Global Information Assurance Certification Paper

Shedding some light on Voice Authentication

Dualta Currie

GSEC- V1.4b

Abstract

Biometric authentication technology and development has grown over the last 6 years from being something we have seen on Science fiction television shows into a reality where we can now purchase the systems and implement them both in our business and private lives.

In this paper I will attempt to explain, in non-technical language, the technologies behind one particular type of biometric authentication, voice authentication. I will look at the human voice, how this is captured by technology, and how this can then be used to verify that the person is who they claim to be.

Introduction:

People have always kept secrets and protected their possessions. It has evolved from the physical to the technological, where now technology is used to restrict access to our resources and user authentication is the 'key to the door'. Authentication techniques have developed with it, and now who can access our resources is controlled using three main methods: [1]

- Knowledge-based authentication is based on information authorised individuals will know, and unauthorised individuals will not. E.g. A PIN or a password. information.
- Object-based authentication is based on possessing a token or tool that permits the person access to the controlled resource. E.g. Keys, pass cards or a SecureId.
- Biometric-based authentication measures individuals' unique physical or behavioural characteristics. It exists today in various forms such as fingerprint verification, retinal scans, facial analysis, analysis of vein structures and voice authentication.

Reasons for using Biometric Authentication:

Biometric authentication has some key advantages over knowledge and token-based authentication techniques. Biometric characteristics are not easily forgotten, like a password, or lost like a key. One can hardly lend someone your finger nor can someone easily steal your eye. That makes them fairly secure, and convenient. Unfortunately, they've had to wait for technology to catch up to the level where it can support their effective use. Only recently has technology provided the statistical, analytical and data processing techniques to support it properly.

Why choose voice authentication?

For the majority of biometric authentication techniques, sophisticated equipment and the physical presence of the person being authenticated is required. For example, fingerprint scanning, pen signatures and retinal scans – not so with voice authentication, where authentication may be given remotely via a device commonly known as the telephone. Given the use of the correct analytical techniques, a person's voiceprint can be as unique as any other biometric characteristic, but yet can be used for authentication remotely and has the added benefit of being less personally intrusive than say, subjecting the person to a retinal or fingerprint scan.

The concept :

Voice authentication is a fairly simple process. To register, a user records sample(s) of their voice which are stored in the authenticating system and become known as their 'voiceprint'. Then, to access this resource subsequently, they supply a sample of their voice to the system, and it decides if it matches their voiceprint before allowing them access.

The risks of it's application:

When deciding whether or not to employ a voice authentication system it is important to consider the application. If it is to be used to authenticate a user to administer their bank accounts for example, this is a completely different risk than say accessing their voicemail on their mobile phone. Should a false acceptance result in the banking application, the consequences would be considered much more severe.

Remember : Risk = Threat x Vulnerability

The elements of any voice authentication system need to be analysed and it must be ensured that individually and collectively, the probability of a vulnerability arising is low, and the potential for an individual or group to exploit the vulnerability unlikely.

Voice biometrics explained:

Biometric characteristics fall into two broad categories :[2]

- Physiological Biometrics are concerned with the unique physical traits of the individual, for example retinal scans, fingerprints and face geometries.
- Behavioural Biometrics are concerned with the unique way individuals perform certain actions, for example conventional pen signatures and key stroke detection.

In the case of voice authentication, there is both a Physiological biometric component (for example, voice tone and pitch) and a behavioural component (for example, accent). This makes it very useful for biometric authentication.

Voice authentication, identification and speech recognition.[3]

Voice authentication, also referred to as 'verification', is just one of the voice-based technologies. Others include voice identification, interactive voice response (IVR), and degrees of speech recognition. These technologies share base technologies and methodologies, but differ considerably in terms of the extent to which reliance is placed on certain sub-technologies.

