A Speaker Verification Biometric In 40 Bytes

T. C. Phipps and R. A. King

Domain Dynamics Limited
Heaviside Laboratories 12 , Cranfield University (RMCS)
Swindon SN6 8LA , United Kingdom
email: ddl@rmcs.cranfield.ac.uk

ABSTRACT

This paper describes a novel Speaker Verification (SV) architecture that minimises many of the challenging problems faced by SV implementors in the *Card Security* arena. The architecture is unusual in two respects. First, the verification engine is not fed with conventional frequency-domain descriptors. Instead the output of a Time-Encoded Signal Processing and Recognition (TESPAR) coder is used. The TESPAR coder is a low-complexity, vector quantiser capable of converting the speech waveform into a highly informative set of time-domain descriptors. A simple 29-element matrix representation of TESPAR descriptors is described which, when used in conjunction with a Fast Artificial Neural Network (FANN), provides a wide degree of cross-speaker separation and strong same-speaker consistency [1].

The paper focuses on a novel SV engine that secures a robust verification performance by encoding the classification power of 100 orthogonal artificial neural networks into a biometric of only 40 bytes, offering a high level of protection to low-tech Dumb Cards as well as to *state-of-the-art* Smart-Card embodiments. The architectural rationale that enables this is presented, together with indications that a final verification decision can be delivered in under 1 second. The key features of the TESPAR/FANN combination and a system implementation in silicon for Smart-Card and Dumb-Card based systems are also discussed.

1. Introduction

1.1 Background

A reliable speaker biometric must satisfy three key requirements. First, the biometric must be *robust*. This means that regardless of the variability exhibited by the speaker's voice, and despite changes caused by the passage of time, the acoustic environment, and the transducer, the biometric should be able to characterise the speaker at all times. Second, the biometric representation must remain highly *discriminative*, providing enough separation to allow each speaker to be uniquely identified. Third, the biometric must be *secure*, it should be immune to reverse engineering and criminal stratagems and, ideally, useless to criminals if stolen or compromised.
Current conventional approaches to speaker verification typically suffer from a combination of the following problems:

- They are computationally demanding, generating biometrics that consume large amounts of memory.
- They respond slowly during interrogation.
- They are architecturally inflexible.
- They frequently require a complicated input *template* to be time-normalised.
- They involve complex registration and training procedures.
- They are susceptible to so called human benign traumas that distort speaker diction.
- They are vulnerable to background noise and communication channel impairments.
- They utilise technologies at the peak of world-wide billion dollar development "S" curves.

The rationale for this paper is to describe, in contrast, a new (TESPAR/FANN) Massively-Parallel Network Architecture (MPNA) that has the potential to provide high levels of robustness and speaker discrimination using a speaker biometric of less than 40 bytes, without degrading the effectiveness of a highly discriminative biometric model. The proposed strategy dispenses with the need for over-complicated training procedures and offers an interrogation response in under 1 second. The architecture also minimises system complexity, obviating the need for large communication infrastructures and centralised databases currently demanded by conventional world-wide card-based transaction systems.

1.2 TESPAR/FANN For The Smart Card

The integration of *Time Encoded Signal Processing And Recognition* (TESPAR) waveform coding procedures with multiple orthogonal *Fast Artificial Neural Networks* (FANNs) as part of a text-dependent speaker verification system has already been shown to be highly effective [2]. The principal objective of the 16 man-month trial under reference was to produce a reliable biometric for Portable Secure Objects: such as high performance *state-of-the-art* Smart Cards.
Using *supervised registration* procedures, the verification performance obtained from processing a 218-speaker database, comprising 150 males and 68 females, was as follows:

- 0 x False Reject errors out of 4360 interrogations (FRR < 0.023%)
- 4 x False Accept errors out of 2616 interrogations (FAR = 0.153%)

Embodying this verification system as a small set (circa 5-15) of biometric networks enables a multiple FANN architecture to be implemented on a single chip or card in 1-2 kbytes [3]. This performance compares very favourably with that achieved by competitor methods.

For these and other productive reasons TESPAR/FANN technology is being used to provide the biometric functions required in the European Union CASCADE Esprit Smart Card project - the objective of which is to develop a 32-bit RISC processor 20 square mm in area for a new generation of Smart Card and secure Pocket Intelligent Device applications [4].

