

**LE1-1930**

**Verification and Validation of Calling Card Application**

**Deliverable D 7.2 - public deliverable**



**This is the public version of deliverable D 7.2 (confidential)**

**About this Document**

Results from CAVE field trial is reported in two documents:

- D7.1 for the Bank Application, and
- D7.2 (this document) for the Calling Card Application.

**CONTENT**

1.	Introduction .....	2
2	Overview.....	3
2.1	System .....	<b>Error! Bookr</b>
3	Mimic Test.....	5
3.1.	Subjects.....	5
3.2.	Mimic test .....	6
3.3.	Analysis of the user call annotations.....	6
3.4.	Analysis of the questionnaires (subjective analysis).....	6
4	Field test .....	7
4.1.	Improvements of the system .....	7
4.2.	Participants.....	8
4.3.	Running the Field test .....	9
4.5.	Analysis of the booklets .....	10
4.6.	Analysis of the questionnaires .....	10
5	Note on user experience of enrolment.....	11
5.1.	Attitudes on enrolment .....	11
6	Conclusions.....	12

This work has been supported by:

- the Telematics programme of the European Union
- the Office Federal de l'Education et de la Science (in Switzerland)



## 1 Introduction

The focus of the CAVE project was not to develop new applications based on the SV technology, but to find out customer reactions and attitudes towards this new technology. The task of workpackage 7 is verification and validation of the demonstrators developed by CAVE. WP7 is the last of the WP's in CAVE

The banking application was run in Zürich Switzerland. Users got fictive bank accounts and their voices were used to identify the owner and the account, as well as to do speaker verification. After users were verified properly, they could transfer money between their accounts. The language used in the banking application was Swiss German. The Swiss tests seem to show some disappointing results, but were very useful as input to the field-test in Holland.

This application used the calling card service as domain; the users were speaking Dutch and entered the calling card number with their voices. The calling card number was evaluated with speech recognition and the users were verified with speaker verification on the given number. This field-test showed much better results than the field-test in Switzerland, very much dependent of the improvements that was done after the experiences gained in the Swiss tests.

The goal for this workpackage and a main focus of CAVE was to test:

- CAVE demonstrators with end-users in field-trials
- how users react to speech recognition and speaker verification in telematic transaction services
- if users are confident with the perceived security of speaker verification
- if it is perceived to be more user-friendly to use voice instead of PIN-code?

The reactions of users are very influenced by the performance level of the technology and by the quality of the design of the human-machine interface of a service.

User reactions and user-classification in different categories (goats or sheep) depending on their ability to cooperate with the system was also by using information about objective system performance collected in the two field tests.

The two field test systems are described fully in deliverables of workpackages 3 and 6 of the CAVE project. In the applications built speech recognition from the CAVE partner Vocalis was integrated for German (in the banking application), or for Dutch (in the calling card application) and also CAVE speaker verification technology..



## 1.1 System Overview

A fully speech driven calling card application was built and tested.

The enrolment was done by a single call to the system, in which a person had to type (i.e. DTMF) his/her card number and PIN code to enter the enrolment dialogue. In the dialogue the person was asked to speak his 14 digit card number digit by digit and he was prompted for a number of random 4 digit sequences (e.g. "say 4 9 0 6").

This card number/PIN code sequence was repeated a number of times. Two minutes after the completion of the enrolment call the training of the models was finished and the service was ready to be used.

The telephone number could only be chosen from a list of 10 telephone numbers, which had been chosen by the client in advance. This way of voice dialling enhanced speech recognition.

An example dialogue of the service is:

System: Welcome to the voice driven calling card service.

    Please say your card number digit by digit.

User: "89311025010278"

System: One moment, please. (pause)

System: Your data are correct, please say the telephone number.

User: "0703324146"

System: The number is being dialled.

Dialogue 1. Example dialogue.

If something went wrong with the recognition, the user always got a second chance. After a second false recognition an operator fallback is mimicked ("In the real service you would be connected to an operator, in this test the connection is interrupted.").

