SPEAKER VERIFICATION AS A USER-FRIENDLY ACCESS FOR THE VISUALLY IMPAIRED

Els den Os1, Hans Jongebloed1, Alice Stijsiger1, Lou Boves1,2
1KPN Research, P.O.Box 421, 2260 AK Leidschendam, The Netherlands
2Nijmegen University, P.O.Box 9103, 6500 HD Nijmegen, The Netherlands

ABSTRACT

There are few operational services that use speaker verification (SV) as a means to provide secure, yet easy to use access. In this paper we describe the first semi-operational service that is offered by KPN Telecom that has been developed within the framework of the LE-4 project PICASSO. We report on the objective and subjective evaluation of a service that offers free access to directory information in the Netherlands. We concentrate on the enrolment phase, since this is the most critical phase for operational services. Hundred seventy-four persons used this service since the end of 1998; all interaction data are logged for analysis. We have interviewed 26 subjects. We have learned that naive subjects have difficulty in understanding what is expected from them; this holds both for the enrolment and for the access phase. Only half of the subjects finished the enrolment in the minimum number of two calls. The other subjects needed more calls, or did not succeed at all. During access, we observed that more than half of the calls which were refused by SV were calls in which the caller did not behave as (s)he should (screaming, whispering, not entering speech at all, or incomplete utterances). Real false rejects amount to 4% which makes clear that from the technology point of view improvements are needed. Still, even persons who have occasional trouble using the service have a quite positive opinion about the speech-secured service.

I. INTRODUCTION

Speaker verification (SV) is gradually becoming a mature technology. Experiments on databases show Equal Error Rates (EER) close to 0.5% [1]. In controlled field tests of a calling card application, encouraging performance of the technology with 14-digit card numbers was confirmed [2]. However, for real-life applications of SV a number of questions remain to be answered.

One question relates to the behaviour of the users during the enrolment phase. It has been suggested that some may invite other persons to call the system in their place. Another issue has to do with the number of enrolment sessions. Research in the lab shows that multiple sessions, containing several utterances, give the best results [1]. However, in a real-life service multiple sessions are highly undesirable [3]. Last but not least, we do not know whether naive subjects understand what they should or should not do when using the service.

In the framework of the LE-4 project PICASSO, KPN Research has developed a speech-secured version of free access to directory assistance (DA), a service offered by KPN to the visually impaired. We report on a test with 174 subjects who started to use the speech driven access from the end of 1998. The results of this test will affect the decision whether the service will be made fully operational. We concentrate on the first three months of this test, and more specifically on the enrolment phase and subjective evaluation.

II. THE SERVICE

KPN offers free access to DA to a registered group of (mainly visually) impaired users. By default this service is offered by a system based on DTMF input (users have to press digits on their keypad). Customers call an 11-digit toll free number that connects to an Interactive Voice Response (IVR) system, which requires entering an additional 10-digit personal identification number. This identification number is composed of 6 digits encoding the date of birth of the subscriber and 4 digits that are arbitrarily chosen. From interviews with subscribers it appeared that some subjects consider this procedure difficult to use. Since the end of 1998, KPN also offers speech-driven access to a selected group. SV is performed on the subscriber’s 10-digit telephone number. We expect that everyone knows his/her own telephone number by heart and that this digit string is long enough for SV to reach satisfactory performance.

THE SPEECH-DRIVEN SYSTEM

A subscriber has to call an 11-digit toll free number. Since subscribers mostly call from their homes, Calling Line Identification (CLI) is used as the primary means for making an identity claim. If no CLI is available, or if a subscriber calls from a different location, (s)he is asked to enter the private telephone number by means of DTMF. If the number is found in the subscriber database, the caller is asked to speak the number, using only connected digits; otherwise, the caller is told that the number is not authorised. The speech is input to a connected digit recogniser, the output of which is checked against the expected digit string. If no more than two digits are in error, the input is accepted and subsequently fed to the SV system. On successful verification the caller is connected to the normal DA service. If SV fails to verify the claimed identity, the caller is asked to say the telephone number a second time. On a second ‘reject’ output, the caller is asked to enter the four-digit numerical part of the postal code of the home address by DTMF. If the code is correct, (s)he is connected to the DA service after all. Admittedly, this provides a relatively low security level. But since the cost of an impostor break-in is low (less than 0.75 Euro), and the

