The ITB Journal

Volume 3 | Issue 2 Article 7

2002

Longer Than a Telephone Wire - Voice Firewalls to Counter Ubiquitous Lie Detection

Carl Reynolds

Matt Smith

Mark Woodman

Follow this and additional works at: https://arrow.tudublin.ie/itbj

Part of the Digital Communications and Networking Commons

Recommended Citation
Reynolds, Carl; Smith, Matt; and Woodman, Mark (2002) "Longer Than a Telephone Wire - Voice Firewalls to Counter Ubiquitous Lie Detection," *The ITB Journal*: Vol. 3: Iss. 2, Article 7.
doi:10.21427/D70N14
Available at: https://arrow.tudublin.ie/itbj/vol3/iss2/7

This Article is brought to you for free and open access by the Ceased publication at ARROW@TU Dublin. It has been accepted for inclusion in The ITB Journal by an authorized editor of ARROW@TU Dublin. For more information, please contact arrow.admin@tudublin.ie, aisling.coyne@tudublin.ie, vera.kilshaw@tudublin.ie.

Longer Than A Telephone Wire - Voice Firewalls To Counter Ubiquitous Lie Detection

Carl Reynolds[1], Matt Smith[2], Mark Woodman[1]

[1] *School of Computing Science,*
Middlesex University, London UK
[2] *School of Informatics and Engineering,*
Institute of Technology at Blanchardstown, Dublin 15
Contact email: c.reynolds@mdx.ac.uk

Abstract

Mobile computing and communication devices are open to surreptitious privacy attacks using emotion detection techniques; largely utilising work carried out in the area of voice stress analysis (VSA). This paper extends some work in the area of removing emotion cues in the voice, specifically focusing on lie detection and presents the results of a pilot study indicating that the use of mobile phones in situations of stress is common and that awareness of VSA is low. Existing strategies for the removal or modification of emotion cues, based on models of synthesis are considered and weaknesses are identified.

Keywords

Lie detection, privacy, mobile devices, voice firewall, voice stress analysis, emotion detection, speech processing.

1. *Introduction*

Invasions of privacy and security in mobile telecomm equipment are often assumed to come from third parties; this could for example involve the use of scanning devices, line-tapping or eavesdropping. It is possible that the receiving party in a voice communication session may also carry out a voice stress analysis (VSA) of a speaker, thus providing an invasion of privacy. A variety of "emotion detecting" VSA devices are readily available with manufacturers and distributors making considerable claims as to their efficacy, despite the many studies casting doubt on the reliability of VSA devices, even within the restricted context of lie detection. Meyerhoff, Saviolakis, Koening, & Yurick, (2001), and Hollien, Geison & Hicks (1987) provide evidence of this; most researchers suggesting results that provide little better than chance at accurate emotion detection.

There is sufficient research in emotion detection through voice analysis to suggest that although we may not clearly state the accuracy of these techniques in a variety of situations, the sharing of common strategies over many research groups is still indicative of a reliable body of knowledge in this area.

Emotion detection through audio analysis is not restricted to speech, according to a local newspaper, a device “Why Cry” has just become available. This analyses a baby’s crying and has a claimed 98% reliability in detecting if a baby is bored, stressed, uncomfortable, hungry or tired (Miranda, 2002). These unsubstantiated claims indicate a popular interest in this area, and the use of lie and emotion detectors is becoming common in popular television shows.

The area of VSA and thus lie detection is largely informed by research into prosody, a term that describes pitch contours, rhythm and chunking (grouping of syllables). There has been considerable work in the investigation of prosody changes in a variety of emotional states and prosody analysis has been extended to include detailed analysis of the frequencies and noise present in speech. Properties that are of particular interest include the fundamental frequency (F0), which is a measure of the pitch and changes in energy in different frequency bands.

Scherer, Johnstone and Banziger (1998) have identified changes in speaker state, a collective term for different types of stress, emotional outlook and attitude and are looking to develop more robust strategies for speaker verification. In their work they have collated a substantial amount of information about the ways in which emotional, cognitive and physiological stress may modify measurable properties of speech. There work suggests that the movement and range of F0 is important together with its energy. This is supported by other work in this area, such as Li & Zhao (1998).

