Project Number	AC095
Project Title	ASPeCT: Advanced Security for Personal Communication Technologies.
Deliverable Type	Major
Security Class	Public

Deliverable Number	D24
Title of deliverable	Vocal Password-Based User Authentication Report
Nature of deliverable	Report
Document reference	AC095/L&H/W27/DS/P/24/1
Contributing WPs	WP2.7, WP2.4
Contractual Date of deliverable	June 1998 (Y4M6)
Actual Date of Delivery	3 September 1998
Editor	Martine Lapere (Lernout & Hauspie)

Abstract	This document contains a report on the work carried out by the ASPeCT project on the local authentication of users to their UIMs using speaker verification techniques. It includes a summary of the technical approach, a description of the demonstrator and concludes with the results and ideas for further work. provides a report on Deliverable D24. The report on the low storage speaker verification system.
Keywords	ACTS, ASPeCT, UIM, UMTS, Personal Authentication, Biometrics, Speaker verification, compression, smart cards

1 Executive summary

Deliverable D24 is the public deliverable of work package WP2.7 of the ASPeCT project. In association with work package WP2.4, WP2.7 has developed techniques for password-based speaker verification suitable for the mobile communication environment. More specifically, WP2.7 has developed speaker verification algorithms that generate voice prints small enough to be stored on currently available smart cards.

Work Package 2.7 is closely related to Work Package 2.4, which is more generally dealing with the security and functional capabilities of the User Identity Modules (UIM), as reported in [D04], [D11] and [D19].

In this report, we give an overview of the research work done within the scope of this work package. Both single- and multiple-password algorithms are discussed. This research work led to the real-time implementation of a best-choice algorithm. Deliverable D23 [D23] dealt with the speaker verification demonstration developed around this implementation. The demonstrator in question was shown in public at the IS&N 98 conference in Antwerp, Belgium, and was well received by its intended audience.

In addition we mention some extra security issues that cropped up during the scope of the project. These will be tackled in an additional research effort, which will be reported in D20, the final technical report of the project.

2. Table of Contents

3. Background and objectives of the work

The emergence of extensive, affordable, accessible and effective global (mobile) telecommunication services is resulting in an explosive growth in the variety and volume of transactions conducted by electronic means. The safeguarding of the integrity and security of such transactions has a high priority and is being addressed by the developers of such services. ASPeCT is investigating the implementation of security services in general for UMTS (Universal Mobile Telecommunication System), and mutual authentication of user and network, in particular. This includes the authentication protocols of the network and also the authentication of a user to the network.

Currently the smart card acts as a security module and the authentication of a user to the smartcard is carried out by means of a PIN (Personal Identification Number). Many users disable the PIN-code because the entry procedure is considered to be too cumbersome. This is especially true in the mobile communications environment. Other people use the same PIN-code for all their smartcards. Eventually these codes are written down or others when put in trace them. Biometrics is an exciting area of recent technology development that deals with user-friendly automated methods of verifying a person's identity from one or more behavioural or physiological characteristics. Various biometric techniques are currently being developed and researched, including fingerprints, palm-prints, hand geometry, retinal and iris scans, signature capture and facial and vocal characteristics. Of these, vocal biometric authentication looks potentially the most attractive in the context of the mobile communication services. Not only can it use the existing speech sensors and signal processing power but, additionally, it allows a user interface, which naturally fits in with the way the system is operated.

There are two obvious ways to perform vocal authentication – the processing can be performed either locally or remotely. Remote authentication may be appropriate for high security transactions over a telecommunications link. A bank, for instance, is very unlikely to trust the Service Provider to carry out the authentication in case of a fund transfer. In a roaming environment, however, the communication back to a remote authentication server is expensive. Therefore the biometric authentication should be performed locally in the terminal. In addition the proposed authentication mechanism must be portable between different mobile terminals by simply transferring the smart card containing the UMTS user application (the UIM: user identity module) from one terminal to another. This implies that all user specific data must be stored on the UIM. This UIM, being implemented on a smart card, will contain limited memory for template storage and possess limited processing power to perform the comparison.

Possible approaches to vocal authentication:

There are three different approaches to achieve voice authentication in general:

- *Free Speech Input*: In this case, the subscriber simply uses the mobile equipment and the authentication takes place in the background using his speech as the input. This is the most complicated form of authentication and requires large amounts of data storage and processing power.
- *Prompted Text*: In this variant the terminal prompts the user to repeat a randomly generated sequence of words out of a limited vocabulary. The characteristics of his response are compared with the responses expected from the correct user. These systems have medium complexity, and need additional text-to-speech software for the prompting of the tokens.
- *Pass-phrase*: This mechanism relies on only one type of utterance supplied by the user. The advantage of such a system is that the knowledge of the utterance can be used as part of the authentication. The overall protection then stems from what is said and from who says it. Although the storage requirements of these systems are an order of magnitude smaller than in the previous two systems, previous implementations still require too much storage to be stored on a smartcard.

