The International
Journal of
Speech,
Language
and the Law

IJSLL (PRINT) ISSN 1748-8885
IJSLL (ONLINE) ISSN 1748-8893

Article

Charlatanry in forensic speech science:
A problem to be taken seriously

Anders Eriksson and Francisco Lacerda

Abstract

A lie detector which can reveal lie and deception in some automatic and perfectly reliable way is an old idea we have often met with in science fiction books and comic strips. This is all very well. It is when machines claimed to be lie detectors appear in the context of criminal investigations or security applications that we need to be concerned. In the present paper we will describe two types of 'deception' or 'stress detectors' (euphemisms to refer to what quite clearly are known as 'lie detectors'). Both types of detection are claimed to be based on voice analysis but we found no scientific evidence to support the manufacturers' claims. Indeed, our review of scientific studies will show that these machines perform at chance level when tested for reliability. Given such results and the absence of scientific support for the underlying principles it is justified to view the use of these machines as charlatanry and we argue that there are serious ethical and security reasons to demand that responsible authorities and institutions should not get involved in such practices.

KEYWORDS: LIE DETECTOR, CHARLATANRY, VOICE STRESS ANALYSIS, PSYCHOLOGICAL STRESS EVALUATOR, MICROTREMOR, LAYERED VOICE ANALYSIS, AIRPORT SECURITY

Affiliations
Anders Eriksson: Gothenburg University
Francisco Lacerda: Stockholm University
email: anders.eriksson@ling.gu.se

IJSLL VOL 14.2 2007 169–193
doi : 10.1558/ijsll.2007.14.2.169

equinox
equinoxonline

Introduction

Charlatans may be found in all walks of life, especially in activities where there is a possibility of making money, and forensic speech science is no exception. The old disreputable voiceprint technique is still around and used by many private investigators in the US in particular. In Germany, a physics professor specialized in crystallography appears in courts claiming to have invented an automatic speaker recognition method based on methods borrowed from crystallography but refusing to subject his methods to independent testing or revealing exactly how the method is supposed to work. These are just two examples. In the present paper, however, we will limit the study of charlatanry to the two most widely used types of so called 'lie detectors'. We will explain how they work, what principles they are claimed to be based on and how they have performed when tested for reliability.

It should be stated right away that at the present time no method for reliable lie detection is known and it is not even known if it should be possible to develop such methods in the future. There are nevertheless several products on the market claimed to be working lie detectors. They do not always call their products lie detectors but use some euphemism like 'stress analyzer' or 'emotion analyzer', but by looking at how the vendors present their products there can be no question that lie detectors are what they want us to believe their products are. Here are two quotations to illustrate what we mean:

> *Diogenes* Digital Voice Stress Analysis application, was originally used in determining attempts at deception in law enforcement activities. In the world today you may hear the words 'lie detector' in reference to this type of technology, but this type of technology actually detects deception in the human voice. (The *Diogenes* Company home page[1])

> Professionals in the field of lie detection know that there is no 'true' lie detector, as lying is not a unified set of feelings that can be measured. … However, LVA is capable of detecting the intention behind the lie, and by doing so can lead you to identifying and revealing the lie itself. (*Nemesysco* home page[2])

In this paper we will not address the question of whether these products should properly be referred to as lie detectors, emotion detectors or deception detectors, but we will use the term 'lie detector' when we refer to these products. The focus of this paper will be on the discrepancies between the claims the producers and vendors make and what their products are capable of delivering.

Validity vs. reliability

The concepts of validity and reliability are much used in psychology, statistics and many other areas and we would like to use a slightly simplified version of that distinction to guide us when judging the lie detectors. There are many different aspects of validity which in psychological research and statistics appear under different labels (test validity, internal validity, content validity etc.). We need not be concerned with these details, but may content ourselves by observing what they have in common. The validity of a test is the degree to which it measures what it is intended to measure. Reliability on the other hand has to do with precision and consistency. How accurately does the method measure what it is intended to measure? How much will the results vary if the measurements are repeated by a given researcher or by other researchers?

To keep this distinction in mind has methodological implications. It seems reasonable, from a methodological point of view, to begin by determining the validity of a suggested method before it makes much sense to study its reliability. If the method can be shown to lack validity altogether it will as a consequence also be unreliable and carrying out a reliability test meaningless. If the validity is not known it will be a 'black box' whose reliability, if any, will remain unexplained. We must keep in mind, however, that validity and reliability are not all or nothing concepts. A method may be valid to a degree and reliability may range from very poor to almost perfect. At the far end of the negative scale we find things like astrology. It would be a complete waste of time to design experiments to determine how precisely horoscopes may predict future events when we know that the validity of the method is non-existent. At the positive end of the scale we find methods like DNA testing whose validity is solidly supported by scientific evidence and whose reliability is extremely high, albeit not perfect.

With respect to lie detectors (as well as stress, emotion or deception detectors) the starting point when judging them ought to be their validity, the most important question being: Have the basic principles upon which they are claimed to be based been verified in scientific studies? If we look back at the history of lie detector testing, we find that this question is seldom asked and even more seldom studied. There are, on the other hand, scores of reliability tests of the 'black box' type. We find this rather surprising and quite unsatisfactory. We would like to argue that the preferred order of things should be first to examine the validity of the procedures to make sure that the applied methods are valid and only when that has been established proceed to reliability tests.

The focus here will therefore not be on the reliability tests but on a rather detailed analysis of the scientific underpinnings of the methods, i.e. their validity. Based on these considerations we then explain why reliability tests have

shown the studied products to be unreliable. We will be concerned with the two most widely used types of lie detectors, the so called voice stress analyzers (VSA), also referred to as psychological stress evaluators (PSE), and a newer type of analyzer said to be based on a multiple layer analysis of the voice. The latter comes in many different shapes as commercial products but they are all said to be based on what is called layered voice analysis (LVA). We will show that the first type lacks demonstrable validity and that the validity of the latter type is to be found at the astrology end of the validity spectrum.

As we will see later in the paper, producers and vendors of the first type, the voice stress analyzers, claim that their products are based on a neurophysiological theory of microtremor and sometimes cite scientific papers to boost the credibility of their products. We will show that those claims are completely unfounded by consulting a wide range of papers on microtremor and in particular the papers the vendors themselves often make reference to. The vendors of the second type make fantastic claims about how their methods can be used to monitor the brain activity underlying lies and deception by analyzing the voice signal, but never mention any scientific discoveries that lend support to such claims.

