



1 PERSAY TECHNOLOGY EVALUATION RESULTS (2006)

1.1 Introduction

During July and August, the National Centre for Biometric Studies at the University of Canberra undertook an evaluation of Persay's more recent voice authentication technologies and compared the performance of these technologies with the evaluation performed in 2005 to confirm any improvements in performance and to quantify these improvements.

Two 2006 technologies from Persay were evaluated:

- The 2006 release of Persay's VocalPassword™ 5.2.0 standard product;
- and
- VocalPassword™ 6.0 alpha (experimental version).

The National Centre for Biometric Studies applied the same evaluation methodology using the same speech database to enable a direct comparison between the results obtained in the 2005 evaluation and the results obtained in this evaluation. Some alternation have been made to the algorithm used to compute the EER from the false accept and false reject characteristics; leading to slight different EER measures. Appendix A contains information on the EER calculation.

1.2 Evaluation Results

1.2.1 Test Case 1-1 Counting 1to9

1.2.1.1 Baseline Results

Base line results for each of the vendor's engines are given in Figure 1.

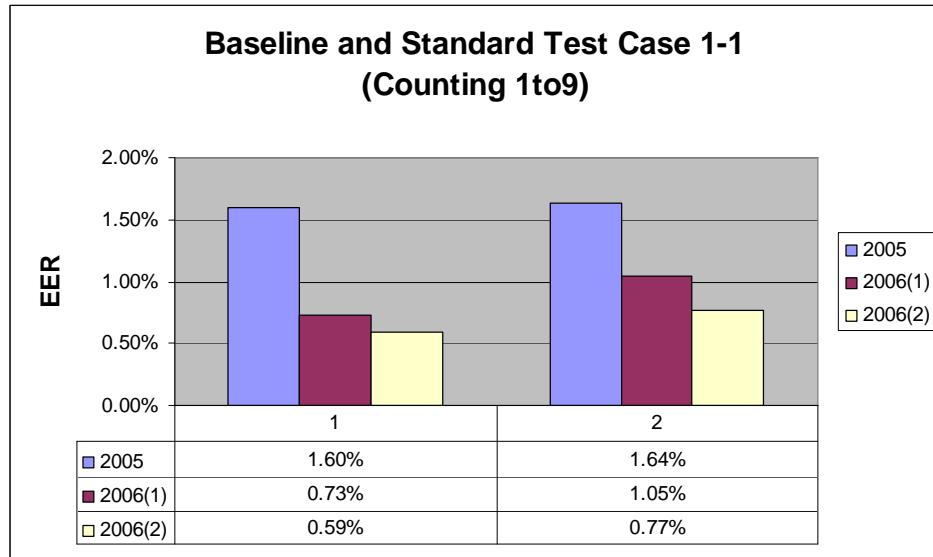
Two baseline results were produced: one for a full 300 by 300 test (1); the other for a standard 300 by 11 test (2). The 300 by 300 is highly redundant, where each enrolled client is tested against every other client. This test provides a complete examination of the performance of the various engines, based on the full speech database available for the test. This involved enrolment of 300 identities verifying against themselves and 299 impostors. In total, this involved processing some 90,000 speech files (that is, 300 clients x (1 client + 299 impostors)). Given the hardware, this took a considerable amount of time, ranging from a few hours for some of the faster engines to a few days for the slower ones.

Exhaustive impostor testing is actually not necessary to obtain an accurate determination of an engine's false accept performance. A 300 by 11 test, involving the same 300 identities being enrolled and tested against their own identity plus 10 randomly selected impostors of the same gender was devised. This test involved computing verification results for 3,300 speech samples, which given the performance of the hardware, provided a tractable solution to the limited computation bandwidth available for the evaluation. This 300 by 11 test was established as a standard test configuration for each of the telecommunications and noise testing scenarios.

Theoretically, both tests should give very similar results. To qualify this, a full 300 by 300 test was performed to create a baseline, followed by the reduced test of 300 by 11 to confirm correspondence.



Figure 1 shows the Equal error Rate (EER) performance of Persay’s 2005 technology, Persay 2006(1) and Persay 2006(2) on the 300 by 300 test and 300 by 11 test.