Voice Stress Analysis uses various cues in the speech signal to ascertain the emotional state of the speaker; there has been considerable research in this area, summarised by Scherer (1995) and presented in a simplified form in Table 1. This paper only considers the 7 factors in Table 1 and five emotional states, although Schuller, Lang & Rigoll (2002) use an 18 dimensional feature vector in order to distinguish between seven emotional states. The neutral user state is assumed in our simplified model, although in practice it would be derived in a calibration period during the operation of a VSA device and as such may not actually be neutral. Many VSA devices only consider stress and neutral state as important (in use as lie detectors). Some research in VSA fails to consider a neutral state and this could lead to some inaccuracies in findings, however as stated previously, the consensus is that the movement and energy of F0 is a key indicator of speaker state, this holds true whether or not a neutral state is used to calibrate results.

It is important to note that this calibration to determine a neutral state is very important. It might be used by an informed speaker as an opportunity to voice the “big” lie at the start of the conversation, and thus to confuse the lie detector’s calibration system. In such a situation it would be possible to take a sample of the same speech from the end section of the speech (preferably during a known “truth statement”) and to stitch this to the beginning. This allows a reliable calibration procedure and the “lie to be exposed”, although not in real time.

Change to Speech	Anger	Fear	Sadness	Joy
Mean F0	Increase	Increase and higher	Decrease	Increase
F0 range	Increase	Increase	Decrease	Increase
F0 variability	Increase			Increase
F0 contour Direction	Downward		Downward	
Mean intensity	Increase		Decrease	Increase
High frequency energy	Increase	Increase		Increase
Articulation rate	Increase	Increase	Decrease	Increase

Table 1: Summary of existing work (from Scherer 1995).

Reynolds, Smith & Woodman (2002) described a number of strategies for blocking VSA that have been summarised in Table 2.

Strategy	Process	Advantages	Disadvantages
Subtractive	Complex waveforms are filtered and amplified to modify the waveform, by subtracting harmonics.	May be subtle, possible to run in real time.	Does not modify changes in F0.
Additive	Simple waveforms are combined to make more complex waveforms.	Easy to implement in real time.	Easily detected, but only modifies Jitter.
Spectral Modeling	Use of transforms to allow manipulation of audio in the frequency domain. In addition to tone control it encompasses audio effects such as pitch tracking and shifting.	Could be undetectable and allow very fine control over a VSA response.	Considerable processing power required.
Avatar	Phones are stored in a look up table and referenced as required giving a substitution for neutral speech.	A complete firewall.	Easily detected and intelligibility only adequate.

Table 2: Strategies for Blocking VSA.

There have been many other studies investigating emotion in speech and some recent approaches, such as that by Schuller, Lang & Rigoll (2002) analyse the content for semantic clues, such as key words, verbosity and non-verbal utterances in addition to consideration of the psychophysiological data. The context of their work is in the area of interaction design and adaptable interfaces, although it may yet play a role in other forms of emotion detection. The work uses acted emotions and this might produce very misleading results, although they publish confusion tables, supporting Reynolds, Smith & Woodman's (2002) observations that "*very different emotions, might elicit the same stress response for a given indicator*". There is evidence that types of stress such as emotional, physiological and cognitive produce different changes in formants or harmonic content in the vowel sounds, with some change being dependent on the speaker's coping strategy (Tolkmitt & Scherer, 1986). Identifying such a strategy in the analysis of small amounts of speech might prove to be difficult.

Many early VSA devices used the presence or absence of micro tremors or jitter (Smith, 1977) to determine stress and anxiety in the speaker. See Figure 1 for a plot of jitter. These reduce with a corresponding increase in anxiety; these manifest themselves as a 3 – 10Hz frequency modulation of the speech signal and may increase in frequency. Prevaricator is an example of a software lie detector that takes this approach to anxiety analysis (Prescott, 1999).

To test the results of emotion research, it is necessary to have a corpus of test data. Many such corpi exist (SUSAS) although the material contained in them may be acted or elicited speech rather than real emotional speech. For the purposes of accurately testing VSA it would seem that real lies would be most appropriate. These are hard to obtain and there may be many questions about the ethical and legal issues. For this work in blocking VSA the accuracy of the original test data is not considered important because we need to provide consistently neutral or predetermined results for any speech input, whether acted or a genuine lie. However the use of acted emotions must cast some uncertainty on the reliability of VSA, when these are solely used as test data, since the stress detected may not be real.