Text-prompted and text-free systems have to capture a wide variety of speaker dependent phonetic events, which lead to complex systems. The storage demands and amounts of enrollment data of these systems are far beyond the quantities that are feasible to consider in the present context.

Therefore, a pass-phrase system with small storage requirements was examined during this project. In addition, we aimed at a system requiring only a short enrollment session. This is to keep the system as user friendly as possible. It is clear that the demands of short enrollment sessions and small storage capacity are directly competing with the achievable speaker separation. Multiple passwords might be expected to give better protection, where again, the storage requirements are competing with the achievable accuracy.

The main objective of WP2.7 can be summarized as follows:

The development and implementation of a real-time demonstration of a best-choice verification algorithm that complies with the limitations of a mobile environment and generates voice templates that are small enough to be stored locally on a smart card.

4. An introduction to speech recognition technology [1].

The problem of automatic speech recognition (ASR) can roughly be described as the decoding of the information conveyed by a speech signal and its transcription into a set of characters. Global recognition is basically a pattern recognition approach in which speech patterns are stored during a learning phase and recognized via pattern comparison techniques. Patterns can be phrases, words, or else sublexical units such as syllables, diphones, or phones. The phonetic approach postulates the existence of a finite set of phonetic units that can be described by a set of distinctive features extracted from the speech signal. Altough it should be possible to recognize speech directly from the analog speech signal, it is common usage to extract features representing the spectral envelope and their delta – the change of the feature vector over time-, for both, the training and recognition process. A speech pattern corresponding to a word or a sentence is made up of a sequence of short-time acoustic vectors. Therefore, when applied to ASR, pattern recognition techniques must be able to compare sequences of feature vectors. A major difficulty associated with this comparison comes from the fact that different occurrences of the same speech utterance, even pronounced by the same speaker, differ in their duration and speaking rate. Since these distortions are mostly non-linear, it is necessary to design efficient time normalization methods to perform reliable and meaningful comparison. In the Dynamic Time Warping (DTW) approach, the trial-to-trial timimg variation of utterances of the same text is normalized by aligning the analyzed feature vector sequence of a test utterance to the template feature vector sequence using a dynamic time warping algorithm. More popular is the use of Hidden Markov Models (HMM's) to model the statistical characteristics of the speech signal. A Markov chain consists of a set of states, with transitions between the states. Each state corresponds to a symbol, and each transition is associated a probability. Symbols are produced as the output of a Markov model by the probabilistic transitioning from one state to another. An HMM is similar to a Markov chain, except that the output symbols are probabilistic: in fact, all symbols are possible at each state, each with its own probability. HMM models consist of several states and represent whole word models or sub-word models such as phonemes. Before a speech system can be put to use, it must be trained. Training is generally conducted by first collecting speech samples from a large number of speakers. Feature vectors are then computed for every predefined time-slice of speech (5-20msec). This information is used with a dictionary containing all the words and their possible pronunciations, along with the statistics of grammar usage, to produce a set of models. At the end of the training process, therefore, there is a set of word or phoneme models, and a dictionary and grammar, all of which make up the recognition database. One of the advantages of phoneme-based recognition is that training data may be shared across words. For example, the [ae] in "cat" and the [ae] in "bat" would both be used for training the [ae] model. When a new word is added to the recognizer, it is not necessary to obtain training data for that word. Instead, one only needs to construct a model for the word by the concatenation of some previously trained phoneme models.

5. An introduction to speaker recognition technology.

Speech is a dynamic acoustic signal with many sources of variation. As the production of different phonemes involves different movements of the speech articulators, there is much freedom in the timing and degree of vocal tract movements. Consequently, depending on a number of conditions, a speaker can modify his speech production while still transmitting the same linguistic message. Speaking styles do differ from spontaneous to read speech, they are influenced by stress or emotion, and are more or less speaker specific. Environment changes also induce intra-speaker variability. Background noise or stress conditions yield an increase in the speakers' effort and a modification of speech production. These modifications at the speech production level produce acoustic-phonetic variations. Physiological differences (length and shape of the vocal tract, physiology of the vocal folds, shape of the nasal tract) are an important source of variation between speakers. These differences induce acoustic variability. For example, it is well known that the vocal tract length and the vocal tract geometry are different among speakers, and thus, the resulting formant frequencies are related in a rather non-linear fashion for different speakers. A shorter vocal tract length yields higher formant values. Articulatory habits also contribute to the inter-speaker variability. They are generally functions of the speaker's personality and differ from the dialect type variabilities, which are specific to a broad group of speakers. In speaker dependent recognition systems, the recognition is done on the same speaker as the one used for training. Sufficient training data should be available for each speaker, however.