1= 300 by 300 Baseline Test Case 1-1

2= 300 by 11 Standard Test Case 1-1

Figure 1. Baseline Performance Counting 1to9

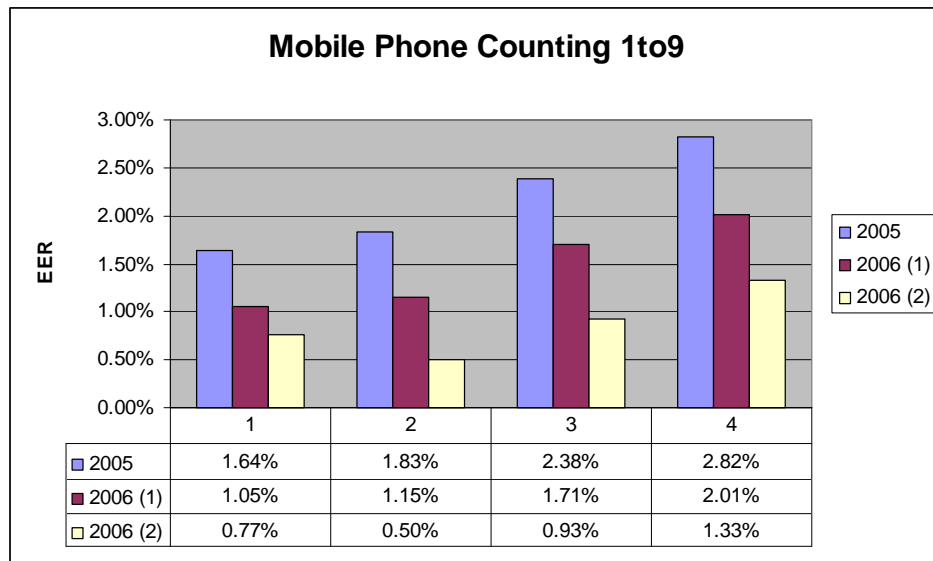
1.2.1.2 Observations

In this examination, Persay 2005 and 2006(1) showed good correspondence between the full 300 by 300 test and the 300 by 11 test, with 2005 changing by 0.04% and Persay 2006(1) changing from 0.73% ERR to 1.05%. In the case of Persay 2006(2) the measured EER increased from 0.59% for the full test to 0.77% for the reduced impostor test.

Examining the relative performance of the 2005 technology compared to the 2006(1) technologies and using the EER performance observed for test 2 (300 by 11) the technology improved from EER=1.64% to EER=1.05% an improvement of 36% on the 2005 performance. Comparing the performance of the 2005 technology to the 2006(2) technology, the performance improvement was EER=1.64% to EER=0.77%, an improvement of 53% on the 2005 technology. In effect 2006(2) technology offers just over twice the performance of the 2005 technology on this task on this database based on EER performance.

1.2.1.3 Mobile Telephone Performance

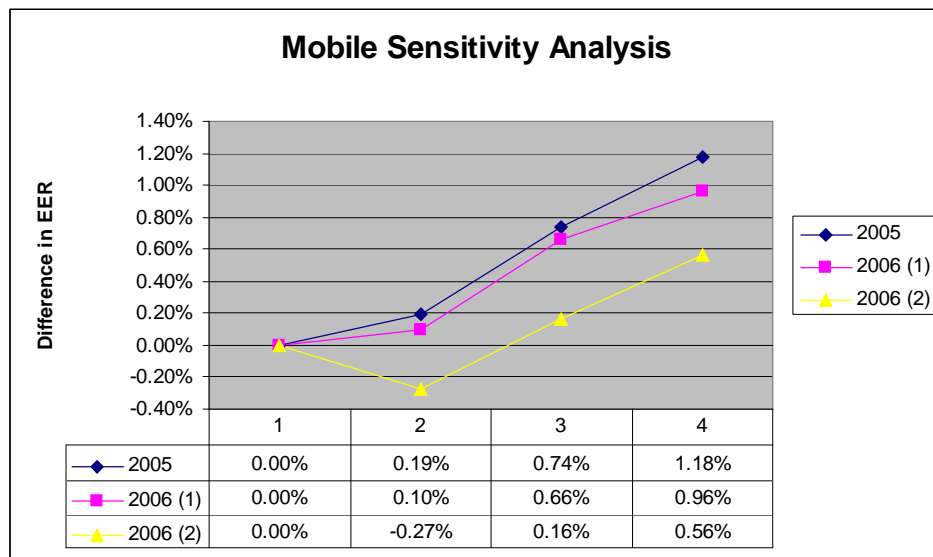
Figure 2 shows the performance of each of the Persay 2005, 2006(1) and 2006(2) engines as the mobile telephone encoding rate was lowered from 12.2 kb/s, 6.7 kb/s to 4.75 kb/s. This was compared against the base rate computed above for land line, which is at a data rate of 64 kb/s.



1=64 kb/s 2=12.2 kb/s 3=6.7 kb/s 4=4.75 kb/s

Figure 2. Performance in Mobile Telephone Communications Networks

The Persay 2005 and 2006(1) engines showed an increase in the EER performance as the data rate used in mobile telephone networks was decreased. Persay 2006(2) showed a small drop in EER performance for mobile telephone encoded speech data. Figure 3 shows the relative performance of each of the engines in this test as the communications data rate is decreased.



1=64 kb/s 2=12.2 kb/s 3=6.7 kb/s 4=4.75 kb/s

Figure 3. Relative Performance in Mobile Telephone Communications Networks

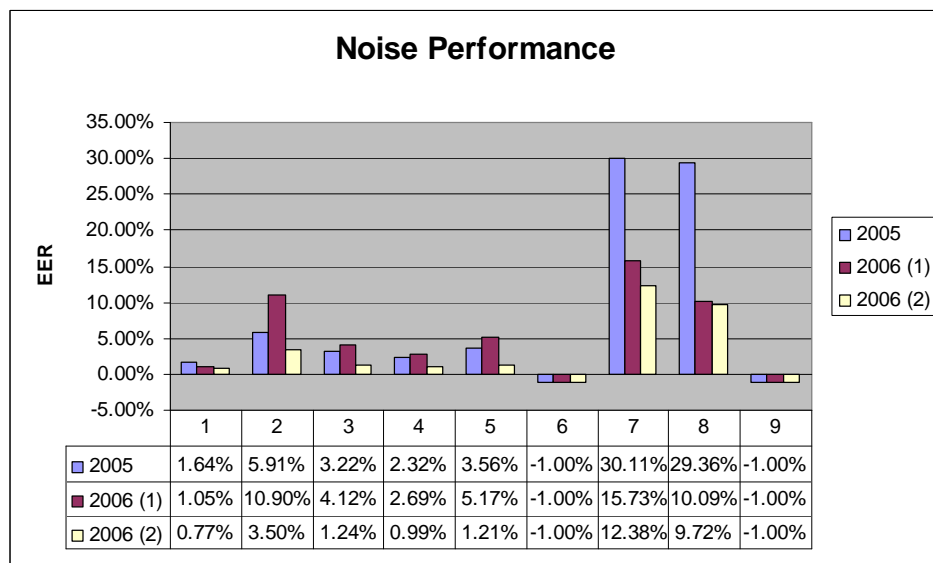


1.2.1.4 Observations

Persay 2005 and 2006(1) engines showed an increase in the EER performance as the data rate using in mobile telephone networks was decreased. Persay 2006(2) showed a small drop in EER performance for mobile telephone encoded speech data. Persay 2006(1) engine shows slightly more robustness in this environment with an increase of 0.96% EER performance between the baseline performance and the lowest data rate of 4.75kb/s, compared to Persay 2005 engine that showed increases of 1.18%. Persay 2006(2) however showed considerable higher levels of robustness when moving from a landline to mobile network with only 0.56% increase in EER performance.

