

URU Plus – a scalable component-based speaker-verification system for BT’s 21st century network

M Pawlewski and J Jones

Identity security is an issue that affects us all. In many everyday transactions, you have to prove that you are you (URU). BT has already established a market position with the award-winning URU product, that provides identity validation. A major extension to URU is URU Plus. This allows users to verify their identity through the use of speaker-verification technology. Following initial identity validation via URU, users are no longer required to produce documentary evidence for identity verification, but can instead verify themselves using their voice. The URU Plus system is a component integrated into BT’s 21st century network authentication capability. It uses state-of-the-art voice verification technology and is accessed via a Web Service API.

1. Introduction

The incidence of identity theft has increased rapidly in recent years, and is an issue that affects everyone. The result is that to complete many everyday transactions, you have to prove that ‘you are you’. BT has addressed this issue with URU [1] which is a text-based identity validation system. URU Plus is a major extension to URU that allows users to verify their identity using speaker verification.

Although speaker verification is a well-established technology, its use is constrained by the inflexibility of stand-alone solutions, where an organisation would have to build and install its own product. However, unlike these conventional systems, URU Plus uses a component-based approach, built using XML-based Web Services, and can be accessed over high-speed networks, rather than requiring a complex and costly in-house installation.

URU Plus has a loosely coupled interface which is language and platform independent. This enables organisations to use it regardless of the legacy systems they already have in place. Users are not tied to a particular interactive voice response (IVR) vendor, or indeed architecture or system. This inherent flexibility enables any business to subscribe to URU Plus and verify their users through voice.

As an XML-based Web Service, URU Plus uses the same infrastructure as conventional Internet-based

systems. It is highly scalable, enabling new applications to be written in a matter of days, in contrast to the months taken to carry out similar tasks on conventional systems.

2. Why voice?

Speaker verification is a technology that has been researched for over four decades. In the early 1960s, Kersta of Bell Labs developed the first computerised speaker-verification system based on spectrographic voice verification. Kersta’s early system used visual pattern matching techniques to provide comparisons between spectrographic representations of speech (see Fig 1). As his system developed he coined the term voiceprint to refer to his spectrograms — a term that is widely used today. Although the voiceprint label persists, modern speaker-verification technology bears little resemblance to visual comparisons between spectrograms. State-of-the-art technology has converged on a stochastic approach that models characteristic features extracted from speech. Most implementations use hidden Markov model or Gaussian mixture model techniques.

BT has been involved with speaker-verification technology since the late 1980s when it ran its first speaker-verification trial with a major high street bank. Although considered successful from a technical standpoint, the early in-house technology was not adopted. Moving on from the 1980s, speaker verification has steadily improved with the advent of

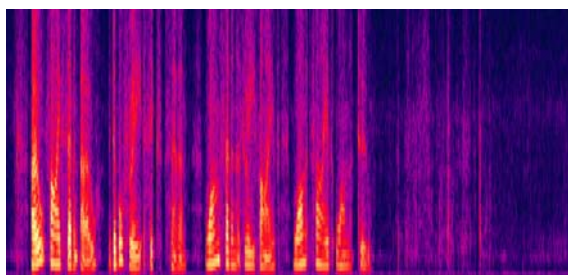


Fig 1 Example of a voiceprint.

new algorithms encompassing the latest innovations in statistical modelling and in signal processing techniques. Equal error rates (EER)¹ in the region of 0.5% are now achievable, though the exact performance depends very much on factors such as the type of service implemented, the user base and the conditions under which the system is operated.

Most commercial security schemes employ two-factor authentication, e.g. password and Secure ID token, PIN and Secure ID token, or personal information and password. Although this provides improved security over single-factor authentication, none of these actually proves that the person making the authentication is genuine — PINs, passwords, security tokens and personal information can all be obtained by impostors.

Biometric technologies, distinct from all other forms of authentication, do actually prove that the user is who they claim to be. These can provide a powerful authentication solution especially when used in conjunction with conventional approaches. This is the URU Plus philosophy — a tool-kit of multi-factor authentication capabilities underpinned by voice biometrics.

A significant drawback of most biometrics is that they require specialised client-side measuring equipment, such as fingerprint readers, iris scanners, cameras with special lighting conditions, etc. A crucial advantage of voice over other biometric technologies is that it does not require specialised hardware for the user interface, the only requirement being a telephone (fixed or mobile) — an everyday piece of kit with which all users are already familiar. Users can enrol and verify remotely via a well-established and pervasive infrastructure.

Potentially the whole of the UK population could immediately use a voice-based biometric service with little or no training and at zero cost for the client-side hardware interface.

¹ Biometric systems use one or more preset threshold values which determine their false acceptance rates and false rejection rates. When the thresholds are set such that these rates are equal, this value is referred to as the equal error rate.