2. TESPAR/FANN Technology

2.1 Coding

TESPAR is the new simplified digital language, first proposed by King and Gosling [5] for coding speech. The process however may be extended to any information bearing entity that can be represented in terms of a band-limited signal. The range of applications so far investigated encompasses seismic signals with frequencies and bandwidths of fractions of a Hertz, to radio frequency signals in the gigaHertz region, and beyond.

TESPAR is based on a precise mathematical description of waveforms, involving polynomial theory, which shows how a signal of finite bandwidth - *band-limited* - can be completely described in terms of the locations of its real and complex zeros. This contrasts with the more conventional approach of linear transformations based on *amplitude* sampling at regular intervals, as has been described by Fourier, Nyquist, Shannon and others. The real and complex zero descriptors of TESPAR and the time-bandwidth data produced by a Fourier transform are mathematically equivalent, and both result in 2TW (the Shannon Number) of digital sample data points describing the waveform [6]. The mathematical underpinnings of this zero-based approach are outlined in Voelcker [7] and Requicha [8].

Given the real and complex zero locations of the signal, a vector quantisation procedure has been developed to code these data into a small series of discrete numerical descriptors, typically around 30 (the TESPAR *symbol alphabet*). Holbeche [9] gives an account of one version of this coding.

2.2 Matrix Formation

The output from a TESPAR coder is a simple numerical symbol stream which may be converted into a variety of progressively informative matrix data structures. For example, the single-dimension vector (or *S-matrix*) is a histogram recording the frequency with which each TESPAR coded symbol occurs in the data stream. A more discriminating data set is the two-dimensional histogram or *A-matrix* which is formed from the frequency of symbol pairs, which need not necessarily be adjacent. Extending this to 3 dimensions would improve the discrimination power still further. Typical A and S matrices are shown in Figure 1 and Figure 2.

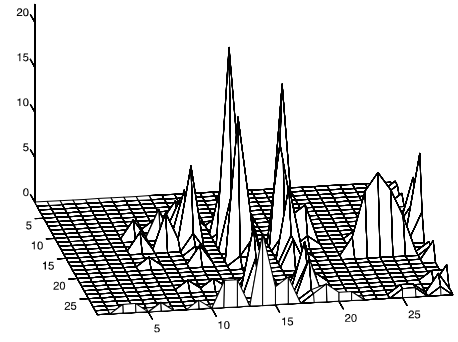

Figure 1 - Un-normalised A-matrix

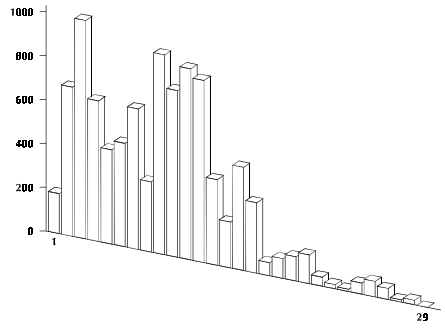

Figure 2 - Un-normalised S-matrix

2.3 Performance Advantages

TESPAR-based verification techniques are presenting significant performance advantages over conventional Fourier based methods. For example:

- They have typically 2 orders of magnitude lower computer processing power requirements, with consequent lower power consumption.
- They use and form simple data structures which are both compact and of known size so that limited memory resources in embodiments such as Smart Cards can be employed efficiently. This has important benefits for data storage and transfer operations.
- The data structures are optimally matched for speech classification methods that use FANN architectures.
- Samples can be obtained direct from low cost analogue sensors such as telephone handset microphones.
- They offer extremely high degrees of discrimination.
- Classification procedures and architectures can, by routine design, enable system errors to be made vanishingly small over a wide range of real world applications and environments.
- Verification speed is minimal, e.g. less than 1 second using current popular microprocessor technology for a single pass interrogation.

3. Massively Parallel Network Architectures (MPNAs)

The new technique of *Massively Parallel Network Architectures* (MPNA) is the product of research conducted by Domain Dynamics Limited (DDL) at the Cranfield University Campus. MPNA embodies the immense power of multiple parallel artificial neural networks and data-fusion decision making to achieve the performance associated with a large number N of trained networks in parallel. A typical range for N is $100 \leq N \leq 1500$.