1.2.1.5 Noise Performance

Figure 4 shows the performance of each of the Persay engines, 2005, 2006(1) and 2006(2) in different noise scenarios (white noise, office noise, city noise and shop noise) and how the increase of noise from 0dB to 20dB affects the performance relative to the baseline



1 = Baseline Performance

2 = 0dB White Noise 3 = 0dB Office Noise 4 = 0dB City Noise 5 = 0dB Shop noise

6 = -20dB White Noise 7 = -20dB Office Noise 8 = -20dB City Noise 9 = -20dB Shop noise

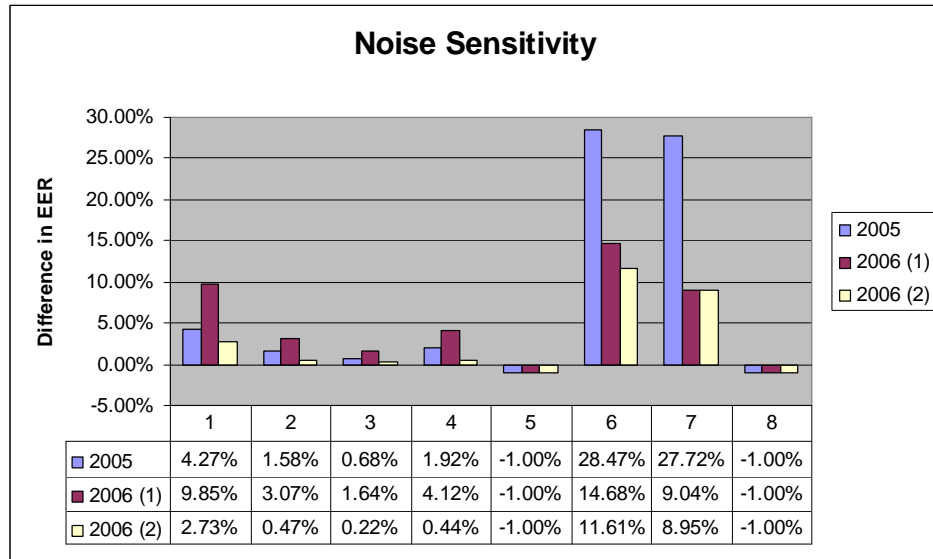
(-1.00% indicates FTA returned)

Figure 4. Noise Performance

All engines showed an increase in the EER with increasing noise. However, the EER performance varied considerably for different engines and noise conditions with the 2005 and 2006(2) engines showing considerably higher levels of robustness at low levels of noise but with Persay 2006(1) and Persay (2) showing higher levels of robustness at higher levels of noise.



Figure 5. shows the sensitivity (i.e. the difference in the EER relative to the baseline performance) as the signal-to-noise ratio is decreased from 0 dB to -20 dB.



1 = 0dB White Noise 3 = 0dB Office Noise 5 = 0dB City Noise 7 = 0dB Shop noise
 2 = -20dB White Noise 4 = -20dB Office Noise 6 = -20dB City Noise 8 = -20dB Shop noise
 (-1.00% indicates FTA returned)

Figure 5. Noise Sensitivity Analysis

1.2.1.6 Observations

All engines displayed an increase in EER performance and some sensitivity to the signal-to-noise ratio, but to differing degrees. An analysis of the noise sensitivity shows that both Persay 2005 and 2006(2) engines were more robust to noise than Persay 2006(1), particularly in the office and city noise environments. At high noise levels (-20 dB), specifically in the white noise test and shop noise test, all engines rejected all utterances submitted, returning a 100% FTA (Fail to Acquire) result. This demonstrates that at high noise levels all Persay engines reject the input signal as outside its processing parameters.

Different engines show quite different sensitivities to the different noise scenarios. Whilst white noise was the most problematic for all engines, the Persay 2005 and 2006(2) engines showed improved performance in office, city and shop noise scenarios. However, in very high noise scenarios Persay 2006(1) and Persay 2006(2) engines showing much higher levels of resilience to non-speech high noise scenarios. This is most likely attributed to the 2006(1) and 2006(2) engines' improved ability to discriminate between speech and non-speech signals.



1.2.2 Test Case 4-1 – Name Verification

1.2.2.1 Baseline Results

Base line results were produced for condition 1 (different speakers, same name). Two baseline results were produced: one for a full 300 by 300 test (column 1 in figure 6); the other for a standard 300 by 11 test (column 2 in figure 6). As with the Ito9 counting test, the 300 by 300 tests provided a complete examination of the performance of the different engines. This involved enrolment of 300 enrolled identities, and verifying against those identities and 299 impostors for each client. In all, these tests processed some 90,000 speech files which took considerable time for the engines to compute. A 300 by 11 test, involved the same 300 identities being enrolled and tested against their own identity, plus 10 randomly selected impostors each saying the same name as the client and of the same gender. This test involved computing verification results for 3,300 speech samples, which given the performance of the hardware, provided a methodology that could be computed with the timeframe of the experiment. This 300 by 11 test was established as a standard test methodology for each of the telecommunications and noise testing scenarios.

