission. The commercial use of any information contained in this document may require a license from the proprietor of that information.

8. REFERENCES

[1] R.E. Kirk: "Experimental design procedures for the behavioural science", Monterey, CA: Brooks-Cole Publishing.1982.

[2] Richards, D.L. "Telecommunications by speech" Butterworths 1973,.

[3] Militello S, Johnston R.D. "A Methodological study for the evaluation of a multilingual IVR user interface", Human Factors in Telecommunications, 17th Symposium, 1999.