URU Plus is a text-dependent speaker-verification system as it is intended for unsupervised verification applications. An unsupervised verification is one where a human agent is not involved with the call and the caller speaks directly to the IVR. A text-dependent system is required for unsupervised applications so that the caller can be 'controlled'. If a text-independent system was used (i.e. the caller can speak any utterance), then an impostor could surreptitiously record the genuine user's voice when they were having any non-specific conversation. The recorded speech could then be played to the unsupervised text-independent system and therefore grant access to the impostor. If, on the other hand, specific speech is requested during a verification episode, a general non-specific recording of the user would be of no value when trying to gain access.

3. Challenges in speaker verification

In the laboratory, the classic speaker-verification system scenario requires calm articulate users to enrol on the system and donate several examples of their speech. Participants are generally expected to provide utterances from various locations and from various telephony handsets. The practice of covering different environments and handsets is intended to cover the range of possibilities under which the user is likely to utilise the system. For best performance, users are generally expected to supply speech over an extended time period, e.g. a few days to a week, or even a month, in order to model subtle changes in voice characteristics.

As with all biometric technologies the performance is, to a certain extent, at the mercy of the user. For a given individual, the single most significant factor that affects performance is the quality and quantity of the enrolment speech. Secondary to this are the environmental factors such as background noise, telephony channel and type of handset used. Variations in the user's voice can occur for a variety of reasons — they may be tired or they may be stressed, they could have a cold, or their voice may have changed over time. For text-dependent systems (where a specific phrase is requested) the user might speak inconsistently during enrolment, thus the voiceprint produced would be a poor representation of the speech. During verification they might speak the wrong phrase which would cause a text-dependent system to reject the genuine user. Modern speaker-verification systems do address these issues, but nevertheless they are all sources of error and do reduce overall performance.

The main challenge for speaker-verification systems is to build robustness in the following respects:

- robustness to voiceprint ageing,

- robustness to intra-speaker variability,
- robustness to background noise,
- robustness to channel effects,
- robustness to play-back attacks.

3.1 *Robustness to voiceprint ageing*

The human voice changes subtly over time. The consequence for speaker verification is that the user's voice will initially match very well to a newly formed voiceprint but will match less well as time progresses. To combat this effect a technique known as adaptation is commonly used. This ensures that the voiceprint always represents the most up-to-date speech of the user.

The adaptation technique involves updating the voiceprint as it is used. If, for example, a user enrolls using three enrolment samples on a single enrolment session and then subsequently verifies their voice a few days later, it is likely that their voice will not have changed significantly over this short time period. When the user is successfully verified with a high degree of confidence, their verification speech is added to the voiceprint so that it now represents a more realistic range of variation of the user's speech. This update process is applied periodically so that at any instant the voiceprint is up to date and truly represents the current characteristics of the speech.

3.2 *Robustness to intra-speaker variability*

Inter-speaker variability, the variation of speech between different speakers, is the effect that makes speaker verification possible. The greater the inter-speaker variability between true speaker and impostor, the more accurate a system is likely to be.

Another variation known as intra-speaker variability refers to the variation of speech from a single speaker. Intra-speaker variability is demonstrated when a speaker pronounces the same word or phrase but cannot repeat the utterance in exactly the same way. Intra-speaker variability is caused by different speaking rates, the emotional state and the speaking environment, e.g. speaking against background noise. The latter, known as the Lombard effect, is the tendency to increase one's vocal intensity and to modify intonation when speaking in a noisy environment. Intra-speaker variability is a source of error in speaker-verification systems.

3.3 *Robustness to background noise*

Acoustic background noise affects the performance of speaker-verification systems. There are two categories of background noise — stationary noise and non-stationary noise.

A noise source is termed stationary if the statistics of the noise do not vary with time, or if they vary slowly with time (quasi-stationary noise), examples of stationary noise being mains hum and fan noise from computer equipment. A simple method for dealing with stationary or quasi-stationary noise is spectral subtraction. Operating in the linear power spectrum domain, the average of the interfering noise is estimated across the spectral shape of the signal and subtracted from it.

The term non-stationary noise refers to fast varying noise. Typical examples include a television playing in the background or the general background noise in a public place such as an airport or railway station. Non-stationary noise is much more difficult to deal with as it varies too quickly to estimate a long-term average.

It is particularly difficult to deal with background speech babble as the speaker-verification system has no way of separating the speech from the genuine user from that background speech prior to verification. Nevertheless, speaker-verification systems can still operate reasonably well in such environments, providing the background noise is not too loud. In general, if the background speech is at a level such that it would be comfortable to have a telephone conversation, then it is likely that the speaker-verification system would verify correctly.

