An Evaluation of "Commercial Off-The-Shelf" Speaker Verification Systems

M. Wagner[1], C. Summerfield[1], T. Dunstone[2,3], R. Summerfield[3], J. Moss[1]

[1]National Centre for Biometric Studies
University of Canberra, Australia
[2]Biometix Pty Ltd, Australia
[3]Centrelink, Australia

t.dunstone@biometix.com, ,

Abstract

An evaluation of commercial off-the-shelf speaker verification systems is reported. The performance of several systems, which were offered for testing, is analyzed against criteria designed to identify strengths and weaknesses that would determine their suitability for the use by government service agencies. Results for three text-dependent systems by Nuance, Persay and Scansoft are presented in this paper.

1. Introduction

An evaluation of commercial off-the-shelf speaker verification systems was undertaken for an Australian government agency in order to facilitate a commercial selection. This paper reports the results of the evaluation for three speaker verification systems ("engines") operating in text-dependent mode. The three engines tested were (a) the Nuance Voice Platform Version 2.0.1 ("Nuance") [1], (b) the Persay VocalPassword Build 5.0.5.0 ("Persay") [2], and (c) the Scansoft SpeechWorks Speaker Verification SDK pro 3.0 ("Scansoft") [3]. The assessment approach provides an analysis of the performance of the engines and their robustness in several areas of sensitivity. The objectives of the assessment are to ascertain the following properties:

1. the ability to handle responses in the form of numeric digits (for example, counting from one to nine), noting that previous work [4] has already reported on this task - these tests are used to validate the testing facility and methodology by comparison with the earlier results;
2. the ability to handle non-numeric responses (for example, enrolment using the user's name, testing using the same name, testing using another name), which should validate the capability of the engine to handle challenge-response-type authentication;
3. the ability to distinguish responses from same-sex siblings;
4. robustness to long-term voice variability (longitudinal testing);
5. robustness to noise with the intent of determining the "break point"[1] of the engine along with some indicative relationships between the noise used during the testing and real-world noise;
6. the ability to handle different communication channels (e.g. land line, mobile telephone and cordless telephone);
7. ease of enrolment; and
8. ease of authentication.

This study only assesses the statistical discriminability of "target speakers" from other speakers in the population ("non-target speakers") and does not attempt to assess system resistance against deliberate impostors. The question of whether it is possible to defeat speaker recognition systems by other means, such as mimicking the voice of a target speaker, is addressed elsewhere, e.g. [5].

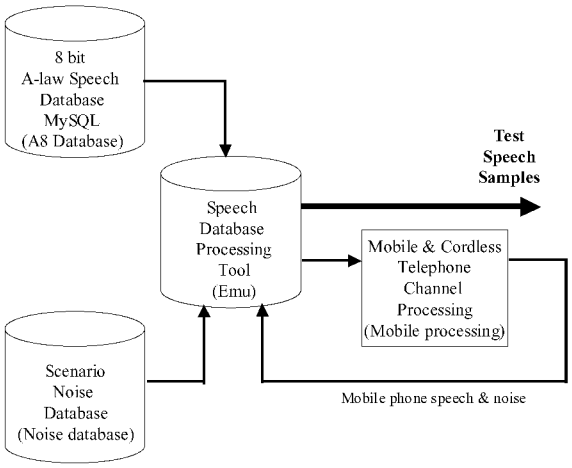


Figure 1. Speech data base for the evaluation

2. Previous Work

In 2004, the International Biometrics Group was commissioned by MasterCard to test Nuance, Persay and Scansoft speaker verification engines for authentication performance and usability relating to ease of use, convenience and confidence [4]. In that study, the engines were set up and tested in accordance with the vendors' instructions. Test subjects were enrolled and tested using numbers for Nuance, pass phrases for Scansoft's (formerly SpeechWorks) T-NETIX engine and random spoken text on Persay's text-independent engine

[1] The "break point" of a system is loosely defined as the signal-to-noise ratio at which the system becomes practically unusable in a given application.