In this technique, an ordered set of N networks, all different, may be generated a priori in non-real time using speech data from a large number of arbitrary speakers. Each individual network may, for example, be trained using a subset of 8 different arbitrary speakers, to provide a single network with 8 target elements in its output layer - see Figure 3.

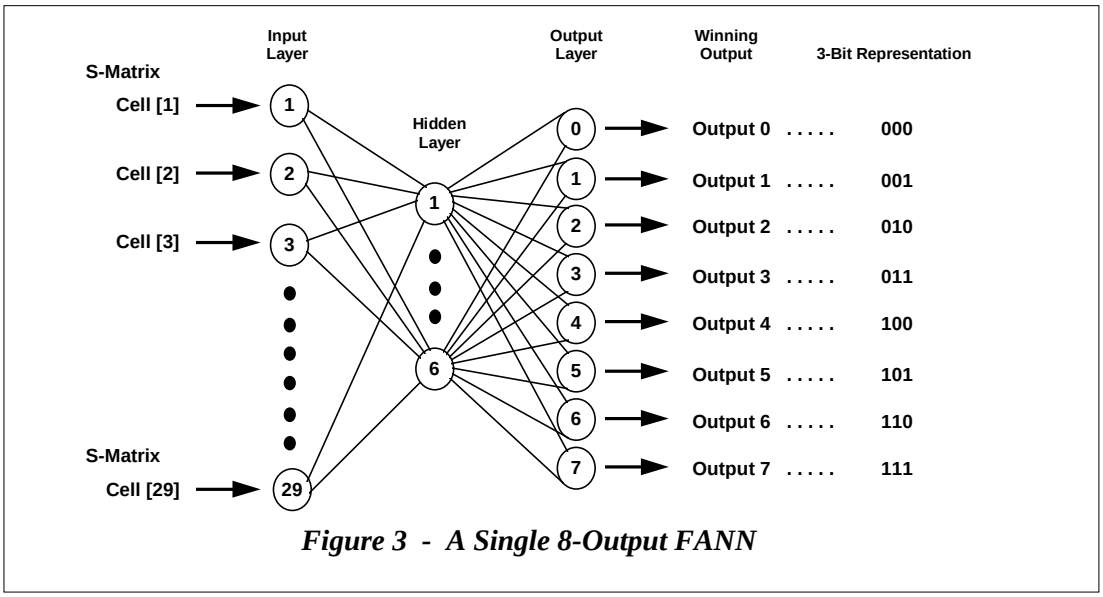

Figure 3 - A Single 8-Output FANN

N of these single networks are then used in an ordered parallel arrangement as an interrogation set, against which all speakers are to be compared, both at registration (enrolment) and subsequent interrogation (verification) - see Figure 4.

When a speaker registers against the N net interrogation set, his/her utterances will be converted to appropriate TESPAR matrices, and compared against each of the N nets in turn. Each net will produce a *winning* output on one of its 8 nodes, indicating, to which of the 8 speakers in that net, the input utterance was closest. The pattern of data outputs across all N ordered nets is then subjected to a variety of mathematical procedures for characterising the speaker. For example, by interrogating an ordered set of N=100 parallel nets, a speaker may be characterised by the resultant data set of 100 three-bit words, i.e. circa 38 eight-bit bytes.

These data describe, to a very high probability, the numerical output profile likely to be generated by the registered user's voice input on subsequent interrogations.

The MPNA so described may be embodied in silicon as a ubiquitous general-purpose low-cost, low-power biometric engine suitable for installations in terminals world-wide.

