LVA-technology – The illusion of “lie detection”[1]

Francisco Lacerda
Department of Linguistics, Stockholm University

Abstract

The new speech-based lie-detection LVA-technology is being used in some countries to screen applicants, passengers or customers in areas like security, medicine, technology and risk management (anti-fraud). However, a scientific evaluation of this technology and of the principles on which it relies indicates, not surprisingly, that it is neither valid nor reliable. This article presents a scientific analysis of this LVA-technology and demonstrates that it simply cannot work.

Introduction

After of the attacks of September 11, 2001, the demand for security technology was considerably (and understandably) boosted. Among the security solutions emerging in this context, Nemesysco Company’s applications claim to be capable of determining a speaker’s mental state from the analysis of samples of his or her voice. In popular terms Nemesysco’s devices can be generally described as “lie-detectors”, presumably capable of detecting lies using short samples of an individual’s recorded or on-line captured speech. However Nemesysco claims their products can do much more than this. Their products are supposed to provide a whole range of descriptors of the speaker’s emotional status, such as exaggeration, excitement and “outsmarting” using a new “method for detecting emotional status of an individual”, through the analysis of samples of her speech. The key component is Nemesysco’s patented LVA-technology (Liberman, 2003). The technology is presented as unique and applicable in areas such as security, medicine, technology and risk management (anti-fraud). Given the consequences that applications in these areas may have for the lives of screened individuals, a scientific assessment of this LVA-technology should be in the public’s and authorities’ interest.

Nemesysco’s claims

According to Nemesysco’s web site, “LVA identifies various types of stress, cognitive processes and emotional reactions which together comprise the “emotional signature” of an individual at a given moment, based solely on the properties of his or her voice”[i]. Indeed, “LVA is Nemesysco's core technology adapted to meet the needs of various security-related activities, such as formal police investigations, security clearances, secured area access control, intelligence source questioning, and hostage negotiation”[ii] and “(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”. Since the principles and the code used in the technology are described in the publicly available US 6,638,217 B1 patent, a detailed study of the method was possible and its main conclusions are reported here.

Deriving the “emotional signature” from a speech signal

While assessing a person’s mental state using the linguistic information provided by the speaker (essentially by listening and interpreting the person’s own description of her or his state of mind) might, in principle, be possible if based on an advanced speech recognition system, Nemesysco’s claim that the LVA-technology can derive "mental state" information from “minute involuntary changes in the speech waveform itself” is at least astonishing from both a phonetic and a general scientific perspective. How the technology accomplishes this is however rather unclear. No useful infor-

[1] This text is a modified version of Lacerda (2009), “LVA-technology – A short analysis of a lie”, available online at http://blogs.su.se/frasse/, and is intended for a general educated but not specialized audience.

mation is provided on the magnitude of these “minute involuntary changes” but the wording conveys the impression that these are very subtle changes in the amplitude and time structure of the speech signal. A reasonable assumption is to expect the order of magnitude of such "involuntary changes" to be at least one or two orders of magnitude below typical values for speech signals, inevitably leading to the first issue along the series of ungrounded claims made by Nemesysco. If the company's reference to "minute changes" is to be taken seriously, then such changes are at least 20 dB below the speech signal's level and therefore masked by typical background noise. For a speech waveform captured by a standard microphone in a common reverberant room, the magnitude of these "minute changes" would be comparable to that of the disturbances caused by reflections of the acoustic energy from the walls, ceiling and floor of the room. In theory, it could be possible to separate the amplitude fluctuations caused by room acoustics from fluctuations associated with the presumed “involuntary changes” but the success of such separation procedure is critically dependent on the precision with which the acoustic signal is represented and on the precision and adequacy of the models used to represent the room acoustics and the speaker's acoustic output. This is a very complex problem that requires multiple sources of acoustic information to be solved. Also the reliability of the solutions to the problem is limited by factors like the precision with which the speaker's direct wave-front (originating from the speaker’s mouth, nostrils, cheeks, throat, breast and other radiating surfaces) and the room acoustics can be described. Yet another issue raised by such “sound signatures” is that they are not even physically possible given the masses and the forces involved in speech production. The inertia of the vocal tract walls, velum, vocal folds and the very characteristics of the phonation process lead to the inevitable conclusion that Nemesysco’s claims of picking up that type of "sound signatures" from the speaker’s speech waveform are simply not realistic. It is also possible that these “minute changes” are thought as spreading over several periods of vocal-fold vibration. In this case they would be observable but typically not “involuntary”. Assuming for a moment that the signal picked up by Nemesysco’s system would not be contaminated with room acoustics and background noise, the particular temporal profile of the waveform is essentially created by the vocal tract’s response to the pulses generated by the vocal folds’ vibration. However these pulses are neither “minute” nor “involuntary”. The changes observed in the details of the waveforms can simply be the result of the superposition of pulses that interfere at different delays.