1-4244-0472-X/06/$20.00 ©

(Persay also provide a text-dependent engine). The engines were tested for both land-line and mobile data, with longitudinal test data being collected six weeks after the enrolment data. The testing resulted in Nuance performing very well with numbers, and in Persay and T-NETIX providing similar performance with spoken text where Persay performed slightly better than T-NETIX for a delayed test.

3. Methodology

3.1. Approach

The approach taken in this study reflects the application requirements. The application requires users speaking a reference number, their name, possibly a PIN and the answer to a secret question.

A corpus of speaker data was collected with the following fields: a customer reference number, the numbers from 1 to 9, the user's name, a "friend's name" (to provide the non-numeric common material) and other information. This corpus provided the reference information to benchmark the performance of each of the speaker verification engines. It enabled the raw performance of each of the engines to be determined in the context of the given application.

3.2. Test Configuration

There are three components to the evaluation environment:

- the speaker data corpus;
- the verification engine under evaluation together with its generic speaker models; and
- the verification results data base.

The speaker corpus was collected from government agency staff. The voice samples were recorded over the public switched telephone network using interactive voice response (IVR) equipment and speech samples were compiled into a digital format (8kHz 8-bit A-law format). To validate testing, some samples were also collected from mobile-telephone and cordless-telephone handsets.

The majority of speech samples were collected from call centre agents. Of a total of 322 participants, 231 (72%) were female and 91 were male (28%), the proportions approximately reflecting the typical call distribution for the government agencies in question. Callers provided an initial-enrolment set consisting of three repetitions of the same material recorded in a single session, a second data set recorded one week later and, for a subset of 74 speakers, a third data set recorded about 9 months later. Samples were also collected from same-sex siblings, from father-son and mother-daughter combinations, and from one pair of female identical twins. Demographic information about each speaker was collected, including sex, age range and ethnic background.

Of the data recorded, the numbers from 1 to 9, the user's name and the friend's name were each used for the text-dependent testing, which is reported here.

The data were processed as required and deposited on a server in a directory structure ready for the speaker verification engines to pick up and process, as shown in Figure 1. The engines were provided with separate command scripts for the enrolment and verification phases, which de-identified the voices being used and enabled a matrix test arrangement between enrolment and verification data.

3.3. Test Data

A matrix of tests was prepared for each of the test procedures, using the numbers from 1 to 9, and the friend's name as common samples that were provided by all speakers. Noise was added to the data at signal-to-noise ratios (SNR) of 0dB and -20dB. Different noise sources were used, including recordings of street noise, shopping-centre noise, office noise as well as computer-generated white noise. The matrix of samples was further processed through different GSM mobile-phone codecs, operating at 12.2kb/s, 6.7kb/s and 4.75kb/s.

Each text-dependent engine was employed to enroll speakers using the three enrolment samples and then to test speakers according to the matrix of test data, including the 9-month-delayed test data and the sibling test data. Baseline tests with "clean", i.e. noise-free land-line, data in a matrix of 322 speakers against themselves and against all other speakers of the same sex were conducted first and detection-error trade-off (DET) curves plotted for the combined male and female false-acceptance (false alarm) and false-rejection (miss) errors.

One series of tests was conducted using the digit set 1-9 for enrolment and testing, and another using the common friend's name for enrolment and testing. All these tests fall within the text-dependent speaker verification paradigm. Two additional test were conducted, one in which speakers enrolled with their own names, but were tested against other speakers who spoke *their* own different names, and another test in which speakers enrolled with their own names and were then tested against themselves speaking the friend's name. Both these additional tests were designed to reveal whether the different engines could be used in a prompted-response authentication setting.

Since a test run for a single noise and channel condition required several hours of computing time on our system, the baseline tests were repeated with a subset of 10 same-sex non-target speakers (NS) for each target speaker (TS). Each NS subset was selected from the full set of (231-1) female or (91-1) male NS, respectively, by using a pseudo-random number generator. It was established that the tests using non-target speaker subsets yielded similar results for both the numbers test and the names test for the operating range of interest. Subsequent tests, including the noise and time-delay tests, were then conducted with the 10-NS subsets, reducing computing time for a single-condition test run to approximately 20 minutes.