The same holds for the verification, if this fails the first time, the customer gets a second chance:

System: Welcome to the voice driven calling card service. Please say your card number digit by digit.

User: "89311025010278"

System: One moment, please. (pause)

System: Something is wrong with the recognition of your voice. Please repeat your card number digit by digit.

User: "89311025010278"

System: One moment, please. (pause)

System: Your data are correct, please say the telephone number.

User: "0703324146"

System: The number is being dialled.



Dialogue 2. Dialogue in if the user is rejected at the first time.

More technical details on the system can be found in the deliverables of workpackage 6..

### **FR/FA versus system-FR/FA.**

This set up differs from the bank system, in which the verification is done only once. Doing so, the standard definitions of False Acceptance (FA) rate and False Rejection (FR) rate are no longer valid.

The dialogue example given above would have one FR and one Correct Acceptance (CA).

For the calling card system we therefore define the system False Acceptance as the case in which an impostor succeeds to use the service with the card number of someone else.

A system False Rejection only occurs, when the true speaker is rejected twice.

### **Tests**

The system was subjected to three major tests: Lab test, Mimic test, and the final Field test. Here, a short overview of these three tests is given. The Mimic and the Field test will be described in more detail in sections 2 and 3.

- **Lab test:**

A small lab test was performed to see if the first version of the system worked. This version still used a TCP/IP connection to the speech recognizer.

This was not acceptable from the users point of view. Also, it was used to enable a small number of expert users to evaluate the overall design of the application and the dialogue structure. The participants were members of an involved research group.

- **Mimic test:**

A larger test was done to evaluate a preliminary version of the system, the test set up and the questionnaires. In the Mimic system, the SCx bus was integrated, which enabled real-time speech recognition.

In parallel two versions of the system were evaluated, one with PIN code and one without (in contrast to the Swiss system where everyone had a PIN code). Participants were asked to use their own account at least 8 times and break in on other accounts (4 times). To encourage test persons, the service (and redirected calls of 10 minutes maximum) were offered for free.

During the test, a preliminary version of the verification program was used.



- Field test:

Some modifications were made to the test set up, the questionnaires and the dialogues according to findings in the Mimic test. Also, an updated version of the verification software was used, compatible with the results of WP 4.

The enrolment was extended to 8 repetitions of the card number and 7 random 4 digits sequences. Again the test persons were divided into “with-PIN” and “without-PIN” groups. Participants were all subscribers to the current calling card service. And also here the users were offered free calls and they were asked to do some impostor attempts.

As well, they were encouraged to ask friends and family to try to break into their number. To evaluate the system, they had to log all calls in a small booklet (date, time, what happened

The updated version of the system turned out to work very well, even when it was used from more ‘difficult’ environments (e.g. airports, phone booths).

Most of the test subjects were very satisfied with the system and its performances. Still further research turns out to be necessary, for example on cellular calls.

## 2 Mimic Test

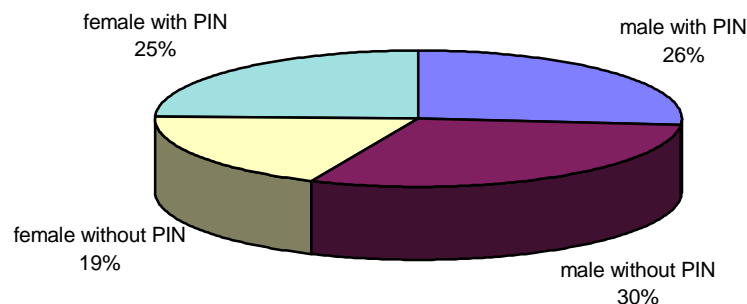
### 2.1 Subjects

In total 53 people (23 female and 30 male) participated in the test. 33 participants filled in an annotation table (to log all calls) and the questionnaire and returned them. The participants were equally distributed over the two versions of the system: one which asks for a PIN code and one which skips the PIN code.

This gives the possibility to measure the perceived security of both versions.

