The ITB Journal

Volume 5 | Issue 1 Article 20

2004

Emotion Authentication: A Method for Voice Integrity Checking

C. Reynolds
Middlesex University, School of Computing science, Bramley Rd, Oakwood, N14 4YZ UK., c.reynolds@mdx.ac.uk

L. Vasiu
Middlesex University, School of Computing science, Bramley Rd, Oakwood, N14 4YZ UK.

M. Smith
Institute of Technology Blanchardstown, Blanchardstown Road North, Dublin 15, Republic of Ireland.

Recommended Citation

Reynolds, C.; Vasiu, L.; and Smith, M. (2004) "Emotion Authentication: A Method for Voice Integrity Checking," *The ITB Journal*: Vol. 5: Iss. 1, Article 20.
doi:10.21427/D75F20
Available at: https://arrow.tudublin.ie/itbj/vol5/iss1/20

Emotion Authentication: A Method for Voice Integrity Checking

C. Reynolds[1], L Vasiu[2] and M. Smith[3]

[1] Middlesex University, School of Computing science, Bramley Rd, Oakwood, N14 4YZ UK.
[2] Middlesex University, School of Computing science, Bramley Rd, Oakwood, N14 4YZ UK.
[3] Institute of Technology Blanchardstown, Blanchardstown Road North, Dublin 15, Republic of Ireland
Contact email: c.reynolds@mdx.ac.uk

Abstract

When people communicate with telephone type systems, it is often assumed that the listener would notice any modification of the speaker's voice. It is possible however to introduce small changes that would not be noticed by a listener but could modify the reading of a Voice Stress Analyser, popularly referred to as a lie detector. Existing approaches to checking the integrity of voice require significant amounts of processing or are able to detect only non-subtle modification such as change of speaker. With the advent of real time voice modification using software and hardware based signal processing, we argue that it is not possible at present to easily determine if a received unencrypted voice message has been modified in some subtle way. This is particularly important in the current climate of biometric and psychophysiological analysis. We critically investigate alternative approaches to a voice integrity check in the light of the subtle changes that might be made to a speaker's voice and propose a method that enables voice integrity checking for digital communications in a variety of scenarios.

Keywords: Voice Integrity, Emotion, Voice Stress Analysis, Voice Communication.

1.0 Introduction

Current digital communication systems are becoming integrated with local network services in a move towards greater mobility and flexibility. For example, Voice Over Internet Protocol (VOIP) is used locally within networks and assists in reducing costs. Within this new environment new technologies such as speaker identification systems are developing. These technologies are becoming more robust and the gathering and usage of biometric and psychophysiological information is increasing.

In many voice communication systems the voice information is encrypted using robust techniques and this may act to prevent third party real-time modification of the voice communication. With some new developments such as VOIP implementation, secure encryption adds to the processing overheads and the latency inherent in the communication. Therefore much of the communication traffic may run as unencrypted data to keep latency to a minimum especially where transmission delays are an issue.

It is possible to access VOIP communications using packet-sniffing approaches. This gives rise to the possibility of interception and transmission of the voice data packets, in an attack commonly called the man-the-middle-attack. Until recently the modification of a voice signal would be difficult to achieve in real time or would involve such delays or gross distortion of the original signal that it would be likely that such an attack would be detected. It is now possible to modify voice parameters subtly that give rise to possible privacy and security risks.

The low cost of secondary storage space required to store digital data is enabling long-term storage of communications sessions. Often this may be for prevention of fraud or training purposes, but in many instances, data is kept as a matter of course and may find use in consumer profiling.

With Voice Stress Analysis (VSA) becoming ubiquitous as a method of determining truthfulness and currently popular in the U.K. insurance market, it is important that any data gathered that might find later use has not been tampered with by a third party or those storing data. We propose that using the emotion cue content of a speech provides a robust method for checking message integrity and can be used to check for third party modification of the voice. Speaker Identification and authentication would not form the basis for message integrity, as it is possible to make subtle changes to the voice that would not affect the parameters used for speaker identification. In the instance of jitter or microtremors, (Smith, 1977. Ruiz, Legros, & Guell, 1990) these are often ignored by voice authentication systems, but find extensive use in VSA devices.

