Tests of Basic Voice Stress Detection Techniques

Piotr STARONIEWICZ
Chair of Acoustics and Multimedia,
Wroclaw University of Science and Technology,
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland,
piotr.staroniewicz@pwr.edu.pl

Abstract

The modern speech processing techniques enable new possibilities of potential applications. Besides speech and speaker recognition, also the information about speakers' physical condition, emotional state or stress can be detected in speech signal. Since emotional stress can occur during deception, its detection in speech could be used for law or security services. The paper presents the comparative tests of two voice stress detection techniques: one based on trials of microtremors detection relying on an iterative EMD method (Empirical Mode Decomposition) and the second one based on the statistical analysis of fundamental frequency and MFCC parameters. The preliminary tests were carried on the group of 12 speakers (6 males and 6 females) answering yes/no to the list of a few dozen personal questions. The presented research revealed the speakers' very high personal influence on the obtained results.

Keywords: Voice Stress Analysis, Empirical Mode Decomposition

1. Introduction

The speech processing technology makes it possible nowadays not only to recognise speech and speakers, but also to identify complex information about a speaker's state or condition. The techniques of speaker emotion recognition [1, 2] are dynamically developing and, in some cases, can be even more efficient in proper emotion classification than humans [3].

From the possible applications' point of view a very interesting method is the technique of stress detection in speech, also known as VSA (Voice Stress Analysis). The stress can be caused by external (i.e. physical) or internal (i.e. psychological) factors. Since deception is one of the possible internal factors, the detection of it could be a valuable application for law or security services. Using the classical polygraph encounters an important obstacle which is the necessity of physical connection to the subject. Hence on the one hand the techniques which do not have such a connection (e.g. face thermal vision or VSA) let us reduce other stress factors for the examined subject, but on the other hand, and can be more discreet or even unnoticeable by the subject. The basis of the numerous commercial VSA applications is the controversial Lippold [4] theory from the 1970s of microtremors, i.e. the reaction of muscle tension of vocal chords during the stress of around 8-12 Hz. At the same time, there are numerous commercial applications on the market, whose working algorithms are not disclosured for the understandable reasons and there are still very few scientific reports letting us assess the real usability of the VSA techniques.

2. Speech database

For the voice stress detection purposes a special speech database was designed. Similarly as in the polygraph tests it includes individuals' yes/no answers to the list of a few dozen questions. As the database speakers the couples which were in a relationship for a certain time (at least one year) were recruited. During the recordings the decisive part was played by the content of the asked questions. It was supposed to make the speaker abandon his/her personal "comfort zone" and force him or her to answer some questions deceitfully. During the recordings, the partners were sitting on the opposite sides and asking each other questions from the list. Each partner had a different list of questions. Three kinds of questions were used:

- *Relevant questions*, which were significant for obtaining the information from the object. These questions were asked directly and they concerned the relation between the partners. The purpose of those questions was to evoke stress, for example during the speaker's deceit.
- *Irrelevant questions* were used as a buffer between the relevant questions. The purpose of those questions was to introduce some break and the speaker's relaxation and they were not related to the topic of the discussion.
- *Control questions* reveal the truth and their purpose is to demonstrate the comparison to the relevant question.

The recordings were carried out in good acoustic conditions in a quiet room with the dynamic microphone Shure SM 58 SE, the acoustic mixer Behringer Eurorack MX 802A and the external analogue-digital converter Creative Sound Blaster Audigy 2 NX. The signals were recorded with the sampling rate of 44,100 samples per second with 16 bit resolution.

3. Tested stress detection techniques

Two stress detection techniques were applied. The first one was based on the trials of microtremors detection and the second one was based on the statistical analysis of chosen voice parameters.

The first technique relies on the iterative EMD method (Empirical Mode Decomposition), which was proposed by Huang in 1993 [5]. The EMD method allows to present the analysed, nonstationary signal as a sum of stationary signals called IMFs (Intrinsic Mode Functions).

The four methods were tested for that technique of stress detection. The simplest one (denoted as VSA1 below) extracts the microtremor from the signal. If the microtremor is in the range of 8-12 Hz, the result is "true", otherwise the application recognises the utterance as "false". This simple test was then developed into a method (VSA2) where the distribution of component frequencies was examined. The method recognises the "true" when the ratio of component frequencies inside the investigated band (8-12 Hz) to the sum of all component frequencies exceeds a chosen value. The third method (VSA3) is very similar but it demands the calibration process for each speaker, where the answers to the irrelevant questions are used. The fourth method (VSA4) also demands the calibration process with the answers to the irrelevant questions. During the classification the algorithm compares the most significant component frequencies of the recognized sample to the most significant component frequencies of the reference samples.