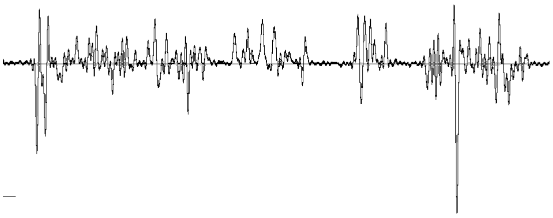


Figure 1: Amplitude/time plot of jitter (normalised).

This work extends strategies for blocking emotional cues and may have an impact on intelligibility, especially in the presence of competing audio streams. It could be claimed that the audio cues that allow auditory scene analysis to take place include the presence of jitter or micro tremors and the movement of F0. This seems likely when considering the work of Bregman (1990) and plays a significant role in our understanding of the cocktail party effect (Cherry 1953 and Arons 1992). In considering strategies for the removal of VSA cues it may be important to consider the retention of as much prosody information as possible, especially where monophonic devices with limited audio bandwidth are concerned. Removal of VSA cues may also lead to the removal of the cues that aid formation of audio streams, this may be of importance where phone conferencing systems are used. No work has yet been carried out to determine the impact of voice stress cue modification on the ability to form coherent primitive audio streams.

In addition to this, removal of cues for emotion may lead to confusion over meaning. This may be noticeable in the case of irony, sarcasm and when telling jokes. With many chatrooms using emoticons, this concept could be extended to mobile voice communications with the addition of specific aural clues.

2. Developing the Strategies for "Voice Firewalls"

The previously considered strategies for preventing the surreptitious use of lie detectors have been refined and the work carried out as follows:

Subtractive

This approach is largely effective against the use of Jitter (see Figure 1) or micro tremors as a means of determining the veracity of a speaker. Setting a high pass filter at 20Hz has no noticeable effect on speech quality and no detection has been indicated in listener tests (Jitter is transmitted by restricted bandwidth devices such as mobile phones). Such filtering easily prevents Prevaricator from producing a high stress result. It was suggested that a subtractive approach be used for adjusting energy in frequency bands, however if a "Voice Firewall" is to be developed against more sophisticated detectors that analyse prosodic cues, then the processing would be better carried out in the frequency domain.

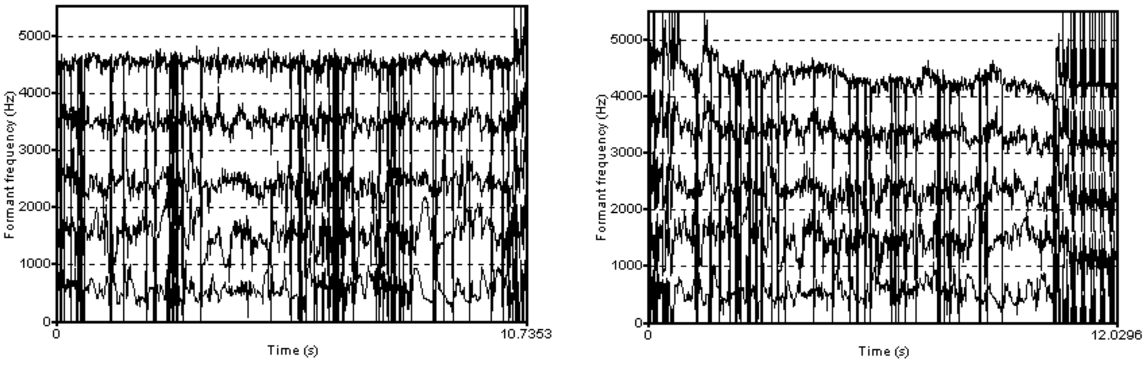
Additive

This is best used with the subtractive approach and replaces the Jitter pattern in the voice signal with a constructed or prerecorded one containing false emotional cues. It can be used to fool simple detectors but is less effective against an emotion detection approach that considers variation in the mean F0 of the voice. The jitter signal remains inaudible, even when normalised.