1 We could have used ASR to enter the number. However, we were not convinced that the digit recognition available today is sufficiently accurate for this purpose.
likelihood of somebody knowing the toll free number, the private telephone number of a subscriber and the attendant’s postal code is small, the security is considered adequate.

Each new subscriber has to go through at least two ‘concealed’ enrolment sessions. If the calling number (identified through CLI or DTMF) is in the database, and there are no speaker models trained for this number, the caller hears a welcome prompt saying that this is the first (or second) time (s)he calls the system. The caller is then asked to say the 10-digit telephone number; next (s)he is asked to say the four-digit postal code; then the telephone number must be spoken a second time. If the telephone numbers are correctly recognised, the call is connected to the DA service, no further questions asked. During the enrolment calls, when only ASR is active, subjects are given three chances to have the number recognised for each of the two times the number must be spoken. Originally, we considered asking the subjects to enter a PIN code during the enrolment sessions, but in a usability test this appeared to be extremely confusing. Since new customers are asked to make both enrolment calls from their home phone, the procedure was considered as sufficiently secure.

If the first call is completed, two tokens of the spoken telephone number are stored, and a counter for this account is set to 1. In the next call, the customer is welcomed with the message that this is the second time the service is accessed, and the exact same procedure as the first time is followed. If two additional valid tokens of the telephone number can be recorded, the speaker model is immediately created.

III. METHOD

We invited 500 users of the DTMF version to participate in the test of the speech-driven free access service. For efficiency reasons, we sent the invitation letter and the instruction letter at the same time. Unfortunately, this caused some confusion, because not all subjects understood that they had to wait for two weeks after their returning the registration form before they could start using the service. This experience shows that it is essential to keep the period between registration of new customers and the availability of a service as short as possible.

The instructions how to use the service were kept as brief and simple as possible. Subjects were requested to make the first two calls from their home phone. We assumed that none of the potential customers would have disabled CLI. Prospective users were given the telephone number of a help desk that is operated by the Council of the Handicapped.

All interactions between the callers and the system were logged, and the speech was stored for later analysis. The data in the log files formed the basis for the objective analysis of the behaviour of the system and the clients. The logs are by necessity incomplete in one crucial respect: there is no way in which the system can distinguish between true clients and impostors. We can only log the system’s decision to accept or reject an identity claim. In order to get at least some insight into the proportion of false rejects we made auditory checks of the speech in the calls where the SV system yielded a ‘reject’ decision.

We have conducted semi-structured interviews with 26 clients who indicated that they wanted to participate in the test. These interviews form the basis for the subjective evaluation presented in section V.

IV. RESULTS – OBJECTIVE ANALYSIS

1. Enrolment:

Up to now 174 subjects started enrolment. Their ages range from 18 to 81 years. For 155 subjects (89%) CLI was available; the remaining 19 clients had disabled CLI for their home number, or they called from another location.

In order to check whether the same person called during both enrolment sessions we listened to all recordings. For four subjects there is some doubt about the identity of the callers. In one case we are certain that two different speakers called, since the speaker in one session is male, and in the other female. Thus, there were relatively few attempts to tamper with the instructions for enrolment.

Table 1 summarises the log data pertaining to the enrolment calls. ‘Completed enrolment’ means that for this account speaker models are available; therefore the subject successfully completed two sessions. ‘Half enrolment’ means that the subject completed one session successfully.