Speaker recognition is the process of automatically recognizing who is speaking by using speaker-specific information included in speech waves. This technique can be used to verify the identity claimed by people accessing systems; that is, it enables access control of various services by voice. Applicable services include voice dialing, banking over a telephone network, telephone shopping, database access services, information and reservation services, security control for confidential information, secure billing and remote access of computers.

Speaker recognition can be further classified into speaker identification (SI) and speaker verification (SV). Speaker identification is the process of determining from which of the closed set of registered speakers a given utterance comes. Speaker verification is the process of accepting or rejecting the identity claim of a speaker. The fundamental difference between identification and verification is the number of decision alternatives. In identification, the number of decision alternatives is equal to the size of the population, whereas in verification there are only two choices, accept or reject, regardless of the population size. Therefore, speaker identification performance decreases as the size of the population increases, whereas speaker verification approaches a constant, independent of the size of the population.

All speaker verification systems use acoustic characteristics extracted from an utterance spoken by the person whose identity needs to be checked. These acoustic characteristics are compared with a model previously trained by the reference speaker during the enrollment session. A way to distinguish between systems is the freedom of the utterance and the way in which the acoustics are modeled. In our context, the utterance will be a password or phrase that was defined during the enrollment session (respectively password systems and text-dependent systems).

Speaker recognition methods can be divided into text-dependent and text-independent methods. The former require the speaker to provide utterances of key words or sentences that are the same text for both training and recognition, whereas the latter do not rely on a specific text being spoken. The text-dependent methods are usually based on template-matching techniques in which the time axes of an input speech sample and each reference template or reference model of the registered speakers are aligned, and the similarity between them is accumulated from the beginning to the end of the utterance. There are a lot of applications in which predefined keywords cannot be used. Therefore, text-independent systems have recently attracted more attention. Text-independent systems, however, need extensive training sessions, and do in general not reach similar performance as text-dependent systems. Figure 1 gives an overview of the specifics of each system.

	Enrollment	Technique	Complexity	Remarks
password	**password(s) 3X**	**Confidence value. SR score**	**medium**	**quick enrollment** **(taping of owners voice)**
text prompted	**phonetic rich text**	**Confidence value** **SR score**	**high**	**(extensive enrollment)** **(tts. required)**
free text	**phonetic rich text**	**no SR**	**very high**	**(extensive enrollment)** **(reliability)**

Fig 1

6. Background art of Lernout & Hauspie with respect to the project.

From its industrial concern, L&H has paid attention in the past to compact representations for the prompted approach. The password approach needs to model only a specific word as opposed to the "full" acoustic model required in the text-prompted and text-free methods. L&H has previously developed an algorithm that generates speaker specific models out of acoustic only input. The speaker specific passphrase models are generated out of a simple say-in of the utterances. Since the models are build out of the acoustic realizations of a password only, we can build a text-dependent speaker verification system in which the password can be freely chosen by the user. Free vocabulary passwords give added security, since not only the match of the test speaker towards the trained speaker specific model is available, but we can also check the correctness of the passphrase itself. The speech recognition engines available at L&H are equipped with "so-called" garbage scoring. The contrast score of the speech matched towards the speaker dependent model and the garbage model checks for the correctness of a spoken password.

The password are encoded into a few tens of bytes. This algorithm used to make coded speaker specific models is referred to as the "baseline userword algorithm" or BUA. Short enrollment sessions in which the password is uttered – say – three times are sufficient. The storage requirements of the BUA strongly contrast with methods that involve full acoustic models and which require a thousand-fold more of speaker specific storage. It is clear that these lower resolution models cannot get the same accuracy as the full models. In preliminary tests the accuracy of these low resolution models seemed to reach less than half that of corresponding L&H high resolution speaker specific models. The BUA runs on top of one of the L&H proprietary engines, running at various sample frequencies and tuned towards a typical office, car or telephone environment. The coded speech segments are sub-sets of the phonetic-based world models used in these engines. The phonetically balanced world models were trained on large databases, including hundreds of speakers.

The presence of this BUA algorithm was a key element for the project: it justifies further research and development for speaker verification based on this algorithm, not only from a scientific and technical point of view, but also from an industrial and economic perspective, given its speaker specific storage requirements.

Average spectral measures are known to be speaker specific and are a cheap way of performing text-independent speaker verification from extremely long enrollment and verification sessions. L&H has previously developed a mean cepstral feature which operates on small speech segments.

