FIELD TEST OF A CALLING CARD SERVICE BASED ON SPEAKER VERIFICATION AND AUTOMATIC SPEECH RECOGNITION

Els den Os[1], *Lou Boves*[1,2], *David James*[3], *Richard Winski*[4], *Kurt Fridh*[5]

[1]KPN Research, [2]KUN , [3] Ubilab, [4]Vocalis, [5]Telia
P.O. Box 421, 2260 AL Leidschendam, the Netherlands
E.A.denOs@research.kpn.com

ABSTRACT

In this research we have studied several human factors problems that are connected to the deployment of speaker verification technology in telecommunication services. We investigate the perception of the safety of a calling card service when it is protected by speaker verification on the 14 digit card number, and compare it to the perceived safety of speaker verification and PIN. Moreover, we compare a voice based interface to the service with a DTMF based interface. The results are crucial for guiding the introduction and deployment of speaker verification technology in actual applications.

1. INTRODUCTION

Of all Speech Technology fields Speaker Recognition has received least attention over the past decade. Up to now hardly any research results have been published in the open literature on the performance and acceptance of speaker verification technology in telephone services for the general public. The LE project CAVE (Caller Verification in Telecommunication and Banking, LE1-1930) under the Telematics Application Programme of the EU was set up to help remedy this situation [1]. The CAVE project is focused on technology development and user requirements research related to Speaker Verification. Thus, the research is limited to applications where a caller claims an identity, and the task of the automatic system is to verify this claim.

The lack of solid knowledge on real world performance and especially on user acceptance of Speaker Verification (SV) technology is now recognised as one of the factors which may slow down the large scale deployment of Interactive Voice Response (IVR) and Electronic Commerce (EC) applications. IVR applications are developing from information services into transaction services. Many information services will become much more appealing if the caller can immediately act upon the information by making transactions. Moreover, many business sectors nowadays see a development towards the deployment of call centres, where most of their (service) operations are concentrated. One obvious example here is retail banking: most large banks are trying to replace expensive main street branch offices by less expensive call centres. Up to now, security in these services is based on four or five digit PIN numbers associated with card or account numbers. If the service is completely automatic, customers are usually requested to enter the complete PIN via the DTMF keypad. If an operator is involved some companies prompt for a random selection of digits from the PIN (so as to prevent the operator from finding out combinations of account numbers and PINs). Other companies, like Dutch PTT Telecom in its *scope* service (the brand name for the calling card of PTT Telecom) have operators prompt for the full four digit PIN.

Calling Card services are one of the rare examples where an SV system is known to be operational (viz. the US Sprint Foncard^{TM} service). Yet, it appears to be extremely difficult to obtain dependable information on the actual use and performance of this SV protected service. This was one reason why we decided to set up a series of experiments to investigate the technological and human factors issues involved in such an application.

Fraud in telematics services is undoubtedly a major cost factor, both for the service provider and the customers. In principle, fraud can be reduced to almost zero, by introducing ever more security measures. However, such measures incur very high costs themselves, both in terms of equipment and/or operators, and in terms of human factors problems, because security measures complicate the procedures for accessing these services. For both reasons developers of Calling Card, Voice Mail and Home Banking services show their interest in speech recognition and speaker verification: not only to reduce fraud, but also to improve the user friendliness of the service interfaces.

In our paper we present the details of the design of an experiment with SV in the *scope* service. We compare the experimental design with the approach taken in a companion experiment, also in the framework of the CAVE project, that is under way in Ubilab, the research lab of Union Bank of Switzerland. Comparison of the two experiments, in different settings, in different countries and with different services should allow a better understanding of the fundamental issues involved in the deployment of SV technology.

2. SPECIFICATION OF THE SERVICE

In this paper we will only describe the functionality of the automatic version of the Dutch *scope* service. The functionality and the user interface of the system under test were copied as closely as possible from the existing *scope* service, offered in the Netherlands by

PTT Telecom. Subscribers to the regular service receive a 'credit card' with a 14 digit card number and a four digit PIN code. To obtain access to the service the customer dials a toll free (country direct) number. Depending on the country from which the call originates the customer is connected with an IVR system or with an operator. Customers can always obtain access to the operator, by not responding to the IVR prompts (in which case that system assumes that the caller has no DTMF available). In order to gain access to the service the customer must provide card number and PIN. Access is only granted if the card number is valid, and if the correct PIN is given. The IVR system accesses a database to check the combination of card number and PIN for validity and conformity. If after a second attempt card number or PIN are still not correct, the caller is connected to an operator. The operator prompts for the same data for the same database query. If either card number or PIN is not correct, there is nothing that the operator can do to help the customer, since operators have no access to the data in the database.
Scope provides additional services (e.g. a message service, the possibility to send presents to anybody within the country, etc.), but it was decided not to implement this functionality in the service as it is tested. Adding additional functionality to the test system would have changed the design of the test completely, since it would no longer have been reasonable to assume that all subjects in the test have prior experience with all relevant aspects of the service.