Model adaptation was not permitted as it would skew the evaluation results due to the relatively small number of speakers. It was confirmed that no model adaptation was taking place in the engines under evaluation by processing speakers in different orders and comparing the corresponding results.

3.4. Failure-to-Enroll and Failure-to-Acquire

Since the objects of the evaluation were commercial off-the-shelf systems, the experiments were conducted without ac-

cess to the "inner workings" of the systems, essentially treating the systems as "black boxes", which take an identity claim and a speech file as input and produce a numerical "score" as output. While input specifications were identical for the three systems under evaluation, namely speech waveforms in a certain format, the output specifications were not completely identical. In particular, the Nuance system produced a "failure-to-enroll" (FTE) output, indicating a refusal by the system to build a speaker model in response to enrolment-speech input, and both Nuance and Persay produced a "failure-to-acquire" (FTA) output, indicating a refusal by the system to make an accept-reject decision in response to test-speech input.

Nuance produced an FTE response to two of the three digits-one-to-nine enrolment files for one male speaker. Nuance also produced FTE responses for either one or two of the three mobile-phone enrolment files for the same male speaker and for two female speakers. In all these cases speaker models were built nonetheless and any potential influence on the results was ignored.

The situation is somewhat different for FTA responses. For an equitable comparison of systems, an FTA response to a TS test may be interpreted as a false rejection (miss) and an FTA response to an NS test may be interpreted as a false acceptance (false alarm). Where DET curves are shown in the following sections, the curves are drawn as if no FTAs had occurred and the percentages of FTAs are noted separately in Table 2 of the Appendix.

These figures can then be interpreted as a corresponding upward shift of the DET curve for a proportion of TS-FTAs and as a corresponding right shift of the DET curve for a proportion of NS-FTAs.

4. Results

4.1. Tests on the Digits 1 to 9

The results for the baseline test on the digit sequence 1-9 for the full set of non-target speakers are shown in Figure 2. These tests relate to Objective 1 of Section 1. It can be seen that over an operating range near the equal-error point, where the DET curve intersects the diagonal, the Nuance engine outperforms the Persay engine with the Scansoft engine a somewhat distant third. The combined equal-error rates vary from 0.91% for Nuance to 4.24% for Scansoft.

In comparison, Figure 3 shows the corresponding results for the subset of 10 non-target speakers and it can be seen that over the operating range of interest in the vicinity of the equal-error point, the DET curves for the three engines exhibit similar behavior as well as similar values. The combined equal-error rates vary from 0.93% for Nuance to 4.31% for Scansoft. The equal-error rates for both the NS full set and the NS subsets are shown in Table 1.

4.2. Tests on Names

To examine Objective 2 of Section 1, three test sets were performed using name utterances. The first of these performed a baseline test on the friend's name recorded by each of the test speakers. Figures 4 and 5 show the DET curves for the full set of NS and for the NS subset, respectively. For this test the three engines perform very similarly throughout the operating range of interest.

The comparison between the full NS set and the NS subset shows once again that the NS subset yields very similar error rates for an operating range in the vicinity of the equal-error point. The closeness of the two result sets provided justification to conduct the time-consuming test series for the different channel and noise conditions on the same NS subsets as were used for the two baseline test comparisons.