Figure 6 shows the EER performance of Persay 2005 and Persay 2006(1) on the 300 by 300 test (results 1) and Persay 2005, Persay 2006(1) and Persay 2006(2) the 300 by 11 test (results 2).

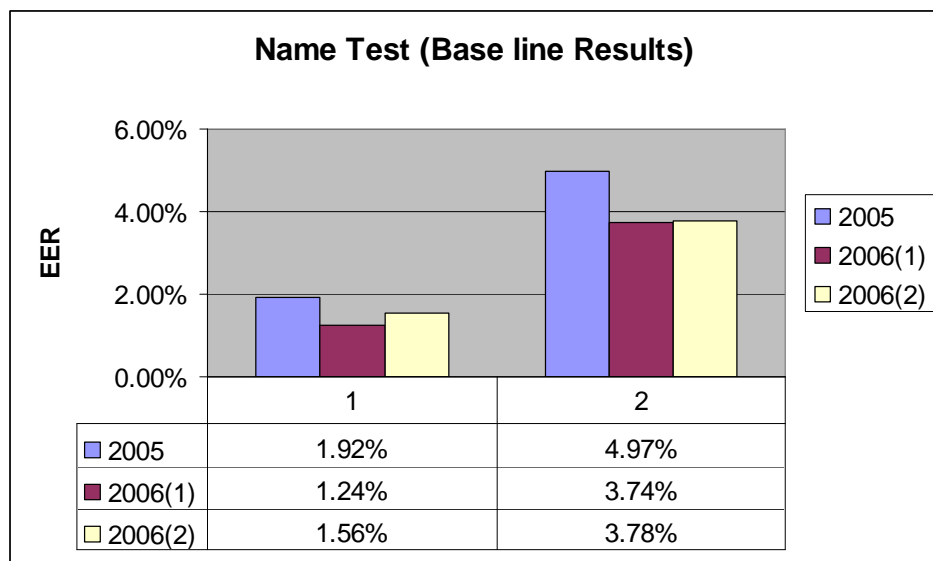


Figure 6. Baseline Tests 300 by 300 (1) and 300 by 11 (2) name speech samples

1.2.2.2 Observations

Persay 2006(1) and Persay 2006(2) showed much better performance than Persay 2005, with both technologies showing 25% increase in performance over the performance of Persay 2005.

It is also interesting to compare the EER performance of each of the engines for name verification against the corresponding Ito9 counting verification test (1-1 series) and the variation from 300 by 300 test and the 300 by 11 test. Table 1 provides a summary of this comparison, with the region in grey showing the difference in EER rates between 300 by 300 tests and 300 by 11 tests.



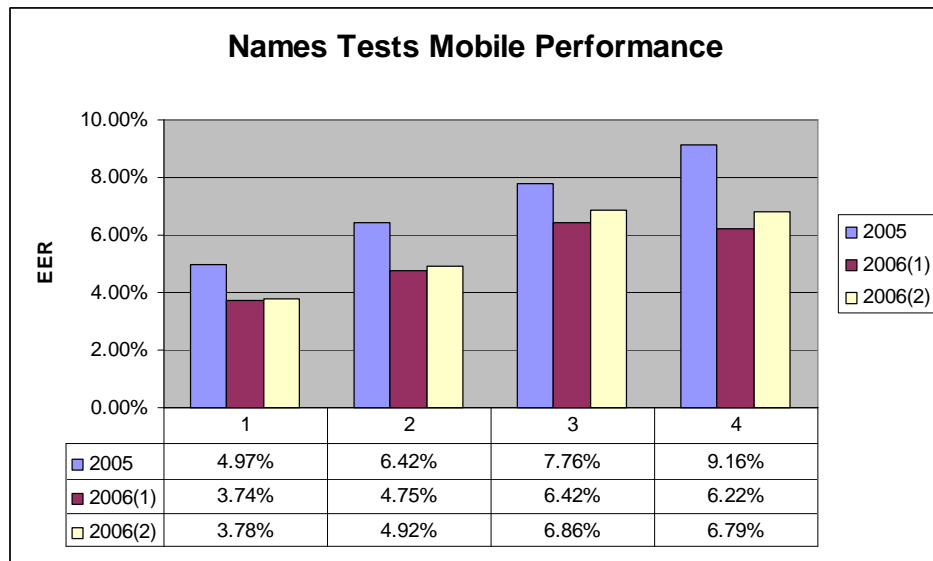
Engine	300by 300 Counting	300by11 Counting	300by300 Names	300by11 Names
2005	1.60%	1.64%	1.92%	4.97%
2006(1)	0.73%	1.05%	1.24%	3.74%
2006(2)	0.59%	0.77%	1.56%	3.78%
2005		0.04%		3.05%
2006(1)		0.32%		2.50%
2006(2)		0.18%		2.22%

Table 1. Comparison of performance of verification engines on counting 1to9 and name verification

This comparison between the 300 by 300 test and 300 by 11 tests shows that in both Counting 1to9 and Names tests the EER performance measured in the 300 by 300 tests were better than the 300 by 11 tests. This may be attributed to the sensitivity of the technologies to discriminate between different genders. In the 300 by 300 tests, enrolled speakers were tested against each other irrespective of gender. In this test male speakers are “imposed” against female speakers and vice versa. In the 300 by 11 tests, however, only same gender speakers were selected as impostors. Thus these tests are more demanding, as the technology is required to discriminate on speech samples produced by the same gender. Where a larger amount of speech is available (as in the Counting 1to9 tests) to performance the technology varied only slightly between the 300 by 300 and 300 by 11 tests, whilst in situations where only a short sample of speech was available (Names tests) the variation between the 300 by 300 and 300 by 11 tests were a lot wider. An analysis of relative performance (see Appendix A) shows that when you compare relative performance of the technologies, the relative performance shows a lower degree of variation, indicating the performance improvement of the technologies relative to each other was relatively stable between the 300 by 300 and 300 by 11 tests.