In general, the company’s descriptions of the methods and principles are circular, inconclusive and often incorrect. This conveys the impression of superficial knowledge of acoustic phonetics, obviously undermining the credibility of Nemesysco’s claims that the LVA-technology performs a sophisticated analysis of the speech signal. As to the claim that the products marketed by Nemesysco would actually be able to detect the speaker’s emotional status, there is no known independent evidence to support it. Given the current state of knowledge, unless the company is capable of presenting scientifically sound arguments or at least producing independently and replicable empirical data showing that there is a significant difference between their systems’ hit and false-alarm rates, Nemesysco’s claims are unsupported.

How LVA-technology works

This section examines the core principles of Nemesysco’s LVA-technology, as available in the Visual Basic Code in the method’s patent.

Digitizing the speech signal

For a method claiming to use information from minute details in the speech wave, it is surprising that the sampling frequency and the sample sizes are as low as 11.025 kHz and 8 bit per sample. By itself, this sampling frequency is acceptable for many analysis purposes but, without knowing which information the LVA-technology is supposed to extract from the signal, it is not possible to determine whether 11.025 kHz is appropriate or not. In contrast, the 8 bit samples inevitably introduce clearly audible quantification errors that preclude the analysis of “minute details”. With 8 bit samples only 256 levels are available to encode the sampled signal’s amplitude, rather than 65536 quantization levels associated with a 16 bit sample. In acoustic terms this reduction in sample length is associated with a 48 dB increase of the background noise relative to what would have been possible using a 16-bit/sample representation. It is puzzling that such crude signal representations are used by a technology claiming to work on "details". But the degradation of the amplitude resolution becomes even worse as a "filter" that introduces a coarser quantization using 3 units' steps reduces the 256 levels of the 8-bit representation to only 85 quantization levels (ranging from -42 to +42). This very low sample resolution (something around 6.4-bit/sample), resulting in a terrible sound quality, is indeed the basis for all the subsequent signal processing carried out by the LVA-technology. The promise of an analysis of "minute" details in the speech waveform cannot be taken seriously. Figure 1 displays a visual analogue of the signal degradation introduced by the LVA-technology.

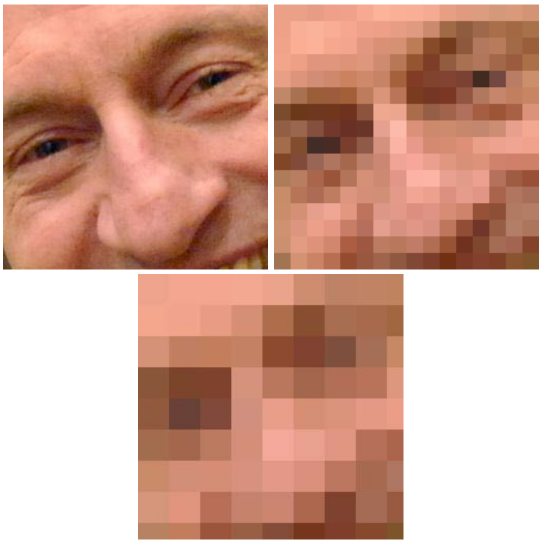

Figure 1. Visual analogs of LVA-technology's speech signal input. The 256×256 pixels image, corresponding to 16 bit samples, is sampled down to 16×16 pixels (8 bit samples) and finally down-sampled to approximately 9×9 pixels representing the ±42 levels of amplitude encoding used by the LVA-technology.

The core analysis procedure

In the next step, the LVA-technology scans that crude speech signal representation for "thorns" and "plateaus" using triplets of consecutive samples.

"Thorns"

According to Nemesysco's definition, thorns are counted every time the middle sample is higher than the maximum of the first and third samples, provided all three samples are above an arbitrary threshold of +15. Similarly, a thorn is also detected when the middle sample value is lower than the minimum of both the first and the third samples in the triplet and all three samples are below -15. In short, thorns are local maxima, if the triplet is above +15 and local minima if the triplet is below -15. Incidentally this is not compatible with the illustration provided in fig. 2 of the patent, where any local maxima or minima are counted as thorns, provided the three samples fall outside the (-15;+15) threshold interval.