The second applied technique was based on the analysis of voice parameters:

- Mean F0 value,
- Range of F0,
- Jitter,
- MFCC parameters (Mel Frequency Cepstral Coefficients).

Since the used parameters are from various domains, their values were normalized. The block scheme of the technique was presented in Figure 1. Two methods were tested for that technique: with mean (denoted as VSA5) and median (denoted as VSA6) values calculated for each parameter. The range of each parameter was then determined on the basis of standard deviations. If the value of the parameter for the answer to the relevant question was not in the determined range the "false" was recognized by the algorithm.

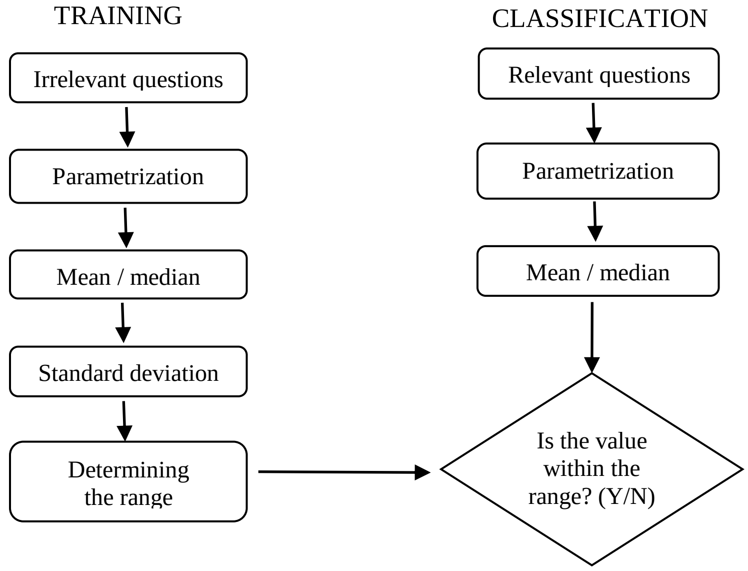

Figure 1. Scheme of stress detection based on statistical analysis of voice parameters

3. Results and discussion

The correctness of deception detection was defined as the ratio of the correctly identified negative and positive data samples to all the data samples. The singular results of the deception detections for each tested speakers were presented in Table 1, where the first four columns of the results (VSA1-VSA4) were obtained for the technique based on EMS detection and the last two columns (VSA5 and VSA6) for the technique based on the statistics of voice parameters. It can be noticed that the differences between the results obtained for the singular subjects can be very substantial: from a very high detection correctness of around 80% (e.g. subject F1) to quite low, even around 40%, for the voices of some male subjects (e.g. subject M6).

Table 1. Detection of deception results for 12 speakers database and six tested stress detection techniques

Speakers		VSA1	VSA2	VSA3	VSA4	VSA5	VSA6
Male / female	Speaker's number						
Females	F1	70.7%	41.3%	70.7%	35.4%	64.8%	70.7%
	F2	81.4%	75.1%	75.1%	87.6%	76.6%	76.6%
	F3	47.2%	41.3%	64.8%	82.5%	64.8%	47.2%
	F4	58.9%	58.9%	47.2%	41.3%	70.7%	70.7%
	F5	56.4%	62.6%	68.9%	62.6%	58.9%	58.9%
	F6	47.2%	58.9%	52.9%	76.6%	76.6%	64.8%
Males	M1	47.2%	58.9%	70.7%	41.3%	58.9%	47.2%
	M2	47.2%	47.2%	64.8%	70.7%	56.4%	50.1%
	M3	41.3%	52.9%	47.2%	64.8%	64.8%	64.8%
	M4	58.9%	59.6%	58.3%	53.4%	81.4%	56.4%
	M5	58.9%	76.6%	64.8%	58.9%	52.9%	47.2%
	M6	37.6%	62.6%	43.9%	31.4%	73.4%	53.4%

The mean values of deception detection for male and female voices for all the tested methods are presented in Figure 2. All the obtained mean results exceed 50%, however, such results rather disqualify the tested techniques for serious law or security applications, at least as the stand-alone ones (a typical polygraph uses techniques from several domains at the same time). As it can be noticed, the best mean value much over 60% was obtained for considerably the simplest method with the statistical analysis of voice parameters (VSA5). Rather disappointing are the results of all the tested techniques based on the EMD method, especially that this method was considered as a very promising one by some researchers for VSA applications [6, 7].