1.2.2.3 Mobile Telephone Performance

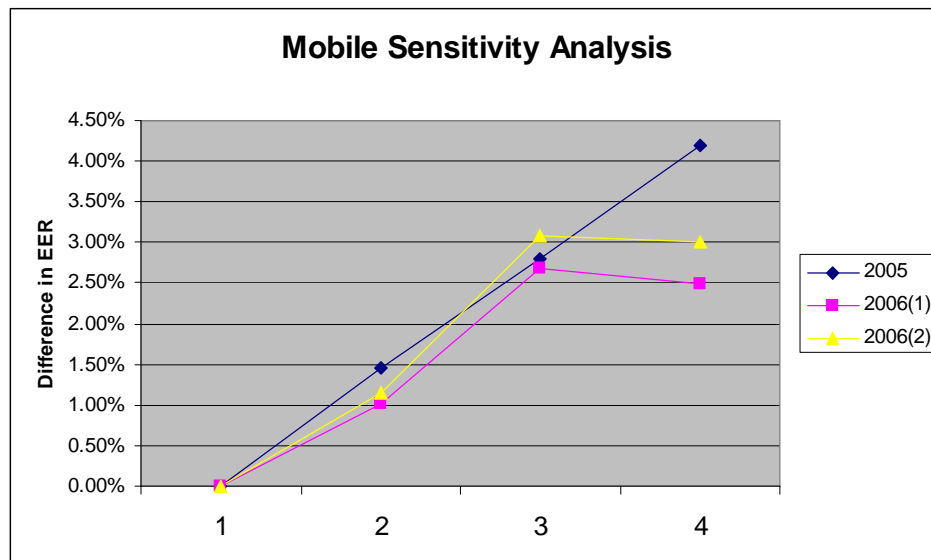
Figure 7 shows the performance of each of the Persay 2005, 2006(1) and 2006(2) engines with the names test condition 1, (different speaker; same name) as the mobile telephone encoding rate was lowered from 12.2 kb/s, 6.7 kb/s to 4.75 kb/s, (conditions 2, 3 and 4). This was compared against the baseline performance for land line computed above, which is at a data rate of 64 kb/s (condition 1)



1=64 kb/s 2=12.2 kb/s 3=6.7 kb/s 4=4.75 kb/s

Figure 7. Mobile telephone sensitivity tests

Figure 8 shows the sensitivity of the engines as the mobile communications data rate is decreased relative to the baseline performance.



1=64 kb/s 2=12.2 kb/s 3=6.7 kb/s 4=4.75 kb/s

Figure 8. Mobile Telephone Sensitivity Analysis

1.2.2.4 Observations

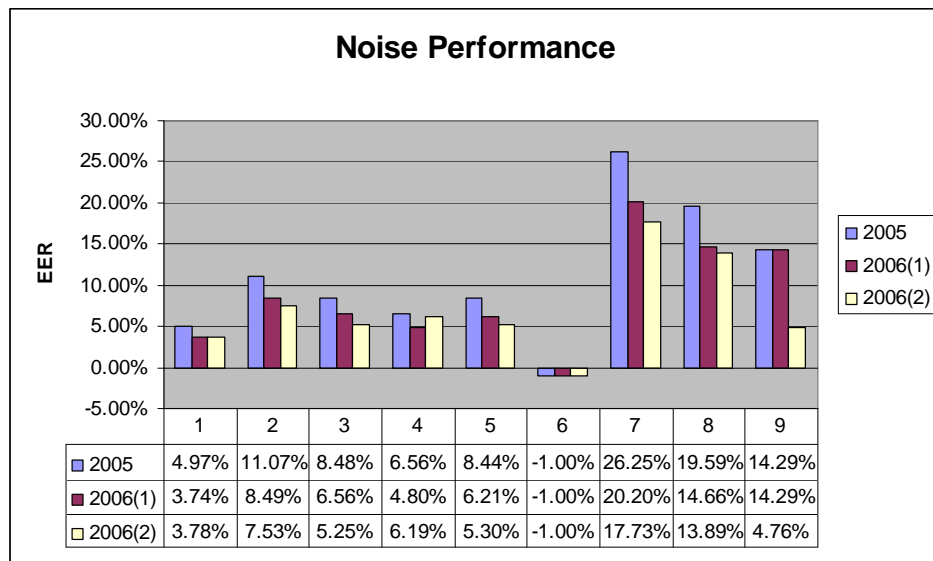
The analysis shows that all engines exhibit increased EER performance as the mobile telephone data rate is reduced. All engines showed comparable robustness in mobile telephone environments except at the lowest data rate, in which case Persay 2006(1) and



Persay 2006(2) engines exhibited a significantly higher level of robustness compared to Persay 2005.

1.2.2.5 Noise Performance

Figure 9 shows the performance of each of the Persay engines in different noise scenarios (white noise, office noise, city noise and shop noise) and as the noise is increased from 0 dB to 20 dB relative to the baseline performance.