Even though we maintain that the examination of the so called lie detectors should have started by asking questions about their validity we will take advantage of the fact that reliability studies exist. We have examined many such studies, but in this paper we will mainly refer to two recently published reports (Hollien and Harnsberger 2006 and Damphousse et al. 2007). They are both of excellent quality and carried out under the assumption that the tested products should be given a fair chance to demonstrate their reliability. That means that the authors cannot be accused of being biased by any prejudice about the validity of the principles behind the lie detectors as any corresponding study we would have performed might have. Hollien and Harnsberger make their unbiased intentions explicit:

> As stated, the primary objectives of this project were to carry out highly controlled research that would at once be 1) impartial to all sides of the prior VSA controversies – i.e., those which led to the need for this research and 2) rigorous enough to address questions concerning the validity and sensitivity of the systems involved. (p. 7)

Damphousse et al. do not state this explicitly, but there is no doubt that their approach has been equally unbiased.

These studies are very useful complements to what we will have to say. They show that these lie detectors perform at chance level as far as reliability is concerned and we will explain why this is so. In our view, however, no more reliability studies of these two types are needed. The case should be considered

closed from that point of view; the evidence against them is just too overwhelming to motivate any more reliability tests.

The Voice Stress Analyzer (VSA)

VSA or PSE are not brand names but a type of analyzer marketed under different brand names by several different companies. As far as we are aware, however, they all claim to be based on the same underlying principle: detecting and monitoring so called microtremors in the voice production mechanism by analyzing the speech signal. Here is a typical quote from the home page of one of the companies that sells a VSA analyzer:

> Micro tremors are tiny frequency modulations in the human voice. When a test subject is lying, the automatic, or involuntary nervous system, causes an inaudible increase in the Micro tremor's frequency. (National Institute for Truth Verification, home page[3])

The first device claimed to be based on this principle, the Psychological Stress Evaluator (PSE), was produced and sold by a company (*Dektor*) formed by three former police officers (Bell, Ford and McQuiston) in the early 1970s. This and subsequent analyzers are presented as applications of scientific discoveries made by a group of researchers, primarily Lippold, Redfearn and Halliday, at the University College London. Here are two typical descriptions:

> In 1971, British physiologist Olaf Lippold discovered the muscle microtremor. Lippold found that voluntary muscles in the arm generate a physiological tremor or micro-vibration at about ten per second when the subject is relaxed. When the subject is aroused or stimulated, the microtremor tends to disappear. Lippold's theory relates to the voice in that muscles in the throat and larynx show microvibration that diminishes with stress through the vagus nerve. (Ridelhuber and Flood 2002, abstract)

> The *Diogenes* Digital Voice Stress Analysis program … has been produced to detect, process and display changes in voice pattern using the 'Lippold Microtremor'. …The Microtremor in laryngeal muscles has been shown to reflect the level of stress being experienced by the individual due to deception. (*Diogenes* home page[4])

Are the VSA/PSE lie detectors really based on the discoveries made by Lippold and his colleagues?

The source most often cited by the vendors is an article published by Lippold in the *Scientific American* in 1971. The article, *Physiological Tremor* (the term used in their publications), is a summary of work begun in the early 1950s.

Its main focus is a description of research aimed at determining the origin and function of physiological tremor. Lippold describes several experiments he and his colleagues have performed in order to verify or discard various possible theories. At the time of the publication they seem to have settled for the idea that the function of microtremor is as part of a feedback system by which voluntary muscle movement is fine-tuned. Lippold compares it to a mechanical servomechanism or a thermostat. Nowhere is there any suggestion that these discoveries may be used in lie detectors or similar applications. All the experiments described are concerned with muscles that control body movement, primarily leg, arm and finger muscles. Psychological stress is never mentioned. The experiments are concerned with the effects of physical tension, muscle temperature or blood flow. There is thus no obvious link between ideas on which the VSA/PSE detectors are based and the results cited in the Lippold article. One or two other papers appear as references in the VSA/PSE promotional materials. There was thus a slight possibility that some other paper might have hinted at such applications. In order to be sure that we had not missed anything we have gone over a large number of papers published between 1952 and 1983 by the London based group of researchers of which Lippold was a member but have not found a single mention of anything pointing in that direction. We may therefore say with confidence that there is no obvious connection between their research and the subsequent construction of tremor based lie detectors.

Microtremor and voice production

As we have mentioned above, the work by Lippold and colleagues was exclusively concerned with studies of muscles that control body movement. It is mentioned in passing by them and others that physiological tremor may be found in all voluntary muscles, but that is an assumption and not based on an extensive testing of various types of muscle. It is nevertheless possible that if corresponding experiments had been performed on the muscles that control voice production the same results would indeed have been found. This possibility has been explored in an experimental study performed by Shipp and Izdebski (1981). In their experiment they used one young male subject. Hooked-wire electrodes were placed in the cricothyroid and the posterior cricoarytenoid muscles and EMG signals were recorded during conversational speech and during sustained phonation. A technical description of the method may be found in Shipp et al. (1970). To verify the system's capability to detect microtremor, EMG activity was also sampled from the biceps muscle where such tremor is known to occur. The authors had to limit their study of the larynx to the recordings made during sustained vowel

phonation since 'EMG activity during conversational speech changed so rapidly … that at the present sampling rate no Fourier analysis could be made of these signals' (p. 504). For the recordings made during sustained phonation, the analysis 'failed to reveal any periodic component in the frequency band from 1 to 20 Hz; the electrical energy was randomly distributed throughout the spectrum'. In contrast, the reference recording taken from the biceps muscle to ensure the appropriateness of the method 'revealed a prominent energy peak at 9 Hz, indicating periodic contraction within the range of normal physiological tremor rate'. That is, applying identical measurement methods to both the larynx muscles and the biceps muscles failed to reveal any prominent spectral peaks in the 10 Hz region in the larynx muscles while the recordings from the biceps were in perfect agreement with results from other comparable studies.