Spectral Modeling

The advantage of a spectral modeling approach is that it utilises the frequency domain. This enables the use of complex filters and formant tracking and modification. Simple pitch correction techniques can be applied; these are able to shift all formants including F0 to enable some of the natural qualities of the voice to be maintained. It is possible to make small changes in real time, although the latency of such an approach has not been calculated. Figure

2 shows the change to formants when pitch changes in real time are applied. This example did not use the original speech parameters to control the pitch shift; this was predetermined by settings.



(a) Original Formants **(b) Pitch Shifted Formants**

Figure 2: Formant change with real time pitch changes.

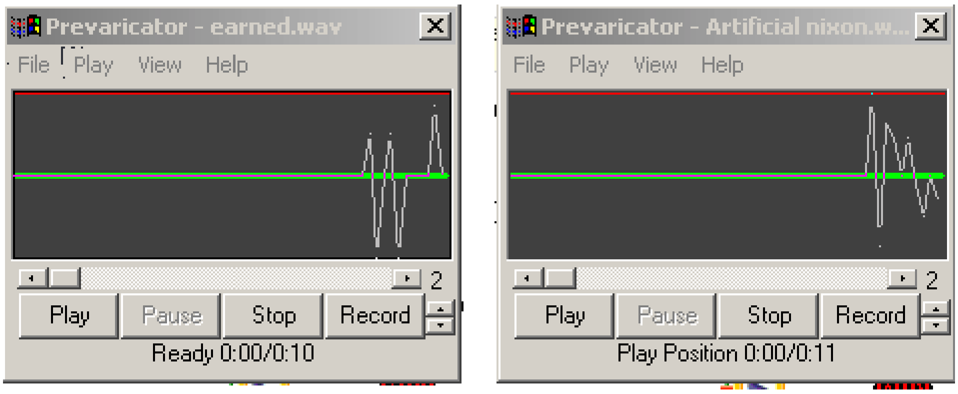
Avatar

The avatar method makes use of a virtual speaker or voice avatar that copies the original. Such a speaker could be a neutral clone of the speaker with no voice stress cues, another person, or a fictional character. This strategy utilises technology that exists in speech recognition and in text to speech software. We can utilise software to convert the user's voice to text and then read this text with an artificial voice. This provides a complete voice firewall solution. It includes the stage of providing a textual output, which involves a requirement to ensure that the text makes sense and that words are correctly spelled. This would not be required for a voice firewall, because the listener performs the required cognitive processing but does not receive the voice stress cues thus making VSA futile. In the proposed avatar approach the spoken words are analysed and parsed to extract the phoneme content (phonemes being basic sounds that form the basis for words). The identified phonemes can reference a look-up table that stores a digital copy of a voice stress neutral equivalent of the phoneme which can be sent to the output. A phoneme palette could be stored in firmware.

It could still be possible to extract some prosodic elements such as chunking from the speech data, to assist in increasing intelligibility.

Many text to speech software packages now have more realistic virtual speakers, although earlier criticism of such systems is that the results are mechanical (Hardman, Sasse, Handley

& Watson, 1995) and that intelligibility is only adequate. The avatar strategy was tested by using a text to speech package and capturing the output. This was then analysed by two lie/emotion detectors. The results illustrated in Figure 3 are for the Nixon "I am not a crook" speech.



(a) Read by Nixon **(b) Read by virtual speaker**

Figure 3: Nixon's speech

The results indicate that the speaker is under stress, and therefore the virtual speaker's voice characteristics in this instance are unexpected. The development of a neutral phoneme database is therefore indicated as important in the context of this work.

3. Analysis of the proposed strategies

In analysing the strategies we have come to the conclusion that spectral manipulation provides an effective form of voice firewall whilst retaining the original speakers voice characteristics. Also the avatar approach offers complete isolation from the psychophysiological cues that may be present in speech. The subtractive and additive approaches have been evaluated with success against simple lie detectors and these can be used surreptitiously. These approaches are not robust enough for blocking emotion detection where sophisticated analysis, such as F0 tracking takes place. The most isolating approach is the avatar, however this needs to be developed to ensure that the phoneme database contains neutral phonemes. These could be based on the speaker's own voice or the voice of another and may have implications for voice recognition.