1 = Baseline Performance

2 = 0dB White Noise 3 = 0dB Office Noise 4 = 0dB City Noise 5 = 0dB Shop noise

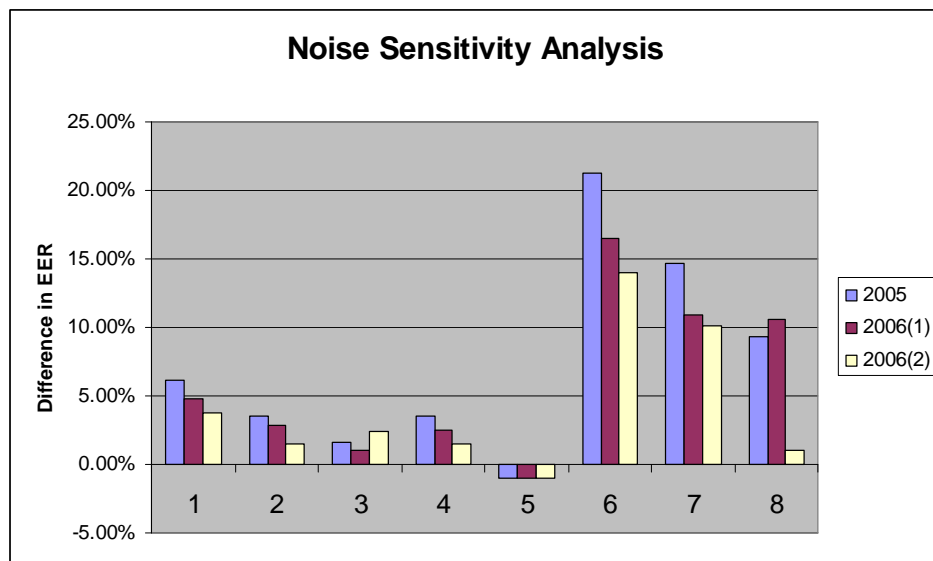
6 = -20dB White Noise 7 = -20dB Office Noise 8 = -20dB City Noise 9 = -20dB Shop noise

(-1.00% indicates FTA returned)

Figure 9. Noise Performance for name verification

1.2.2.6 Observations

All engines showed an increase in the EER with increasing noise. However, under all conditions Persay 2006(1) and 2006(2) exhibited higher level of robustness against all noise conditions compared to Persay 2005 engine. In very high noise conditions (condition 6) all engines exhibited a fail safe mode, returning 100% FTA (Failure to Acquire) errors. Figure 10 shows the sensitivity, that is the difference in the EER relative to the baseline performance; as the signal-to-noise is decreased from 0 dB to -20 dB. (Except for condition 3 (office noise condition) Persay 2006(1) and 2006(2) engines were less sensitive to the noise conditions, compared to Persay 2005 engine. Further, except for condition 3, Persay 2006(2) was less sensitive to noise conditions than Persay 2006(1).



1 = 0dB White Noise 3 = 0dB Office Noise 5 = 0dB City Noise 7 = 0dB Shop noise
 2 = -20dB White Noise 4 = -20dB Office Noise 6 = -20dB City Noise 8 = -20dB Shop noise
 (-1.00% indicates FTA returned)

Figure 10. Noise Sensitivity Analysis

1.2.3 Test Case 4-2 and 4-2A – Name Verification

Using condition 1 for names (Test Case 4-1 – Name Verification) above as a baseline, the performance of the engines was examined for condition 2 and 3 name verification. Conditions 2 and 3 are defined as:

Condition 2: Impostors say different names to those enrolled (different speakers; and different names)

Condition 3: Impostors as the enrolled speakers say different names to those enrolled (same speaker, different name)

Figure 11. shows the performance of Persay 2005, 2006(1) and 2006(2) engines.

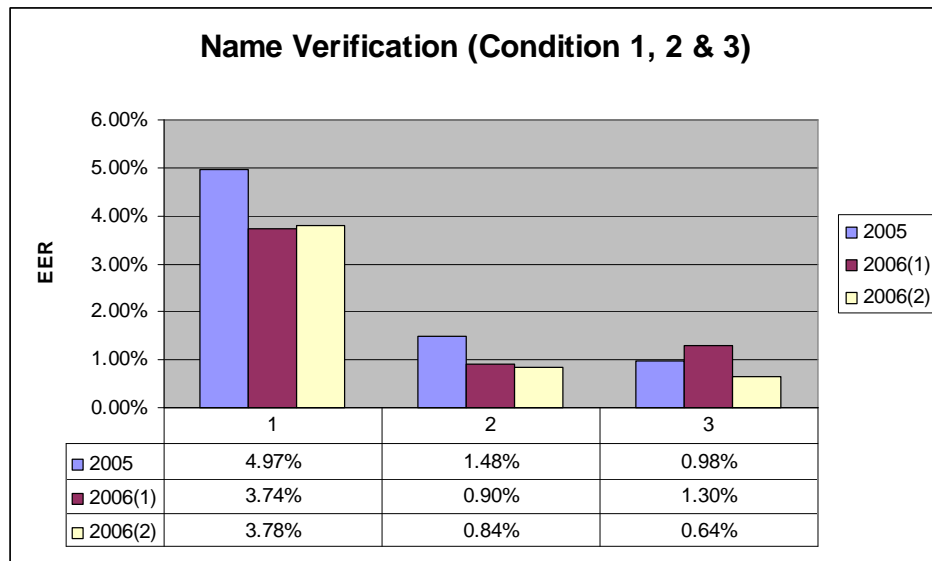


Figure 11. Performance for Name Test Conditions 1, 2 and 3.