**male / female // with PIN / without PIN**

**N = 53**





## **2.2 Mimic test observation**

During the Mimic test, because of the heavy load of the system, several awkward bugs in the software were found. It turned out that for a system of this complexity, some failures could only be detected by running a test with a sufficiently high number of calls.

One of the major problems was that the enrolment (performed off-line after the enrolment session was finished) failed for some persons. Although during the enrolment call recorded speech was checked with the speech recogniser, the off-line enrolment function failed due to too short silences at the beginning or the end of the utterance.

### **Analysis of the calls (objective analysis)**

It is very useful to analyse all calls to see what happened. Most interesting is of course the number of false accept/reject decisions. For this the help of the user was needed. If someone is rejected twice the system asked if the user was an impostor. If the answer was 'yes' (implemented by saying '1'), then the system had taken a correct reject decision. If the answer was 'no' (implemented by saying '3'), then the system had taken a false reject decision. Approximately the same idea is used for the accept decisions.

Performance details might be obtained from consulting the CAVE consortium partners.

## **2.3 Analysis of the user call annotations**

Each user was asked to note briefly what happened during each call. 33 forms with these annotations were sent back. As a lot of subjects called much more than they were asked to, they did not fill in all the calls. Furthermore, they tended to mark especially the unsuccessful calls, which explains the high difference between the objective call analysis and the user call annotation analysis.

Results can be obtained through the consortium partners.

## **2.4 Analysis of the questionnaires (subjective analysis)**

The evaluation of the perceived quality of the service was based on the returned questionnaires of 33 subjects (54 % of the participants). Because of the heavy influence of the non optimal system performance and availability, only a number of relevant results are mentioned here.



It was clear to most users that the system has to learn their voice before they could use it. The enrolment procedure itself was judged as rather short and easy. In this version of the service, one call with 3 repetitions of the card number and 3 random 4 digit sequences was necessary to complete the enrolment. This took about 2 minutes.

The ease of use and the call set up time were judged equivalent to the existing calling card service. People agreed that in general the service was easy to use and made a friendly impression. Furthermore, the instructions and prompts were assessed as neither too long nor too short.

### **3 Conclusions**

The Mimic test proved that the test set up was applicable for the field test in its current form. Thanks to the heavy load, several bugs in the system could be found and corrected. Furthermore, the prompts, questionnaires and instructions were slightly changed according to the remarks of the users. Concerning the technology of speaker verification, the Mimic test showed that it is necessary to have a reliable algorithm, otherwise people will immediately lose confidence in the security as soon as something goes wrong (False Reject or False Accept).

### **4 Field test**

#### **4.1 Improvements of the system**

According to the findings in the Mimic test and the Swiss field test a number of improvements could be made to the system:

- A priori threshold setting is inevitable for a real-life service. Although some work has been done on speaker dependent threshold, it was decided to use speaker independent thresholds. This was supported by an off-line test, in which a number of speaker dependent threshold setting algorithms were compared. It appeared that with a speaker independent threshold only a few speakers required a more optimal threshold.
- The PIN codes were chosen according to a grammar that avoided digits that were more likely to be confused by the speech recogniser. This would enhance the speech recognition on PIN codes.
- For the telephone numbers sex dependent models were used in parallel, instead of sex independent models. This should increase the speech recognition performance on the telephone numbers.
- To verify if a rejection decision was correct a yes/no question was asked instead of '1' for an impostor '3' for a false reject.



- The latest version of the SS1 verification software from VOCALIS was used.
- RING delivered the newest version of their software, in which some bugs concerning the 'line busy' detection for outgoing calls were corrected.
- The enrolment prompt now mentioned that the random strings consisted of 4 digits.
- The enrolment was extended to 8 utterances of the card number and 7 random 4 digit repetitions. This took about 4 minutes.
- The time stamp to measure the average call set up time corresponds to the time of the first ring in the outgoing call, instead of the moment when the B subscriber answered the call (which was the case in the Mimic test). This gave a more accurate average call set up time.
- All participants received a small card on which the card number, enrol and access telephone numbers as well as the personal 10 telephone numbers were printed.