To allay any confusion between the various technologies, I will briefly describe the differences between them, before focusing particularly on the issue of voice authentication:

Voice (or speech) authentication attempts to verify that the individual speaking is, in fact, who they claim to be.[14] This is normally accomplished by comparing an individual's voice with a previously recorded "voiceprint" sample of their speech.

Voice identification attempts to identify the individual's voice. This normally involves comparing an individual's voice with a number of previously recorded samples of speech, in an attempt to ascertain which, if any, it closely resembles.

Speech recognition does not attempt to give any information as to the identity of the speaker, but instead attempts to determine what they are saying.

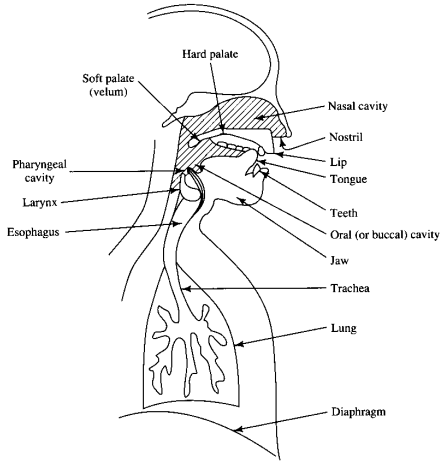
These technologies converge depending on the application, where often speech recognition is employed in conjunction with identification and authentication, hence the confusion.

Speech [4,5,6,15]:

Now that we are aware of the technologies voice authentication employs, we need to look at how the voice is produced, the characteristics that allow us to extract meaning from it, and the method by which it can be converted into a form that can be handled by computer systems. In doing this particular attention will be paid to those characteristics of the voice that render it unique for each individual, therefore allowing their identification. To do this, we can examine the physiological component of human speech, which is produced by the human voice tract.

In simple terms, the voice is created by air passing over the larynx or other parts of the vocal tract. The larynx vibrates creating an acoustic wave, essentially a hum, which is modified by the motion of the palate, tongue and lips. Sounds produced by the larynx are called *phonated* or *voiced* sounds. Examples of voiced sounds would be the *m* in "mud" or the *r* in "ram". Simultaneously, other sounds are produced by other parts of the vocal tract, for instance whispered sounds are created by air rushing over the arytenoids cartilage at the back of the throat. Sounds not originating in the larynx are called *unvoiced* sounds. Examples of these would be the *f* in "fish" or the *s* in "sea".[7]

Figure 1: The vocal tract[8].



All sounds produced are, at the same time, fundamentally influenced by the actual shape of the vocal tract. This shape is brought about both as a consequence of hereditary and developmental factors, and of environmental factors.

In parallel to these physiological characteristics, speech contains a behavioural component. This manifests itself as the accent of the voice, and affects how quickly words are spoken, how sounds are pronounced and emphasized, and what other mannerisms are applied to speech.

Together these physiological and behavioural factors combine to produce voice patterns that are essentially unique for every individual, and are difficult or impossible to duplicate short of recording the voice.

When analysed using modern technology, human speech appears to be rather inefficient, in terms of time and energy expended to transfer information. Speech is constructed out of various sounds, termed *phonemes*. Common English usage utilises around 40 phonemes, analogous to the characters of the phonetic alphabet seen in most dictionaries. For instance the word "mud" uses three phonemes denoted /m/ (the mmm sound), /u/ (the uh), and /d/.

Almost all the information in each phoneme can, in fact, be deduced from only a small fraction of the entire phoneme sound. For example, the *n* sound in the word "man" may take one or two tenths of a second to say. Yet, for analytical purposes, only the first 20 or 30 milliseconds and the last 10 or 20 milliseconds of the sound is

vital information. Once this information is known, the remainder, middle part of the sound is redundant; its sound, volume, duration and all other characteristics are inherently defined by the beginning and end. This is also true to a greater or lesser degree of all other phonemes. This means that each phoneme, and therefore speech itself can be reduced to a finite number of characteristic sounds.

