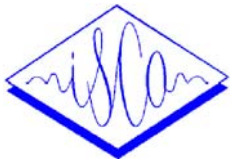

EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

PRELIMINARY RESULTS OF A MULTILINGUAL INTERACTIVE VOICE ACTIVATED TELEPHONE SERVICE FOR PEOPLE-ON-THE-MOVE

Fulvio Leonardi Giorgio Micca Sheyla Militello Mario Nigra
CSELT - Centro Studi e Laboratori Telecomunicazioni
Via G.Reiss Romoli, 274
I-10148 Torino (Italia)
(leonardi,micca,militello,nigra)@cselt.it

ABSTRACT

The EURESCOM P502 project, Multilingual Interactive Voice Activated (MIVA) telephone services, launched in 1995 for a three-years term, aimed at designing and experimenting on an automatic multilingual telephone assistant for people-on-the-move, that provided them with instructions about the use of most important telephone services in the country they are traveling. The core information provided by the system is: emergency services, international and national calls, card calls. Six European Telecom research laboratories were involved in the project: CNET, the project leader; British Telecom, Deutsch Telekom, KPN, Portugal Telecom and CSELT. The final prototype has to include a language selection module and a menu-driven procedure, using a common structure of the information contents in all languages. Several factors are currently being investigated, such as the impact of a talk-through capability, the effect of the cellular network as well as the usage of different national networks on the ASR performance, and the optimization of the dialogue strategy at the system interface level. The prototypes are in the process of being tested within the individual national research units, and cross-country tests will follow. As a further benefit, the potential savings which can be obtained by sharing the costs of development of ASR-based multilingual telephone services, will be estimated successively. A final field trial of the national implementations of the systems has to be carried out starting October '97 for a thorough evaluation of the multilingual services.

1. INTRODUCTION

When designing and implementing automatic, voice-activated telephone applications, one is often faced with the choice of the speech technology and human-machine interface to be used. There is no doubt that successful applications of speech technology need a careful dialogue design. It is still a matter of study how these different technologies may be balanced for applications devoted to inexperienced subjects. In order to study the impact of different technical features on the user behavior [1] we have performed a comparative experiment by using three different information inquiry systems applied to the same domain. In fact, in spite of the availability of the cut through technology, there is a lack of empirical studies about the effectiveness of this kind of functionality.

The service variants which were experimented were based on different technologies: the first one, called Baseline version (BSL-IC) did not employ cut-through and it was based, as well as the cut-through version (CT-IC), on a phonetic recognition model, while the third version (CT-EC) worked with whole-word models of speech. CT-EC used explicit confirmation at each step of the dialogue. All the systems have been tested by inexperienced subjects. In this paper we present and discuss the objective and subjective data collected in experimenting the system in Italian, trying to answer to the following questions:

1. Is there any relationship between the type of service experimented and the subjective measurements ?
2. Is there any difference between order of trials (learning effect) ?
3. Are transaction times and number of turns affected by the type of service ?
4. Which is the relationship between time/performance and subjective evaluations?
5. The type of problem that is proposed to the subject, does it affect the subjective evaluations?

In section 2 we give an overall view of the system, while section 3 briefly describes the recognition component. Section 4 illustrates the measures and evaluations of the service. Section 5 presents the data obtained from the experimentation, followed by a discussion in Section 6. Finally some conclusions are drawn in Section 7.

2. THE MIVA SYSTEM

The Italian inquiry service MIVA, developed within the framework of P502 "EURESCOM" project, communicates with the caller using the isolated word recogniser AURIS®.
The Dialogue Server (DS) is responsible for controlling the entire information system. This engine runs on an IBM-compatible personal computer under MS-DOS. The system is multi-channel. The controller was developed using OMNIA (Object oriented MaNager for Interactive Applications), a telephone service creation environment developed by CSELT. The engine has four major functionalities including telephony, recognition, message prompting and dialogue control. MIVA is designed according to a system-driven dialogue strategy.