<table>
<thead>
<tr>
<th>Table 1. Number of subjects who attempted enrolment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total # of subjects who started enrolment</td>
</tr>
<tr>
<td>CLI present</td>
</tr>
<tr>
<td>CLI absent</td>
</tr>
<tr>
<td># Subjects who completed enrolment</td>
</tr>
<tr>
<td># Subjects with half enrolment</td>
</tr>
<tr>
<td># Subjects failing to complete first session</td>
</tr>
</tbody>
</table>

We looked at the data of the subjects who failed to complete at least one enrolment session. Three persons made a large number of calls (between 6 and 12), but they disconnected during the session, or they failed to speak the digit string in the way expected by the system (either too slow, or they started speaking before the prompts were completed). Three subjects made one call, four called three times and one tried four times.

Part of the problems encountered during enrolment were due to inconsistent combinations of telephone number and postal code in the system’s database. Apparently, it is necessary to very closely monitor the enrolment process in services protected by SV. Software must be designed that tracks enrolment calls, and that generates problem reports for the service supervisor.

In previous field tests [2] we never observed subjects who failed to speak their 14-digit card number in the time slot allotted by the system. Apparently, we have
been misguided by the ease with which sophisticated subjects, used to modern technology, deal with instructions of automated services. This finding corroborates observations made in other services which were extensively piloted with early adopter volunteers. Invariably, the volunteers obtained a substantially higher performance than the real customers when the service was rolled out.

Two enrolment sessions suffice to record the four tokens of the telephone number that we needed to train speaker models\(^2\). Fig. 1 shows the real number of calls that the clients needed to complete two enrolment sessions. Just over half the subjects enrolled successfully after the minimum number of two calls. An additional 25% needed three calls, while the remaining 25% needed more than one additional call.

An enrolment call with CLI present requires three actions of the caller: the telephone number, plus two DTMF inputs, for the telephone number and the postal code. Even for the subjects who needed only two calls to complete enrolment things were not completely ideal. For these subjects the average number of actions in the enrolment calls was 3.7 (71 subjects) for the calls with CLI and 5.8 for the 8 subjects without CLI.

![Figure 1: Number of enrolment calls needed to complete enrolment (number of subjects = 146)](image)

### 2. Access Calls

Table 2 summarises the data on the calls placed for telephone numbers for which speaker models were available. The distribution of the calls over the subjects is very uneven; some are heavy users, while most clients use the service only occasionally.

‘Successful calls’ are calls in which the SV system initiates a transfer of the call to the DA service. The only way to make sure that none of these ‘successful’ calls actually come from impostors would be auditory checks of more than 1000 speech files. So far, this has not been done.

In 44 calls the client failed to enter the postal code using DTMF twice in a row. Some of these calls may have come from impostors, but we have reasons to believe that it is more likely that clients have difficulty in using DTMF. In 63 calls the digit recogniser failed to recognise the spoken input as the intended phone number. In many of these cases the caller did not manage to produce the number in the time slot offered by the system. This corroborates the conclusion that naive subjects may not be able to comply with the requirements of present day speech systems.

Ideally, an access call with CLI requires only one action to be performed by the caller: say the telephone number. In actual practice, the average number of actions is 1.3. For the calls without CLI present these numbers are 2 and 2.4, respectively.

### V. RESULTS – SUBJECTIVE EVALUATION

At this time 26 interviews have been held with persons who have responded positively to the invitation to participate. Of these, 3 never called the system, 3 made only a single enrolment call, 10 completed enrolment in only two sessions, and 10 needed more than two sessions to complete enrolment. It appears that at least in