1.1 Background

There has been a great deal of voice emotion research by groups such as the Geneva Emotion Research Group. Scherer pointed out the possibilities of resynthesis in better understanding emotion in the voice and a move towards a more theoretical approach to work in this area (Scherer, 1989. Scherer, Johnstone & Bänziger, 1998. Scherer, 1995. Johnstone & Scherer, 1999. Schuller Lang & Rigoll 2002).

The types of cues that might be analysed could include:

- **Pitch information**, including the mean pitch, its amplitude, its variation and rate and direction of change, in addition to the formants. These are strong harmonics that give a voice its characteristic sound.

- **Amplitude variation** during and between utterances as a whole and in different frequency bands. It is possible to consider a general envelope and also more specific peaks in different formant amplitudes.

- **Low frequency changes** otherwise known as jitter or micro tremors (Smith, 1977. Ruiz, Legros, & Guell, 1990).

VSA manufacturers suggest that the changes in these parameters during speech provides psychophysiological evidence for a speaker's emotional state and can be used to determine truthfulness of a statement by looking for stress cues. Although many researchers currently consider VSA devices be very unreliable (Janniro, M. J., & Cestaro, V. L. 1996. Meyerhoff, Saviolakis, Koening & Yurick, 2001) such devices still find extensive use in the insurance industry in the UK. It is also possible that more accurate and robust techniques for Voice Stress Analysis (VSA) will exist in the future and this provides a motivation for developing voice integrity checks.

1.2 Watermarking and Fingerprinting Strategies

Watermarks and fingerprints are information that is hidden in a file. Watermarks are often copyright messages that are file specific and fingerprints are often unique messages, for example that might identify the file's owner. This type of message hiding is known as steganography. Providing message integrity might be achieved in a number of ways. These include

- **Fragile Watermarking.** This approach to steganography is not particularly robust, and often relies on changes in either the amplitude or frequency domain to least significant bits. This is often considered as unsuitable for voice data, as the watermarks are often unable to survive the compression process.

With traditional watermarking, the purpose is often to prove ownership and methods exist to remove the watermark data, using tools such as Stirmark (Petitcolas, Anderson & Kuhn, (1998)) often restricting the effectiveness of the watermark. In this instance the watermark is to show integrity of the file and any hacker would be required to modify the data without affecting the watermark. This is far more difficult and would require knowledge of the watermarking process.

- **Semi-Fragile Watermarking.** These schemes accept some changes to a file or to a data stream but work with respect to a threshold. This could allow basic tone control or amplitude changes, but not changing the length of the file for example. By only placing data across frequency bands that are not quantised or lost in lossy compression such as MP3 it is possible to retain the watermark even when compression takes place. It may be difficult to guarantee the integrity of data with this approach as it is possible to make changes below the threshold.

- **Robust Watermarking.** These watermarks are designed to withstand attacks and are suitable for embedding copyright information. This might be in conjunction with a web spider in order to check for marked files.

- **Content Based Watermarking.** Wu and Kuo (2001) have considered a semi-fragile watermarking scheme based on semantic content. In this instance semantic is taken to mean higher level detail such as pitch. This approach adopts a semi-fragile approach, which provides clear indication of dramatic changes in the voice as a result of modification, but would not identify changes to the high frequency energy or the addition or removal of jitter. The tests carried out involved replacing one speaker's voice with another.

2.0 Hacking

There are many opportunities for subtle modifications of emotional voice cues. This modification may be subtle and difficult to identify.

Different VSA devices use different voice emotion cues to decide on the veracity of statements made by a subject. However it is very simple to remove jitter using a very steep high pass filter or to add jitter either by modulation of the voice signal or by adding a very low frequency waveform between 8 and 15Hz. For many traditional VSA approaches this is sufficient to modify the readings of the VSA device. Devices might also use other voice stress cues such as frequency changes in the second formant or changes in high frequency energy.