In view of the fact that there is only one study directly testing the microtremor hypothesis, no matter how convincing the results may seem, there is always the possibility that a differently designed experiment might have resulted in a different result. It is, for example, possible that the negative result was due to the fact that they were looking for tremor in the wrong frequency region. This is not meant as criticism of their study. The objective of the study was to verify or falsify the claims made by the lie detector proponents that there is microtremor in the 10 Hz region, and from that perspective it was obviously the right thing to do. But if we were to search for microtremor in general we would have to be open to the possibility that it could be found in a completely different frequency range. Ocular microtremor (OMT), for example, seems to occur at much higher frequencies. A recent study by Bolger et al. (1999) found tremor in the 70–103 Hz range with an average of 84 Hz. Several earlier studies cited in their paper have reported tremor rates ranging from 30 Hz to 100 Hz. In a paper by McAuley and Marsden (2000) summarizing a large number of studies we find reported frequencies between 1 Hz and 100 Hz. What frequencies we should expect from muscles involved in speech production, if any at all, is not possible to say based on these studies. We have found one study that has some bearing on speech, however. In a study by Smith and Denny (1990) EMG measurements were used to study the activity of the diaphragm during speech and breathing. Activity was registered in the 20–110 Hz range in deep breathing. In speech, however, the 60–110 Hz range was significantly reduced.

There are a few EMG studies where the larynx muscles have been studied. None of them mention microtremor, and two of them (Hirano and Ohala 1969 and Hirano et al. 1970) cite no frequency measures at all. The study by Faaborg-Andersen (1957) has a section on frequency data for four major larynx muscles during silence and phonation. Firing rates for single motor units were

found to range between 5 and 50 Hz for different muscles and different phases of phonation. We may also observe that there were substantial differences in the behaviour of the studied muscles (cricothyroid, arytenoid, posterior cricoarytenoid and vocal muscles). If these signals were to have any effect on the resulting auditory signal (which of course we know nothing about) we should expect frequency components distributed over the whole frequency range and rapidly varying spectra.

Even based on this brief summary we may thus conclude that the 10 Hz peak found for the muscles studied by Lippold and others is by no means universal. Anything in the 1–100 Hz range at least is possible and only specific experimental studies can determine what the frequencies are in any given case.

We may summarize these findings by saying that the only scientific study explicitly involving the larynx muscles found no microtremor at all. The two most likely explanations for this finding are that 1) there is no microtremor at all in these particular muscles, or 2) microtremor in them does not occur in the 10 Hz region like in the much larger muscles controlling body movement, but in some other, probably higher, frequency region and that we are unlikely to find any stable frequency peaks since the larynx muscles are typically in rapid motion.

Conclusions concerning the question of microtremor

Based on the literature survey we have made, we feel confident in saying that there is no scientific evidence to support the idea that microtremor in the 10 Hz region occurs in the muscles involved in speech production. After a thorough study of relevant scientific papers published during the past 50 years we have not found a single study whose results lend support for this idea. Claims like 'The Microtremor in laryngeal muscles has been shown to reflect the level of stress being experienced by the individual due to deception' (*Diogenes* Company home page) are thus simply untrue.

But even if we speculate that there may be microtremor also in these muscles, there is a very long way to go before we arrive at a voice stress detector. First we must demonstrate that such tremor is possible to detect in the speech signal, which is of course not necessarily the case and must be demonstrated separately. Secondly it must be shown that the tremor is affected by psychological stress and that tremor fluctuations as a function of stress are rapid enough to be detected within the time frame of single utterances, not to speak of the single words or syllables that are often claimed to contain enough information for vocal microtremor analyses. Studies involving other muscles tend to show that similar effects apply over much longer durations. Fatigue, for example, may have an effect over a day or more. Restricting the blood flow to

a muscle damps physiological tremor but it takes between 30 and 60 seconds before the tremor is fully damped. Such time windows are hardly useful to detect stress in single, often short, utterances like 'yes' or 'no'. It is also worth pointing out that while physiological tremor has mostly been found and measured in muscles under static tension, the speech organs responsible for voice production are typically in constant motion. The period time for the assumed microtremor, 100 milliseconds, is in fact longer than most speech segments in connected speech. This means that during one single period of the supposed microtremor frequency, multiple adjustments of the larynx muscles are often needed to produce continuous speech. Thirdly it remains to be demonstrated that the effects of stress caused by lie or deception may be reliably separated from stress caused by other factors. Regardless of method, this last requirement is at present not possible to meet and we do not even know to what extent it ever will be.

Finally we will cite two comments we found on one of the *Diogenes* home pages containing references to 'Studies Validating Voice Stress Analysis'. For a paper by Lippold et al. (1957) the comment reads: 'Lippold, Redfearn and Vuco begin exploring the correlation between muscle activity and stress'. The actual paper reports a study of variation in the grouping of action potentials in the calf muscle as a function of contraction, stretching, fatigue and cooling and the word 'stress' is never even mentioned. A study by Lippold (1970) is said to be the study where: 'Lippold first discovers the physiological tremor in the human voice in the 8–12 Hz range' whereas the paper itself is exclusively concerned with the study of physiological tremor in the left hand middle finger.

In summary, as our survey has shown, the VSA approach completely lacks demonstrable scientific validity. In fact, all available evidence indicates that its validity is non-existent.

Reliability studies of commercially available VSA-based lie detectors

As we have said above we hold that before it is really meaningful to test the reliability of a method it should be demonstrated that we have reasonable grounds for assuming that the method is valid. There are nevertheless a large number of studies where the assumed reliability has been investigated. It is not our intention in this paper to discuss these studies in any detail. The papers are easy enough to find for those who want to consult them. A comprehensive list of references may be found in Damphousse et al. (2007). It comes as no surprise to us, of course, that the lie detectors almost invariably perform at chance level when tested by qualified researchers. Here are two typical conclusions, one from an early and one from a recent study:

> Both trained and untrained analysts were unable to … sort the voice-stress patterns consistently, at a greater than chance level (Lynch and Henry 1979: 91)

> The findings generated by this study led to the conclusion that neither the CVSA [Computerized Voice Stress Analyzer] nor the LVA were sensitive to the presence of deception or stress. Several analyses of subsets of the data were undertaken to explore any possibility that either system could perform under even more controlled conditions, but no sensitivity was observed in any of these analyses either (Hollien and Harnsberger 2006: 41)

We will end this part of the paper by reporting some of the results found in the study by Damphousse et al. mentioned above. This study differs from most otherwise similar studies in that the veracity of the tested statements was decided by comparison with irrefutable physical evidence:

> … we interviewed arrestees in the Oklahoma County jail about their recent illicit drug use. Answers by the respondents were compared to the results of a urinalysis to determine the extent to which they were being deceptive. Then, their 'actual deceptiveness' was compared to the extent to which deception was indicated by the VSA programs (Damphousse et al. 2007: 26)

We have chosen to present this particular study for two reasons. First, it represents the situation today because the equipment tested was the most recent model. That eliminates excuses that the results were obtained on an outdated model and that significant improvements have been made on later models. Second, the test situation comes as close as one can get to a real field work situation. But of course none of this helps if the equipment is an invalid product.