"Plateaus"

Potential plateaus are detected when the samples in a triplet have a maximum absolute amplitude deviation that is less than 5 units. The ±15 threshold is not used in this case but to count as a plateau the number of samples in the sequence must be between 5 and 22. The number of occurrences of plateaus and their lengths are the information stored for further processing.

A blind technology

Although Nemesysco presents a rationale for the choice of these "thorns" and "plateaus" that simply does not make sense from a signal processing perspective, there are several interesting properties associated with these peculiar variables. The crucial temporal information is completely lost during this analysis. Thorns and plateaus are simply counted within arbitrary chunks of the poorly represented speech signal which means that a vast class of waveforms created by shuffling the positions of the thorns and plateaus are indistinguishable from each other in terms of totals of thorns and plateaus. Many of these waveforms may even not sound like speech at all. This inability to distinguish between different waveforms is a direct consequence of the information loss accomplished by the signal degradation and the loss of temporal information. In addition to this, the absolute values of the amplitudes of the thorns can be arbitrarily increased up to the ±42 maximum level, creating yet another variant of physically different waveforms that are interpreted as identical from the LVA-technology's perspective.

The measurement of the plateaus appears to provide only very crude information and is affected by some flaws. Indeed, the program code allows for triplets to be counted as both thorns and plateaus. Whether this is intentional or just a programming error is impossible to determine because there is no theoretical model behind the LVA-technology against which this could be checked. In addition, what is counted as a plateau does not even have to look like a plateau. An increasing or decreasing sequence of samples where differences between adjacent samples are less than ±5 units will count as a plateau. Only the length and the duration of these plateaus are used and because the ±5 criterion is actually a limitation on the derivate of the amplitude function, large amplitude drifts can occur in sequences that are still viewed by LVA-technology as if they were flat. Incidentally, given that these plateaus can be up to 22 samples long, the total span of the amplitude drift within a plateau can be as large as 88 units, which would allow for a ramp to sweep through the whole range of possible amplitudes (-42 to +42). This is hardly compatible with the notion of high precision technology suggested by Nemesysco. Finally, in addition to the counting of plateaus, the program also computes the square root of the cumulative absolute deviation for the distribution of the plateau lengths. Maybe the intention was to compute the standard deviation of the sample distribution and this is yet another programming error but since there is no theoretical rationale it is impossible to discuss this issue.

Assessing the speaker's emotional state

The rest of the LVA-technology simply uses the information provided by these four variables: (1) the number of thorns per sample, (2) the number of counts of plateaus per sample, (3) the average length of the plateaus and (4) the square root of their cumulative absolute deviation. From this point on the program code is no longer related to any measureable physical events. In the absence of a theoretical model, the discussion of this final stage and its outcome is obviously meaningless. It is enough to point out that the values of the variables used to issue the final statements concerning the speaker's emotional status are as arbitrary as any other and of course contain no more information than what already was present in the four variables above.

Examples of waveforms that become associated with "LIES"[2]

Figure 2 shows several examples of a synthetic vowel that was created by superimposing with the appropriate delays to generate different fundamental frequencies, a glottal pulse extracted from a natural production.

After calibration with glottal pulses simulating a vowel with a 120 Hz fundamental frequency, the same glottal pulses are interpreted as indicating a "LIE" if the fundamental frequency is lowered to 70 Hz whereas a raise in fundamental frequency from 120 Hz to 220 Hz is detected as "outsmart". Also a fundamental frequency as low as 20 Hz is interpreted as signalling a "LIE", relative to the 120 Hz calibration.

Using the 20 Hz waveform as calibration and testing with the 120 Hz is detected as "outsmart". A calibration with the 120 Hz wave above followed by the same wave contaminated by some room acoustics is also interpreted as "outsmart".

The illusion of a serious analysis

The examples above suggest that the LVA-technology generates outputs contingent on the relationship between the calibration and the test signals. Although the signal analysis performed by the LVA-technology is a naive and ad hoc measurement of essentially irrelevant aspects of the speech signal, the fact that some of the "detected emotions" are strongly dependent on the statistical properties of the "plateaus" leads to outcomes that vaguely reflect variations in F0. For instance, the algorithm's output tends to be a "lie" when the F0 of the test signal is generally lower than that of the calibration. The main reason for this is that the program issues "lie"-warnings when the number of detected "plateaus" during the analysis phase exceeds by a certain threshold the number of "plateaus" measured during calibration. When the F0 is

[2] The amplitudes of the waveforms used in this demonstration are encoded at 16 bit per sample.