In a practical sense this fact is of limited use in normal speech. Even if the human voice were capable of producing a shortened version of all phonemes, and the human ear able to perceive them, social and aesthetic factors would probably still prevent us from using them. Nevertheless, the fact that there exists a great deal of redundant information in human speech is important because it means that human speech can be successfully compressed. All it requires is the technology to remove the redundant components at the transmission side, and reintroduce them at the receiving side.

For instance, uncompressed human voice over a standard phone connection will require 64,000 bits of information per second (64kbps) to transmit. In comparison, a GSM phone using data compression needs only 13kbps. This would imply that up to 75% of human speech is unnecessary information. This has a bearing on our ability to process speech in real time. Rather than analyse the entire speech pattern, in effect we need only analyse the significant 20% to 25% of it.

The human vocal tract can produce sounds with frequencies of between around 100Hz and 8,000Hz. This range can be exceeded, on occasion by persons with trained voices. For example, operatic basses can produce notes of 75Hz and lower, while alto sopranos can occasionally exceed 8kHz. Most phonemes however use frequencies between 300Hz and 3400Hz, so telephones are designed only to transmit sounds in this frequency range.

In fact certain phonemes, particularly the soft *c* and *s*, and *f*, often use frequencies higher than 3400Hz but these are not transmitted over the line. The system still works because the human brain can unconsciously insert the missing sounds when necessary. This restricted frequency range is also the reason why music, which uses a much broader frequency range than voice, does not sound well over a telephone.

This range is restricted for technical reasons. It allows phone companies to frequency *multiplex* or “stack” multiple phone conversations on trunk phone lines between exchanges. If they were to allow a larger frequency range to each telephone, they could stack fewer conversations on a line, so they would need to lay more cables.

In practice the loss of the higher pitched phonemes is of little consequence to voice verification either. In fact, it is the lower pitched, voiced phonemes, the ones that are most dependent on the physical structure of the vocal tract, that are of the greatest use in voice verification. These phonemes are least affected by factors such as coughs, colds or mouth injuries.

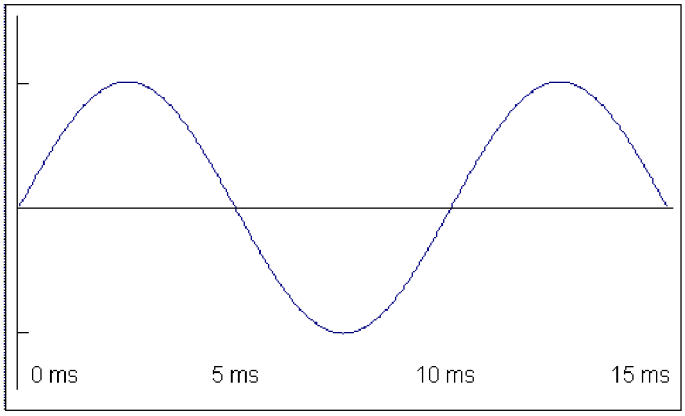
Digitising human speech.

The human voice produces a highly complex acoustic wave, which, fortunately, the human ear and brain have evolved to interpret effectively. In technical terms the voice is an *analogue* signal. An analogue signal is defined as one with a continuously variable physical value. A single frequency tone, such as the dial tone

on a telephone will be a simple analogue signal. Such a simple signal will take the form of a Sine Wave such as the one shown in Figure 2.

In explaining how analogue signals can be converted to digital data, I will use the example of the Sine Wave. This is appropriate because of a basic rule of signal theory called Fourier's principle. In essence, this rule states that any signal, even one as complex as a voiceprint, is in fact a combination of many simple sine waves, of varying frequencies (cycles per second) and amplitudes (strengths), added together. This means that any process that works for a Sine Wave, will work for any other kind of signal as well.