> All previously published research conducted in a lab setting has failed to find support for VSA theory or technology … Our research therefore complements previous research by failing to find support for the VSA products in a real world (jail) setting. In addition, the programs do not seem to have very high inter-user reliability even though the programs were relatively easy to learn and implement (Damphousse et al. 2007: 89)

A technical note on the Voice Stress Analyzers

A few studies have looked at the VSAs from a technical point of view and the 'sophisticated hardware' some of them advertise seems to be no more than simple low pass filters (VanDercar et al. 1980: 176–179). Today most of that is done by software, but the basic principles are likely to be the same. The low pass filtered signal is then supposedly subjected to a frequency analysis. For at least one of the products this is made explicit. 'Micro tremors are tiny frequency

modulations in the human voice' (*National Institute for Truth Verification, NITV* home page). The *CVSA* (of the *NITV*) has not been tested for frequency modulation sensitivity but a similar product, the *Diogenes* Lantern has. Haddad et al. (2002) produced synthetic signals with fundamental frequencies of 80 Hz or 160 Hz. These were then frequency modulated with frequencies varying from 1 Hz to 25 Hz and the resulting signals were tested using the Lantern VSA detector. It turned out that the VSA analyzer was almost completely insensitive to variation in frequency modulation. The authors drew the conclusion that: 'Since there was no variation of indicated stress from different input signals, it can be assumed that the systems tested do not use microtremors as indicated in their claims' (p. 11).

Interestingly, they also found that the sensitivity of the system was not, as one might expect, tuned to the frequency range claimed to be crucial in the microtremor analysis 8–12 Hz but to a rather different range:

> It was determined, late in the testing phase of this project, that the *Diogenes Lantern* System measures the energy change of the spectrum envelope between 20 Hz and 40 Hz. This is what the *Diogenes* Lantern System claims to be microtremors. It is the change of energy in the speech envelope. (Haddad et al. 2002: 11)

It thus seems as if at least this particular VSA analyzer does not analyze frequency changes at all and is not even operating in the claimed frequency region. Results like these certainly raise the question as to whether we may trust any of their claims. Not even the technical specifications seem to be correct.

The LVA analyzer

LVA is an acronym for Layered Voice Analysis. The company that manufactures the LVA is called *Nemesysco* and their product line comprises many different products, but they are all basically utilizing the same technology. Here is one example of how the LVA technology is presented:

> LVA uses a patented and unique technology to detect 'brain activity traces' using the voice as a medium. By utilizing a wide range spectrum analysis to detect minute involuntary changes in the speech waveform itself, LVA can detect anomalies in brain activity and classify them in terms of stress, excitement, deception, and varying emotional states, accordingly. (*Nemesysco* home page[5])

All this is possible, according to *Nemesysco* since 'every 'event' that passes through the brain will leave its traces on the speech flow'.

And here is how the extraction of brain events is accomplished:

> LVA has two basic formulas comprised of [sic] unique signal-processing algorithms that extract more than 120 emotional parameters from each voice segment. These are further classified into nine major categories of basic emotions. Depending on the goal of the analysis, up to eight 'final analysis' formulas can also be applied to the emotional parameter data. These include: Lie stress analysis, Arousal level, Attention level, Deception patterns match, and additional methods for veracity assessments. (*Nemesysco* home page[6])

One would assume that such extraordinary discoveries must be widely published and discussed, but the fact of the matter seems to be that they are completely unknown to the scientific community. The solution to this apparent mystery will become clear in the following paragraphs where we describe in detail what the LVA technology is all about.

Another aspect of the LVA technology that is highlighted in all their promotional materials is how extremely technically advanced it is.

> Layered Voice Analysis (LVA) is the most sophisticated truth detection technology available today … LVA is based on the technology of vocal stress analysis calculated from a series of sophisticated algorithms that detect states of stress. (*V-solutions* home page[7])

LVA fiction meets reality

Contrary to the claims of sophistication – 'The LVA software claims to be based on 8,000 mathematical algorithms applied to 129 voice frequencies' (Damphousse et al. 2007: 15) – the LVA is a very simple program written in *Visual Basic*. The entire program code, published in the patent documents (Liberman 2003) comprises no more than 500 lines of code. It has to be said, though, that in order for it not to be possible to copy and run the program as is, some technical details like variable declarations are omitted, but the complete program is unlikely to comprise more than 800 or so lines. With respect to its alleged mathematical sophistication, there is really nothing in the program that requires any mathematical insights beyond very basic secondary school mathematics. To be sure, recursive filters and neural networks are also based on elementary mathematical operations but the crucial difference is that these operations are used in theoretically coherent systems, in contrast to the seemingly ad hoc implementation of LVA.

Let us begin with a short technical description. In the verbal description of the program for the patent documents, the author describes the program as 'detecting emotional status of an individual based on the intonation information'. But whereas intonation in phonetics means variation in pitch encoded by fundamental frequency (albeit almost always accompanied by other prosodic

factors) the author of the LVA mistakenly believes that what he calls 'thorns' and 'plateaus' represent intonation. A 'thorn' is defined in the following way (Liberman 2003):

> A 'thorn' is a notch-shaped feature. For example, the term 'thorn' can be defined as: a. a sequence of 3 adjacent samples in which the first and third samples are both higher than the middle sample.
>
> or
>
> b. a sequence of 3 adjacent samples in which the first and third samples are both lower than the middle sample.

And plateaus are defined in the following way (Liberman 2003):

> A 'plateau' is a local flatness … A sequence may be regarded as flat if the difference in amplitude between consecutive samples is less than a pre-determined threshold.

Thorns and plateaus are illustrated graphically in Figure 1.