## 4.2 Participants

Most of the subjects were employee-users of the existing calling card service. In addition participants of the Mimic test and some outside (real world) participants were recruited. Letters were sent out to 300 persons including a form for participation. The response rate was about  $90/300 = 30\%$ . The potential participants were asked to indicate the 10 telephone numbers they wanted to use during the test. All received the small card, a booklet to log all calls, a sheet with instructions, and a questionnaire.

Deelnemer:	Man/vrouw	Kaartnummer:	Pincode:	Yellow Card:(totaal, h			
				#Eigen	#And	#And	#And
A.v.d.Molen	man	89 3110 2506 4051	0	8	8	8	2
M.Kneppers	man	89311025063095	0	6	5	5	1
A. van Heering	man	89311025061040	0	8	7	7	4
Nils Vergeer	man	89311025097481	1				
P.B. Huygen	man	89311025089017	0	7	7	7	3
Erwin Drenth	man	89311025083481	1	8	8	8	4
v. Drimmelen	man	89311025103800	0	8	8	7	4
A.R. Faasse	man	89311025065751	0	8	8	8	4
L.Palm	man	89311025082285	1	1	1	1	
J. vd. Dulk	man	89311025066387	0	8	6	6	4
Ad Roovers	man		0	8	7	6	3
Bert v/d Horst	man		1	8	8	6	4
Hans Lingema	man		1	8	8	8	3
C Brouwer	man		1	8	5	3	4



### **4.3 Running the Field test**

#### **4.3.2 Observations**

No major problems occurred during the field test.

Some 'bad blocks' on the hard disk caused the system to crash once.

#### **Analysis of the calls**

75 (84%) of the subscribed users actually enrolled to the service and used it one or more times (19 times on average, 1413 normal calls).

Almost all enrolled account numbers have been subject to impostor attempts. (3.72 for each account on average, 380 in total).

For the FR/CR annotation we rely on the co-operation of the user.

Impostors calling a valid number from the list of telephone numbers were annotated as CA's (which is in a way logical, given the definition of impostor: "some who uses the service on some else's account without his/her permission").

Because of the possible big influence of the 85 not annotated rejections, it was decided to check the audio files manually. This resulted in 33 extra FR cases.

From the booklets (returned by 50 (67%) of the enrolled participants) it was found that many FR were caused by 'fooling the system'. People changed their voices on purpose, or used a cellular phone. (Although the system seems to work rather good with cellular phones).

23 FR's came from the same person, but were not observed until the analysis of the not annotated rejections. Otherwise it would have been possible to manually adapt this person's threshold.

At the utterance level we can make a comparison with the results from the laboratory. It is very clear that the high number of false rejections is a point of attention.. The second chance enables the system to decrease the relative number of rejected true speakers, whereas the relative number of accepted impostors is increased. Further research for a priori set thresholds and adaptive thresholds can improve on both false rejections and false acceptances.

#### ***Twins, Sisters, Brothers.....***

were the largest cause of the FA. 5 FA's were caused by a pair of twins of which both participated. The others were a sister, resp. brother who were able to enter the service. One participant mentioned that he wouldn't care if it was only his brother who was able to use the service.



#### 4.4 Analysis of the booklets

Each user was asked to note briefly what happened during each call. As some subjects called much more often than they were asked to, they did not fill in all the calls. As with the Mimic test they tended to mark especially the unsuccessful calls, which explains the difference between the call and the booklet analysis. Moreover the booklets covered most participants who had ran into FA's.