3.4 *Robustness to channel effects*

Variability in telephony channel can be a source of error. This is referred to as the cross-channel effect. For speaker verification, an unknown speech signal is compared with a reference voiceprint — the voiceprint itself having been constructed from previously acquired speech. Successful verification is achieved when there is a suitable match between the two. For telephony applications, the unknown signal is received over a telephony channel. The characteristic of that channel has a filtering effect which distorts the speech signal, as does the handset microphone. In cases where the enrolment speech was collected over the same type of channel and handset microphone, the distortion effect is roughly the same. Consequently, unless the distortion is particularly acute, the comparison between voiceprint and speech signal is not usually adversely affected. If, on the other hand, the enrolment speech is collected on a different type of channel from the subsequent verification speech, e.g. GSM network and fixed network, then performance does degrade.

There are several approaches to address this problem. The best and most effective approach (though perhaps not the most practical) is for the system to employ multiple voiceprints for each user, e.g. one generated with speech from fixed network calls, one

from GSM and one from voice over Internet protocol (VoIP).

Alternatively, speech from the various channel sources can be incorporated into a single voiceprint. This technique can have a blurring effect such that the voiceprint now models an averaged fixed/GSM/VoIP channel. Nevertheless, the stochastic modelling approach, the essence of most speech-based systems, facilitates the incorporation of mixed channel data to a certain extent, but this is not as accurate as using separate models for each channel.

There are also several techniques for compensating for the telephony channel effect [2—4]. Many of these techniques amount to removing the long-term average channel value which is effectively a multiplicative constant in the frequency domain.

3.5 Detection of recordings — playback attacks

Detection of recordings is an issue that affects all speaker-verification systems and there is no fail-safe solution for this.

The root of the problem lies in the opposing requirements:

- make the system robust to variations in telephony channel, i.e. ensure that the system accepts a genuine caller's speech irrespective of the channel,
- make the system robust to the playback of recordings, i.e. ensure that the system rejects a genuine caller's speech if a recording of their voice is transmitted through playback equipment, which is, from the speaker-verification system point of view, just another channel.

There are various strategies that can be employed to reduce the possibility of playback attacks on unsupervised text-dependent systems. The simplest and most effective technique is to vary the verification speech that is requested. This reduces the chance of an impostor being able to use any speech that they may have recorded from the genuine user.

Another more sophisticated approach is in the detection of identical utterances. Paradoxically, the fact that humans cannot produce identical utterances, even in quick succession, can be put to advantage in the detection of recordings. If a speaker-verification system hears an utterance that has previously been used on the system, it is possible to detect the fact that the utterance is identical to a previous utterance. Thus, if an impostor were to secretly record a genuine user speaking to the system, the system would already be aware of that particular utterance and would therefore

be able to flag a repeat of that utterance as being a recording.

4. The URU Plus system

Speaker-verification systems are commonly coupled with applications where the voice verification functionality is closely coupled with the IVR. The two typically run on the IVR platform within banks or call centres and are generally only used in that particular context (see Fig 2).

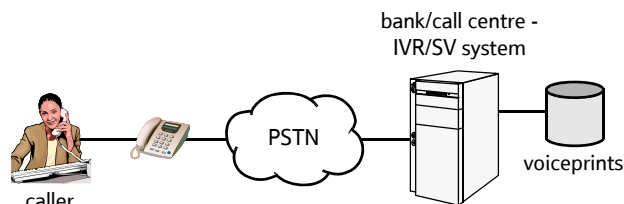


Fig 2 Conventional speaker-verification system.

As part of the BT 21st century network architecture (21CN) authentication capability, the philosophy for URU Plus is quite different from the conventional approach. The idea here is to provide an effective reusable speaker-verification component that can be encapsulated in many different applications. Applications running on external customer IVRs and those running on BT-hosted IVRs share the same speaker-verification component and can also share the same voiceprint store (see Fig 3).

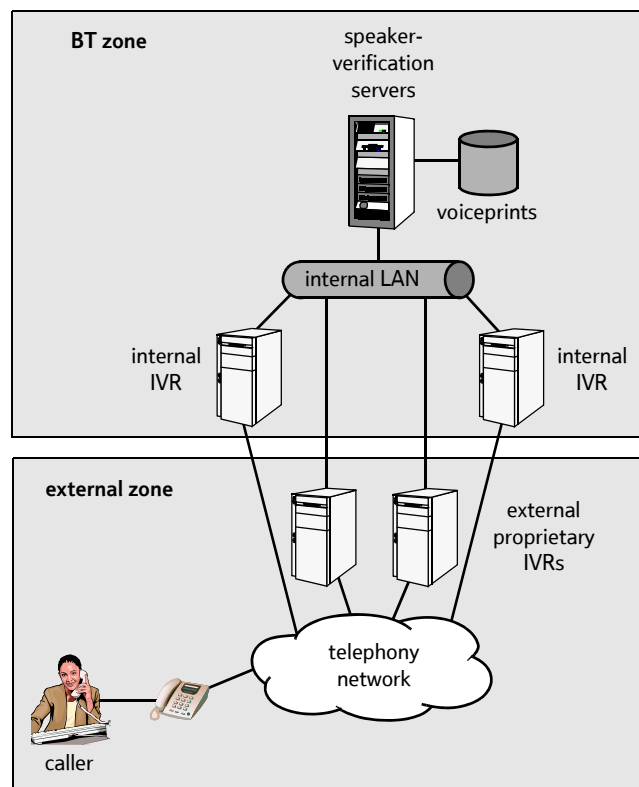


Fig 3 URU Plus speaker-verification system.