1.2.3.1 Observations

All engines showed considerable improvement in EER performance from test condition 1 to test condition 2 where impostors were saying different names to that of the clients, with Persay 2006(1) and Persay 2006(2) showing the stronger discrimination between client and impostor compared to Persay 2005. Test condition 3 also showed that all the engines were able to discriminate when the client was specified as the impostor but said a different name from that enrolled by that identity. This shows that all the engines in this test exhibit an ability to discriminate when a client says a different name to that enrolled with Persay 2005 and Persay 2006(2) engines showing stronger performance. This test showed that all Persay engines have increased performance in discriminating when a client says the “wrong name”, compared to an impostor saying the “correct name”.

1.2.4 Test Case 3-1 Longitudinal Speech Data “Counting 1to9” and “Names”

Two longitudinal tests were performed, one “Counting” and saying “Names” shown in figures 12 and 13. Results for Persay 2006(2) engine are not available for the “Counting 1to9”.

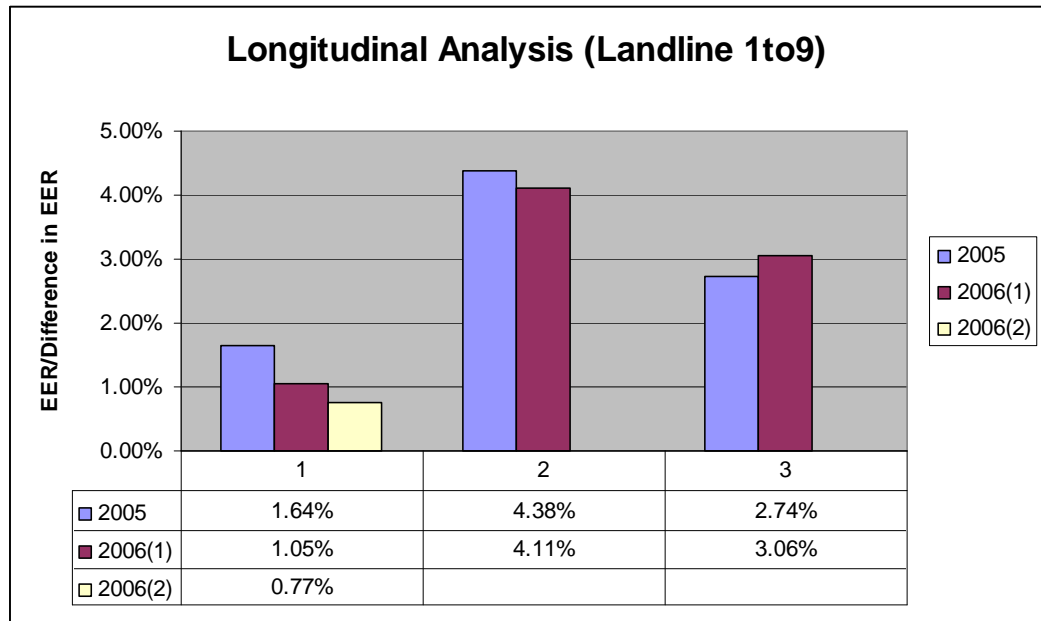


Figure 12. Longitudinal Test Results for landline data, Counting 1to9

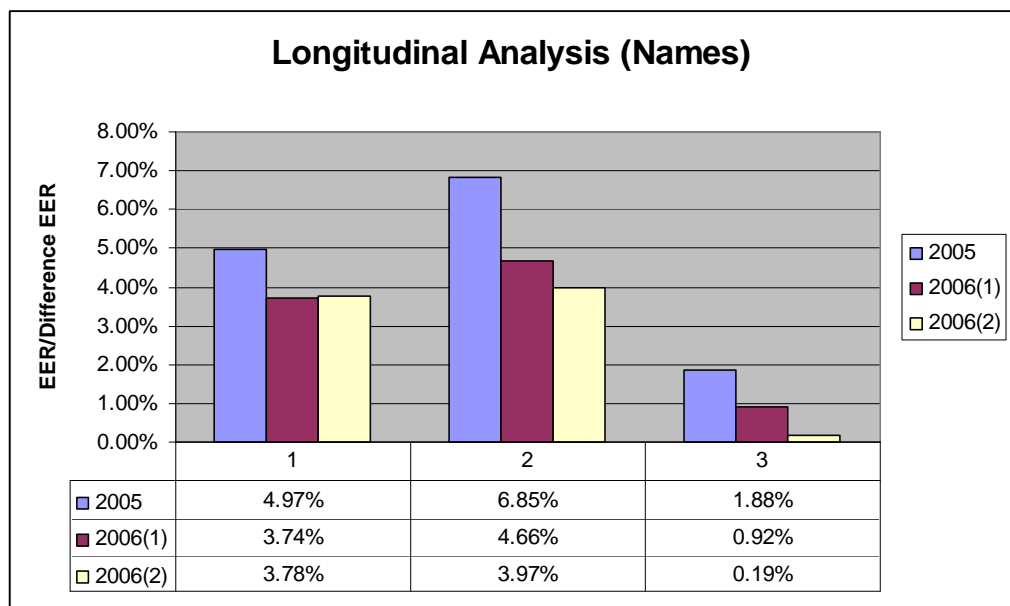


Figure 13. Longitudinal Test Results for landline data, Names

1.2.4.1 Observations

Longitudinal testing for counting 1to9, graphed in Figure 12, showed an increase in EER performance of both Persay 2005 and Persay 2006(1) engines. In the case of the Name tests shown in Figure 13, the EER of all engines increased, but with Persay 2006(1) and Persay 2006(2) showing considerably smaller increases than Persay 2005, suggesting that Persay 2006(1) and Persay 2006(2) are less sensitive to longitudinal effects than Persay 2005.



