ISCA Archive
http://www.isca-speech.org/archive

6th International Conference on Spoken
Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

# SPEAKER VERIFICATION IN OPERATIONAL ENVIRONMENTS – MONITORING FOR IMPROVED SERVICE OPERATION

*Yong Gu*[3]*,Hans Jongebloed*[1]*, Dorota Iskra*[1]*, Els den Os*[1]*, Lou Boves*[1,2]
*email: {h.a.jongebloed,d.j.iskra,e.a.denos, l.w.j.boves}@kpn.com, yong.gu@vocalis.com*

[1]KPN Research, KPN Royal Dutch Telecom, Multi-Media Department,
P.O.Box 421, 2260 AK, Leidschendam, The Netherlands
[2]Nijmegen University, Department of Speech and Language, The Netherlands
[3]Vocalis Ltd., Cambridge, UK

## ABSTRACT

There are very few, if any, published accounts of practical experience with Speaker Verification as a means to provide secure access to telematics services. Yet, there is no reason to expect that Speaker Verification is very different from speech recognition, for which many deployed services have shown the need for close and intensive on-line monitoring during the time when the service becomes operational. In this paper we present our experience with a monitoring scheme for Speaker Verification during the field test of a financial investment game. Many of the issues that were monitored were suggested by our experience with a semi-operational service, viz. free access to Directory Assistance for visually impaired. A newly developed enrolment procedure, that can flag potentially weak speaker models, is an essential part of the monitoring procedure.

## 1 INTRODUCTION

Operational services that use speaker verification for secure access have already been predicted for at least a decade, but few successful large-scale telecommunication services have yet materialised. Field tests in the European LE projects CAVE [1] and Picasso [2] show that there is still a significant performance gap between operational services and controlled experiments on pre-recorded databases like SESP [1] (even though recorded in real life conditions). Part of the performance problems is due to the failure of the human computer dialog to elicit the speech material needed by the system. Pre-recorded databases seldom contain such 'invalid' utterances, and certainly not in the subset destined for model enrolment. In an operational service the risk of non-compliant user behaviour is largest during the first interactions with a system. Experience shows that clients tend to stop using a service when they experience problems [3]. The aim of the paper is to test the feasibility of an on-line monitoring procedure that can indicate potential problems early enough to allow successful corrective measures to be taken. A newly developed enrolment procedure, that is able to flag potentially poor speaker models is an essential part of the procedure.

In 1999, KPN ran a one-year field test with the speech secured Free Access to Directory Assistance (FADA) service [3]. This service allowed authorised visually impaired customers to use the operator based directory assistance service for free. A helpdesk was set up for the clients that they could contact in case of problems. After three months an evaluation was carried out based on the recorded dialogs. This showed that the *enrolment* into the service already caused problems, mainly related to the failure of the callers to adhere to the (tacit) expectations of the system (e.g. the need to complete the digit sequence in the required time slot) and partly to system errors. About 16% of the clients failed to complete the required two successful calls for enrolment and stopped using the service. These clients did not contact the helpdesk, although they had serious problems using the service. Apparently, the help desk should be more proactive: if 'common' problems are detected, the help desk should contact the clients to assist them in the proper use of the service.

The overall analysis of *access* calls showed that the false reject rate for individual clients were very high for a substantial proportion of the callers [4], [5]. It turned out that these errors were related to the quality of the speaker models after enrolment [6]. This model quality is strongly dependent on consistent behaviour of the user during enrolment. A measurement of the model quality thus yields input to the monitoring process to prevent individual performance problems due to low quality models.

This paper reports on an experiment with on-line monitoring in a telephone based speech driven investment game. The next section presents the service, the monitoring and the model quality assessment. Section 3 presents the results of the field test, followed by the discussion and conclusions in Sections 4 and 5.

## 2 METHODS

This section describes the speech driven investment game, the monitoring process and the model quality check. Finally, the design and implementation of the field test are described.

### 2.1 Investment game

The application was a speech-driven investment game, that allowed the participants to obtain the latest stock information, trade stocks and enquire about game scores. The game had an on-line connection to the Amsterdam Stock Exchange. The task of the participants was to maximize the value of their portfolio by selling and buying stock through the SV secured system.

The speaker verification and recognition engine used was the Vocalis system, an improved version of the one used in the FADA service [9]. The callers had to identify themselves by saying a seven-digit account number, on which the verification was carried out as well. As an additional security measure customers calling from the fixed network had to say a five-digit PIN-code. Mobile users could only use the service with their private handset, so that calling line identification (CLI) information could be used as an extra security check. To increase

ASR robustness an error correction mechanism was used which allowed for one wrongly recognised digit in the account number.