4. Evidence of Requirement

During 2002 a pilot survey of mobile phone use was conducted of 32 London commuters. The following results were obtained:

1. Do you use a mobile phone or other mobile voice communication device for business purposes?

   YES **14** NO **18** *(2 had no phone)

2. Do you ever use your mobile device to transmit sensitive information or decisions?

   YES **5** NO **9**

3. Are you aware of any security or privacy risks when transmitting sensitive or confidential information?

   YES **7** NO **7**

4. Have you ever been in stressful situations when transmitting such information?

   YES **11** NO **3**

Although a relatively small, informal survey, making no claims for statistical significance, these results indicate that it is not unreasonable for us to suggest that mobile phones are:

- Used for business purposes, and
- Used to transmit sensitive information.

It is also reasonable to suggest that many users would be unaware of security or privacy issues and that the majority were likely to use their phone in stressful situations. Those who were aware of privacy risks considered eavesdropping as the main risk followed by scanning. No users considered the use of emotion detection. Two of the users who claimed to have never used a mobile phone under stress admitted to answering their phone while driving and did not consider this to be stressful.

5. Conclusions

Four strategies for blocking the use of emotion/lie detectors have been reduced to two, with the subtractive and additive approaches only being effective against the most primitive of stress/lie detectors and these may be considered a subset of spectral modeling.

The techniques developed, although still not fully implemented in a stand–alone form are capable of blocking lie detectors, but there are trade-offs on intelligibility, as yet not fully measured.

The following conclusions can be made:

- The avatar strategy may not be able to use an "off the shelf" phone database, as these may not be emotion neutral. This suggests that future work should include the generation of emotion neutral phoneme databases.
- The spectral modeling strategy could allow some elements of prosody to remain, although the extent to which the intelligibility of the speech remains unaffected has not been measured.
- The proposed strategies may be implemented by modifying existing algorithms developed from models of synthesis.
- It is possible to confuse voice stress analysis based lie detectors in many ways. This is of value not only with the need to retain privacy, but also because these detectors often do not appear to have the accuracy claimed or that the accuracy claimed is given for specific experimental conditions.
- Mobile communications devices do catch their users unguarded, as they often are involved in other tasks that are not related to the issues involved in the communication itself. The pilot survey carried out illustrates that mobile users are answering their phones in stress situations, even when not recognised as such. It also illustrated a lack of awareness of privacy issues, and use of mobile communication equipment as a business communication tool. A more extensive survey is required in order to understand the impact that psychophysiological tools might have on the way we disseminate information in an increasingly "mobile" world.

6. References

B. Arons (1992). A review of the cocktail party effect. *Journal of the American Voice I/O Society*, 12 (July)

A. S. Bregman (1990). *Auditory Scene Analysis: The Perceptual Organisation of Sound.* The MIT Press.

E. C. Cherry (1953). *Some experiments on the recognition of speech with one and with two ears.* Journal of the Acoustical Society of America, *25, 975-979*

V. Hardman, A. Sasse, M. Handley & M. Watson (1995) *Reliable Audio for Use over the Internet.* Presented at INET '95

H. Hollien, L. Geison & J. W. Hicks Jr. (1987) Voice Stress Evaluators and Lie Detection, *Journal of Forensic Science* 8703 Vol 32 Iss2 405-418

Y. Li & Y. Zhao (1998) Recognising Emotions in Speech Using Short-term and Long-term Features. Proceedings of the ICSLP. pp. 2255-2558.

J. L. Meyerhoff, G. A. Saviolakis, M. L. Koening & D. L. 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.

E. Miranda (2002) So that's what the baby's trying to say when it cries. *Evening Standard* Monday 14 October 2002

M. C. Prescott (1999) *Prevaricator* Software v1.04 Freeware.

C. Reynolds, M. Smith & M. Woodman (2002) "*Your Nose is…*". In Proceedings of SCI 2002, the Sixth World Mutliconference on Systemics, Cybernetics and Informatics, Orlando Florida, USA, July 2002.

K. R. Scherer, T. Johnstone & T. Bänziger (1998) (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).

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

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

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

SUSAS (2002) SUSAS *Speech Under Simulated and Actual Stress*. Web reference located at: http://www.ldc.upenn.edu/ldc/news/release/SUSAS.html

F. Tolkmitt & K. R. Scherer (1986) Effects of experimentally induced stress on vocal parameters. *Journal of Experimental Psychology: Human Perception and Performance,* 12, 302-313.