Figure 2: Sine Wave



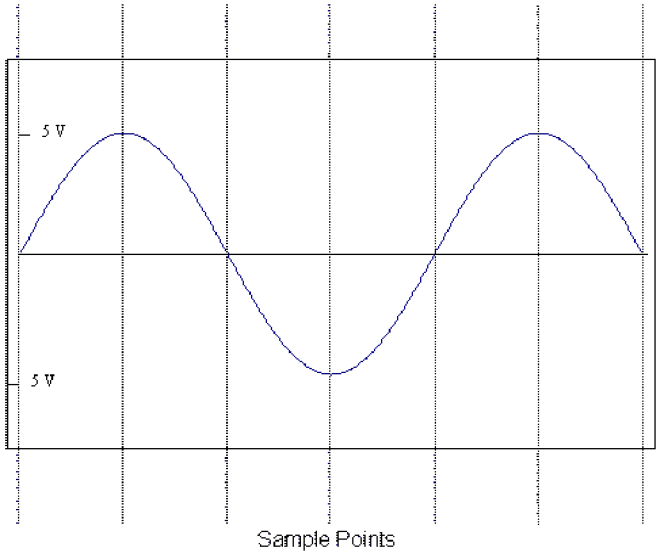
The initial problem is to convert this analogue sound signal into a digital signal (i.e. a sequence of numbers that a computer can input and manipulate). The first stage, converting it from a sound to an electrical wave is simple, using a microphone.

The second stage involves converting the signal from a continuous wave to a series of discrete voltage measurements. This is done by a process called sampling. Sampling involves measuring the voltage of the signal at regular intervals, many times per second.

If we use the sine wave of Figure 2 as an example: Say the wave has an amplitude of 10 Volts: that is to say its peak positive voltage is 5 Volts and its correspondingly lowest negative voltage is –5 Volts. We see that the wave goes through a complete cycle every 10 milliseconds (ms), this means that it completes 100 cycles every second, so it has a frequency of 100Hz. (Hz, *Hertz*, is the measure of cycles or measured events per second.)

To effectively sample this signal we need to measure its voltage every 5 ms, a "sampling frequency" of 200Hz. We know that we must use at least this sampling frequency because of a fundamental rule of communications called Shannon's Law, which says that in order to fully sample any signal, without losing any part of it, you must have a sampling frequency of at least double the highest frequency contained in the signal. Figure 3 shows the sampling of the signal.

Figure 3: Signal Sampling.



In our example, sampling gives us a series of voltage measurements 0V, 5V, 0V, -5V, 0V, 5V. In fact with a 200Hz sampling frequency we get 200 of these voltage measurements a second.

These sample voltages are measured and fed into a device called an analogue to digital converter. This device assigns a value to each measured voltage level. In telephone applications this assigned number is of the value 0 to 255, so that each voltage measurement is turned into one byte of data.

So in this case when the measured voltage is 5V the A to D will produce a value of 255, when the voltage is 0 the A to D will output 128 and when the voltage is –5V the A to D will generate a value of 0.

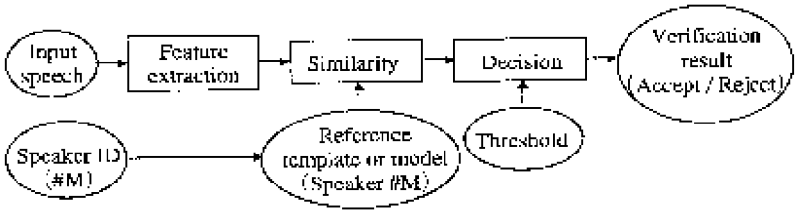
Our series of voltage measurements will therefore be turned into a sequence of numbers: 128, 255, 128, 0, 128, 255, 128

In the case of a voice signal, the number sequence will be by no means as predictable as in the case of a sine wave. Additionally, because the maximum frequency of the telephone voice signal is 4000Hz, the sampling frequency must be 8000Hz. Because each sample is one byte of data, we can see that digitising a telephone voice signal generates data at 8 kilobytes per second, or 64 kilobits per second.