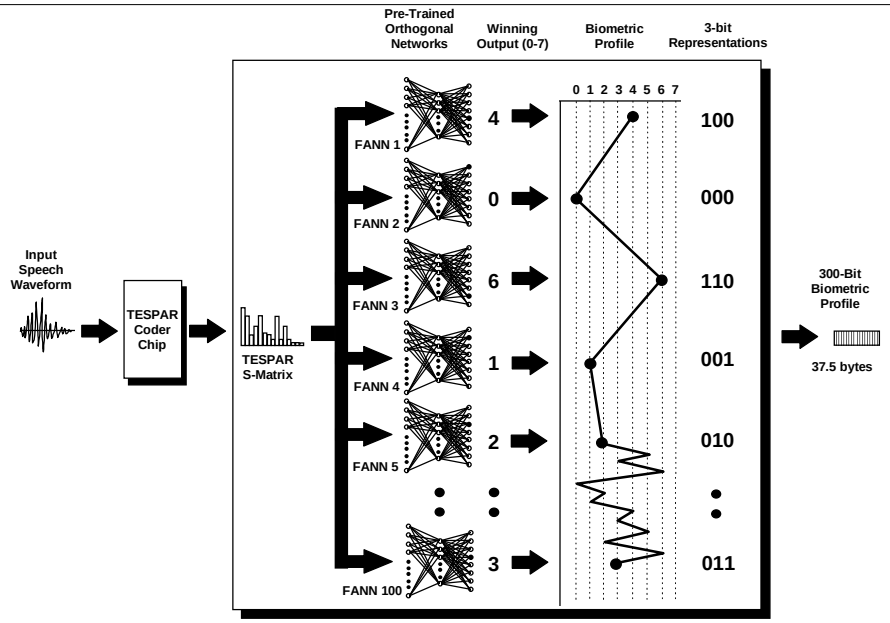

Figure 4 - A Massively Parallel Network Architecture (MPNA)

3.1 Implementation Issues

The TESPAR coding and vector quantisation process is already available both as a software algorithm, and in a low power ASIC silicon design. Beyond this, TriTech Microelectronics of Singapore are in the process of producing a range of very low-cost, low-power TESPAR embodiments in silicon which offer a high degree of flexibility for integration into a wide range of potential high-volume TESPAR applications.
In association with this activity, work is in hand in collaboration with King's College and University College London to adapt their pRAM Neural Network architecture to the task of classifying TESPAR data structures [10, 11, 12]. pRAM technology provides Neural Networks that *can be*

trained on the silicon itself. Thus the realisation of complete TESPAR/FANN single chip solutions is in sight, capable of training in situ and adaptable to widely differing low cost, high volume applications.

4. System Implications

The relevance of the proposed architecture in the commercial arena may, very simply, be exemplified by reference to the following diagrams which compare the features of three principal conventional system options with the new MPNA configuration.

4.1 Verification Performed At A Remote Central Site

- Figure 5 shows the configuration of a typical conventional verification system with a central biometric database.

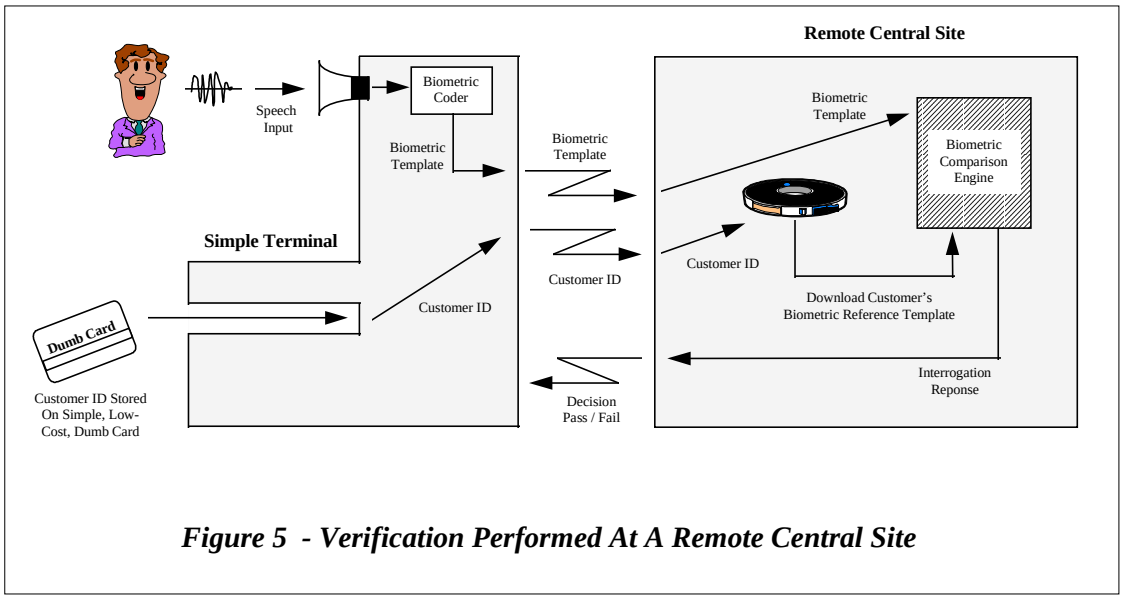

Figure 5 - Verification Performed At A Remote Central Site

Key disadvantages of this method are:

- World-wide registration and usage is complicated
- The configuration is prone to congestion and delays caused by many terminals seeking simultaneously to access the remote central site.
- Data transfer may be vulnerable to criminal interception.
- Essential back up is costly, involving complex administrative and communication system configurations.