<table>
<thead>
<tr>
<th>Table 2. Summary data on access calls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total # of access calls</td>
</tr>
<tr>
<td>CLI present</td>
</tr>
<tr>
<td>CLI absent</td>
</tr>
<tr>
<td>Total # successful calls</td>
</tr>
<tr>
<td>Accepted by SV</td>
</tr>
<tr>
<td>Accepted by postal code</td>
</tr>
<tr>
<td>Total # unsuccessful calls</td>
</tr>
<tr>
<td>Twice failed utterance verification</td>
</tr>
<tr>
<td>Twice wrong DTMF postal code</td>
</tr>
<tr>
<td>Twice unsuccessful SV</td>
</tr>
<tr>
<td>Unexplained user hang up</td>
</tr>
</tbody>
</table>
some cases clients in the groups who did not complete enrolment failed to understand the invitation or the instruction. Also, most of the persons in this group were satisfied with the existing DTMF access. They tend to access the service from their home phone: problems with entering the PIN are avoided by programming the sequence under one of the speed-dial buttons on the handset. These persons typically use the DA service to obtain numbers they call for the first time. In 20% of the cases the DA service is used because the customer could not remember a number. On average this group makes about six phone calls per day. The DA service is called three times per week.

There were no striking differences between the answers of the two groups of clients who completed enrolment. Apparently, encountering some problems during the initial use of an SV-protected service does not by necessity cause people to abandon the service altogether, even though subjects kept the alternative of accessing the DA service via the DTMF-based system.

The enrolled subjects estimated that they make approximately 5.5 calls per day and that they access the DA service five times per week. With few exceptions the DA service is called to obtain numbers that were not previously dialled. One or two persons said they still use the DTMF-based access along with the SV-protected version under test. Eight out of ten respondents say that the SV-protected access to DA is easier to use than the older DTMF access. On a scale from 1 to 10 the SV system is given an 8, compared to a 7 for the DTMF system. Seventy percent of the respondents think that the service is reasonably secure. The remaining subjects did not have clear opinions about the safety of the system. One potential disadvantage of the SV-protected version is that it is no longer possible to allow partners to also access the DA service for free. However, none of the respondents found this a problem. They said that they never disclosed their PIN-code for the DTMF service, because they feared that the free access would be discontinued if unauthorised use would be established. Other customers live on their own.

Over 80% of the respondents found the new service easy to learn. The same number found the invitation letter and the explanation of the service clear and easy to understand. Nine out of ten respondents seemed to understand the difference between enrolment and access sessions, and to understand the need for enrolment. This is somewhat surprising, because these technical issues were at best alluded to in the written explanation. Some 70% of the subjects are aware that they were taking part in a test. With few exceptions respondents knew that their voice was being recorded.

Over 80% of the subjects find the service fast. Some said that they did occasionally disconnect before getting access to a DA operator, e.g., because they remembered the number they were looking for, or because the call was only meant to show the service to somebody. Only seven out of ten subjects said that they find it easy to say their telephone number in the form of a digit string, although things become easier with practice. Eighty percent find the service friendly, and the remaining 20% finds it businesslike, but o.k.

A low 10% of the respondents reported errors in their interaction with the system. Moreover, many said that the existing services did not always work exactly in the way one would expect either. For instance, DA sometimes has long waiting queues. A full 90% of the respondents is satisfied with the new service.

At the end of the interview subjects were asked to make suggestions for improvements. Some said they prefer to use their name or a much shorter number for SV.

VI. CONCLUSIONS AND RECOMMENDATIONS

From the results of this study it appears that truly naive subjects have much more difficulty in dealing with speech-driven services than we anticipated on the basis of previous field tests with early adopters. In the service at hand, we might have alleviated some of the problems if barge-in were available. Also, it may be necessary to play explicit examples of how to speak the digit sequence, at least during the first enrolment session. As soon as the ASR system reports compliant user behaviour, this explicit prompting can be abandoned, in favour of the much shorter prompts used in our system.

We have noticed that the way in which subjects speak their telephone number evolves over time. Therefore, it may be necessary to replace the initial models at some point in time by models trained with speech from sessions that are more representative of normal use. This should help in reducing the proportion of false rejects.

An SV service requires special procedures and techniques for setting up and maintaining the underlying databases. These techniques should be the focus of future research. Also, an SV-based service requires close monitoring of the interactions, especially for new customers.

Last but not least, the customers are very satisfied with the SV-protected service. They find it easy to learn and easier to use than the DTMF alternative. Moreover, they find it pleasant to use.

REFERENCES