BT's 21CN capabilities are usually exposed as Web Services, although other interfaces may be used. The overall objective behind the capability approach is service-based reuse. Rather than creating a complete application and support infrastructure for each product, products are composed of 21CN element aggregations. Thus, within the 21CN authentication capability, a speaker-verification component is available to any 21CN BT product proposition.

4.1 Scalability — the loosely coupled approach

Speaker-verification servers are connected in a loosely coupled manner, using Web Services (SOAP, simple object access protocol over HTTP). This is complemented by using Web Services security to sign and identify aspects of the data exchange which are publicly exposed. This approach to the system design allows the same voiceprint store to be used in a range of contexts without a specific system dependency such as the IVR (see Fig 4).

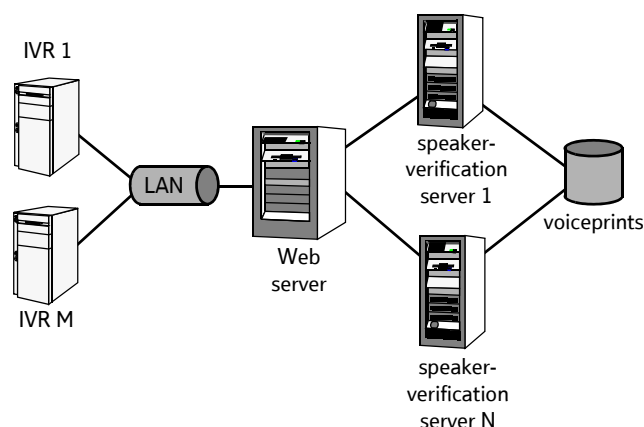


Fig 4 URU Plus speaker-verification system.

The Web server hosts a number of 'Web sites and virtual directories', in this case hosting the Web Service interfaces that provide access to the speaker-verification software. Since standard Internet components are used, such as the Web server, standard approaches to scalability can be taken. No state or session information need be preserved between calls to the speaker-verification functionality. This simplifies scalability — just add more servers when required and distribute the load using network load balancing. Each server acts as an isolated instance, connected to a database cluster to maintain access to a common voiceprint store.

4.2 Speaker-verification component

The speaker-verification component is implemented using Microsoft's .NET technology. This facilitates straightforward integration into the Web Services environment. The speaker-verification component is text dependent and language independent. It does not

require knowledge of language as it operates directly on the sound of the utterance irrespective of language or accent. This property is extremely important for a generic component which will have multiple applications across different languages and accents. The text-dependent, language-independent property is emphasised here as not all speaker-verification vendors provide this functionality.

As an example, an alternative approach for providing text-dependent speaker verification would involve the use of speech recognition to enforce text dependency. This would run on top of a text-independent system. Such an approach can have advantages, e.g. the ability to switch between text-dependent and text-independent applications, but the downside is that it requires knowledge of the languages spoken in order to set up the appropriate speech recognition lexica and grammars for the speech recognition aspect of the task. For a specific stand-alone system using English, this would be less of a problem, but for multiple applications, allowing multiple languages (more than 300 languages are spoken in London), this approach would be problematic.

5. Secure Web site access — URU Plus WebCheck

The URU Plus component-based approach enables the development of sophisticated applications. The following application demonstrates secure Web site access.

In this scenario the user accesses a secure Web site using speaker verification. The user makes an identity claim at the log-in screen by keying in their account number and password. The system then displays a 'call me' button. When the button is clicked an IVR system calls the user on their preferred telephone number (mobile or fixed). The IVR asks the user to speak a specific phrase for verification. They are subsequently accepted to the Web site if their voice matches the voiceprint corresponding to their claimed identity. Every time a user's identity is checked, they are notified by e-mail and SMS (see Fig 5).

The system provides a number of desirable security features compared with conventional password authentication. The user has an independent channel for content and speaker verification. This provides secure three-factor authentication (see Fig 6):

- password via the Internet,
- speaker verification via the telephony network,
- possession of a specific telephone number (token).