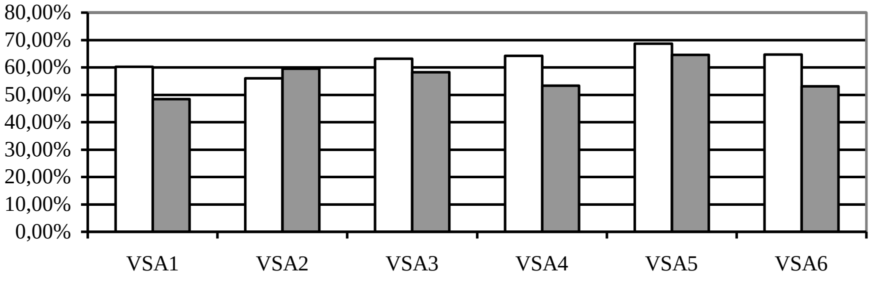

Figure 2. Mean values of deception results for female (white) and male (gray) voices for six tested stress detection techniques

3. Conclusions

The usage of VSA techniques in law or security applications is a very controversial problem. On the one hand the manufacturers of such software put pressure on introducing such applications. On the other hand, its effectiveness was not confirmed objectively so these techniques are not considered as reliable [8].

The results obtained from the carried out tests are significantly much lower than that which are declared by the producers of commercial devices for the detection of deception in the speaker's voice. Moreover, the scores depend very highly on the speaker's individual characteristics. It is evident that some subjects react to stress in the way that can be easier detected than others. The most promising results of over 60% were obtained for considerably simple techniques with statistical analysis of voice parameters, which still gives a promise that the scores of VSA techniques can be improved in the future with more sophisticated algorithms. The problem is even more difficult because of very short utterances that have to be used in such a detector (yes/no answers), which in speech emotion analysis using pitch changes can be a significant obstacle. Also, it is worth mentioning that in such experiments no database can be considered as a hundred percent reliable, even after an anonymous deception validation by the subjects, which was carried out in our case.

There exists a fundamental problem in all the techniques of detecting the deception. The stress is a human reaction to the deception, but it can also be caused by other stimuli. Therefore it would be beneficial to avoid intrusive techniques which demand physical connection of the subject to the detectors. It evokes additional stress which can potentially change the results. Other very big advantage of the non-intrusive techniques is that there is no necessity to make the subject aware of being tested.

References

1. P. Staroniewicz, *Considering basic emotional state information in speaker verification*, Proc. 4th International Conference on Biometrics and Forensics (IWBF), Limmasol, Cyprus, 3-4 March 2016, IEEE 2016.
2. P. Staroniewicz, *Automatic recognition of emotional state in Polish speech*, *Toward autonomous, adaptive, and context-aware multimodal interfaces: theoretical and practical issues*, Lecture Notes in Computer Science, Springer, **6800** (2011) 347 – 353.
3. P. Staroniewicz, *Recognition of emotional state in Polish speech – comparison between human and automatic efficiency*, Lecture Notes in Computer Science, Springer, **5707** (2009) 33 – 40.
4. O. Lippold, *Physiological microtremors*, Scientific American, **224**(3) (1971) 65 – 73.
5. N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N. C. Yen, C. C. Tung, H. H. Liu, *The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis*, Proc. Roy. Soc. Land. A, (1998) 903 – 1005.

6. J. Z. Zhang, N. Mbitiru, P. C. Tay, R. D. Adams, *Analysis of stress in speech using Adaptive Empirical Mode Decomposition*, 2009, Conference Record of the Forty-Third Asilomar Conference on Signals, Systems and Computers, IEEE 2009.
7. N. Mbitiru, P. Tay, J. Z. Zhang, R. D. Adams, *Analysis of Stress in Speech Using Empirical Mode Decomposition*, Proceedings of The 2008 IAJC-IJME International Conference.
8. C. S. Hopkins, D. S. Benincasa, R. J. Ratley, J. J. Grieco, *Evaluation of voice stress analysis technology,* Proceedings of the 38th Hawaii International Conference on System Sciences, IEEE 2005.