The memory capacity of the conventional low-cost Dumb Card is very small - circa 80 bytes - leaving little or no opportunity for applying a conventional voice biometric capability on the card.

4.2 Verification Performed At A Local Terminal

Figure 6 shows the configuration of a typical *local* verification system suitable for use by small groups or communities viz. key medical staff and administrators.

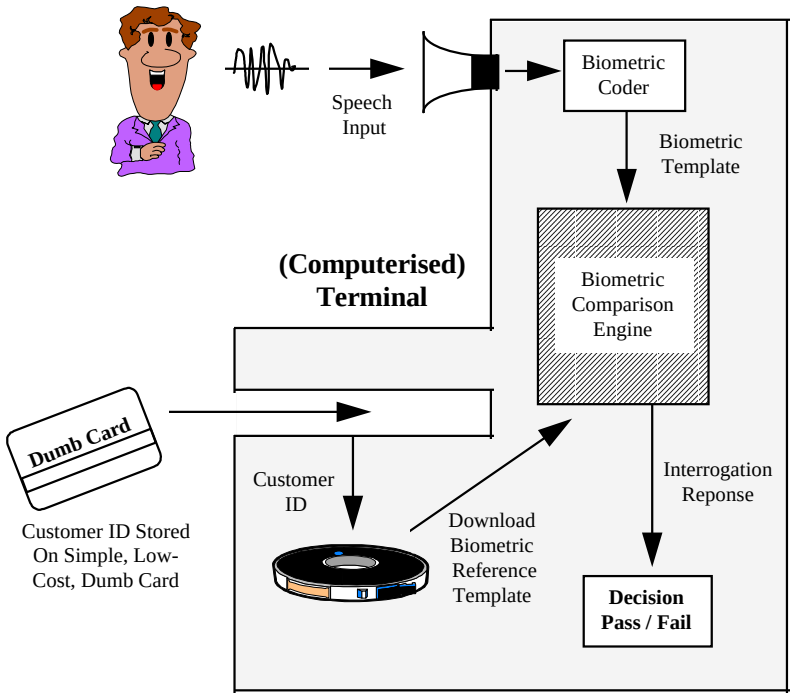

Figure 6 - Verification Performed At A Local Terminal

The major disadvantages of this method are:

- As with Section 4.1 , the memory capacity of the low-cost Dumb Card is very small - circa 80 bytes - leaving little opportunity for storing a conventional voice biometric reference template on the card.
- The finite capacity of the biometric store at the local terminal severely restricts the number of users that can be accommodated by the verification system.

4.3 Verification Performed On The Smart Card

Figure 7 shows an *ideal* configuration of a verification system in which the biometric engine and biometric reference template are stored on the high-tech card.