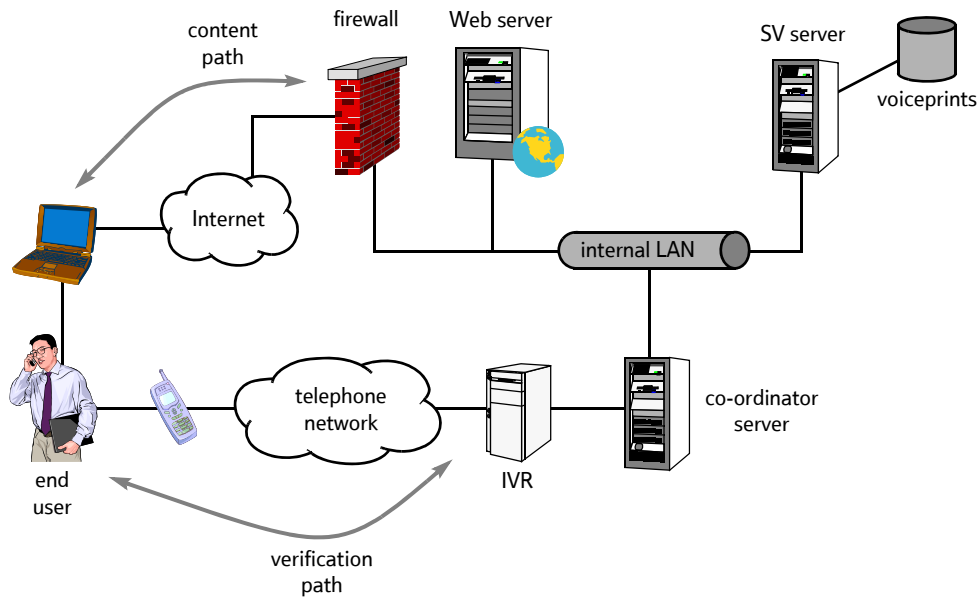


Fig 5 URU Plus WebCheck.

In addition to the three-factor authentication, the system also notifies the user, via SMS and e-mail, every time their identity is checked. This audit facility acts as an additional security control in detecting possible abuse of the system.

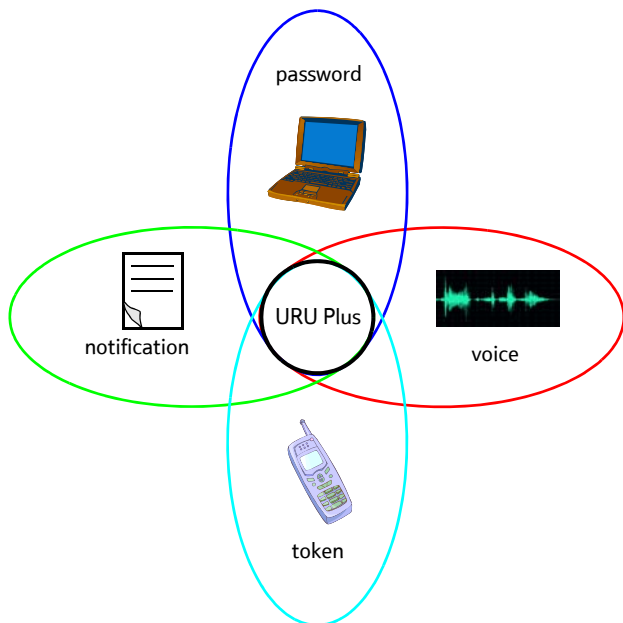


Fig 6 Three-factor authentication with additional security via the notification audit facility.

5.1 Enrolment

Before a user can verify via the system, they have to enrol. The enrolment process involves validating the claimed identity of the applicant and then collecting the reference voice samples that are used to generate the voiceprint for use in subsequent verifications. The initial

validation stage is very important as it is essential to enrol the correct identity at the outset. This is achieved using BT's text-based validation service — URU [1].

Following a check of the credentials of the user, the sequence of messages exchanged for enrolment is shown in Fig 7. The sequence starts with the user navigating to an enrolment page within a Web-based application. Details such as the telephone number and enrolment ID are collected.

The Web-application makes a Web Service call to the co-ordinator object, passing the enrol request information. The co-ordinator object checks that the identity is not already enrolled and a response is returned immediately. This contains a reference for the request. The co-ordinator places the enrolment on the request queue (proc tab). The request queue is processed by a background process within the co-ordinator (cmservice). The request is then passed on to the IVR. The IVR makes an outbound call to the enrol subject. The enrol subject answers the call and a prompt script is played — this provides guidance on the enrol process. The enrol subject provides their first utterance, the IVR passes this to the co-ordinator object, which then passes the utterance on to the voice-verification system (URU Plus). This sequence is performed three times. Providing that the utterances are sufficiently consistent, the user enrolment will succeed. If the enrolment fails, this is communicated to the IVR and the Web application.

5.2 Verification

The sequence of messages exchanged for verification is shown in Fig 8.