An early test measured achievable error rates of 10 to 15% on systems using only 1 second of text. A demo including the BUA algorithm and the mean cepstral separation was given to the consortium on February, 19, 1996.

7. Research work done within the scope of the ASPeCT project

Task 1:

Single password speaker verification

Single password verification is achieved by training a word and speaker specific password by means of the Basic Userword Algorithm. In a normal situation, the occasional impostor is not aware of the right password, so the gross protection of the system comes from the rejection of wrong passwords. Only spoken utterances which pass this first password check will be submitted to a second intrinsic speaker specific check.

Maximization of speaker-dependence of the BUA.

In the enrolment session the BUA algorithm generates speaker-specific password models from three acoustic repetitions of a word. The separation power of the BUA was further optimized within the scope of this project .

The expectation value of the score of a test utterance of the bona fide user is estimated out of the scores available at training time. Better estimation formulae for this expectation value led to an increase in separation from 73% to 80% for the BUA The influence of the length of a transcription was studied: longer transcriptions give better rejection of out-of-vocabulary words, while shorter transcriptions give better speaker rejection. A best compromise was chosen. The BUA separation power was further studied on different L&H proprietary engines. These studies include the influence of huge versus small basic world models, the influence of basic feature extraction, and the influence of the sample frequency (11 versus 8 kHz). The BUA was adapted in order to get rid of initial transients.

Spectral characteristics.

Another measure that is known to be speaker specific is the long term spectral average. Looking for speaker specific features that require only limited amount of storage and are compatible and complementary to the information contained in the normalized score over the password, are the word specific mean cepstral coefficients. The separation power over the word-based average cepstra was further optimized in the scope of the project. Again an expected profile is calculated out of the training session. The mean cepstral feature used is a 12-th dimensional feature vector. Some dimensions are more important than others, however. After applying appropriate weighting to the different coefficients, the separation power increased from 87% to 89%. Application of an L&H proprietary spectral normalization technique made this spectral feature more robust towards spectral biasing, while keeping the same separation power.

Combinatory logic for multiple criteria.

The statistical dependence of the two criteria explored in 2.7.1.1 and 2.7.1.2 was explored. The two features do not only have another intrinsic equal error rate, but as well are of different dimensionality. A theory was developed in order to combine both features in an optimal way. It makes a multi-dimensional decision for uncertain data. In principle, for each of the selected features a fuzzy matching score is calculated. The two fuzzy scores are modified in order to take into account the different dimensionality and the different basic separation power,. Th two modified features are further combined in order to give the fuzzy match of the utterance.

With the application of this theory, we found as separation power for the combined features: 10% for known passwords and 1.2% for unknown passwords.

Anti-password models for utterance rejection

Low cost out-of-vocabulary word rejection is often based on the comparison of the acoustical match with the expected word(s) and with a model that represents “any word”. In the case of password verification, there is only one single expected word, so the general model should be modified towards a so called “anti-password”. This means a model that represents any word except the password. General speech anti-password models could be trained a priori. We did notice however that it is favorable to have the same number of HMM states in the general model and in the password model. In our special case where the passwords can be freely chosen and are trained on-line, a dual anti-password should be trained in the meantime. Generation of anti-passwords models by discriminative training of a general speech model is out of question, since this model would require too much storage. In assembling anti-password models by the selection of models out of a pre-trained pool of general speech HMM models, it was possible to keep the storage need quite low. Not only the change of acceptance of bad passwords, but also the intrinsic speaker verification equal error rate decreased by the introduction of these anti-passwords. The equal error rate dropped from 1.2% up to 0.3% for unknown, and from 10% up to 6.7% for known passwords. .

Real-time implementation and robustness check

Test of a real time implementation of this system showed some crucial robustness lacks. More precisely the system clearly suffered from initial transients, and the environmental adaptation procedure needed to be modified in order to be able to work on short utterances only. This gave major improvements on robustness. The system was implemented on different L&H proprietary engines, of which an 8kHz engine was selected for implementation. First of all, the CPU load is 25% lower than for its 11kHz version counterpart. Second, the frequency response of the microphones embedded in the mobile terminal is likely to be of a lower standard than this limit.

In order to test the robustness of the system against mobile terminal and microphone swap, a simple simulator was build that models different microphone characteristics and additive noise. The typical frequncy response is approximated by applying a numeric filter with a number of poles and zeros ranging from 0 to 5. Five different frequency transfers were picked up in random sequence in order to test the robustness against microphone swapping. It is clear that these different transfers lead to more dissimilarity in the utterances. Consequently, we noticed a doubling of the false rejection rate.