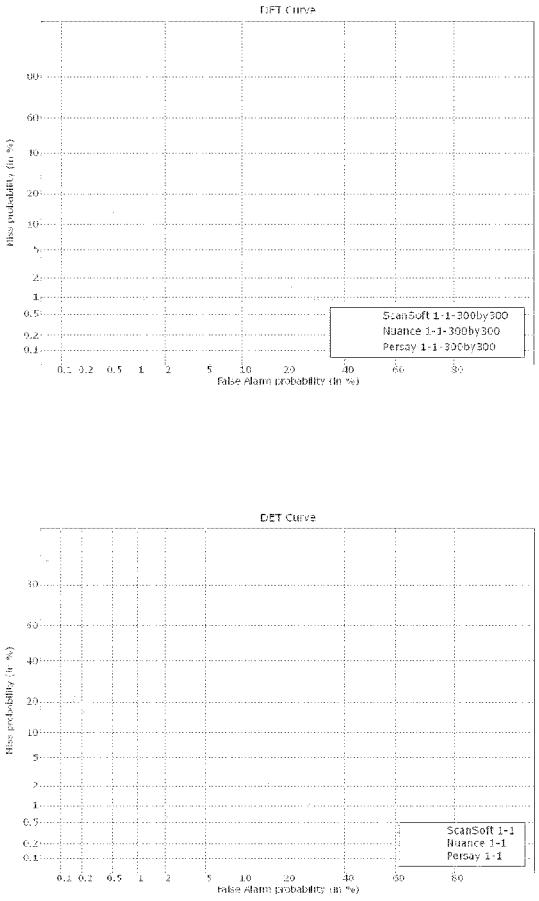


Figure 3. Digits through landline, NS subset

Table 1: Equal-error rates for the baseline digit tests

Condition	Nuance	Persay	Scansoft
Female NS Full Set	0.84%	1.71%	4.60%
Male NS Full Set	1.11%	1.12%	3.30%
Combined NS Full Set	0.91%	1.86%	4.24%
Combined NS Subset	0.93%	1.73%	4.31%

A different test, which is somewhat text-independent, has the NS saying a different name from the TS (their own names). Figure 6 shows the DET curves for that test, which confirm that an "impostor" who will use the wrong phrase for authen-

tication will be rejected with greater certainty than an impostor who uses the correct phrase.

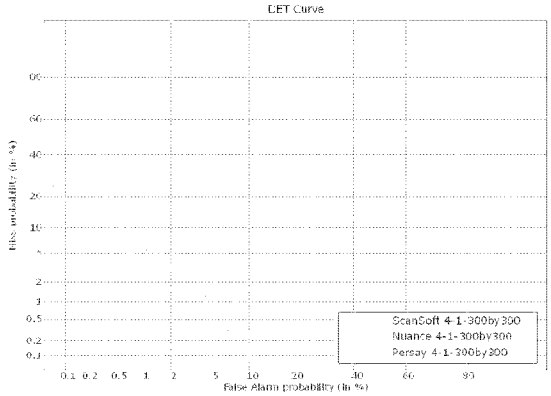
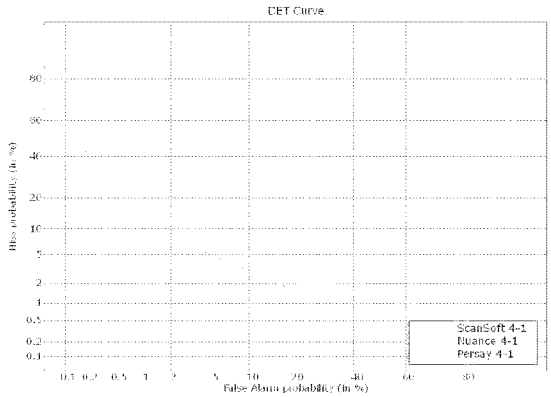


Figure 4. Name through landline (TD), NS full set.



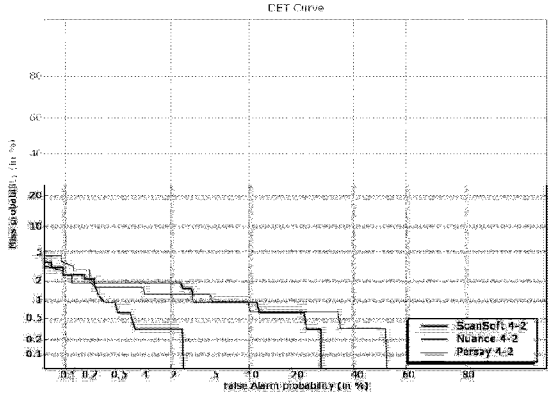


Figure 6. Distinct name through landline (TI), NS subset.