To demonstrate a simple hacking approach we have developed tools such as a jitter generator (JitGen4) that can add jitter in a variety of ways (Fig 1.). Including the generation of low frequencies and the modulation of the speech signal. These have been developed using graphical modular synthesiser development tools such as SynthEdit (McClintlock, 2002) and run in AudioMulch (Bencina, Ross 2004), a VST (Steinberg) real time processing environments.

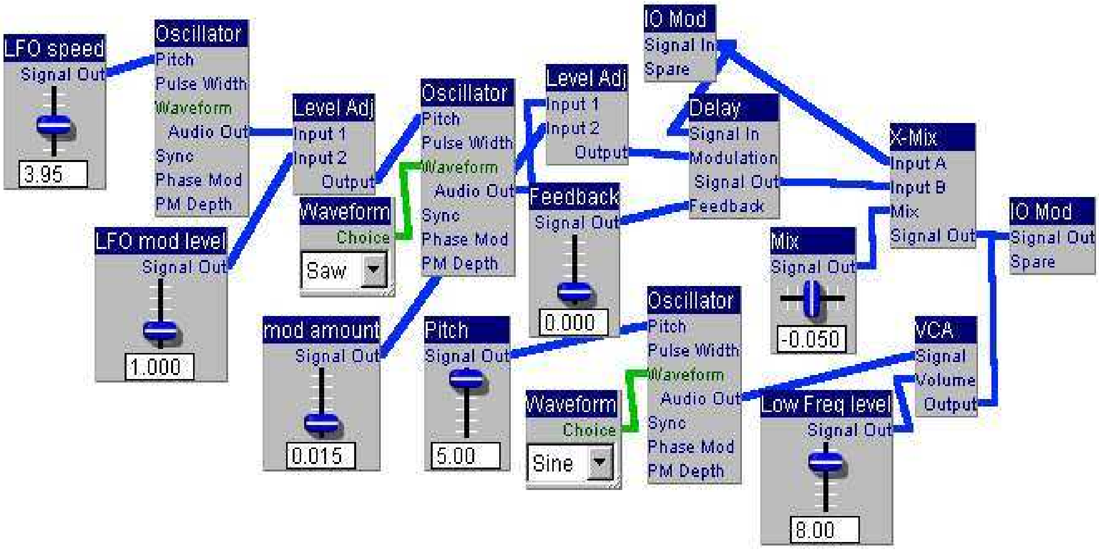
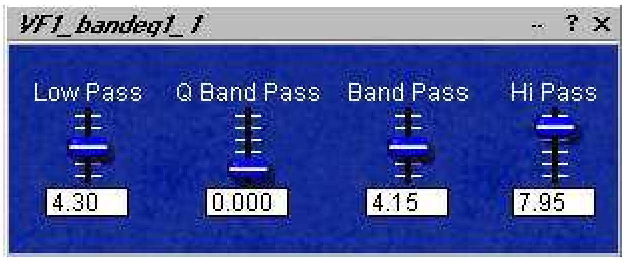


Figure 1: JitGen4 Structure.

We have developed simple filters (Fig 2.) that can also be used in the same environment together with off-the-shelf audio compressors to limit the dynamic variation for a given frequency range. If used in conjunction with expanders to increase the dynamic variation in a different frequency range, the overall envelope does not appear to change but the processing can have a significant impact on the voice emotion cues.


VF1_bandeq1_1

Low Pass 4.30

Q Band Pass 0.000

Band Pass 4.15

Hi Pass 7.95

Figure 2: Filter front end showing available control.

Although these devices are crude, they demonstrate the potential of "Emotion Hacking" and associated risks to privacy. Third party use could lead to erroneous judgements made about the speaker and the semantic content of their message. This effectively leads to a breakdown in trust.