Tests on the real-time implementation revealed as well the need to check for abnormal signal conditions. Abnormal conditions, being too loud, too quiet, or signals having a bad signal to noise ratio, will all have severe impact on the performance of the system. The abnormal signal detection module which was build in is based on a signal peak and background noise tracker. Additional checks were included that check appropriate length of the passwords. Too short passwords give bad verification results, while too long passwords would require too much storage capacity.

In the real-time implementation Internal deliverable, the required password models are synthesized from the simple speaking of the passwords. This enables the user to change his password locally in an autonomous and flexible way in order to avoid possible misuse by a third party. Since no written or typed input is required, this process can run in a completely hands-free and voice-controlled manner.

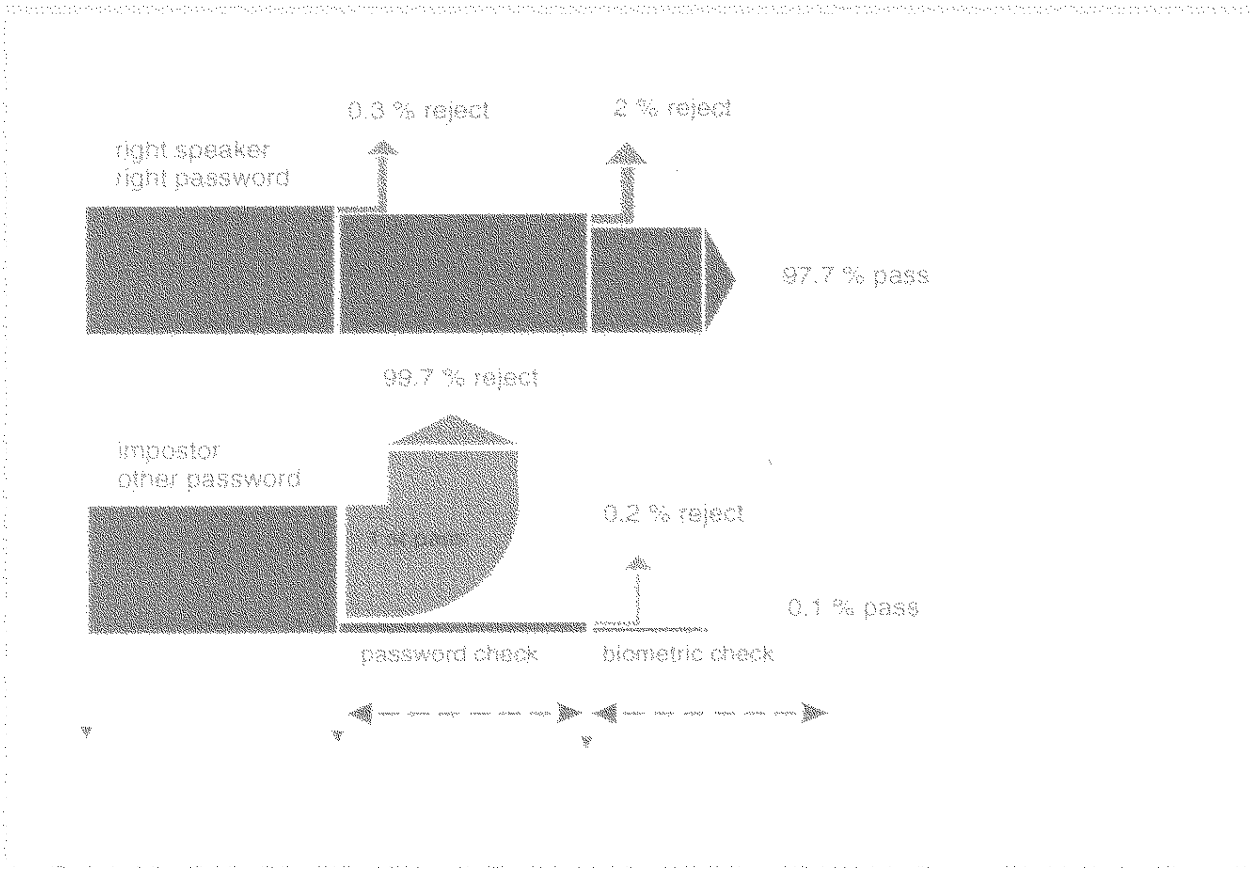


Fig 2

Figure 2 gives the functional diagram of this implemented single password system. By the cascading of the password and speaker check, we were able to construct a system where the bona fide users get accepted with a 97.7%, while impostors who are not aware of the right password get in with only a 0.1% chance. Impostors, aware of the password, would however still get easy access to the system. The multiple password system should give better results here.