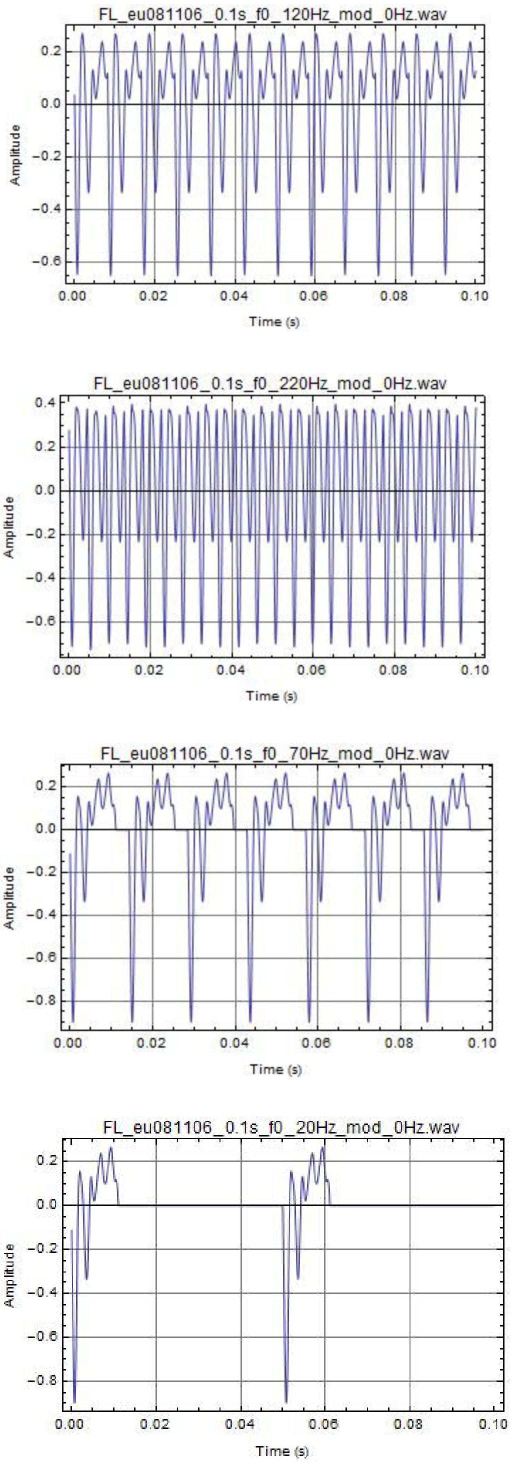

Figure 2. The figures above show synthetic vowels constructed by algebraic addition of delayed versions of a natural glottal pulse. These waveforms lead generate different “emotional outputs” depending on the relationship between the F0 of the waveform being tested and the F0 of the “calibration” waveform.

low, the final portions of the vocal tract's damped responses to more sparse glottal pulses will tend to achieve lower amplitudes in between consecutive pulses. Given the technology's very crude amplitude quantization, these low amplitude oscillations are lost and the sequences are interpreted as plateaus that are longer (and therefore fewer within the analysis window) than those measured in speech segments produced with higher F0. Such momentary changes in the structure of the plateaus are interpreted by the program's arbitrary code as indicating "deception". Under typical circumstances, flagging "lie" in association with lowering of F0 will give the illusion that the program is doing something sensible because F0 tends to be lower when a speaker produces fillers during hesitations than when the speaker's speech flows normally. Since the "lie-detector" is probably calibrated with responses to questions about obvious things the speaker will tend to answer using a typical F0 range that will generally be higher than when the speaker has to answer to questions under split-attention loads. Of course, when asked about events that demand recalling information, the speaker will tend to produce fillers or speak at a lower speech rate, thereby increasing the probability of being flagged by the system as attempting to "lie", although in fact hesitations or lowering of F0 are known to be no reliable signs of deception. Intentionally or by accident, the illusion of seriousness is further enhanced by the random character of the LVA outputs. This is a direct consequence of the technology's responses to both the speech signal and all sorts of spurious acoustic and digitalization accidents. The instability is likely to confuse both the speaker and the "certified examiner", conveying the impression that the system really is detecting some brain activity that the speaker cannot control[3] and may not even be aware of! It may even give the illusion of robustness as the performance is equally bad in all environments.