The first five valid calls (i.e. providing a correct PIN/CLI and account number) were used for enrolment. The enrolment and access dialogs comprised the same prompts to the user (Figure 1). The speaker independent verification threshold was set to minimise the false accepts with an acceptable amount of false rejects, based on the results of an off-line test with comparable data. When recognition mistakes were made or the customer was rejected due to verification, he/she received a maximum of two additional attempts within one call to access the service.

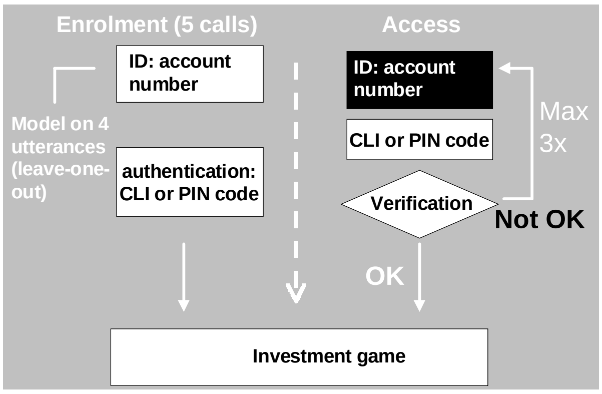

**Figure 1:** Flow of the identification/authentication sub-dialog of the investment game application. The left part shows the dialog for enrolment, the right part for the speech-secured access (after the model was trained).

## 2.2 Monitoring

The monitoring process comprised three levels in which system log data were converted into signs of potentially erroneous system behaviour. The different levels were:

**logging**: **A**ll system events like dialog turns, customer database query results, system errors, and outputs from ASR and SV were stored in a database. All events were tagged with a unique dialog identity string. Each dialog was also labelled by the user ID of the client, derived from the speech recognition results or the CLI. If multiple account numbers were recognised without granting access, the user ID was set to “unclear”.

**interpretation**: The raw logging data were analysed to detect problem situations. Database queries for certain system events (e.g. caller rejected) can yield indications of recurrent problems for individual clients. The queries were designed with the most common problem observed in the FADA field tests in mind. If the proportion of problematic calls exceeded a pre-defined threshold, the client was marked as “to be checked” (e.g. multiple rejects might indicate a bad model quality or inappropriate usage of the service).

**action list generation**: To solve the problems signalled on the interpretation level, a number of observation-action rules were formulated. The actions are either diagnostic, to check the exact nature of the error (e.g. are the rejects true or false) or aimed at solving the problem. If necessary, clients can be contacted and instructed to adapt their behaviour to match the system’s expectations (e.g. “please speak after the beep”).

All three levels were active on-line. As soon as a dialog was finished, all dialog events were added to the database and the calculations for the problem levels were carried out. Several examples where this monitoring was beneficial will be presented in Section 3.2 and 3.4.

## 2.3 Model quality check

In the FADA service it appeared that for too large a group of speakers the false reject rates were unacceptable [6]. In [7] it was shown that careful selection of the enrolment material improves the model quality and therewith overall system performance. In the present system a leave-one-out method was implemented that should yield optimal models and at the same time flag potentially weak models.

The **leave-one-out** method requires *N* utterances of which *N*-1 are used to train the model. The remaining utterance is tested against the model, yielding a SV score $S_i$. Scores below a certain threshold indicate a good match between the model and the utterance. This training and testing is done for each of the *N* possible permutations yielding *N* scores. The model that yields the *highest* score $S_i$ on the test utterance is the model with the most consistent utterances. If one or more scores $S_i$ exceed the threshold, because utterance(s) *i* do not closely match the model, the trained model is assigned a ‘low quality’ tag and the corresponding utterances are tagged as ‘outliers’.

The output of the leave-one-out (number of outliers and the verification score(s)) was used as extra input for the monitoring. In the case of one or more outliers the speech data were checked manually to decide if impostor material was involved or other factors that led to the inconsistency. Results of the leave-one-out algorithm are presented in Sections 3.3 and 3.4.

## 2.4 Field Test

With the game a field test was carried out which lasted five weeks. The participants were all high-educated employees of a financial institution. The group was divided in mobile users who could only access their account by their own mobile phone and users who could use (mainly) a fixed line telephone. A proactive helpdesk was set up that, besides receiving feedback and error reports from users, actively motivated clients to call the service regularly and provided clients with feedback on their problems reported including hints to adapt their user behaviour.

When the test was completed, the participants were asked to fill in a questionnaire in which they were asked about several aspects of the test, the technology used and the possible service opportunities. In order to establish a False Accept rate for the system (where an impostor is granted access) 15 new participants were invited to make controlled impostor attempts at the end of the field test.