Internal deliverable [D21] describes the single password research work.

D22: Multiple password verification

In the search for a better intrinsic speaker separation, we looked at a multiple password system. It is clear that multiple passwords should give better separation than single password systems. It was not clear however, to what extent results of single password verifications were independent, and how to make the best combination of single results.

Limiting the vocabulary to a fixed predefined set and using a more speaker-specific basic model set seemed to reduce the intrinsic speaker verification error rate. In the above development, sub-word units of passwords of other speakers model the passwords of one speaker. A methodology was developed in order to make an optimal combination of single password verification results.

This was achieved by associating a certainty measure to each single password trail. This certainty measure is high for scores that are mainly achieved for bona fide users, and varies to low for scores only found with impostor trials. In between these extremes lies a fuzzy zone where the decision is hard to make. A multiple trial system is shown in figure 3. The system gives better separation while not giving in on user friendliness. In general, impostors will need more trials and will therefore be penalized for being inconsistent. Multiple trials of different passwords will in addition be combined in order to make the final decision.

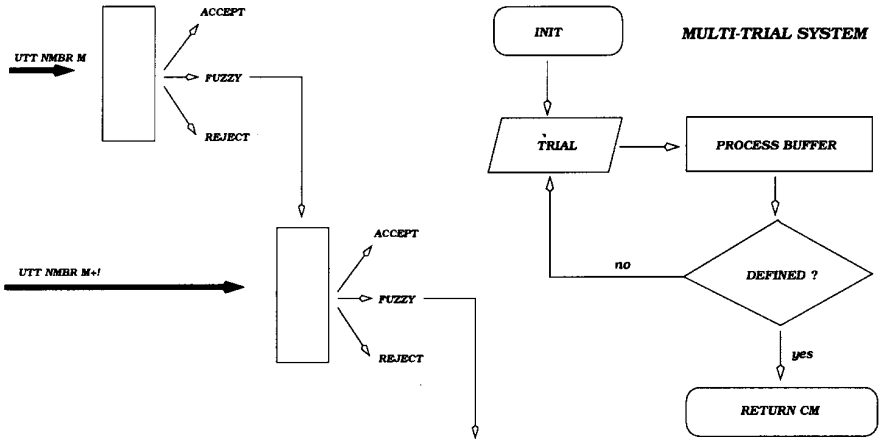


Fig 3

Internal deliverable [D22] describes this combinatory methodology.

For a fixed vocabulary system the intrinsic speaker verification error (same passwords) of a triple trial, dual password system is as low as 1%. For a triple password system it dropped to 0.5%.

This performance worsens towards a 2.3% for a triple trial dual password, where “free passwords” are allowed instead of passwords restricted to a fixed vocabulary.

The influence of microphone swap on the multiple password system was tested as well. In the multiple password mode the microphone swap is responsible for a 50% increase in error rate (3.3% versus 2.3%).

D23: Low storage speaker verification demo

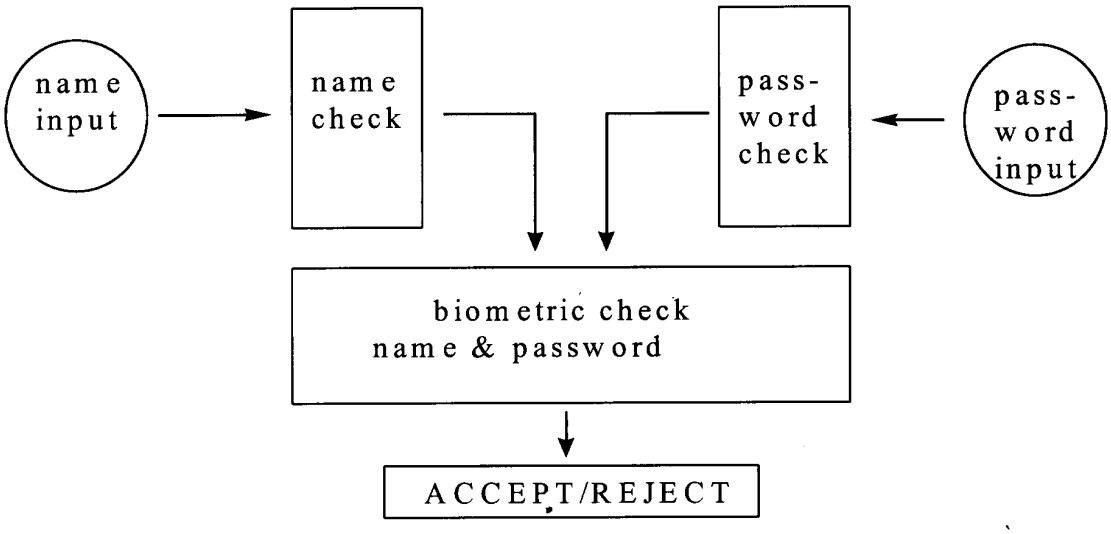
The implementation described in [D23] demonstrates the real-time implementation of the multi-password system of [D22]. The demo includes a full training session, the storage of the voiceprints onto the smart card, the retrieval of data from the smart card, and the verification process.