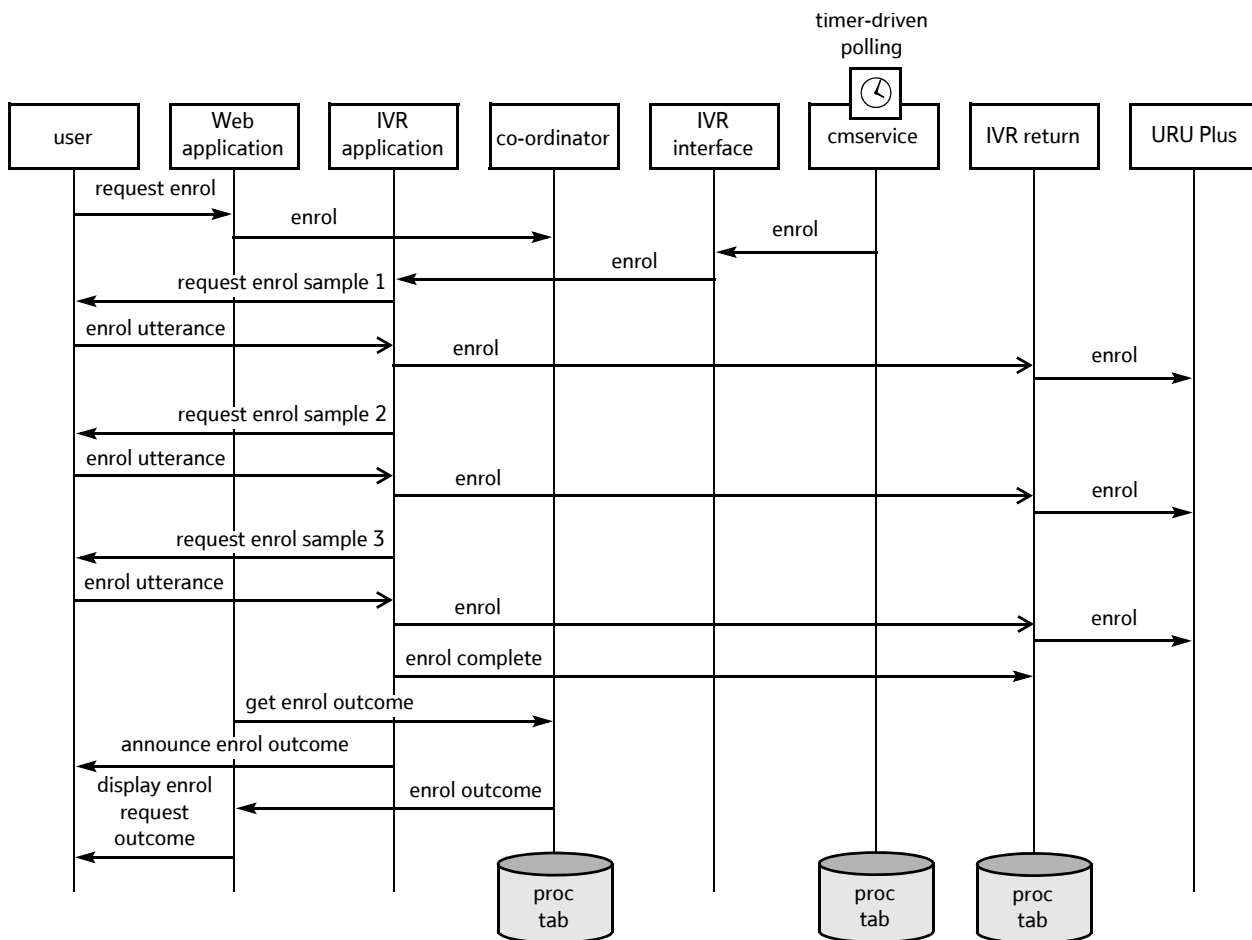


Fig 7 UML enrolment messaging sequence.