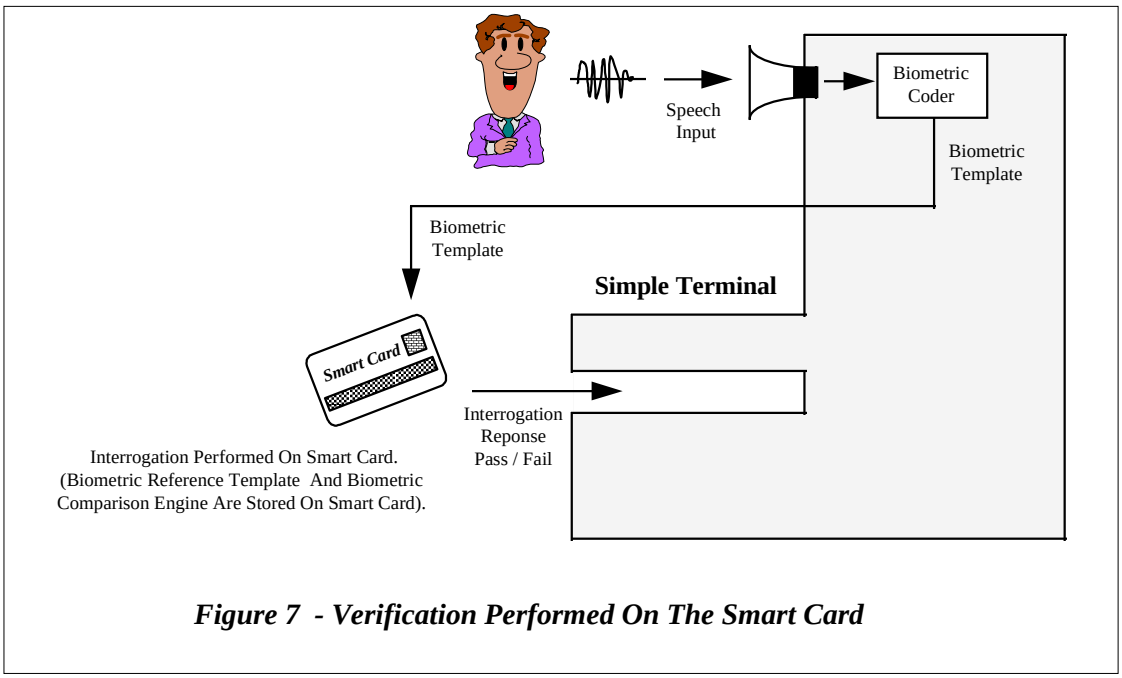

Figure 7 - Verification Performed On The Smart Card

Key limitations of this method are primarily commercial:

- It requires the use of a high-tech relatively costly Smart Card to support the necessary processing power required by the biometric comparison engine.
- To store the biometric reference template, the Smart Card must posses a sufficiently large amount of on-card memory.
- ***These technical requirements have prevented the billions of Dumb Cards manufactured to date, from providing a voice biometric capability. The proposed MPNA configuration opens up this important market.***

4.4 The New TESPAR / FANN MPNA Configuration

Figure 8 shows the proposed TESPAR/FANN MPNA configuration.

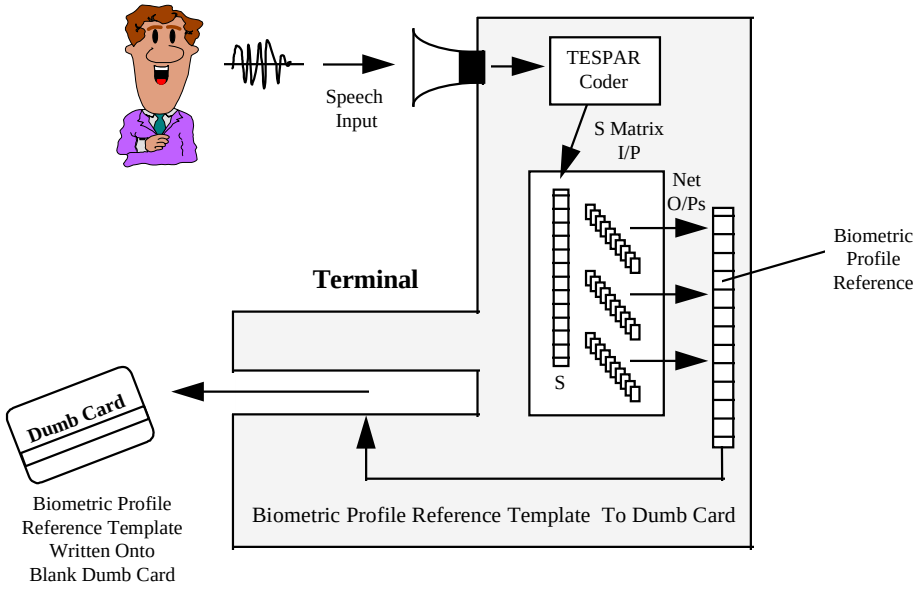

Figure 8 - The New TESPAR / FANN MPNA Configuration

Key advantages of this method are:

- The system is equally compatible for use with both Smart Cards and Dumb Cards.
- The configuration may be used to provide coverage on either a *local* or a *world-wide* basis.
- The system is almost infinitely flexible. Thus, commercial and market segmentation is simple to achieve.
- The system allows new markets and products to be rapidly catered for as they emerge.
- The scale and volumes associated with the introduction of a common world-wide ubiquitous silicon biometric engine will significantly minimise system installation and running costs.
- The proposed system may be introduced piecemeal and progressively.