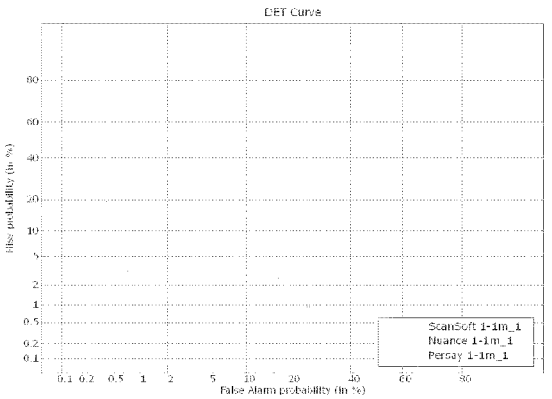
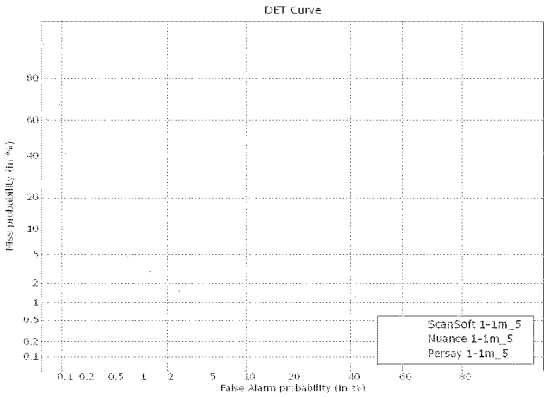


Figure 7. Digits through GSM (12.2kb/s), NS subset.



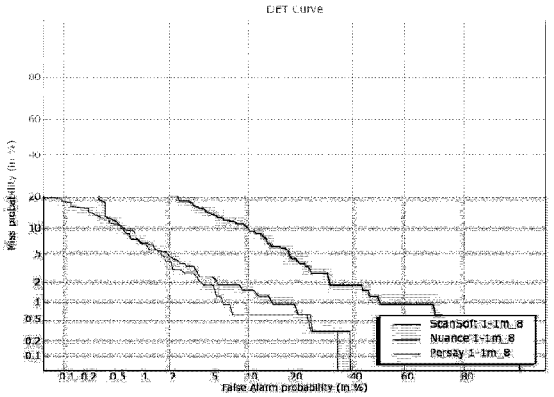


Figure 9. Digits through GSM (4.75kb/s), NS subset.

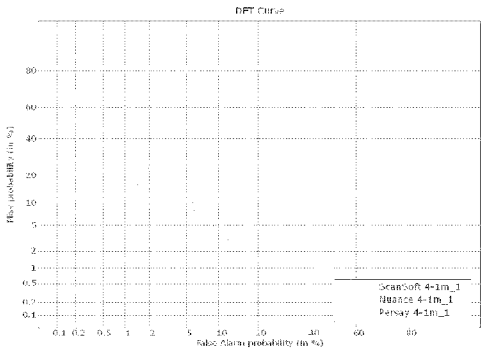


Figure 10. Names through GSM (12.2kb/s), NS subset.

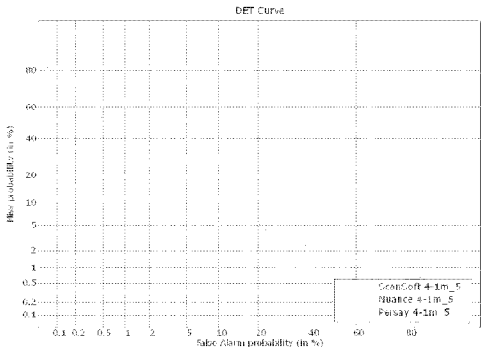


Figure 11. Names through GSM (6.7kb/s), NS subset.

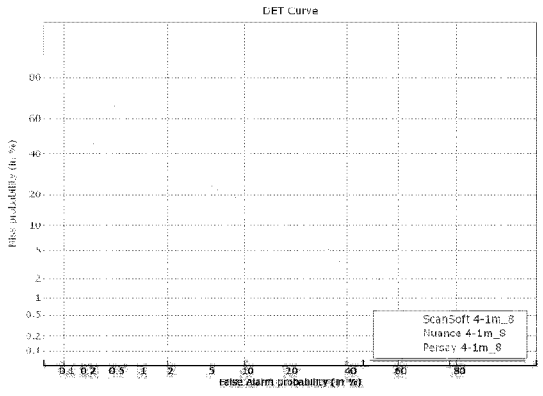


Figure 12. Names through GSM (4.75kb/s), NS subset.