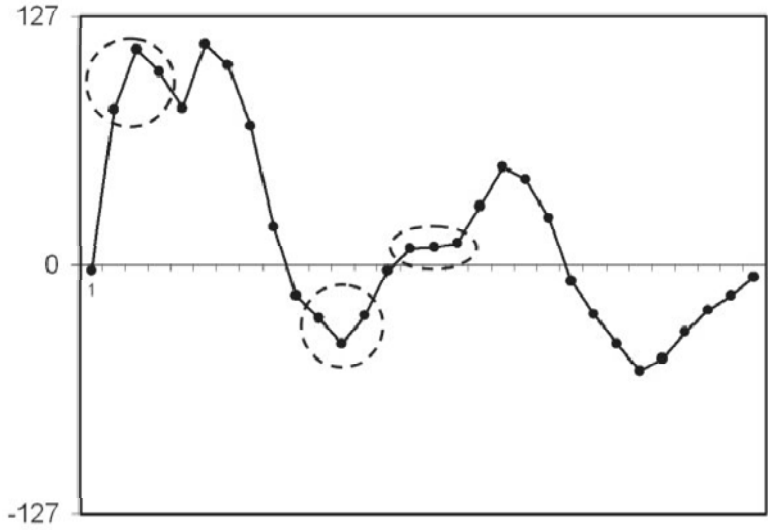

Figure 1. The diagram above is based on a corresponding diagram in the LVA patent documents. It illustrates a portion of a digitized signal and what is meant by 'Thorns' (marked with circles) and a 'Plateau' (marked with an ellipsis).

The speech signal consists of pressure oscillations relative to the ambient atmospheric pressure, so inevitably there will be all sorts of amplitude variation due

to the very nature of the signal. These variations capture a mixture of aspects related to both the voice source and the characteristics of the frequency and phase transfer function of the vocal tract (along with sub-glottal cavities). In the transmission process the signal is further influenced by ambient acoustics, and if the signal is recorded, factors like the type of microphone used and linearity of the recording system will also modify the signal. Producers of the LVA and other voice based detectors show no signs of being aware of this complexity.

When an analog signal is digitized the complex continuous variation found in the original signal is replaced by a simplified discrete representation. How closely this representation matches the original depends on the sampling parameters but the match will never be perfect. It is in the digitization process that the 'thorns' and 'plateaus' are created. There is obviously an indirect relationship between thorns and plateaus and the original waveform, but the number of thorns and plateaus, which is the very basis for all computations in the LVA, depends crucially on sampling rate, amplitude resolution and the threshold values defined in the program. It is therefore correct to say that these computations are basically no more than statistics based on digitization artefacts.

The lie detection is performed in two steps. The first step is called 'calibration'. In this step, speech samples meant to represent neutral emotion are recorded. Based on these recordings a baseline is calculated. The baseline is no more than a simple statistical summary of the number of thorns, the number of plateaus, the distribution of plateau lengths (they are allowed to vary between 3 and 20 samples) and the range of variation in these factors. When this step is completed, the actual emotion detection may begin. In this step the statements to be tested are recorded and the statistics for these statements are compared with the baseline. Based on the deviations from the baseline the various emotional states are computed. Here is an example copied from the patent documents:

> A crLIE value to 50 may be deemed indicative of untruthfulness, values between 50 and 60 may be deemed indicative of sarcasm or humor, values between 60 and 130 may be deemed indicative truthfulness, values between 130 and 170 may be deemed indicative of inaccuracy or exaggeration, and values exceeding 170 may be deemed indicative of untruthfulness. (Liberman 2003)

And that is all there is. There is nothing special with these computations, except that there is no theoretical basis for them or independent motivation for the proposed ranges. First of all, the creation of digitization artefacts is completely

independent of what the recorded sound represents. The program would analyze any sound the same way, be it a man speaking, an idling car engine, a dog barking or a tram passing by. Secondly, the number and distribution of thorns and plateaus depend crucially on a number of factors that have to do with how the digitization is performed. Different sampling frequencies and amplitude resolutions would produce different results. Exactly at which moment in time the sampling begins can also have an effect.

We initially intended to use the code published in the patent documents to make a running copy of the program, but the code is rather messy and not particularly well structured and we decided it would not be worth the time and effort to clean up the code in order to convert it into a running program. The Damphousse et al. group report that the program crashed repeatedly during their experiments so it is obviously rather unstable too. It is rather easy, however, to reconstruct what the program is supposed to do by following the code together with the verbal comments and explanations in the patent documents. We therefore decided to simulate the program in *Mathematica* instead to get better control and be able to monitor the computations more closely.

The presentation of analysis results in the program is modelled on what we, in the beginning, jokingly called 'the horoscope principle', a description we have come to regard as more and more accurate in the course of our work. Let us exemplify what we mean by a short example based on our simulations. In one of our tests we used an interview with a well-known Swedish politician. Using the threshold settings suggested in the patent documents we got the following result for the main output labels:

> Untruthfulness, Low stress, Thinking less than in the calibration,
> Normal excitement

Looking at it superficially this is not an implausible profile for a politician. But as we will explain in the following paragraph, the combination of output labels and variables is not motivated and anyone of the logically possible permutations of variables and labels would work equally well. Choosing the same input but a different variable/label combination produced the following analysis: 'Truthfulness, High stress, Thinking less than in calibration, Normal excitement' which is an equally plausible description as is 'Truthfulness, Normal stress, Thinking more than in calibration, Low level of excitement' and so on. And all these results are, of course, completely detached from anything that has to do with the mental state of the speaker.

The output of an analysis is structured much along the same lines as horoscopes. By presenting the result as a combination of several statements it is possible to achieve an overall description that seems reasonably balanced

and plausible – a little bit of negative information and a little bit of positive or neutral information, and general enough not to seem implausible. Rather large intervals for each emotional degree makes extreme combinations less likely, further enhancing the apparent reasonableness of the output. For this insight into human psychology at least we must give the author of the program some credit.

In our simulations we have used the output labels thresholds suggested in the patent documents, but as we will explain in the following paragraph, the correspondence between the labels and what they represent (e.g. emotional stress level vs. average number of thorns), is perfectly arbitrary. We have also noticed that they are given slightly different wordings in different applications, and radically different wordings in special applications like the so called 'love detector'.[8] From the producers' point of view this makes a lot of sense. Why waste time and energy inventing a new program when all that is needed to build a love detector for example is to rename 'cognitive stress' as 'sexual attraction'?