# 3 RESULTS

109 customers called the service at least once. 20 customers (18%) were given mobile accounts. A total number of 2662

calls were made to the system, 30% of which came from mobile users. Thus, the mobile users are slightly over-represented. However, no significant performance differences were found between mobile and fixed network users. 113 calls (4%) had unclear identity (i.e. in one call more than one attempt was made to enter the account number and different numbers were recognised each time). The number of calls per user ranged from 1 to 192, with an average of 22 calls. The record number of 192 calls came from a mobile client.

Impostor attempts were performed on the accounts for which a sufficient number of calls were made by the true client. A total of 1230 impostor verification attempts were made for 50 accounts.

The remainder of this section presents the results for different aspects of the trial. Table 1 summarizes the results.

| | **Number** | **% of total** |
|---|---:|---:|
| Customers | **109** | |
| Enrolled customers | 98 | 89% |
| Low quality models | 18 | 18% |
| Calls | **2587** | |
| False Rejects | 55 | 2% |
| Impostor calls | **1230** | |
| False Accepts | 7 | 0.6% |

**Table 1:** Overall results from the field test.

## 3.1 Field test monitoring

During the field test the monitoring relied on human supervision to take actions in case of problems. On a daily basis an overview of the system events was generated. It was formatted in such a way that it was easy to see the performance for individual clients. Clients for whom the FR rate exceeded 10% of a sufficiently large number of calls were marked as “suspicious”. For these clients the recordings of the enrolment and access utterances, and occasionally also the SV scores, were checked. The helpdesk reported problems detected by customers, which were then analyzed using the reports generated by the monitoring procedure. This way, new observation-action rules evolved (e.g. “check immediately with the client if the mobile CLI does not correspond to the account number”).

## 3.2 Enrolment

In the enrolment phase, a dozen callers failed to obtain access because the system recognized digit strings that did not match the existing account numbers. Listening to the files revealed that the customers confused the account number and the PIN. So it turns out that even highly educated users may fail to read or comprehend the instructions from the service provider [3]. A single round of active feedback to the customers by the help desk alleviated the problem.

98 of the 109 participants made at least five valid calls (where the identity of the caller was established with certainty) which means that for 89% of the callers a model was trained. Thus, despite our attempts to be as pro-active as possible, still 11% of the first time customers stopped using the service at the initial stage.

## 3.3 Model quality

The leave-one-out strategy signaled potential problems for 18 models (18%). For these questionable models one of the five utterances (in one case two and in one case three) received an outlier tag (the verification score for this utterance was higher than the threshold). These cases were analyzed by listening to the in total 21 outliers as well as the remaining enrolment utterances. In two cases an utterance with extra non-digit (background) speech was left out of the model, in another it was an utterance made clearly by an impostor speaker. The remaining 18 outliers had several different caused: poor or inconsistent quality of the telephone line (6), different speaking tempo (4) and different pause locations (8) in the digit string between enrolment and access. The last two causes were due to the fact that those users adapted their behaviour (i.e. speaking in a more natural way), as they became more familiar with the system.

## 3.4 Verification - False Rejects

In 58 of the 2587 calls the customers were refused access as the speaker verification rejected them in the third access attempt within the same dialog. Auditory comparison of the speech files with other files of the same client showed that 55 were False Rejects (FR), giving a 2% False Reject rate on dialog level.

19 clients (19%) were responsible for all FR’s. For the clients who called more than 20 times FR rates ranged from 1-10%. The large range in individual false reject rates suggest that the majority of the problems is due to user behaviour rather than failures of the system. Two clients encountered FR rates exceeding 10%. One of them tried to test the system for its robustness (e.g. by making hands free calls while driving at high speed). Inspection of the data of the other client showed that the rejects could not be explained. For both customers the verification thresholds were slightly increased to reduce the chance of future FRs. By the end of the field test, their individual FR rate had decreased below 10%. However, the “test driver” still caused 19 (35%) of all false rejects.

After listening to the remaining 36 rejected utterances it appeared that about 30% were probably due to background noise. About another 30% were due to varying speaking tempo and pauses made at different locations in the digit string than in the enrolment utterances. Another 30% rejects were due to poor line quality, sudden changes of the volume in an utterance, or background speech. The remaining 10% could not be explained.

Since we were interested in the extent to which FRs can be predicted by the model quality, the rejected utterances were related to their speakers’ model information. It appeared that 64% of the FRs were associated with “low quality” models. Of the speakers who encountered FRs 8 (42%) had poor quality models.