Voice Authentication: how is it done?

Now that we understand how we can turn the sound wave that is produced by the vocal tract into a stream of data comprehensible to a computer, we must continue by looking at how this is used in authentication. Basically, how the sample of speech taken for authentication gets compared to the 'voiceprint' taken at registration.

Figure 4. Basic structure of Voice Authentication System[9]



There are a number of techniques used in registering a voiceprint, depending on the level of accuracy required from the authentication process. It may be that a single voiceprint sample is given. For example the user is required to speak a particular phrase or sentence into the system a single time. Typically this phrase will contain as many low pitch (frequency) phonemes as possible, these are the least susceptible to change. The recorded voiceprint is then stored as the template against which all that users' authentication attempts will be made. When the user wants to authenticate, they repeat the same sentence they used to register, which is then compared with the template voiceprint. The system will then decide if the users' voice and the template are sufficiently similar to authenticate the user.

In more sophisticated systems, the user might be required to repeat the same registering phrase a number of times. This will allow the system to construct a template composed of a range of voiceprints. Having more than one voiceprint is useful because people do not always say the same sentence the same way. Many people are not comfortable with being required to repeat words into a machine – in essence they suffer from "stage fright". They tend to become increasingly comfortable the more often they perform the task. Hence it is a good idea to take number of prints to ensure that the system can identify those vocal characteristics present in the voice which are there regardless of how uncomfortable or otherwise the speaker is while creating the voiceprint.

The disadvantage of systems where users have to register and authenticate with a given sentence is that it raises the possibility of a third party using some means to record an authentication attempt. Such a recording then need only be played back to the authentication system in order to gain access.

In order to overcome this vulnerability, more sophisticated systems use the registration process to generate a "general" voiceprint of the user. In this case the voice template is created from either a long registration phrase, or a number of different ones. This creates a voice template that is large and sophisticated enough that the system will have a reasonable chance of making an accurate authentication

regardless of the authenticating phrase. In such a system users will be required to speak a random phrase to authenticate themselves, which will be compared with the general voice template generated at registration.

The disadvantage of this is that the more general voice template produced by this approach precludes a voiceprint comparison technique called “Template Matching”. Because of the nature of the mathematical techniques used to compare templates with authentication phrases, the general voice template cannot be compared with a random authentication phrase with the same degree of reliability as identical registration and authentication voiceprints. This means that the comparison is less reliable for the general voiceprint approach than for the specific registration phrase method.

A compromise is possible though if users register with a number of registration phrases, but authenticate with only one of them, randomly chosen each time a user want to gain access to the protected system. This greatly lessens the chance of recording equipment being used to defeat the system, as recording any particular authentication attempt would still not guarantee that the authentication phrase would be the same at any later attempt. It also means that both template matching, and more accurate comparison techniques can be used.

A simple way of implementing this is to register users by getting them to count slowly and clearly into the system from, say, one to ten a number of times[10]. The system can then easily separate the ten spoken numbers from one another to create ten separate voice templates. To authenticate, the system then prompts the user to repeat three or four numbers, randomly chosen each time. For example “One Nine Six Two”. This approach has the advantage of making registration quick and straightforward, while still providing a very large number of potential authentication phrases.

It is worth noting that non-telephonic voice-verification systems have much less trouble with this particular threat. Most recording techniques, while producing high grade sound duplication at audible frequencies, also leave sound “artefacts” at frequencies above and below the audible spectrum. These artefacts can be identified by basic analysis of the signal. Unfortunately, the telephone system, by filtering-out very high and very low frequency sounds, usually prevents phone-based authentication systems from making use of this characteristic.

Some manufacturers of voice authentication systems claim that their products can differentiate between real and recorded voices. This is probably true in the case of low quality recording equipment and media, but as the quality of the recording improves, the chance of detecting a recording is likely to decrease significantly.