The UK's DWP's evaluation of LVA

The UK's Department of Work and Pensions has recently published statistics on the results of a large and systematic evaluation of the LVA-technology[iii] assessing 2785 subjects and costing £2.4 million[iv]. The results indicate that the areas under the ROC curves for seven districts vary from 0.51 to 0.73. The best of these

[3] Ironically this is true because the output is determined by random factors associated with room acoustics, background noise, digitalization problems, distortion, etc.

results corresponds to a d' of about 0.9, which is a rather poor performance. But the numbers reported in the table reflect probably the judgements of the "Nemesysco-certified" personal[4] in which case the meaningless results generated by the LVA-technology may have been overridden by personal's uncontrolled "interpretations" of the direct outcomes after listening to recordings of the interviews.

Tabell 1. Evaluation results published by the UK's DWP.

	N	Low risk cases with no change in benefit	High risk cases with no change in benefit	Low risk cases with change in benefit	High risk cases with change in benefit	AUC of ROC Curve
		True Negative	False Positive	False Negative	True Positive	
Jobcentre Plus	787	354	182	145	106	0.54
Birmingham	145	60	49	3	33	0.73
Derwentside	316	271	22	11	12	0.72
Edinburgh	82	60	8	8	6	0.66
Harrow	268	193	15	53	7	0.52
Lambeth	1101	811	108	153	29	0.52
Wealden	86	70	7	8	1	0.51
Overall	**2785**	**1819**	**391**	**381**	**194**	**0.65**

Conclusions

The essential problem of this LVA-technology is that it does not extract relevant information from the speech signal. It lacks validity. Strictly, the only procedure that might make sense is the calibration phase, where variables are initialized with values derived from the four variables above. This is formally correct but rather meaningless because the waveform measurements lack validity and their reliability is low because of the huge information loss in the representation of the speech signal used by the LVA-technology. The association of *ad hoc* waveform measurements with the speaker's emotional state is extremely naive and ungrounded wishful thinking that makes the whole calibration procedure simply void.

In terms of "lie-detection", the algorithm relies strongly on the variables associated with the plateaus. Given the phonetic structure of the speech signals, this predicts that, in principle, lowering the fundamental frequency and changing the phonation mode towards a more creaky voice type will tend to count as an indication of lie, in relation to a calibration made under modal phonation. Of course this does not have anything to do with lying. It is just the consequence a common phonetic change in speaking style, in association with the arbitrary construction of the "lie"-variable that happens to give more weight to plateaus, which in turn are associated with the lower waveform amplitudes towards the end of the glottal periods in particular when the fundamental frequency is low.

The overall conclusion from this study is that from the perspectives of acoustic phonetics and speech signal processing, the LVA-technology stands out as a crude and absurd processing technique. Not only it lacks a theoretical model linking its measurements of the waveform with the speaker's emotional status but the measurements themselves are so imprecise that they cannot possibly convey useful information. And it will not make any difference if Nemesysco "updates" in its LVA-technology. The problem is in the concept's lack of validity. Without validity, "success stories" of "percent detection rates" are simply void. Indeed, these "hit-rates" will not even be statistically significant different from associated "false-alarms", given the method's lack of validity. Until proof of the contrary, the LVA-technology should be simply regarded as a hoax and should not be used for any serious purposes (Eriksson & Lacerda, 2007).

References

Eriksson, A. and Lacerda, F. (2007). Charlatanry in forensic speech science: A problem to be taken seriously. *Int Journal of Speech, Language and the Law*, *14*, 169-193.

Liberman, A. (10-28-2003). *Layered Voice Analysis (LVA)*. [6,638,217 B1]. US patent.

[4] An inquiry on the methodological details of the evaluation was sent to the DWP on the 23 April 2009 but the methodological information has not yet been provided.

[i] http://www.nemesysco.com/technology.html

[ii] http://www.nemesysco.com/technology-lvavoiceanalysis.html

[iii] http://spreadsheets.google.com/ccc?key=phNtm3LmDZEME67-nBnsRMw

[iv] http://www.guardian.co.uk/news/datablog/2009/mar/19/dwp-voice-risk-analysis-statistics