4.3. Mobile Telephone Tests

In this series of tests, the test samples for both digits and names were channeled through different GSM modems [7] before being tested against the speaker models that were constructed from the landline data described in sections 4.1 and 4.2.

In decreasing order of mobile channel quality, the data were filtered though a 12.2 kb/s modem, a 6.7 kb/s modem and a 4.75 kb/s modem.

Figures 7, 8 and 9 show the results for the digit data after passing through the 3 different mobile phone quality filters. These curves need to be compared with the digit result for full landline quality in Figure 3.

As is expected, the error rates increase with diminishing signal quality. In the two higher-quality conditions, the Nuance engine performed slightly better than Persay, while in the lowest-quality condition Persay is slightly ahead of Nuance. In all three GSM conditions, the Scansoft engine performs significantly worse than both of the others.

Figures 10, 11 and 12 show the results for the names data after passing through the 3 different mobile phone quality filters. These curves need to be compared with the digit result for full landline quality in Figure 5. In this case, the error rates increase only slightly with diminishing signal quality and the performance of all three engines is fairly close with Scansoft only slightly behind the other two.

4.4. Noise Tests

The performance of the three systems in noise was analyzed by adding to the test samples synthesized white noise, recorded office noise, city noise and shop noise, for signal-to-noise ratios of 0dB (i.e. noise at the same level as the signal) and -20dB (i.e. noise at a 100 times higher level than the signal). Figures 13 to 16 show the performance of the three systems when the four noise types are added to the test samples at 0dB. The results show that white noise is a worst case scenario compared with real-life noise environments. The results also show the Nuance and Persay engines rejecting many of the samples with -20dB-SNR white noise, while the Scansoft engine provides questionable results instead. These results suggest that without appropriate FTA rejection the Scansoft system may be vulnerable to an attack using loud

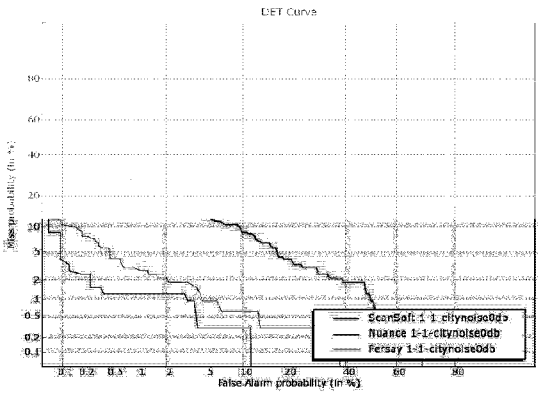
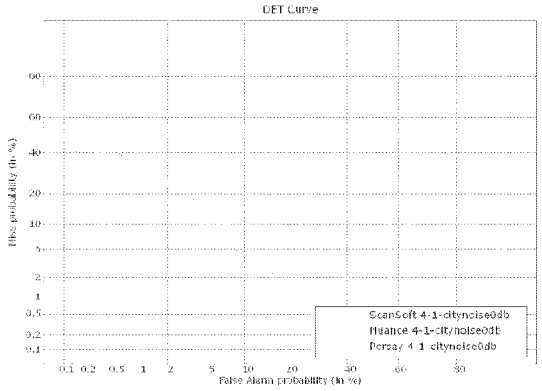
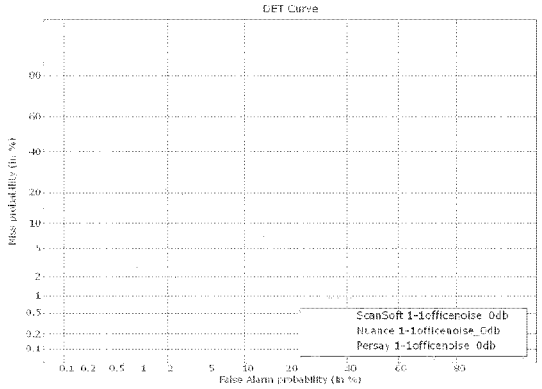


Figure 13. Digits + city noise (0dB), NS subset.