3. BUILDING THE MIMIC SYSTEM

For its basic functionality (i.e., making calling card calls) *scope* offers a type ahead facility in its IVR interface. Customers can type card number, PIN code and telephone number (separated by asterisks and terminated by a pound sign) without having to wait for the prompts. Alternatively, they have to select the calling card functionality via a selection in the main menu, after which they are prompted for card number, PIN and telephone number, always in that order. We decided to copy the latter version of the IVR interface, except for the initial selection: since the automatic version only offers basic calling card services, that has become redundant. Still, using this design it was possible to create an interface that is at once clear for the customer and capable of being implemented with ASR limited to connected digit recognition. Moreover, customers should find it natural to identify themselves by means of their *scope* card number, since this is what they are used to in the present operational version of the service. Names are never asked (and seldom given).

The interface of the *scope* service as implemented here differs in at least one essential way from the design in the Sprint FonCard^{TM} service, where customers have to enter a checksum protected version of their social security number. By using the *scope* card number for the caller to identify her/himself we do not mix concepts of different worlds and services. Also, regular *scope* users know that their card number is mandatory to obtain access to the service.

As in the operational *scope* service customers are given two opportunities to enter a correct combination of card number and PIN (in our experiment half of the subjects does not need the PIN). If after the second attempt card number and/or PIN are still not valid, the caller hears an error message, saying that access has failed; in the operational service (s)he would be passed to an operator, but since that would be too expensive for our experiment the caller is now disconnected.

For lack of proven echo cancellation technology integrated in the telephony server we did not allow barge-in. To speed up the service as much as possible great care was taken to design very short, yet unambiguous prompts.

To test the technical performance of the system and at the same time investigate several human factors issues, we have built a Mimic system that implements the basic Calling Card service. The flow charts of the two parts of the system (enrolment and the actual service) are shown in Fig. 1. To that end a connected digit recogniser was trained for Dutch, and implemented on the standard Vocalis platform. Training material was taken from the Dutch POLYPHONE corpus [2]. The ASR is used to recognise card numbers, PINs and telephone numbers.

In order to limit undesirable interactions between ASR and SV performance to a minimum heavy use was made of the syntactic limitations in card and telephone numbers. Especially card number, which should adhere to a rigid ISO standard, obey strict syntax rules. By relying on these rules it was possible to obtain ASR performance sufficient for the service. Recognition of PINs was enhanced by constructing four digit PIN codes in which the last digit served as a checksum for the first three digits. This procedure is different from the operational *scope* service (where customers can select and change their PIN at will), but it was considered acceptable in the framework of the present experiment.