Comparing Voiceprints:

Up to this point we have simply assumed that the registration template and the authentication voiceprint are compared, and a yes or no answer produced. The comparison process is actually a lot more complicated than this.

The first thing to be noted about the voiceprint comparison process is that it never delivers an absolute positive or negative outcome. Any comparison will only give a probability, under particular conditions, that the registration template and the authentication voiceprint are from the same person. At best the system will be able to reliably state that there is an N% chance, in normal operating conditions, that the

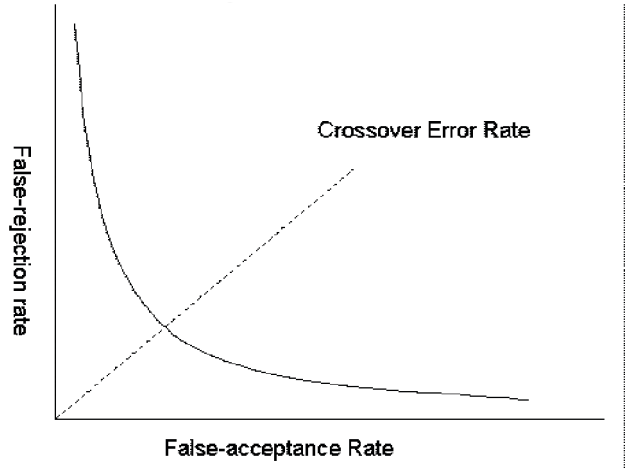
registration template and the authentication voiceprint were produced from the same person's voice. It is up to the individuals responsible for the resource protected by the voice authentication system to decide the percentage probability at which it is acceptable to permit a user access. In determining this "authentication threshold" two metrics need to be borne in mind:

- The False-acceptance Rate (FAR) is the percentage of invalid voiceprints incorrectly authenticated as valid users.
- The False-rejection Rate (FRR) is the percentage of valid users whose voiceprints are incorrectly rejected.

The FAR and FRR are inversely interdependent to one another. Generally, attempts to reduce the FAR will result in an increase in the FRR. That is to say that as the authentication system demands a higher degree of confidence that invalid voiceprints will not be accepted, the probability that valid voiceprints will be rejected increases.

It is a characteristic of the mathematical techniques used to compare voiceprints, and the nature of the voiceprints themselves that efforts to decrease the FAR will not in fact increase the FRR by the same extent. A typical FAR vs. FRR relationship diagram is shown in Figure 5. It can be seen that if a very low FAR is required, it is necessary to tolerate a very high FRR, and that at very low FARs small reductions in the false-acceptance rate can only be bought at the expense of large increases in the FRR. Setting acceptable FAR and FRR is invariably a compromise decision for the operator of the authentication system.

Figure 5: FAR vs. FRR.

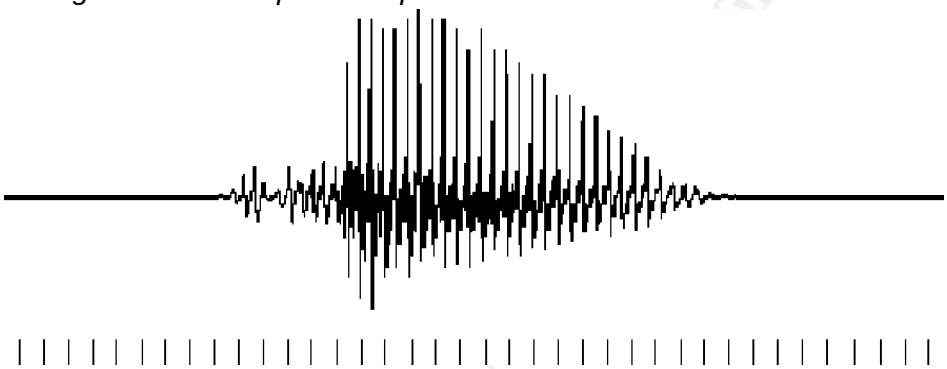


A third comparison metric, the Crossover Error Rate (CER) can be found for any voice verification system. (Indeed, the FAR, FRR and CER can be applied to any biometric authentication system.) The CER is defined as the error rate where the FAR equals the FRR. The lower the CER is, the more accurate and reliable the authentication system is likely to be.