To sum up by saying that there is absolutely no scientific basis for the claims made by the LVA proponents is an understatement. The ideas on which the products are based are simply complete nonsense.

Definitions of some fundamental variables in the LVA program

Haddad et al. (2002: 23) present a table which summarises variable descriptions for *Diogenes* Lantern (VSA) and the LVA based Vericator by *Nemesysco*. We have adapted their list and added in column three the definitions of the variables in the program (Table 1). A comparison between the two sets of definitions is a telling illustration of the discrepancy between LVA fiction and reality. It should be obvious even to the technically less advanced reader that the assumption of a correlation between for example 'emotional stress level' and 'the average number of thorns' is completely arbitrary.

To further illustrate the randomness of the approach we would like to make the reader aware that there is nowhere in the program or the text in the patent document any motivation why, for example, the average number of thorns should represent 'emotional stress level' and not 'global stress level'. The assignment of interpretations to variables is also completely arbitrary and the program would work perfectly well producing the same type of analysis if we changed the combination of variables and interpretations around in any of the possible permutations. So why was the particular combination listed above chosen? You will not find the answer in any of the documents and our personal guess is that there is no particular reason.

Table 1: Variable descriptions (modified from Haddad et al. 2002)

Variable	As described to the user	As defined in the program
SPT	A numeric value describing the relatively high frequency range. Vericator associates this value with emotional stress level	The average number of thorns
SPJ	A numeric value describing the relatively low frequency range. Vericator associates this value with cognitive stress level	The average number of plateaus per sample
JQ	A numeric value describing the distribution uniformity of the relativity low frequency range. Vericator associates this value with global stress level	The standard error of plateau length
AVJ	A numeric value describing the average range of the relativity low frequency range. Vericator associates this value with thinking level	The average plateau length

Reliability studies of the LVA

The LVA is much newer than the VSA-based machines so there are only a few reliability studies. We will again quote the ones by Hollien and Harnsberger (2006) and Damphousse et al. (2007):

> The performance of LVA on the VSA database … was similar to that observed with CVSA. That is, neither device showed significant sensitivity to the presence of stress or deception in the speech samples tested. The true positive and false positive rates were parallel to a great extent. (Hollien and Harnsberger 2006: 40)

> Although the LVA instrument tended to perform better than the CVSA instrument, both programs failed consistently to correctly identify respondents who were being deceptive. (Damphousse et al. 2007: 88)

As was the case with the VSA, the vendors of the LVA equipment complained that the negative results in the above cited studies were due to the fact that the research teams had not properly followed the required procedures. But in both studies, the research teams had been trained by following the in-house training provided by the vendors. Also the test procedures had been thoroughly discussed with the vendors.

We would like to counter a possible objection to our description of the LVA program right away. The vendors might object that we have based our verdict only on the patent documents and not on the 'real thing' and that the present version is vastly improved. We would not be particularly worried by any objections along those lines. First of all the *Nemesycso* Company makes reference to the patent in all its documentation. There is no mention anywhere that what is described in the patent documents is not what the current machines are based on. Quite on the contrary, the company is seeking patents in more countries all the time using the same description. Secondly, we have read the correspondence between the Hollien group and the LVA vendors regarding their complaints about the methodology. In these documents there is a rather detailed technical discussion mentioning functions used in the program, suitable thresholds and so on. All this information is in perfect agreement with what we have found in the patent documents and so is the brief function description in Haddad et al. There is thus no indication that the code has been changed in any substantial way. It will surely have been updated with respect to graphical interface and other details but the basic principles are most certainly the same.

Who is Mr Liberman?

We might as well have asked: Who is *Nemesysco*, the company behind the LVA products, because Mr Liberman and *Nemesysco* seem to be one and the same. Damphousse et al. (2007: 14) report as follows: 'The LVA was developed in Israel by Amir Lieberman [sic] who applied mathematic algorithm science to voice frequencies', giving the impression that the program is based on some advanced mathematical theory. As we have pointed out, this is far from the truth. When we first became aware of the LVA, in connection with an attempt in 2004 to introduce the LVA on the Scandinavian market, we too were given the impression that Mr Liberman was indeed a high ranking Israeli mathematician. We do not know the origin of these rumours. It has been said that the information once appeared on the *Nemesysco* home pages but we have not been able to confirm this. Screening the *Nemesysco* home pages we became highly suspicious of these claims, however. To acquire more information about the person behind the products we consulted an Israeli colleague who is an active speech science researcher and asked him if he knew of a mathematician by that name. He did not. A controversy arose between us and the Scandinavian representatives of the LVA whom we, after a careful study of the LVA claims, accused of trying to peddle a bogus

product. This controversy, partly fought in a newspaper, caught the interest of a journalist, Arne Lapidius, who was working in Israel for the Swedish daily *Expressen*. After some research he managed to locate Mr Liberman, a 32 year old (in 2004) businessman in a small office in the town of Natania. The business appeared to be a one-man operation. Mr Lapidus interviewed[9] Mr Liberman about his academic background and was told that he basically had none. He has no degree (never had time to get one, he explains) but has taken some courses in marketing at an Israeli open university. As we have explained above, the LVA is a simple program written in rather amateurishly used Visual Basic. Given what we now know about Mr Liberman, that is about what one would expect rather than '8,000 mathematical algorithms applied to 129 voice frequencies' (Damphousse et al. 2007: 15). What still remains for us to understand is how insurance companies, security agencies, police departments can be willing to invest hundreds of thousands of dollars, pounds, and euros in equipment without ever asking who are behind the products, what are their qualifications, what are the scientific principles upon which the products are based. The program code is part of the patent documents and may be downloaded from patents on-line. Any qualified speech scientist with some computer background can see at a glance, by consulting the documents, that the methods on which the program is based have no scientific validity. Why did those who so willingly invested huge amounts of money not even bother to look? For us this is the real puzzle.

Sales figures for some of the cited products

While, as we have seen, the voice stress detectors are not of any real use as the lie or stress detectors they are claimed to be, they have certainly not been without success in other areas. One such area is making money for the vendors. The programs are sold pre-installed on (usually) laptop computers. The *National Institute for Truth Verification*, for example, sells their program (CVSA) pre-installed on a Dell laptop computer for $US9,995. And this is the least expensive option. A quick price search on the Internet shows that the computer itself can be found for around $US2,000. To be a 'certified examiner' one is required to go through a training programme organized by the vendor at the cost of $US1,440 per student. This company alone claims to have sold their products to over 1,400 agencies throughout the US. Even if each agency has bought only one laptop of the cheapest kind and sent only one person to the training we are talking about a gross income of more than $US16,000,000. The LVA is even more expensive. They charge around $US25,000 for a comparable laptop/training package.