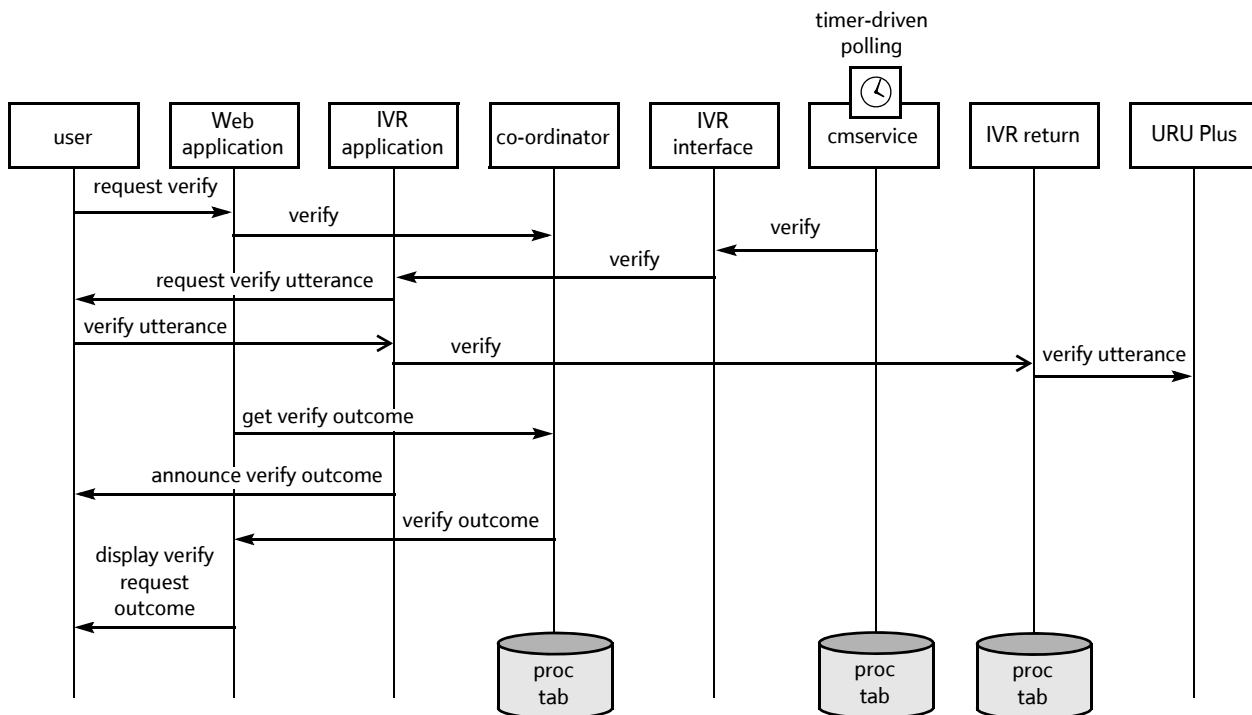


Fig 8 Verification messaging sequence.

The sequence starts with the user navigating to a 'verify page' within a Web-based application. The user makes an identity claim. The Web application makes a Web Service call to the co-ordinator object, passing the verify request information. The co-ordinator object checks that the identity is enrolled on the system and a response is immediately returned containing a reference for this request.

The co-ordinator places the verify request on the request queue (proc tab). The request queue is processed by a background process within the co-ordinator (cmservice). The request is then passed on to the IVR. The IVR makes an outbound call to the user. The user answers the call and a voice prompt script is played. This provides guidance on the verify process. The user provides their utterance which must contain the same phrase used during enrolment. The IVR passes this to the co-ordinator object which then passes the utterance on to the voice verification system. The voice verification system returns the verify outcome. The outcome is announced as success or failure by the IVR system (failure is actually announced as a system error by the IVR, thus avoiding giving too many hints to an impostor).

5.3 Speaker verification accuracy and tuning

The decision to accept or reject a caller is based on the quality of the match between the caller's voice and the voiceprint corresponding to that of the claimed identity. The system returns a probability score indicating the extent to which speech matches the model. The actual decision to accept or reject depends on whether or not the probability score exceeds a predetermined threshold.

Data for the calculation of a performance curve is obtained via a calibration procedure. This entails the collection of representative speech samples from a group of speakers. Calibration is performed during system development, with speakers who may not be users of the eventual system. Off-line processing allows each user's enrolment speech samples to be matched against their own verification speech samples, to generate a true user score distribution histogram (see Fig 9). Each user's speech is also matched against every other user's speech (within male and female groups) in order to generate an impostor score distribution histogram (see Fig 10). It can be seen that true users tend to obtain high scores, whereas impostors tend to obtain low scores.

A typical performance curve (in this case for callers speaking a 16-digit credit card number over a variety of telephony channels) is illustrated in Fig 11.

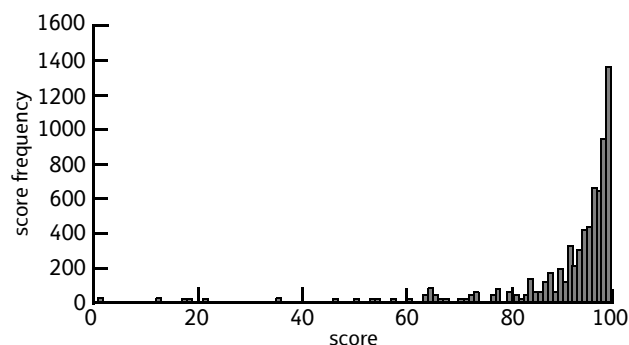


Fig 9 True user scores.

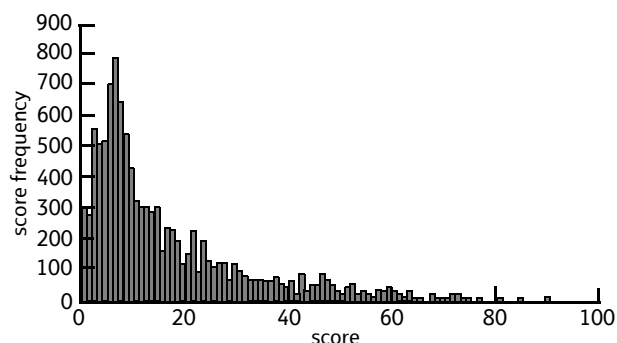


Fig 10 Impostor scores.