3.0 Requirements

The requirements of integrity checking are very different from the need to retain information within a file. This implies that the embedded data needs to be suitably immune to transfer through a communications network but not robust enough to survive changes to harmonic content or jitter. General requirements are listed:

- **Blind detection:** This is where the file contains data that self validates the integrity and does not rely on an original file being available, this is of particular importance when the integrity check may need to occur in real time upon receiving a spoken message.
- **Low Distortion:** The impact of adding a fingerprint does not degrade the quality of the transmission.
- **Low Latency:** The delay caused by adding the fingerprint, encryption and validating the data on reception should not have an impact on the communication.
- **Low Data Rate:** The amount of data added by the fingerprinting process, should be small.

Low latency can be partly achieved by only having very small amounts of data that have to be encrypted. If we consider common hacking techniques such as:

- **Ciphertext only attack**. If we know the structure of a file then file predictability allows us to know when we have successfully attacked an encrypted message.
- **Known plaintext attack**. If we know part of a message, then it is much easier to attack by attempting to encrypt the known message fragment and comparing to the encrypted file.
- **Chosen plaintext attack**. This is where we encrypt a message with the public key and try to decode. This is very time consuming and is often described as a brute force attack.
- **Man-in-the-middle attack**. Intercept and pass on.
- **Timing attack**. The time taken to encrypt is used as an indication of the key size, this is a sophisticated attack and difficult to perform.

A small amount of encrypted data sent with no encrypted header is difficult to attack, in terms of breaking the encryption.

4.0 Proposed Emotion Profile Integrity Checking

To ensure low latency only a small amount of encryption is carried out. The encrypted data is a randomly generated seed that is used to generate a pseudorandom sequence. The sequence generated is used to indicate the parameters used to perform the integrity check, and at what intervals in the file. The encrypted seed is sent to the receiver who is also able to generate the same sequence and can extract the voice parameters in order to check them. The use of a seed to generate the numbers has a number of advantages. It needs no encrypted file header and is relatively small. This makes it difficult to attack using standard cryptanalysis techniques. The proposed approach is illustrated in Fig.3 at the end of the paper.

Although we are of the opinion that transmission without compression is impractical due to the high bandwidth required, we feel that as part of a communications system any insertion and extraction of a fingerprint for integrity checking should take place in the compressed file, directly prior to, and after transmission. This removes the need for the fingerprint to be robust enough to survive lossy compression.

An authentication system needs to embed a fingerprint that is both unique and that allows the integrity of the data to be checked. It needs to be able to work without user intervention and in near real time. That is, with a low enough latency to be undetectable.

We propose a voice integrity checking method based on emotional cue features. Unlike existing techniques we consider not a semantic based approach but a psychophysiological approach that could possibly be combined with a fingerprint. The approach taken is developed to indicate any modification that might have taken place. Previous work by Wu and Kuo (2001) has suggested that changes to the brightness (they might mean tone in this context) or amplification should not damage the watermark. We suggest that the listener carries out any required modification to the sound data in real time and that stored files should remain faithful to the received data. This would prevent accidental modification of emotional cues that might later be used for analysis. To ensure that this does not occur, we would suggest that fingerprinting files for integrity checking remain fragile.

If a seeded pseudorandom selection of emotion characteristics is used to generate encrypted data in either packet headers or in least significant bits, this can be extracted simply at the other end. Not every feature needs to be sent simultaneously and this makes it difficult to hack such a system, as it would identify parameters in the voice with high risk, when it comes to the use of stress analysis.

Examples of possible parameters for the integrity checking include:

- Pitch variation
- Mean energy in given frequency band
- Jitter levels
- Jitter frequency
- Distance between formants
- Ratios between formant amplitudes

5.0 Conclusion

The current available techniques for voice authentication and message integrity checking are suitable for detecting many of the changes in the voice but may not be strong enough to detect the very subtle changes that can be used used for modifying voice emotion cues. The proposed approach allows a combination of fragile watermarking and a secure emotion cue content-based system to provide security for the listener. Security for the speaker would require additional fingerprinting to show that the listener had not retrospectively modified the file and the fingerprint.