Success stories and moral issues

If we consult the home pages of the lie detector vendors we will find a completely different picture of the reliability of their products. Here are two examples concerning police investigations.

> The subject was shown the deceptive charts and, following several hours of interrogation confessed
>
> The subject was then confronted with the results of the CVSA as well as other information connecting him to the crime and he gave a full confession
>
> (*National Institute of Truth Verification* home page[10])

But lie detectors have also been used with great success by insurance companies:

> A car insurer which introduced phone lie detectors says a quarter of all vehicle theft claims have been withdrawn since the initiative began. Admiral started using Digilog voice stress analysis technology in May, in an attempt to stamp out fraudulent claims. When policy holders call, they are told they are being recorded and their voices are being analysed. (BBC, 30 October 2003[11])

and banks:

> Fraud detection savings have increased six fold since the introduction of voice recognition analysis software in 2003, Halifax Bank of Scotland General Insurance reported this week. In 2005, of the total claims referred for investigation, 39% were assessed using the DigiLog VRA technology, which identified 44% as High Risk prompting further assessment. Claimants withdrew their claims voluntarily in half of these cases. (Post online, 28 September 2006[12])

And there is more to come:

> Lie detectors will be used to help root out benefit cheats, Work and Pensions Secretary John Hutton has said. So-called 'voice-risk analysis software' will be used by council staff to help identify suspect claims. It can detect minute changes in a caller's voice which give clues as to when they may be lying. The technology is already used by the insurance industry to combat fraud and will be trialled by Harrow Council, in north London, from May. (BBC, 5 April 2007[13])

Insurance companies seem to favour an application of the LVA called 'voice-risk analysis' which judging from the description on the *Nemesysco* home page is

basically the lie detector with a different name. The insurance companies do not define success in terms of confessions like the police, of course, but in terms of increased benefits and reduced costs. How are we to explain these success stories in the face of what we have said above about the complete lack of both validity and reliability of these products? The explanation lies in what in the scientific literature is referred to as 'the Bogus Pipeline Effect'; 'The expectation is that subjects will answer more honestly if they believe that the truth can be tested for accuracy' (Damphousse et al. 2007: 82) or 'no one wants to be second-guessed by a machine' to put it in the words of the originators (Jones and Sigall 1971: 349).

This hypothesis has been confirmed in many studies. A short but clear description and useful references may be found in Damphousse et al. (2007: 82). Their investigation includes an attempt to quantify the Bogus Pipeline Effect in a lie detector study. We are not aware of any other comparable study where that has been done. Their own experimental investigation did not include a study of the Bogus Pipeline Effect but material from a previous experiment carried out at the same prison made such a study possible. Conditions in the earlier study were basically the same except for the absence of a lie detector. In both studies subjects were informed about the use of urinalysis and in the Damphousse et al. study they were also informed that their answers were to be analysed by a lie detector. By comparing the two studies it was possible to isolate the influence of the Bogus Pipeline Effect caused by the lie detector information. It turned out that the effect was substantial. In the Damphousse et al. study only 14% lied about recent drug use compared to 40% in the study where no lie detector was used or mentioned. The authors conclude that 'Arrestees who thought their interviewers were using 'lie detectors' were much less likely to be deceptive when reporting recent drug use'. It is important to point out that the remarkably strong effect is the effect of informing the subjects about the use of a lie detector only. Whether the lie detector actually does anything or is even physically present is irrelevant. Telling the subjects that a lie detector will be used, but without actually using one, will have the same effect as long as the subjects believe that a lie detector is used. This is the important message to keep in mind when judging reports by the police about how the use of a lie detector helped them get a confession or when insurance companies inform us about how much money they have saved by a decrease in insurance fraud. Bringing down false claims from 40% to 14% is likely to correspond to millions of pounds or dollars.

So if U.K. insurance companies claim they cut their costs by millions of pounds by using 'lie detectors', or US police officers say they can make suspects confess by showing them the results of the 'Voice Stress Analysis', or social security administrators say they may bring down benefit fraud, we have

no reason to question this. The question in these cases is not about reliability but about moral principles. We know from the reliability tests reported above and from our own study of the scientific validity of these gadgets that they have no ability to reveal lies and deception as such. To inform customers or suspects that you have a lie detector capable of distinguishing between deception and the truth is simply untrue, a lie if you wish, and that raises a number of moral questions to consider. Should we accept that insurance companies increase their profits by lying to their customers? Is the use of lies acceptable if it makes a suspect confess? Do we want councils to bring down social benefit costs by lying to their clients? We find no reason to answer these questions by providing our own personal views, but we think that anyone who has an insurance policy, who applies for social benefits or is concerned with the methods used in criminal investigations, should ask these questions and pose them to those responsible for deciding the policies in the respective area.

We would like to make one more observation related to confirming evidence of the type: 'and then they confessed'. Haddad et al. (2002) fall into the same trap of counting confessions as proof of test reliability: 'Both suspects confessed and were subsequently convicted of murder' (p. 16). The study also contains two reports by police officers using the Lantern VSA in their work. One of them reports: 'I have found in several cases that a person 'fails', if you will, on all relevant/crime questions, but has been found to have not committed the crime'. Statements like these are seldom heard, but when we look at the reliability tests we learn that false positives are about as common as true positive (Hollien and Harnsberger 2006: 37) which is precisely what one would expect from a machine that operates at chance level. Conveniently forgetting about the false positives will of course boost the reliability figures in all cases where confessions are used as the criterion. The Bogus Pipeline effect in combination with 'forgetting the false positives effect' goes a long way in explaining reports like 'Over the past three years I would say that I have achieved a success rate of about 97 percent on tests vs. confessions' (Criminal Investigator, Michael G. Adsit, cited in Haddad et al. 2002: 17).