5. Development tools

All the work described so far has been conducted using a Domain Dynamics' proprietary PC-based development system, the TADS-XS 50. The system includes an extensive library of both conventional and TESPAR signal processing and data analysis software, operating under the popular MATLAB™ graphical user interface. FANN classification architectures are created, trained, tested and interrogated within the system using the proprietary FasTEST software suite. This development facility is proving extremely valuable in enabling third parties to evaluate TESPAR/FANN architectures in a wide range of real world classification tasks.

6. Conclusions

A new Massively-Parallel TESPAR/FANN Network Architecture (MPNA) has been presented that has the potential to provide significant levels of robustness and speaker discrimination using a speaker biometric of less than 40 bytes without degrading the effectiveness of a biometric model utilising some 100 artificial neural networks. The key advantages of the proposed strategy are:

- The need for a costly centralised management system and massive data storage requirements is dispensed with thus making it available to the populations of both Smart Cards and Dumb Cards.
- Complicated training procedures are minimised and an interrogation response of under 1 second is proposed.
- The configuration may be used to provide coverage on either on a *personal*, *local* or a *world-wide* basis.
- Commercial and market segmentation is simple to achieve.
- New markets and products may be rapidly catered for as they emerge.
- A common world-wide ubiquitous silicon biometric engine will significantly minimise system installation and running costs.
- The proposed system may be introduced piecemeal and progressively.

7. Acknowledgements

Thanks are due to:

- Domain Dynamics Limited for their permission to publish this paper and for their support and funding of the research work under which the TESPAR/FANN MPNA technology has been developed.
- The Principal of Cranfield University (RMCS) for his permission to publish this paper.
- Annabel Clifton and Ashley Elkins for their contributions to this document.

8. References

[1] M. H. George and R. A. King. “A Robust Speaker Verification Biometric” , Proceedings IEEE 29th Annual 1995 International Carnahan Conference On Security Technology, pp. 41-46, UK , October 1995

[2] R.A. King, “TESPAR/FANN: an effective new capability for voice verification in the defence environment”, presented at the Royal Aeronautical Society Conference on The Role of Intelligent Systems in Defence, London, March 1995, p. 5.1-5.8

[3] S R Timms and R A King. “Speaker Verification Utilising Artificial Neural Networks and Biometric Functions Derived From Time Encoded Speech (TES) Data.” , IEE Second International Conference on Private Switching Systems and Networks, pp. 59-64, London, June 1992

[4] CASCADE Esprit Project EP8670 Data Sheet, 1995

[5] R.A. King and W. Gosling, Electronics Letters, vol. 14 (15), pp. 456-457, 1978

[6] C.F. Shannon, Communication in the Presence of Noise. Proceedings of the IRE, January 1949

[7] H.B. Voelcker, “Toward a unified theory of modulation”. Proceedings of the IEEE, vol. 54 (3), pp. 340-353; and vol. 54 (5), pp. 735-755, 1966

[8] A.A.G. Requicha, “The zeros of entire functions. theory and engineering applications”. Proceedings of the IEEE, vol. 68 (3), pp. 308-328, March 1980

[9] J. Holbeche, R.D. Hughes, and R.A. King, Proceedings of the IEE International Conference on Speech Input/Output: Techniques and Applications, pp. 310-315, 1986

[10] T.G. Clarkson, C.K. Ng and J. Bean, “A Review of Hardware pRAMs”, in the Proceedings of the Weightless Neural Network Workshop ‘93, University of York, April 1993

[11] D. Gorse and J.G. Taylor, “A review of the theory of pRAMs”, in the Proceedings of the Weightless Neural Network Workshop ‘93, University of York, April 1993

[12] D. Gorse, D.A. Romano-Critchley and J.G Taylor, “A Modular pRAM Architecture for the Classification of TESPAR-encoded Speech Signals”, in the Proceedings of Neuro Fuzzy ‘96, Prague, April 1996