The Mathematics of Voiceprint Comparison

Figure 6 shows a sample voiceprint. This particular voiceprint shows the total signal strength, over all frequencies, varying over a period of around half a second. (Each "tick" along the bottom represents the passage of 10ms.) We could, however, extract from this voiceprint, more specific prints taken at individual sound frequencies. In other words, the print is essentially a three dimensional entity, it varies both in terms of signal strength, over a spectrum of frequencies, and over a period of time. Together these three dimensions come together to form a complex and unique vocal "fingerprint."

Figure 6. A Voiceprint sample



There are two standard methods for comparing such voiceprints to permit voice authentication: Template Matching and Feature Analysis.[11]

Template matching (TM) is the simpler of the two techniques and the more accurate when used properly. Unfortunately it also has the most limitations. TM compares the digitised version of a voiceprint against a digitised template, without performing any significant modifications to either print. It attempts to work out the probability that one voice print is the same as another voice print based on comparisons of the amplitude of the voice signal at various frequencies at various times over the entire period of the authentication phrase.

In theory this works well, and under ideal circumstances it is very effective, producing a low CER. The technique is, however, very susceptible to the presence of background noise both during the creation of the template and at the time of authentication. Because the technique does not refine the voiceprint to any great degree, it cannot distinguish between the background noise and speech. Hence, if the background is noisy at registration, in order for authentication to occur, the background must also be noisy at authentication. Not only that, but the background noise must be the same type of noise.

Template matching is not completely unusable, however. As long as the signal to noise ratio is high, both at registration and authentication, the technique is effective. If a relatively high FAR is tolerable, then there is no reason why it should not be used

in voice verification. In applications where a high FAR is not acceptable though, it is preferable to use the other technique, Feature Analysis.

Feature Analysis (sometimes called Feature Extraction) does not really use any characteristic of speech that can be described in simple terms. It is simply a facet of human speech that when digitised and subjected to certain mathematical techniques, can be reduced to a series of mathematical values. These values cannot be said to have a direct bearing on any physical characteristic of speech. What can be said is that they have more to do with the way sounds change into one another both inside and between phonemes. When they are generated from human speech, particularly speech containing the lower pitched phonemes, they are an extremely reliable way of comparing one voiceprint with another.

Feature Analysis is useful in that it is far more resilient in the case of weak signal strengths or significant background noise. The mathematical techniques used in FA tend to isolate and amplify the particular mathematical features of human speech, while ignoring or lessening those of less structured sounds, typical of background noise. Hence, the majority of voice authentication systems use feature analysis as their primary means of comparing voiceprints.

This is not to say that template matching does not have its uses in this application, however. As a technique it is better at comparing some of the higher frequency phonemes and some other features of voice that FA might not necessarily pick up.

Because of this, some voice authentication systems use both techniques in parallel. It is understood that in real life situations that any template matching system would need to be set with a relatively high FAR, to compensate for its inability to cope with background noise. Nevertheless, in parallel, there is still the potential that a feature analysis system might miss a tell tale difference, which the template matching system picked up. In this way the system would at least have a "second line of defence."

Is Voice Authentication being used?

Business-level voice verification systems have been available for around six years. Large numbers of reputable companies have implemented systems based on computer voice technologies, such as Chase Manhattan Bank, Prudential Securities, Charles Schwab, Visa, AIB Bank and Trintech [10]. Indeed, voice authentication systems have also been used by US and UK police forces to keep track of individuals on bail, parole or curfew orders.[12]. Many companies that employed these systems for one application later extended it's use to others, finding that it's effectiveness and the cost savings it generated were too good to ignore.