3. RECOGNITION COMPONENT

3.1. Recognition technology

The speech recognition component used in the laboratory experiments was based on word or phonetic level, gaussian mixture CDHMMs, trained by means of a variable resolution K-means algorithm [2]. Models with both fixed (26 states) and variable (6 X n. of phonemes) number of states with skip transitions were tested in the experiments, with a maximum number of gaussians per state ranging from 1 to 4. The training speech database consisted of a minimum of 400 repetitions per each of the 74 application words, tests were carried out both with PSTN and GSM data. Rejection models were based on a

general 26-state noise model plus an average 40-state model previously obtained from a different phonetically balanced speech database. Two head & tail models were optionally included at the beginning and at the end of each word to wipe out breaths and other extralinguistic phenomena. To this purpose we used either a short 5 state model or a longer 26-state model corresponding to the noise model itself. The recogniser used an N-best Viterbi decoding algorithm with beam search for improving the search. The Front-End processing used 12 cepstral and 12 delta-cepstral parameters computed along a MEL frequency scale in the telephone band. A 0.95 pre-emphasis factor was adopted with a 8 Khz sampling frequency. MEL frequency grouping was carried out on FFT 256 sample energies with a 10 ms shift and a Hamming window of 256 samples. Overall frame energy and delta-energy parameters were added to complete the feature vector. High pass RASTA filtering is finally applied to remove low frequency channel variations.

3.2 Laboratory vocabulary tests

Recognition results obtained by testing the recogniser on a vocabulary test consisting of the 26 country names included in the application are reported in Tab. 1. The test corpus consisted of 8174 utterances from about 600 speakers for PSTN data, and of 4600 utterances from about 300 speakers for GSM data. Utterances were automatically end-pointed by means of an adaptive threshold, energy based End Point Detector.

	PSTN			GSM		
	WR	**SUB**	**DEL**	**WR**	**SUB**	**DEL**
26-states h/t 26 sts.	98.59	1.10	0.30	98.24	1.00	0.76
6 X-states h/t 26 sts.	98.56	0.99	0.45	97.63	1.21	1.15
26 states h/t 5 states	97.09	1.88	1.02	97.00	1.43	1.56
6 X-states h/t 5 states	96.88	1.76	1.35	96.17	1.30	2.52

Tab. 1 Recognition results for different recognition models

4. DIALOGUE SYSTEM EVALUATION

4.1 Overview of the service variants

Three versions of the application were built: the service and the kind of information provided are described in Fig. 1.
The baseline variant was commonly adopted by all the partners of the project in order to share a unique test bed for the different implementations.
All the three variants are briefly described in the following:

BSL-IC: No cut through capability, implicit confirmation, access to country name from a list, contextual help, phonetic recognition models.

CT-IC: Cut through capability, implicit confirmation, access to country name by direct pronunciation, contextual help (explicitly requested or automatically provided in case of error or time-out), phonetic based models.

CT-EC Cut through capability, explicit confirmation, access to country names by direct pronunciation, contextual help (implicit in case of error or time-out of the response), generic help with a flat list of topics (under explicit request), whole-word recognition models.