The hardware configuration consists of a PC system with a built-in, or add-on sound card of the Soundblaster type, an external microphone, and an external SmartCard Terminal with the corresponding SmartCards. The SmartCard terminal is connected to the PC via a serial port.

A top layer script was developed invoking the verification engine for the training and test sessions and performing the multi-password scoring. It is interfaced on one side with the Soundblaster card, for the acquisition of samples, and on the other side with the cardreader, for the storage and retrieval of the voiceprints.

The demonstration is equipped with a vocally driven user interface, including prompts to insert the cards, and to start a training or a verification session. During the training session, a vocal feedback replays the user's input to give feedback on the recording quality. The quality of the speech recordings is checked as well, and if necessary the user is prompted to change his speaking volume.

To keep the system as user friendly as possible, we opted to incorporate a dual password verification system in this demo. More specific, the user will be prompted to give his/her name, together with a user defined password system.



Within the system, however, the user's name will be treated in completely the same way as the user's password. At training time, the user has the complete freedom to chose the 'so called' name input. He/she can use his/her proper name (first + last name), or use a nickname. A minimum length constraint of 0.8 seconds is used on the speech input. The voiceprints of the name response and the password response will be stored on the Smart Card. The typical storage is 336 bytes per voiceprint.

At verification time, the user will again be asked to prompt his/her name (or nickname), and his/her password. For each password the user will be given three trials, with possible exit functions after the first and second trial. This in order to enhance the user friendliness. In general bona fide users will need less trials then impostors. Maximum three trials will be used to check a single password (being the users name or password), and the scores of the different passwords of the complete session will be combined in order to make the final accept/reject decision.

To conclude, the demonstration illustrated the feasibility of the implementation of a multiple password voice verification system, generating voiceprints small enough to comply with the storage constraints of currently available SmartCards. All functionality of the verification system was illustrated, going from on-line training of the voiceprints, to the actual storage on the Smartcard, over the retrieval of a voice print from the card and the user identity check.

8. Results from the public demonstration.

The demonstration was shown in public at the IS&N98 Conference in Antwerp. Even in the adverse conditions as found on a technology fair, the system seemed to perform quite well. People were astonished to see the system work, despite the high level of background noise in the demo boot. The audience was very enthusiastic about the idea of enrolling their voice on a smartcard, and seemed very pleased with the free choice of passwords. During this test, non of the intruders managed to impersonate the system, even if they assisted in the training session, and thus did not only have full knowledge of the password, but were also able to imitate the bona fide speaker's speaking style.

9. Security Aspects and pointers to add-on research.

Although the applicability of a vocal biometric authentication was successful demonstrated in the scope of the project, a major security issue is still holding the system from commercial application. The fact is that the smartcard is at the moment only used as a storage device, while

the verification process and ultimate decision is done locally in the terminal. The terminal can, however, in general not be regarded as a trusted third party. This is particularly true in a mobile communications environment. A secure operation can only be guaranteed if the ultimate decision is taken by the smart card itself. Therefore it was decided to extend the study package to include the feasibility of splitting of the algorithm within a client-server architecture. In this new structure the terminal would be the client and the card the server. The algorithm would be split in such a way that the ultimate decision is taken by the card and not by the terminal, since the card is the only trusted party for the service provider.

The results of this work will be reported in the final project technical report, D20.

References:

Speech Recognition:

[1] J.C. Jugua & J.P. Haton, Robustness in automatic speech recognition, Kluwer Academic Press. (Overview of speaker recognition based on this work)

Speaker Recognition:

[1] J.C. Jugua & J.P. Haton, Robustness in automatic speech recognition, Kluwer Academic Press.

[2] B.S. Atal : Automatic recognition of speakers from their voices. Proc IEEE, vol 64, no 4, pp 475-487,1976

[3] A.L. Higgings, L.Bahler and J.Porter Speaker verification using ramdomized phrase prompting, Dig Sig Proc, vol 1, pp 869-872, 1986

[4] A.E. Rosenberg & Al. The use of cohort normalized scores for speaker verification. Proc 1992 ICSLP oct 1992, pp 599-602.

[5] F.K.Soong, A.E.Rosenberg,L.R.Rabiner and B.H.Juang. A vector quantisation approach to speaker verification. IEEE 1985, pp. 387-390.