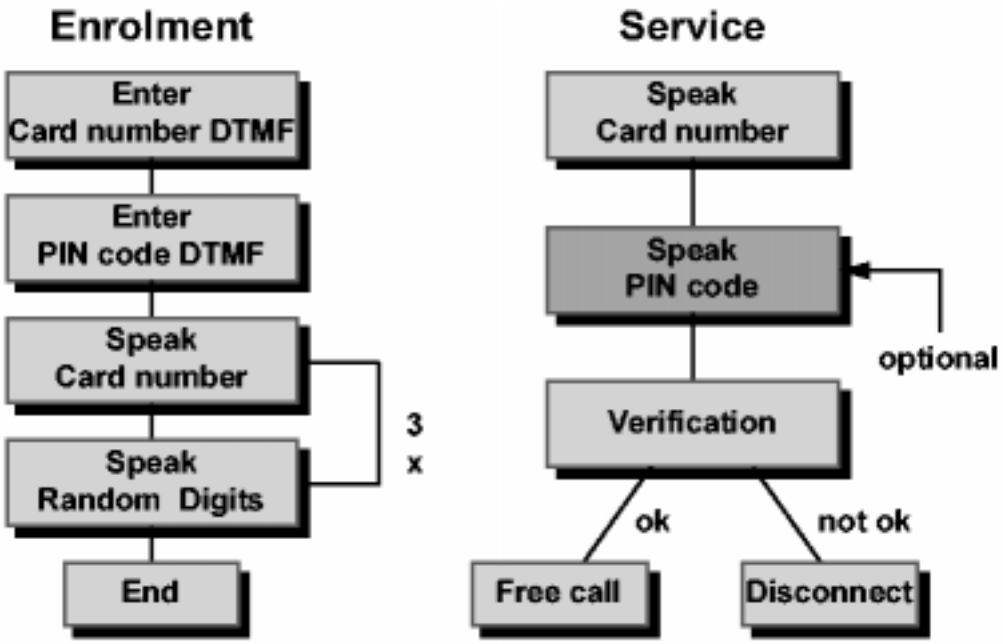
The subjects are prompted to speak all numbers as connected digits. A pilot experiment was performed to determine the prompt formulation that would yield the maximum number of error free tokens (no hesitations, only connected digits). Two formulations were compared, viz.

- Say your card number, digit by digit
- Say your card number as 'eight nine three zero ... '

Although the first prompt formulation yielded slightly more responses containing non-digit material (especially with area codes in telephone number, where 010 and 020 were pronounced as *Oh ten* or *Oh twenty*, we opted for this formulation because it is much shorter.

A basic text-dependent speaker verification algorithm has been implemented on the Vocalis hardware platform. The algorithm is based on left-to-right HMMs, with speaker models for all digits which appear in the card number. The number of states in each model is equal to four times the number of phonemes in the canonical phonemic transcription of the word. A single Gaussian distribution is trained for each state.

Figure 1. Flow diagrams of enrolment and service



This is the topology that appeared to perform best in extensive tests carried out in the technical part of the CAVE project [1]. Speaker verification is attempted on the 14 digit card number only. Therefore, the card number is used both for the identity claim and the identity check.

4. TEST SET-UP

The field test was set up to investigate whether a voice interface was accepted by the *scope* customers as an alternative for the DTMF interface in situations where DTMF is not available. In addition, we wanted to know whether SV on the card number only, without the caller having to give the PIN code, was perceived as safe enough to accept it as an alternative for the DTMF interface. In addition, we wanted to know whether the customers find the version without PIN easier to use than the version with PIN. Therefore, two versions of the application were built, one which exactly copies the DTMF interface (prompt for card number, PIN, and telephone number) and another one in which the prompt for the PIN was omitted.

4.1. Subjects

Eighty subjects (40 females and 40 males) were recruited from the staff of PTT Telecom, all of whom were frequent users of the *scope* service. Therefore, all subjects knew the interface of the service, as well as its basic functionality. Half of the subjects were assigned to the with-PIN group, whereas the remaining 40 subjects were assigned to the without-PIN group. The subjects received a temporary *scope card*, with an attendant PIN code (if they belong to the 'with PIN' group.

The requirement that all subjects know the service is considered of primary importance. Experience shows that many experiments with speech based services run into difficulties because the subjects do not understand the service under investigation. *Scope* makes large investments in customer education, for instance by means of a quarterly magazine, in which the service and its interface is always explained. These educational activities will remain necessary until it is feasible to market services based on intelligent spoken dialogues and intelligent agents that know how to adapt very rapidly to the experience that a customer has with the service.

4.2. Experimental procedure

After having completed the enrolment, subjects had to call the system at least five times during a one week period under their own identity. The service access dialogues of all calls have been recorded for further analysis. All subjects received card number and PIN of three other subjects, whom they did not know. They were asked to try and break into the account of these victims, in order to get a feeling for the safety offered by the SV protection. Subjects were not suggested to invite relatives to break into their accounts by using their card number and PIN.

In our experimental setup subjects are encouraged to call the system, because each time they obtain access they are allowed to make a ten minute domestic call for free. This design should guarantee that all subjects do call at least five times.

With their invitation and instruction letter subjects received a booklet with preprinted forms, designed to record their calling behaviour. For each call they were asked to record whether it was made as a true customer or as impostor, and whether it was successful. In case of failure, they were requested to record the cause of the failure (no correct verification of their ID, failure of ASR to recognise card number, PIN or telephone number, etc.). The main function of the booklets was to support the subjects in filling out the

questionnaire, or in answering the questions when they were called by the experimenter.