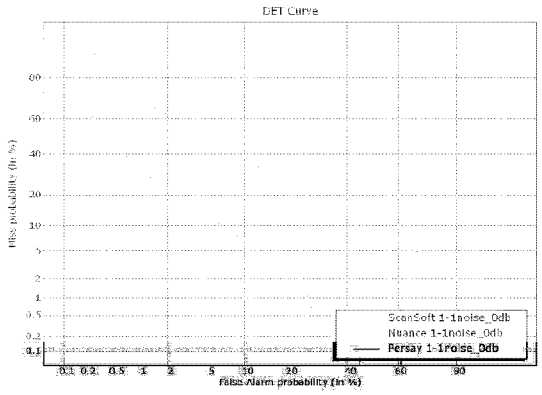


Figure 16. Digits + white noise (0dB), NS subset.

For noise added to the test samples at a signal-to-noise ratio of -20dB, the Nuance system retains good performance while the performance of the Persay and Scansoft systems deteriorates markedly as is shown in Figure 17.

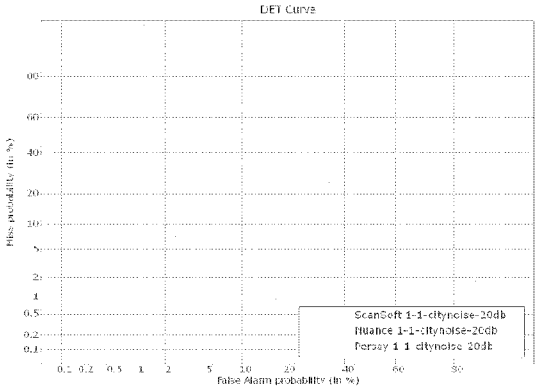


Figure 17. Digits + city noise (-20dB), NS subset.

In general, the results show a high level of resilience to background noise with Nuance consistently outperforming Scansoft for office, city and shop noise. Persay performed rather well with 0dB SNR, but slightly worse than Nuance for -20dB SNR.

4.5. Longitudinal Tests

The results for the longitudinal tests on the digits 1-9, spoken by a subset of 73 speakers some 9 months after enrolment, are shown in Figure 18. These DET curves need to be compared with Figure 3. The Nuance and Persay speaker verification engines performed better than the Scansoft engine, but the performance of all three engines deteriorated compared with the baseline test, which used data recorded between several days and one week after the initial enrolment sample.

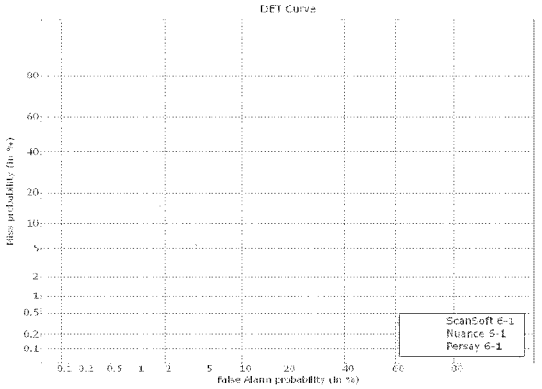


Figure 18. Digits for TS subset after 9 months, NS subset.

It should be borne in mind, however, that the longitudinal results are based on a smaller sample of speakers.

4.6. Enrolment Tests

To meet Objective 7, reduced-enrolment tests were performed. The Scansoft system does not perform in that condition and Figure 19 shows the DET curves for two-session enrolment for the Nuance and Persay systems. Similarly, Figure 20 shows the DET curves for one-session enrolment. These DET curves need to be compared with the baseline three-session enrolment results, which are shown in Figure 3. Nuance performed similarly for one and three enrolment sessions, but a little worse for two enrolment sessions, while Persay performed very similarly for all three enrolment conditions.