#### 4.5 Analysis of the questionnaires

50 participants returned the questionnaire (67%). A number of graphs can be found on the next pages. Following are the most interesting points of the subjective evaluation:

- enrolment:  
The enrolment turns out to be easy (70%), and not too long (66%). One subject suggested to mention in advance the number of times someone will be prompted for the card number.
- User-friendliness:  
the system is evaluated to be easy to learn (95%) and friendly to use (80%). Most of the subjects would like to have operator fallback after two failures of the system.
- The system performance is rated rather good:  
The recognition of the various numbers is rated 'rather good' by 80% of the participants (on average: card number 4,0; PIN code 4,2; Telephone numbers 4,5). The Likert scale ranges from 'very bad (1) to very good (5)', so people might fill in 4, even if it works rather good (one small failure and 'very good' is no longer applicable).
- The Speaker Verification  
This holds as well for the perceived quality of the Speaker Verification: The True Speaker (TS) acceptance is rated 4,0 ('the system recognises my voice rather good') and the False Acceptance (FA) is rated 3,9 ('it's difficult to break in'): 75% of the respondents rate 4 or 5 for the TS acceptance and 75% do this for the FA. This is in contrast with the objective performance of the system; it seems that people still have the feeling that it is not really secure, although they never ran into mistakes. Another surprising observation is made. The participants that tested the version of the system with the additional PIN code gave the same rate for the safety of the system as did the users of the non-PIN version. Apparently, the PIN code, generally accepted as a good security tool, was completely overruled by the feelings for the safety of the verification technique.
- The limitation to 10 numbers ...is good enough for a test, but most of the test persons point out that it wouldn't be enough for a real application. Although



most people tend to use the service only for their family (mostly during holidays), they either want to be able to change the telephone numbers or type them in by DTMF..

- Acceptance:  
80% would use voice controlled calling card services (if it works well enough) instead of the current service.

Next to the questions a lot of (useful) improvements are suggested:

- shorter card numbers is desired by almost everyone.
- name dialling
- possibility to change personal 10 number list
- for extra security: type in the PIN code instead of saying (impossible for non-DTMF service)
- DTMF telephone number (impossible for non-DTMF service).

## 5 Note on user experience of enrolment

Services using Speakers Verification differs from most other services because they require *enrolment*, and it is therefore of high interest to understand how users will react to the procedures that are needed for that, and the different CAVE tests do provide some information about this.

### 5.1 Attitudes on enrolment

The users were quite understanding regarding the need of enrolment, and had generally more positive attitudes than could be expected (as can be seen from the table below). Enrolment is regarded as simple, and the procedures are not too long. However it must be remembered that all users in these cases were recruited within the own organisation, and may therefore have been positive biased. But in other cases, people from within have been more critical than the average user.

<i>Test</i>	<i>Numb. of calls</i>	<i>Enrolment procedure</i>	<i>Numb. of digits</i>	<i>Time to do</i>	<i>User attitudes</i>
Bank	2	9 digit customer code + 4 digit PIN + 6 times sequences of 4 digits	(2 * 37) 74 digits	?	Die Stimmregistrierung war einfach (94%) Zwei sessionen sind zufiel (16%)
Card: Mimic test	1	3 times 14 digit card number + 3 times sequences of 4	54 digits	2 min	"Rather short and easy" Enrolment is easy (94%) Enrolment is not to long



		digits			(87%)
Card: Field trial	1	8 times 14 digit card number + 7 times sequences of 4 digits	140 digit	4 min	Enrolment is easy (70%) Enrolment is not too long (66%)

Table: Enrolment in the different tests

The result should then not be taken as a definite judgement, but more as an indication of that the users are initially positive towards this technology, but they may very well change opinion later depending on circumstances or when more information is available.

## 5.2 Conclusions

A selection

- The Dutch field test shows that the technique is very promising. Both objective and subjective results show that the technique can be used in operational services.
- Nevertheless, further research is needed on better a priori threshold setting methods.
- Participants rather evaluate the verification technique instead of the service in which the technique is used.
- This is illustrated by the way the with-PIN users evaluate the safety of the system. They seem to judge the system only by the verification thereby overruling the general accepted safety of a PIN code.
- Real users in a real-world operational service have to be used to reveal real-life behaviour which could not be tested in the field test environment.
- Another point for future research is the verification for cellular phone calls, which caused extra FR's in this field test.