As VSA devices become more sophisticated and with Moore's law (Moore, 1965) still holding true and providing the potential for very low latency complex processing, the requirements of integrity and authentication checking will change. Our proposal anticipates some of these changes and provides a novel and robust approach to voice integrity checking.

The advantages of our approach include better risk analysis in deciding which aspects of the voice signal should be ignored in any stress analysis and a robust integrity check that requires less processing overheads than current strategies that use encryption.

5.1 Future Work

Work needs to be carried out to test the proposed method in real communication scenarios. We also need to ascertain the extent to which real time processing of media makes integrity checking necessary. This may be achievable by developing the tools that might be used to modify the voice and investigating their impact on VSA devices. The work in this area will hopefully lead to more rigorous methods for Voice Stress Analysis and more reliable techniques for preventing such analysis when required.

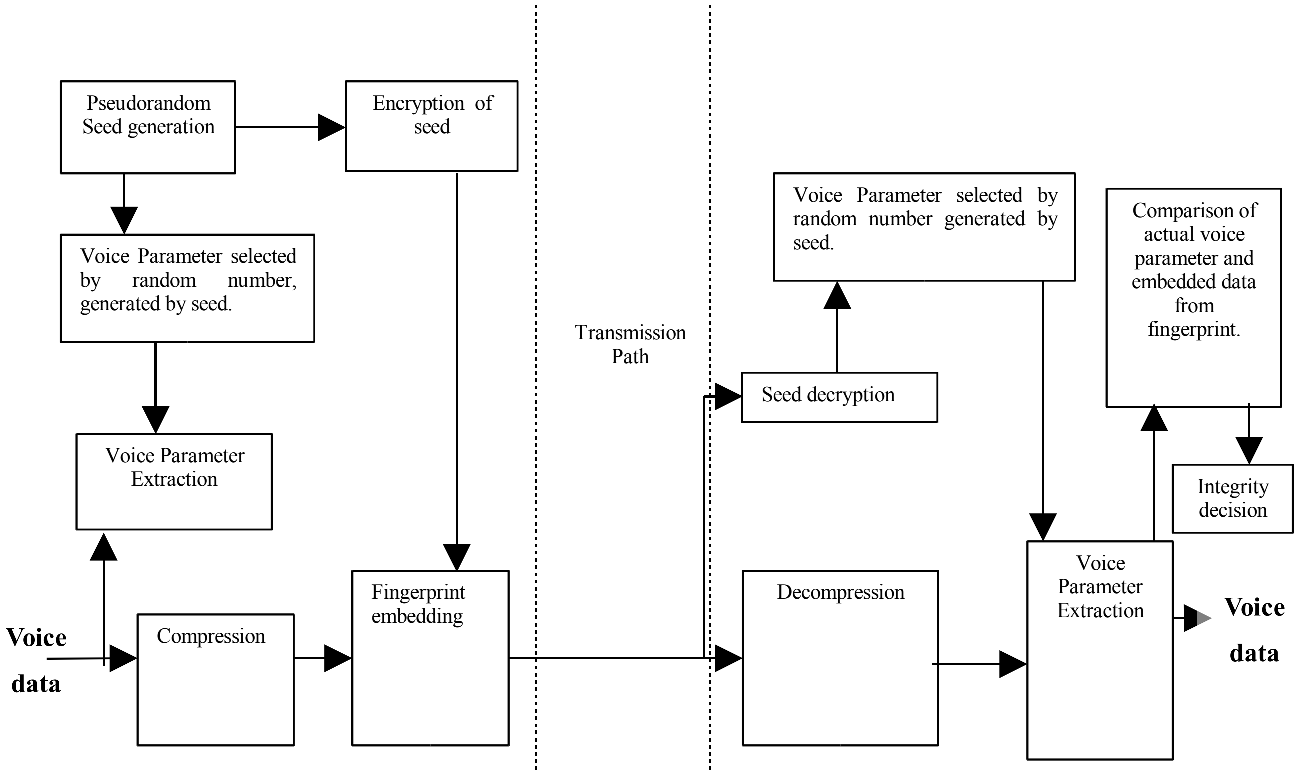

Pseudorandom Seed generation
Encryption of seed
Voice Parameter selected by random number, generated by seed.
Voice Parameter Extraction
Voice data
Compression
Fingerprint embedding
Transmission Path
Voice Parameter selected by random number generated by seed.
Seed decryption
Decompression
Voice Parameter Extraction
Comparison of actual voice parameter and embedded data from fingerprint.
Integrity decision
Voice data

Figure: 3 Diagram showing integrity checking process.

6.0 References

Bencina, Ross (2004), Audiomulch ver 0.9b15 2/2/2004 available from www.audiomulch.com

Fabien A.P. Petitcolas, Ross J. Anderson, and Markus G. Kuhn. Attacks on Copyright Marking Systems

David Aucsmith, Ed., *Second workshop on information hiding*, in vol. 1525 of *Lecture Notes in Computer Science*, Portland, Oregon, USA, 14{17 April, 1998, pp. 218{238. ISBN 3-540-65386-4.

Janniro, M. J., & Cestaro, V. L. (1996). Effectiveness of Detection of Deception Examinations Using the Computer Voice Stress Analyzer. (DoDPI95-P-0016). Fort McClellan, AL : Department of Defense Polygraph Institute. DTIC AD Number A318986.