As for the individual, products are available which allows private users to use voice authentication to control the extent to which family members can browse the Internet, ensuring children cannot access inappropriate sites. E.g. Deep Space Nine voice print product.[13], or to remove the need for typing by using Dragon Naturally Speaking to name but a few.

The Verdict:

Current opinion is that voice authentication technology is about as reliable as automated fingerprint reading technology. Crossover Error Rates of less than 1% would appear to be the norm. This would imply that over an extended period of use, the FRR would typically be lower than for conventional passwords or PINs, which customers tend to forget at regular intervals.

In the author's opinion, Voice authentication has entered the mainstream as a verification technology. As mobile phone-based "m-commerce" applications and web-based "voice portals" become more prevalent, voice authentication will become increasingly common in many sectors of the economy. In the short to medium term it is not unreasonable to predict that voice authentication will become a standard customer verification technology.

References:

1. A Practical guide to Biometric Security Technology, Simon Liu and Mark Silverman, 2001, IEEE

http://www.computer.org/itpro/homepage/jan_feb01/security3.htm

2. Hendry, M, "Smart Card Security and Applications", Artech House

3. Privacy Laws & Business 9th Privacy Commissioners' / Data Protection Authorities Biometrics as a Privacy-Enhancing Technology: Friend or Foe of Privacy? Workshop, Dr. George Tomko, Sept. 15th, 1998

http://www.dss.state.ct.us/digital/tomko.htm

4. Speech Technology for Telecommunications, F.A. Westhall, R.D. Johnston, A.V. Lewis

http://www.eee.bham.ac.uk/woolleysi/downloads/speech/2.pdf

5. Speech Links, Andrew Hunt

http://www.speech.cs.cmu.edu/comp.speech/

6. Making Speech Sounds – Phonetics, Oxford university press

http://www.oup.com/pdf/elt/catalogue/0-19-437239-1-b.pdf

7. Phonetics and the theory of speech production

http://www.acoustics.hut.fi/~slemmett/dippa/chap3.html

8. Speaker Verification, Biometrics Research dept, MSU

http://biometrics.cse.msu.edu/speaker.html

9. Speaker Recognition, Identifying people by their voices, BRNO University of Technology, Ing. Milan Sigmund, 2000

http://deer.ro.vutbr.cz/nakl/habilit/sigmund.pdf

& Speaker Recognition, Sadaoki Furui, NTT Human Interface Laboratories, Tokyo

cslu.cse.ogi.edu/HLTsurvey/ch1node9.html

10. Visa gets behind voice recognition for e-commerce, Paul Roberts, IDG News Service 10/21/02

www.nwfusion.com/news/2002/1021visa.html

11. Voice Recognition, Jim Baumann, Human Interface Technology Laboratory, University of Washington.

http://www.hitl.washington.edu/scivw/EVE/I.D.2.d.VoiceRecognition.html

12. CWN – News and information for Coventry and Warwickshire, Coventry city council news, 21/03/01, Coventry internet developments ltd.

http://www.cwn.org.uk/politics/coventry-city-council/2001/03/010321-young-offenders.htm

13. Star Trek Deep Space Nine Voice Print Security (advertisement)

http://www.cybertown.com/qvoice/qvoice.html

14. FAQ on Speech Recognitions and voice biometrics, J. Markowitz Consultants

http://www.jmarkowitz.com/ask.html

15. Review and evaluation of Biometric Techniques for Identification and Authentication - Final Report, Infosec, European Communities 2000

http://www.cordis.lu/infosec/src/stud5fr.htm

16. Ammemheuser, Maura, "Banks roll out voice-enabled machines" *Bank Systems & Technology* Aug. 2000.

17. Baker, Glenn "Voice Power" *New Zealand Management* Aug. 2001.