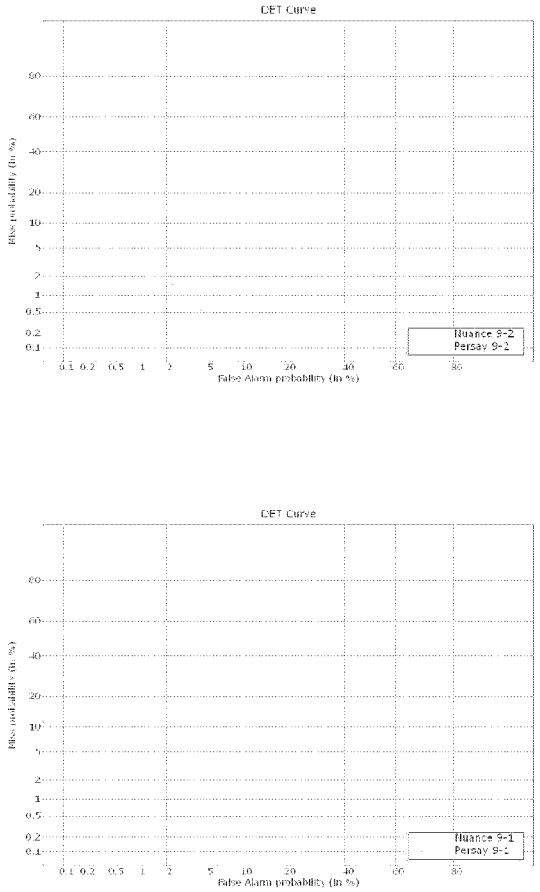


Figure 20. Digits for 1-session enrolment, NS subset.

5. Conclusions

This paper reports on the performance of three text-dependent speaker verification engines by Nuance, Persay and Scansoft in different experimental conditions, including spoken digits and spoken names, landline and mobile channels of differing quality, different noise conditions, different time spans between enrolment and tests and different numbers of enrolment sessions. For most of these conditions, the Nuance engine performed best, closely followed by the Persay engine. An equitable comparison between the three systems was achieved to the extent possible given that the experimenters had no access to the inner workings of the systems under evaluation. An element of uncertainty remains because two of the systems produce Failure-to-Enroll and Failure-to-Acquire responses, which must be assumed to reduce the net error rates for those systems by removing some "difficult" samples from the statistics.

6. Acknowledgements

The authors wish to thank Biometix Pty Ltd for the provision of its Performix biometric analysis tool, and Centrelink for the provision of the speaker data corpus.

7. References

[1] (accessed on 16 April 2006)

[2] (accessed on 16 April 2006)

[3] After the merger of Scansoft and Nuance the web reference is now redirected to where information on the Speechworks speaker verification system seems no longer to be available (accessed on 16 April 2006).

[4] K. Wadhwa, Voice verification: technology overview and accuracy testing results, in Biometrics 2004, International Biometric Group, London, UK, 2004.

[5] Y.W. Lau, M. Wagner & D. Tran., Vulnerability of Speaker Verification to Voice Mimicking, Proc. Int. Symp on Intelligent Multimedia, Video and Speech Processing, Hong Kong, pp 145-148, 2004.

[6] Large scale evaluation of automatic speaker verification technology, Centre for Communication Interface Research, University of Edinburgh, 2000.

[7] Miksoft, Mobile AMR Converter, http://www.miksoft.net/mobileAMRconverter.htm (accessed on 16 April 2006)

8. Appendix

The Failure-to-Acquire (FTA) rates as percentages of the number of tests conducted are tabulated in Table 2 for all DET curves in this paper.

Figure	**Nuance**	**Persay**
Fig. 2	0.6%	5.6%
Fig. 3	0.6%	1.1%
Fig. 4	-	31.8%
Fig. 5	-	3.9%
Fig, 6	-	38.5%
Fig. 7	0.7%	1.0%
Fig. 8	0.7%	1.0%
Fig. 9	0.4%	-
Fig. 10	-	3.1%
Fig. 11	-	3.5%
Fig. 12	-	3.6%
Fig. 13	0.2%	2.6%
Fig. 14	-	4.3%
Fig. 15	0.7%	3.6%
Fig. 16	10.5%	10.2%
Fig. 17	2.7%	8.8%
Fig. 18	-	1.1%
Fig. 19	0.7%	0.9%
Fig. 20	0.6%	8.0%

Table 2. Percentage of FTA responses for the Nuance and Persay systems in the different experiments.