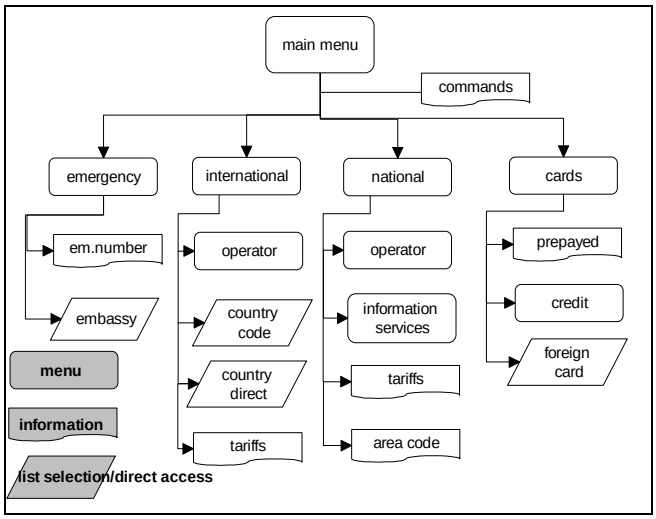

Fig 1: flow chart of MIVA service

4.2 Experimental planning

The independent experimental variables were: Service Variant (3), Type of Phone (2: PSTN/cellular) and Type of Query (2: Answerable and Non-Answerable). The experiments were designed according to the Split-Plot Factorial Design, denoted by SPF-pqr [3]. Six different types of questions were selected in order to cover all the branches of the service. A panel of 24 X 3 subjects was used in the tests. Subjects were requested to place four calls to the system, half on the PSTN, half on the cellular network. The experimental conditions were presented in reverse balanced order, so that the two groups of subjects were split into two halves of 12: the second half had the same questions of the first half, but in reverse order. Both subjective and objective measurements were collected. Subjective evaluations, rated upon a scale of five points, were the following: 1) Ease of Use of the service (EASY), 2) Learnability (UNDER), 3) Pleasantness (PLEAS), 4) perceived Fluency in the Dialogue (SPEED), 5) Effort in the Dialogue (EFFORT), 6) perceived Correct Response of the System (NOERROR), 7) Duration of the Call (DURATION).

An Overall Satisfaction parameter, rated upon a six points scale was also introduced representing the global impression of the user about the service. The objective measurements computed on the panel of 72 subjects were [4]: Transaction Success (TS), Word Accuracy (WA), Correct Deletion (CD), False Deletion (FD), Insertion (INS) and Substitution (SUB); these measures have been adapted to the evaluation of menu-driven applications. Besides, the Total Transaction Time, the average one Turn Time and the average number of Turns for each dialogue were computed.

5. RESULTS

Both objective and subjective measures were processed by means of the Analysis of Variance test and χ^2 test, to evaluate the influence of the above mentioned independent variables on the measures which were defined as evaluation parameters.

5.1. Objective results

The χ^2 test outlined a difference in percentage of TS and Type of Query in CT-IC (χ^2=7.86, p=0.005), this relationship was not found in the other two services BSL-IC (χ^2=2.10,p=0.147 n.s.) and CT-EC (χ^2=0.57, p=0.45 n.s.). The BSL-IC CT-EC variants better led subjects to understand that the information was not available in the service.

The Analysis of Variance performed on objective recognition performance highlighted a significant effect on WA [F=7.45, p=0.001] due to Type of Variant [F=22.30, p=0.001] and Type of Query [F=26.95, p=0.001]: the WA measured in CT-EC (92.03) is significantly greater than the WA in BSL-IC (82.31 %) and in CT-IC (75,10%); the percentage of rejections was higher for cellular phones (2,32 %) than for PSTN phones (0,99%) especially in BSL-IC. We found an effect due to Type of Variant [F2=14,76, p=0.001], where the CT-EC greatly reduced the insertion rate due to the better tuning of the recogniser (whole-word models instead of phone models, rejection models compatible with whole-word models).

Total Transaction Time was affected by the service variant [F2=5.19,p=0.0061]: BSL-IC (198,79") and CT-EC (192,65") were longer than CT-IC (172,23'), so the explicit confirmation reduced the major advantage of the cut-through capability. Also the Type of Query condition affected time [F2=18,31,p=0.0001], with Non-Answerable queries (217,61") significantly longer than answerable ones (173,03"), especially in CT-EC.

	BSL-IC	CT-IC	CT-EC
TS	80.21%	60.42%	82.29%
Total Time	192.65"	172.23"	198.79"
Avg. n. of Turns	18.13	13.31	20.16
Avg. Turn Time	12.95"	11.59"	11.43"

Table 2: Time and TS Performance

	BSL-IC (utt.=718)	CT-IC (utt.=739)	CT-EC (utt.=1091)
WA	82.31%	75.10%	92.03%
CD	8.91%	7.71%	2.80 %
FD	4.04%	0.81%	1.38%
INS	3.06%	12.18%	3.43%
SUB	1.67%	4.19%	0.33%