[6] A.E.Rosenberg, Chin-Hui Lee and Frank K Soong. Sub-word Unit Talker verification using Hidden Markov Models. IEEE1990, pp 269-272.

[7] T.Masui and S.Furui. concatenated phoneme models for text-variable speaker recognition. IEEE 1993 pp 391-394

[8] J. Kuo, C.H.Lee and A.E. Rosenberg. Speaker set identification through speaker group modeling. BAMFF '92'

[9] M.I. Hannah, A.T. sapeluk, D.I. Damper & I.M Roger, (1993) The effect of utterance length and content on speaker verifier performance. Proc Eurospeech pp 2299-2303, Berlin 1993.

[10] M.E. Forsyth, A.M Sutherland, J.A.Elliott & M.A. Jack (1993). HMM speaker verification with sparse training data on telephone quality speech. Speech Communication , Vol 13, pp 411-416.

ASPeCT project deliverables.

[D04] Report on the use of UIM's for UMTS

[D11] Report on limiting smart card constraints on UIM's

[D19] Final trial and demonstrations

[D20] Project final report and results of trials.

[D21] Algorithms for single password speaker verification.

[D22] Algorithms for multiple password speaker verification

[D23] Demonstration of password-based user authentication.

Glossary of terms

Speech Recognition	A group of techniques where the words in spoken utterances are being recognized by a computer program
Speaker Independent/Dependent Recognition	Speaker independent recognition is the ability of the recognizer to allow any speaker to use the system successfully. With the speaker dependent recognition, however, only one single speaker is able to successfully use the recognizer. In that case the models which are used during recognition have been especially tuned towards the voice characteristics of that single particular speaker.
Automatic Speaker Identification	A system where the computer determines the identity of a speaker out of a closed set of possible candidates. The utterance of the speaker is compared to all speaker models available in the system. The best matching model provides the identity of the speaker.
Automatic Speaker Verification	A technique where the user first identifies himself in order to activate the single speaker model corresponding to the claimed identity. Next the computer determines whether or not a test utterance of this speaker matches this model. If the match is good enough, the system decides that the speaker is the person he claims to be, else the system signals that the speaker is most probably not the person he claims to be.
DTW: Dynamic time warping	A technique which allows to comapre feature strings of different length. The best alignment between the strings is the one which yields the minimum distance between both strings. The distance computation and finding the best alignment are combined into a single integrated algorithm. Dynamic Time Warping is essential for speech recognition since multiple utterances of the same word will almost always be of variable length.
Hidden Markov Model (HMM)	A Hidden Markov Model is a stochastic finite state machine in which the states can not be observed directly.
Channel Adaptation	A Technique which makes the extracted features independent of the influence of the recording channel: telephone lines, microphones, etc.
Cepstrum	Small set of parameters equivalent to the smoothed log spectrum
ACT	Acceptance Threshold on a single verification tria
RCT	Rejection Threshold on a single verification trial
FAT	Fuzzy Acceptance Threshold on multiple verification trials
FRT	Fuzzy Rejection Threshold on multiple verification trials
FATMP	Fuzzy Acceptance Threshold on multiple passwords
Preprocessing	Extraction of features from the speech waveform that are useful in the recognition process
LPC Linear Predictive Coding	One of the basic speech analysis techniques
CMS	Cepstral Mean Subtraction. Particular algorithm to filter out the situation dependent characteristics of the recording channel (telephone line, microphone).
LPC	Linear Predictive Coding: One of the basic speech analysis techniques
PLP	Perceptual Linear Prediction
Phoneme	Smallest meaningful unit of spoken language (language dependent) as defined by

	phonologists.
Phonemic Alphabet	The set of phones defined for a certain language
Diphone	a particular sequence of 2 phonemes
Triphone	a particular sequence of 3 phonemes
Speech Unit	A speech unit is the elementary linguistic unit used inside a speech recognition system. A speech unit can be differentiated from other speech units based on acoustic properties
Word	The fundamental linguistic unit (in most languages)
Vocabulary	a list of words known to the system
Vocabulary size	the number of words in a vocabulary
Baseform	The description of a word as a sequence of speech units
Userword	A word that is defined only for a specific user and of which its model was constructed on spoken examples only.
BUA	Basic userword algorithm
DUT	Derived userword transcription
False Rejection Rate	The fraction of the test samples that were rejected by the decision mechanism but should be accepted
False Acceptation Rate	The fraction of test samples that were accepted by the decision mechanism but should be rejected.
EER	Equal Error Rate. The False rejection at the decision point where the false rejection rate equal the false acceptation rate

.