Security *issues*

There have been rumours in the press for some time that lie detectors will be used in security surveillance, for example at airports. The *Nemesysco* Company is marketing a product called the *Gate Keeper* (GK1 is the current model) using the LVA technique that is meant for such applications. They claim that it is already in use at Moscow International Airport (*Domodedovo*) but we have not been able to obtain independent confirmation. If this is true and more

airports will follow, we are not only looking at morally dubious business but at a very real threat against airport security. Since the LVA technique is totally unreliable it would mean that part of airport security will be based on decisions no more valid than throwing a pair of dice. Obviously the GK1 will not replace other systems and procedures, but it is serious enough if it is allowed to divert attention from real security work by letting security personnel waste their time and effort on a completely meaningless task.

Is there anything we can do to prevent charlatanry in forensic speech science?

Charlatanry, fraud, prejudice and superstition have always been with us. If we look back in history and compare with what we see today there is little that gives us hope that progress in science will diminish the amount of superstitious nonsense we see around us. Astrology, for example, seems to be more popular than ever and totally unaffected by how many times astronomers explain that it is complete nonsense. We are therefore somewhat pessimistic about the possibility of efficiently removing charlatanry from forensic speech science. But we hope that responsible authorities like the police and security services will listen to scientifically trained experts in the field rather than to smooth talking and wishful thinking from vendors of bogus lie detectors and similar gadgets. That is probably where we should invest our efforts. We must also take great care when we present our results so that the issue does not appear as a scientific controversy, which it is not. No qualified speech scientist believes in this nonsense so there is absolutely no controversy there, and it is very important that this becomes clear. We have included sufficient detail in this paper to provide the reader with useful arguments in the struggle against charlatanry. We hope that the effort will not turn out to be totally without effect.

Notes

1 http://www.diogenescompany.com/ This web reference, and all others listed in this article, were checked on 28 October 2007.

2 http://www.nemesysco.com/technology-lvavoiceanalysis.html

3 http://www.cvsa1.com/CVSA.htm

4 http://www.diogenescompany.com/vsaprogram.html

5 http://www.nemesysco.com/technology-lvavoiceanalysis.html

6 http://www.nemesysco.com/technology-lvavoiceanalysis.html

7 http://www.vsolutions.org/

8 http://www.love-detector.com/

9 The interview with Mr Liberman (in Swedish) appeared in *Västerbottens-Kuriren* on 17 December 2004. It is not available online.

10 http://www.nitv1.com/realcases.htm

11 http://news.bbc.co.uk/1/hi/uk/3227849.stm

12 http://www.postmagazine.co.uk/public/showPage html?validate=0&page=post_login2&url=%2Fpublic%2FshowPage html%3Fpage%3D346755 (Login required)

13 http://news.bbc.co.uk/1/hi/uk/6528425.stm

References

Bolger, C., Bojanic, S., Sheahan, N. F., Coakley, D. and Malone, J. F. (1999) Dominant frequency content of ocular microtremor from normal subjects. *Vision Research* 39: 1911–1915.

Damphousse, K. R., Pointon, L, Upchurch, D. and Moore, R. K. (2007) *Assessing the Validity of Voice Stress Analysis Tools in a Jail Setting*. Report submitted to the U.S. Department of Justice. http://www.ncjrs.gov/pdffiles1/nij/grants/219031.pdf

Faaborg-Andersen, K. (1957) Electromyographic Investigation of intrinsic laryngeal muscles in humans. *Acta Physiologica Scandinavica* 41, Supplement 140. (147 p.)

Haddad, D., Walter, S., Ratley, R. and Smith, M. (2002) *Investigation and Evaluation of Voice Stress Analysis Technology, Final Report*. National Institute of Justice, NCJRS 193832. http://www.ncjrs.gov/pdffiles1/nij/193832.pdf

Hirano, M. and Ohala, J. (1969) Use of hooked-wire electrodes for electromyography of the intrinsic laryngeal muscles. *Journal of Speech and Hearing Research* 12: 362–373.

Hirano, M., Vennard, W. and Ohala, J. (1970) Regulation of register, pitch and intensity of voice. An electromyographic investigation of intrinsic laryngeal muscles. *Folia Phoniatrica* 22: 1–20.

Hollien, H. and Harnsberger, J. D. (2006) *Voice Stress Analyzer Instrumentation Evaluation*. Final Report, CIFA Contract – FA 4814–04–0011. http://www.clas.ufl.edu/users/jharns/Research%20Projects/UF_Report_03_17_2006.pdf

Jones, E. E. and Sigall, H. (1971) The bogus pipeline: a new paradigm for measuring affect and attitude. *Psychological Bulletin* 76(5): 349–364.

Liberman, A. (2003) *Apparatus and Methods for Detecting Emotions*. US Patent 6638217 B1. http://www.freepatentsonline.com/6638217.html

Lippold O. C., Redfearn, J. W. T. and Vučo, J. (1957) The rhythmical activity of groups of motor units in the voluntary contraction of muscle. *The Journal of Physiology* 137: 473–487.

Lippold, O. C. (1970) Oscillation in the stretch reflex arc and the origin of the rhythmical, 8–12 c/s component of physiological tremor. *The Journal of Physiology* 206: 359–382.

Lippold, O. C. (1971) Physiological tremor. *Scientific American* 224: 65–73.

Lynch, B. E. and Henry, D. R. (1979) A validity study of the psychological stress evaluator. *Canadian Journal of Behavioural Science* 11: 89–94.

McAuley, J. H. and Marsden, C. D. (2000) Physiological and pathological tremors and rhythmic central motor control. *Brain* 123: 1545–1567.

Ridelhuber, H. W. and Flood, P. (2002) Policy review: CVSA Is a valid law enforcement tool. *Law Enforcement Executive Forum* August 2002: 95–100. (http://www.leeforum.com/)

Shipp, T., Fishman, B. V., Morrissey, P. and McGlone, R. E. (1970) Method and control of laryngeal EMG electrode placement in man. *Journal of the Acoustical Society of America* 48: 429–430.

Shipp, T. and Izdebski, K. (1981) Current evidence for the existence of laryngeal macrotremor and microtremor. *Journal of Forensic Sciences* 26: 501–505.

Smith, A. and Denny, M. (1990) High-frequency oscillations as indicators of neural control mechanisms in human respiration, mastication, and speech. *Journal of Neurophysiology* 63: 745–758.

VanDercar, D. H., Greaner, J., Hibler, N. S., Spielberger, C. D. and Block, S. (1980) A description and analysis of the operation and validity of the psychological stress evaluator. *Journal of Forensic Sciences* 25: 174–188.