2 FINDINGS AND CONCLUSIONS

2.1 Finding from Test Case 1-1

Test case 1-1 (Counting 1to9) clearly showed differentiated performance between Persay 2005 and the Persay 2006(1) and 2006(2) engines in terms of baseline performance.

Finding 1.1

Text dependent evaluation involving Counting 1to9 (Test Case 1-1): Persay 2006(1) and Persay 2006(2) engines show between 36% and 53% improvement in performance over the Persay 2005 engine.

Testing using mobile telephone codec encoded speech showed, Persay 2006(1) and Persay 2006(2) engines exhibited higher level of performance compared to both Persay 2005, with Persay 2006(1) engine also showing much higher levels of robustness.

Finding 1.2

Text dependent evaluation involving Counting 1to9 (Test Case 1-1) using mobile encoded speech data: Persay 2006(2) engine showed significantly higher levels of robustness compared to Persay 2005 and Persay 2006(1) engines.

Whilst some Persay engines showed higher levels of noise robustness than others, all engines showed better performance in scenario noise compared to performance in white noise. This implies that the performance against white noise represents a worst case condition for all speaker verification Persay engines in the evaluation.

Finding 1.3

All Persay engines perform poorest in white noise conditions compared to other scenario noise demonstrating that white noise is a worst case condition for all Persay engines involved in this evaluation. However, Persay 2006(1) and Persay 2006(2) engines showed significantly higher level of robustness against noise compared to Persay 2005, with Persay 2006(2) showing between 3 and 20 times better performance than Persay 2005 in the same noise scenarios.

2.2 Finding from Test Case 4-1, 4-2 and 4-2A

Test Cases 4-1, 4-2 and 4-2A (Name verification) showed Persay 2006(1) and Persay 2006(2) engines offered better performance than Persay 2005 engine on the same speech datasets. All Persay engines shows equivalent robustness in mobile communications channels, except at the lowest data rates, where Persay 2006(1) and Persay 2006(2) showed significantly better robustness than Persay 2005.

Finding 2.1

Text dependent evaluation involving Name verification (Test Case 4-1, 2, 2A): Persay 2006(1) and Persay 2006(2) engines performed around 25% better than Persay 2005. All engines shows similar levels of robustness except at the lowest data rate, where Persay 2006(1) and Persay 2006(2) showed much higher level of robustness.



All text dependent engines tested showed an ability to discriminate in the situation when an impostor says the right answer to a “shared secret” question, based purely on the difference in the voice quality of the speaker.

Finding 2.2

Text dependent evaluation involving Name verification (Test Case 4-2, 2A): all Persay engines would have an ability to discriminate between a client and an impostor saying the right answer to a “shared secret” question – indicating that all Persay engines are suitable for implementing this function.

Finding 2.3

All engines showed increased performance on Test Case 4-2 (different speaker, different name) compared to Test Case 4-1 (different speaker; same name) indicating that all Persay’s engines have an increased ability to discriminate in the situation when an impostor says the wrong information to a “shared secret” question.

Test Case 4-2A (Name verification: same speaker, different name) showed that all text-dependent engines have an ability to discriminate between the client speaking information that is different from the enrolled information. This result indicated that these engines would be suitable to implement a “shared secret” solution where the engines could be used to discriminate where an enrolled speaker provides the wrong answer to a shared secret question.

Finding 2.4

The results of the Test Case 4-2A, (same speaker, different name) indicated that all Persay’s text-dependent engines tested are able to discriminate when an enrolled speaker says a different name to the one originally enrolled. This indicates that all engines could be suitable for implementation of a “shared secret” solution when the correctly enrolled speaker provides the wrong answer to a “shared secret” question.



3 APPENDIX A

An updated version of the Performix evaluation tool which provides a more accurate measure of the EER performance was used in these tests. This makes direct comparison with the results reported in 2005 difficult. However, it is possible to translate EER scores using the performance variation between the recalculated EER of the 2005 system and the 2006 systems. By applying these relative changes in performance, it is possible to infer the performance improvement, thus enabling comparing the performance of Persay 2006(1) and Persay 2006(2) technologies with other vendors' technologies reported in 2005 evaluation report.

Table A1 summaries the relative performance improvement of the 2006(1) and 2006(2) technologies compared to the 2005 technology, as well as the translated EER scores of the 2006 systems using the original 2005 EER scores, accordingly.

Test	2005	2006(1) improvement	2006(2) improvement	2005 EER	2006(1) EER	2006(2) EER
1-1-300by300	0%	54.38%	63.13%	1.55%	0.707%	0.571%
1-1	0%	35.98%	53.05%	1.73%	1.108%	0.812%
1-1-noise0db	0%	-84.43%	40.78%	3.55%	6.547%	2.102%
1-1-office0db	0%	-27.95%	61.49%	1.86%	2.380%	0.716%
1-1-city0db	0%	-15.95%	57.33%	2.02%	2.342%	0.862%
1-1-shop0db	0%	-45.22%	66.01%	2.23%	3.238%	0.758%
4-1-300by300	0%	35.42%	18.75%	5.11%	3.300%	4.152%
4-1	0%	24.75%	23.94%	5.33%	4.011%	4.054%
4-2	0%	39.19%	43.24%	1.32%	0.803%	0.749%
4-2A	0%	-32.65%	34.69%	1.67%	2.215%	1.091%

Table A1. Translated Results from 2005 Evaluation Report