JitGen4 (2004) available upon request from C. Reynolds, carl9@mdx.ac.uk.

Johnstone, T. & Scherer, K. R. (1999). The effects of emotions on voice quality. Unpublished research report. *Geneva Studies in Emotion and Communication*, 13(3).

McClintock. Jeff, (2002) SynthEdit rev 0.9507 available at http://www.synthedit.com

Meyerhoff, Saviolakis, Koening & Yurick (2001) DoDPI Research Division Staff, Physiological and biochemical measures of stress compared to voice stress analysis using the computer voice stress analyzer *(CVSA).* (Report No. DoDPI01-R- 0001). Fort Jackson, SC: Department of Defense Polygraph Institute, & Washington, DC: Walter Reed Army Institute of Research.

Moore Gordon E. 1965 *Cramming more components onto integrated circu*its. Electronics, Volume 38, Number 8, April 19, 1965

Ruiz, Legros, & Guell. (1990). Voice analysis to predict the psychological or physical state of a speaker. *Aviation, Space, and Environmental Medicine*, (1990).

Scherer, Johnstone & Bänziger (1998) Scherer, K. R., Johnstone, T., & Bänziger, T. (1998, October*).* Automatic verification of emotionally stressed speakers: The problem of individual differences. Paper presented at *SPECOM'98, International Workshop on speech and Computers*, St. Petersburg, Russia. Geneva Studies in Emotion and Communication, 12(1).

Scherer (1995) Scherer, K. R., "Expression of emotion in voice and music", *J. Voice*, 9(3), 1995, 235-248.

Schuller Lang & Rigoll (2002) Automatic Emotion Recognition by the Speech Signal Björn Schuller, Manfred Lang, Gerhard Rigoll, Institute for Human-Machine-Communication, Technical University of Munich 80290 Munich, Germany, presented at *SCI 2002*. CD-ROM conference proceedings.

Smith (1977) Smith, G. A. (1977) Voice analysis for the measurement of anxiety. *British Journal of Medical Psychology*, 50, 367-373.

Steinberg. VST is a trademark of Steinberg Soft- und Hardware GmbH"

Wu. C hung-Ping and Kuo. C.-C. Jay *Speech Content Integrity Verification Integrated with ITU G.723.1 SpeechCoding.* IEEE International Conference on Information Technology: Coding and Computing (ITCC2001), pp. 680-684, (Las Vegas, Nevada), April 2001