## 3.5 Verification – False Accepts

In only 7 (0.6%) of the 1230 impostor attempts the system granted access to the caller. The leave-one-out enrolment algorithm does not involve tests of the model with (pseudo) impostor utterances (as did [6] and [7]). Thus, no relation with the model quality flag can be expected. The difference in FA and FR rates corresponds to the off-line optimization of the threshold before the field test. Setting the threshold gives the service provider the freedom to choose the balance between FR and FR rates which is most appropriate for the application.

## 3.6 Subjective Evaluation

98 of the 109 clients filled in the questionnaire. They were very pleased with the service, rating it with an 8 on a scale from 1 to 10. They preferred a combination of a PIN code and SV to SV only. They also suggested many opportunities for speech-secured services in the financial sector. In addition 92% of the participants would desire a real operator to backup the system in case of problems. The clients who did not complete the first five calls mentioned lack of time to as the major reason for abandoning the game. When interpreting the results of the questionnaire one should keep in mind that the application did not involve real money.

# 4 DISCUSSION

The results of our experiment show that on-line monitoring of an SV service is feasible and valuable. By keeping track of the performance of individual clients problems could be diagnosed early enough to take corrective measures. Pro-active monitoring also solved the problems some clients had with the instructions. Using the model quality flag as an extra indication of potential problems helped to speed up the diagnosis of FRs.

The question remains, however, how to standardise the actions initiated by the warning flags generated by the monitoring. An example is the sort of action that should be undertaken, either in case of a “low quality” flag or of a high proportion of rejects. Currently only the enrolment and access utterances are checked and the threshold can be adapted manually. An alternative action could be re-training of the model with new utterances or continuous unsupervised adaptation of the speaker model. Experiments in [8] showed, however, that the success of adaptation depends on the quality of the initial models.

From the questionnaire it appears that clients would like to have operator fall-back. If a client is rejected, an operator can still grant access based on additional information. They can also provide extra feedback or instructions, and in this way prevent the customer from stopping to use the service. This is similar to the use of SV in the Home Shopping Network service [10].

The model quality measure implemented in the current system is a good initial indicator of the possible problems for low-quality models. At the same time, however, a large number of false alarms are generated. Many models evaluated as poor do fine during verification. The risk of this method is that outlier tags are attached to incidental poor quality utterances where the rest of the utterances contributing to the model are fine. In this way the aim of obtaining better quality models is still achieved as the low quality utterance is left out, but the whole model receives unjustly a low quality tag.

# 5 CONCLUSIONS

Real-time monitoring as proposed in this article increases the usability of an operational speaker verification service and guarantees the required minimal level of performance. Together with the model quality check it leads to a lower number of False Rejects as the signalled problems can be diagnosed and resolved in a short span of time. Both these mechanisms contribute significantly to the preservation and satisfaction of customers.

# 6 ACKNOWLEDGMENTS

The PICASSO project is partially funded by the European Commission, as part of the Telematics Applications Action Line in the 4th Framework Programme.

# 7 REFERENCES

1. F. Bimbot et al. “Speaker verification in the telephone network: research activities in the CAVE project,” *Proc. Eurospeech*, pp. 971-974, Rhodos, 1997.
2. F. Bimbot et al. “An overview of the Picasso project research activities in speaker verification for telephone applications,” *Proc. Eurospeech,* pp. 1963-1966, Budapest, 1999.
3. E. den Os, H. Jongebloed, A. Stijsiger, and L. Boves. “Speaker verificiation as a user-friendly access for the visually impaired,” *Proc. Eurospeech,* pp. 1263-1266, Budapest, 1999.
4. J. Koolwaaij and L. Boves. “A new procedure for classifying speakers in speaker verification systems,” *Proc. Eurospeech*, pages 2355-2358, Rhodos, 1997.
5. G. Doddington, W. Ligget, A. Martin, M. Przybocki, and D. Reynolds. “Sheep, goats, lambs and wolves: a statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation,” *Proc. ICSLP*, pp. 1351-1354, Sidney, 1998.
6. J. Koolwaaij, L. Boves, H. Jongebloed, and E. den Os. “On model quality and evaluation in speaker verification”, *Proc. ICASSP*, pages 3759-3762, Istanbul, 2000.
7. J. Koolwaaij and L. Boves. “The concept of model check in speaker verification,” submitted to *Speech Communication*.
8. C. Fredouille et al. “Behaviour of a Bayesian adaptation method for incremental enrolment in speaker verification,” *Proc. ICASSP*, pages 1197-1200, Istanbul, 2000.
9. Y. Gu and T. Thomas. “An implementation and evaluation of an on-line speaker verification system for field trials”, *Proc of ICSLP*, Sydney, 1998.
10. http://www.hsn.com/