4.3. Enrolment

All subjects went through a single enrolment session, in which they had to say their card number three times, spoken as a digit string. These utterances formed the enrolment speech. Enrolment was done off-line. The three productions of the card number were checked auditorily. If one or more of the tokens contained errors, the subject was called by the experimenter, and asked to repeat the enrolment session.
It is well known that a single enrolment session is not enough to obtain good estimates of the variation within a customer's speech behaviour, nor of the variation caused by different handsets and acoustic backgrounds. Yet, most service providers believe that a single enrolment session is the maximum they can ask of their customers. The technological problems can be solved by using incremental enrolment, using confirmed access attempts to improve the speakers' models. However, for the Mimic test reported here incremental enrolment was not available.

4.4. The Questionnaire

The questionnaire used in the Mimic test consists of seven sections, four of which contain open questions. The remaining three sections consist of Likert scales, i.e., statements for which the subjects are asked to say to what extent the (dis)agree. Of the four sections with open questions, three are reserved for a subset of ten subjects, randomly selected from the two groups (with and without PIN). The questions define a semi-structured interview, designed to obtain information that should be useful in future experiments.
The first open question section is related to the test itself. Subjects are asked to give a general impression, and to point out aspects that struck them as especially good or bad.
The second set of open questions addresses the speech based service interface and the way in which it has been used by this subject during the experiment. Among other things, it is asked what this subject missed in the service under test, whether attempts were made to break into other subjets' accounts, and whether it seems to be a good idea to leave out the PIN. Of course, the way in which the last question is formulated differently for subjects in the with and without PIN groups.
The third section of questions in the semi-structured interview relates to the present *scope* service. We want to know subjects' experience with the service: what do they use it for and how do they like the DTMF interface.

The first set of Likert scales, completed by all subjects, is related to enrolment. We did not ask whether subjects would have been willing to go through two or more enrolment sessions. The major reason for not asking that question is that we do not expect naive subjects to know how much security would be added by additional enrolment sessions. For the same reason we only ask whether the enrolment session was perceived as being too long and complicated, without the companion statement implying that enrolment was too short and easy.
The second set of Likert scales addresses the actual use of the speech based service. It includes statements about the perceived security of the SV protected service (the same statements have been used for the with and without PIN groups), statements comparing the speed and user friendliness of the speech based interface compared to the existing *scope* service, about the procedures for coping with problems with the speech based interface and about the sufficiency of the information about the new interface (both prior to its use -in the form of written instructions- and during the use -in the form of what little help information that is given).
The last set of Likert scales is about the perception of the performance of the underlying technology, ASR as well as SV.

The last set of open questions, to be answered by all subjects, addresses general aspects of the calling card service and its use. To some extent it overlaps with the questions that have already been described for the semi- structured interview.

5. RESULTS AND DISCUSSION

At the time of this writing the experiment was not yet completed, mainly because of unexpected difficulties in building a stable operational system. Full results will, however, be available at the time of the conference.
One problem we will have to solve in data processing is due to the fact that the subjects were free to call whenever they wanted and from wherever they wanted. As a consequence, there is no incontestable evidence of the identity of the callers.

6. CONCLUSION

In this paper we have described the design of a field test of a calling card service protected by means of Speaker Verification. In a version of the service that uses speech instead of DTMF and SV instead of, or in addition to a PIN, we have investigated a number of human factors aspects of the deployment of a combination of SV and ASR in a service for the general public.

ACKNOWLEDGMENTS

The authors are indebted to Thomas Moser of Swiss Telecom and to Jan van Scheffel and Bart van der Mark of KPN Research for their help in building the Mimic system. Thanks are also due to Angelien Sanderman ITB/KPN Research, for reviewing the questionnaire.

REFERENCES

[1] Bimbot F., Hutter H.-P, Jaboulet C, Koolwaaij J., Lindberg J., Pierrot J.-B. (1997) Speaker verification in the telephone network: An overview of the technical development activities in the CAVE project. *Proceedings EUROSPEECH-97.*

[2] den Os, E.A., Boogaart, T.I., Boves, L. & Klabbers, E. (1995) The Dutch Polyphone corpus. *Proceedings EUROSPEECH-95*, pp. 825-828.