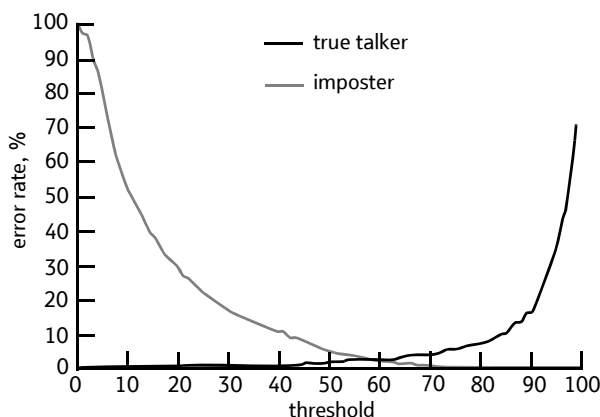


Fig 11 Text-dependent speaker verification performance curve. Enrolment — three versions of a 16-digit credit card number. Test utterances from fixed and GSM networks.

The performance curve is derived from the histogram distributions (true user scores (Fig 9) and the impostor scores (Fig 10)). The true user portion of the curve is calculated by plotting, for each possible score, the percentage of true users whose scores are less than that score. Similarly, the impostor portion of the curve is calculated by plotting, for each possible score, the percentage of impostors whose scores are greater than or equal to that score.

For applications that require high security it would be appropriate to set the threshold relatively high. As an

example, a threshold of 85 would prevent all impostors from gaining access to the system, but would consequently make it relatively difficult for true users to gain access — approximately 8% of true users would be rejected. Conversely for a lower security application the threshold could be relaxed, making it easier for true users to gain access, but at the expense of making it easier for impostors to access the system as well.

The performance of a system is usually quoted in terms of equal error rate (EER). This is the threshold value where the percentage of true users being rejected by the system is equal to the percentage of impostors gaining access to the system. For the URU Plus system the EER value is roughly 1% for real-world mixed channel (GSM and fixed network) applications. This 1% value encompasses errors made by the users, errors due to channel effects and errors due to background noise.

6. Conclusions

URU Plus is a tool-kit of multi-factor authentication capabilities underpinned by voice biometrics.

Telephony-based speaker verification is a pervasive low-cost way of including a biometric check. It provides a very strong binding between the presented credential (voice) and the user. A significant advantage of speaker verification over other forms of biometric is in the fact that the user does not require specialised equipment to use the system.

Although speaker-verification performance can be affected by various human and external factors, it does provide a powerful authentication solution. As an example, if used in conjunction with two (or more) factor authentication, an impostor would have to be in possession of the user's registered mobile phone, have knowledge of their password and be able to reproduce their voice on demand. For very high security applications even more security factors, such as secure ID token, can be introduced to the system.

The URU Plus system is unique in that it uses a loosely coupled, component-based approach. Whereas traditional voice-based systems are tightly tied in with a specific IVR and with specific speaker-verification

software running on that IVR, the URU Plus architecture can handle speech delivered from a variety of sources. URU Plus also provides the capability of sharing voiceprints between applications. This puts BT in a unique position to provide a generic identity verification service. The loosely coupled architecture allows new service applications to be constructed quickly and efficiently. Applications can be hosted by BT, or existing customer IVR applications can easily connect to the URU Plus, speaker-verification infrastructure.

References

- 1 Gahan C J: 'URU — on-line identity verification', BT Technol J, 22, No 1, pp 43—51 (January 2004).
- 2 Pawlewski M: 'Speaker recognition using spectral coefficients normalised with respect to unequal frequency bands', US Patent 5583961 (10 December 1996).
- 3 Pawlewski M and Downey S: 'Channel effects in speaker recognition', Proc of the Institute of Acoustics, 18, Part 9, p 115 (1996).
- 4 Reynolds D A: 'Channel robust speaker verification via feature mapping', Proc ICASSP, pp 53—56 (2003).



Mark Pawlewski joined BT in 1989 after working in the field of sub-sea acoustics for the North Sea oil industry.

Since joining BT he has worked in various research areas including speech technology, computer vision and security research. He has several years' experience in core algorithm development for automatic speech recognition and speaker verification. In his present role he is a Technical Group Leader in the BT Security Research Centre.

He holds six patents, two of which are in speaker verification, and also holds degrees in physics and software engineering.



James Jones joined BT in 1990, working on network management software.

He is currently working in the Security Research Centre on mobility security and is Technical Design Authority for URU Plus.

He obtained a BEng in Electronic Engineering from Manchester Polytechnic in 1989.