Table 3: Recognition performance

5.2 Subjective evaluations

The Analysis of variance performed on the Ease of Use parameter revealed that there was a difference among the Service Variants [F=6.34, p=.002]. The perceived Correct Response of the System was not significantly affected by any condition [F=1.57, p=107, n.s.]. On the Pleasantness parameter a significant difference was found for the Service Variant condition: BSL-IC (3.73), CT-EC (3.51) and CT-IC (3.06) [F=7.52,p=0.0007]; CT-IC score, being significantly lower than the scores of the other two variants, suggested that the adoption of the cut-through functionality without request of explicit confirmation, combined with phonetic recognition models, determined an impoverishment of the feeling of control of the system by the subjects. The perceived Fluency in the Dialogue was significantly affected by the Service Variant variable [F=3.35, p=0.0002], where CT-EC was perceived as the slowest one [F=10.68, p=0.0001] with a mean score of 3.29. The perceived Effort in the Dialogue was also affected by the experimental conditions, with a main effect due to the dialogue variant variable [F=7.52, p=0.0007]. There was also a main effect due to Type of Query close to the significance threshold [F=3.39, p=0.066 n.s.], in fact Non-Answerable queries were perceived as requiring more effort. The perceived Correct Response of the System seemed to be the most diagnostic measure, with a highly significant effect due to Service Variant. [F=16,27, p=0.0001]; for this parameter there was also a noticeable effect due to the Type of Query, [F=18,10, p=0.0001] in fact for Non-Answerable queries the subjective assessment was 3.5 against a mean value of 4.20. The Duration of Call was perceived as not really adequate with a mean score of 2.95 and a significant effect due to the type of service [F=4,96, p=0.007] (CT-EC: 2.61; BSL-IC: 3.13); since the objective duration of CT-EC and BSL-IC were almost the same we can conclude that subjects were bored with the number of confirmations of the CT-EC version. No significant effect was found due to Type of Telephone in any of the experimental conditions. Fig. 2 shows the subjective profile of the three variants: it can be noticed that CT-IC and CT-EC follow the same trend except for the perceived Duration of Call. The cut-through functionality was used by subjects because the Average Turn time was lower for CT-IC and CT-EC variants (Tab. 2); anyway, as it can be elicited from Fig. 3, the Total Transaction Time was not affected by the order of call, meaning that no relevant learning effect was measured for this objective parameter.

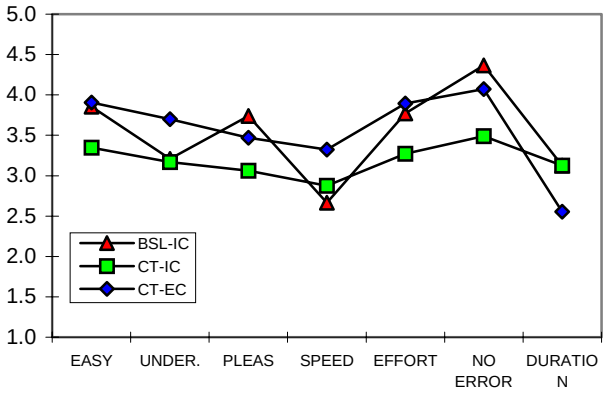

Fig 2: Profile of the three services

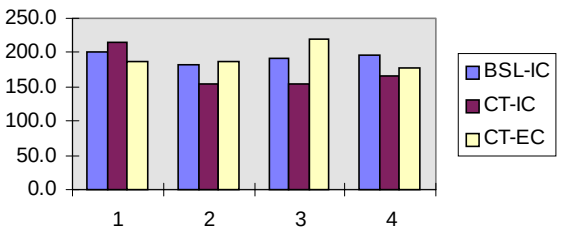

Fig 3: Total Transaction Time for each service as a function of the order of call (1-4).

6. DISCUSSION

The analysis of both the subjective evaluations and of the objective measurements has confirmed the following points:

a) The cut-through capability reduced the average One-Turn Transaction Time, but increased the Difficulty in using the service in case of no explicit confirmation required to the user; this result led to the conclusion that cut-through capability caused a worse recognition accuracy when phonetic models were used, while this capability could be successfully implemented with whole-word models.
b) The larger number of insertions due to the cut-through significantly reduced Pleasantness and increased Effort in Dialogue. The number of insertions decreased by including whole-word models instead of phonetic models and consequently the subjective evaluation of Pleasantness, Effort in Dialogue and Learnability increased.
c) The perceived Fluency in the Dialogue was affected by the presence of request of explicit confirmation by the system.
d) The perceived Duration of the Call was affected by the presence of request of explicit confirmation.
e) Subjects had a stronger perception of Correct Response by the system with the CT-EC variant than without explicit confirmation.
f) The Transaction Success rate was significantly higher for BSL-IC (nearly 80%) than for CT-IC (nearly 60%); this effect was mainly due to a non-optimal design of the error recovery procedure, rather than to technological limitations of the cut-through implementation. In fact the Transaction Success rate improved to 82% by introducing the explicit confirmation as in the CT-EC variant.
g) No significant difference in performance was detected between PSTN and cellular modes.
h) Perception of Correct Response of the System was more difficult in case of Non-Answerable queries. This effect was mainly due to the direct access procedure to name lists implemented in CT-IC; in this procedure, the pronunciation of a country name which was not in the vocabulary could not be perceived by subjects as a lack of information, even if it was correctly rejected by the system. In CT-EC, this effect was reduced by introducing the explicit confirmation and an explicit message: this message was activated in the direct access procedure after two consecutive rejections and suggested the possibility that the required country might not be present in the vocabulary of the recogniser.
i) A remarkable difference in the Overall Satisfaction parameter was apparent among the evaluations of the three service variants: 81% of subjects rated the BSL-IC variant in the positive half of the scale, 86% of the subjects gave a positive evaluation of the CT-EC variant, while only 66.7% of the subjects considered the CT-IC variant satisfactory.
j) There was no request by any subject for help by the flat list in the test of the CT-EC variant: hence no conclusion could be drawn about the advantage of introducing this option in the system.

On the basis of this major results from subjective and objective measurements, variant CT-EC was chosen as the best service structure. However, it was suggested to reduce the number of requests of explicit confirmations, because the presence of these turns in the dialogue showed to significantly affect both Total Transaction Time and perceived Duration of Call and Fluency in the Dialogue. The criterion for dropping some of the explicit confirmation turns should depend on the size of the active vocabulary; in fact with smaller vocabularies, after two or three interactions, subjects are annoyed by repeated requests of confirmation. On the other hand explicit confirmation played an important role in dealing with Non-Answerable questions, in particular when a country-name was not in the database. Therefore the explicit confirmation was left in the system in the mode of direct access to a list.

7. CONCLUSIONS

Cut-through capability resulted to be an important feature, although in this kind of application it did not induce a shorter duration of calls; use of the cut-through capability improved subjective evaluations of Pleasantness and Learnability and reduced the Effort in Dialogue especially for the lack of system's correction requiring the user to "speak after the beep". Even if rejection rate was a bit higher with the cellular phones, subjects did not perceive it and there was no difference at all in their evaluations. The application was quite complex, due to the large amount of information provided; therefore, subjects showed no learning effect, either in Transaction Success rate or in Duration of the Call. Some guidelines can nevertheless be specified, for any typical session of an application of this type: the dialogue should not last more than 180 seconds, it should not contain more than 20 utterances and no more than 4-5 errors, and finally, recognition Word Accuracy should keep above 80%. These guidelines were deduced by a Regression Analysis, applied to study the relationship between objective and subjective experimental data, can be useful in predicting subjective evaluations from objective performance measurements. This approach will be further investigated in successive studies.

The authors are indebted to G.Castagneri for his contribution in the design of the experimental plan, to U. Allisiardi for technical assistance in the implementation of the system and to E. Foti for the recordings of the prompts.

.

REFERENCES

[1] R. Billi, G. Castagneri and M. Danieli “Field trial evaluation of two different information inquiry systems”, IVTT 96, pp. 129-134.

[2] L.Fissore, F. Ravera, P. Laface: “Acoustic-phonetic modeling for Flexible vocabulary speech recognition”. In *EUROSPEECH ’95*, p. 799-802, 1995

[3] R.E. Kirk: “Experimental design procedures for the behavioral science.”, Monterey, CA: Brooks-Cole Publishing..

[4] M.Danieli and E.Gerbino: “Metrics for evaluating dialog strategies in a spoken language system.”, Working Notes of the AAAI-95 Spring Symposium on Empirical Methods in Discourse Interpretation and Generation, Stanford, California, March 1995, pp.34-9.