http://dx.doi.org/10.5755/j01.eie.22.5.16348

Analysis of Closing-to-Opening Phase Ratio in Top-to-Bottom Glottal Pulse Segmentation for Psychological Stress Detection

Miroslav Stanek[1], Milan Sigmund[1]
[1]Brno University of Technology, Department of Radio Electronics
Technicka 12, 61600 Brno, Czech Republic
mirek.stanek@phd.feec.vutbr.cz

***Abstract*—This paper is focused on investigating the differences in glottal pulses estimated by two algorithms; Direct Inverse Filtering (DIF) and Iterative and Adaptive Inverse Filtering (IAIF) for normal and stressed speech. Individual glottal pulses are mined from recorded speech signal and then normalized in two dimensions. Each normalized pulse is divided into a closing and opening phase and further segmented into *n*-percentage sectors in Top-To-Bottom (TTB) amplitude domain. Three parameters, the kurtosis, skewness and pulse area, as well as their Closing-To-Opening phase ratios, are analysed. Designed GMM classifier is trained on speakers from Czech ExamStress database a further applied on other part of ExamStress database and also for English database SUSAS to investigate the independency of presented approach on spoken language and speech signal quality. The results achieved by DIF indicate independency on language and records quality (contrary to methods using IAIF). The best *n*-percentage sectors in the TTB segments can be seen between 5 % and 40 %. In this case, methods based on DIF reached a psychological stress recognition efficiency of 88.5 % in average. The average stress detection efficiency of methods based on IAIF approached 73.3 %.**

***Index Terms*—Analysis of speaker state; psychological stress detection; glottal pulse analysis; closing-to-opening phase ratio.**

I. INTRODUCTION

Current trend is to monitor the actual emotional state of speaker by non-invasive methods like remote analysis of speech signal mostly for the employees of risk professions, *e.g.* pilots, rescuers, *etc.*, to avoid some dangerous or unpleasant situations. Psychological stress can be classified as an emotion, thus the psychological state influences human behaviour and self-confidence. Due to this reason, it is appropriate to recognize the stress of a speaker immediately, especially in situations when the speaker's behaviour is negatively influenced by distress.

Many methods of stress detection exist and are based mostly on directly mined speech features like MFCC [1], pitch [2], formants [3], *etc.* Other publications present methods using a set of chosen features, *e.g.* TEO energy, spectral centroid, pitch range, *etc.* [4] or psychological stress recognition based on plane shapes, so called vowel polygons, created by relevant formant values [5]. The differences in acoustics-phonetics between stressed and normal speech are described in [6], where enhancing the speaker's pitch is mentioned as the most obvious audible change in stressed speech. Psychological stress recognition is not frequently published and observed, but other used methods, features and databases can be partly found in written surveys oriented on emotional speech recognition [7], [8], or in a survey describing research methods and further steps only for speech under stress [9].

Generally, methods of stress recognition by glottal pulse analysis are less applied, which opens the possibilities to uncover novel observations in this field. For example, Iliev *et al.* used glottal and other features with optimum-path forest to emotion recognition [10] as well as Muthusamy *et al.* [11]. According to [12], vocal tract cavities are affected by psychological stress which can be detected from LPCs. Glottal analysis complemented by other speech features was published in [13], where accuracy increases from 75 % to 92 % approximately after adding the glottal feature. Another study [14], based only on statistical analysis of glottal pulses using the glottal pulses' fixation in maximum and overlaying, presents psychological stress recognition of 88 %. Thus, techniques of emotion recognition based only on analysis of glottal features have not been published and presented.

Compared to previously published methods of psychological stress recognition, the presented paper describes an innovative method based only on glottal pulse analysis in amplitude domain. Exactly, the main novelty of this work is to analyse mined glottal pulses as a two-dimensional shape or *e.g.* probability distribution. The fundamental idea in this paper is also based on the assumption that glottal flow is independent on spoken phonemes, which leads to provide experiments of stress recognition on real stress databases containing different languages to prove if glottal flow analysis, and in particular presented methods, can be successfully applied for psychological stress recognition independently of language and phonetic contents.

Manuscript received 30 November, 2015; accepted 3 May, 2016.
Research described in this paper was financed by Czech Ministry of Education in frame of National Sustainability Program under grant LO1401. For research, infrastructure of SIX Center was used.

II. Material and Methods

This section presents and describes the introductory steps and options necessary for our observation of differences between glottal pulses in stressed and normal speech. The applied methods and glottal pulse estimation are based on using the software tool Aparat [15].

A. Used Methods

Many ways of glottal flow estimation exist, but algorithms based of inverse filtering techniques try to achieve the reliable estimation of glottal source wave. In other words, inverse filtering techniques remove the influence of vocal tract directly from speech. Due to reason of trying to obtain the most realistic glottal flow, estimation techniques based on inverse filtering are used in our experiments. The glottal pulses were estimated from the speech signal by two common algorithms [16] – the Direct Inverse Filtration (DIF) and the Iterative and Adaptive Inverse Filtering (IAIF), both applied on originally captured and normalized (NS) records at a vowels' beginning (VB) and centre parts (CP). Obviously, in our research, eight different methods analysing and estimating glottal pulses were applied. Each method is characterized individually as follows:

- Method 1 uses DIF algorithm, VB
- Method 2 uses DIF algorithm, VB, NS
- Method 3 uses IAIF algorithm, VB
- Method 4 uses IAIF algorithm, VB, NS
- Method 5 uses DIF algorithm, CP
- Method 6 uses DIF algorithm, CP, NS
- Method 7 uses IAIF algorithm, CP
- Method 8 uses IAIF algorithm, CP, NS

According to obvious differences between the used methods, the final comparison should uncover their suitability and efficiency for detecting psychological stress.

B. Used Database

Two different databases were used in the presented experiments. Firstly, 12 male Czech native speakers were randomly selected from the previously created database ExamStress [17] where the same speech is recorded during the final oral exams (stress influence) and a few days later (normal state) for each speaker. Due to this reason, the differences between normal state and real psychological stress can be observed.

Secondly, the SUSAS [18] database was used for validating the psychological stress detection efficiency on the English language and bad quality captured records containing high noise levels, voice distortion and signal clipping. Specifically, the part containing real psychological stress captured by 2 apache pilots almost out of fuel was used in the presented experiments.

C. Used Glottal Pulse Features

Differences between stressed and normal speech were observed and further classified by a vector of three glottal pulse features. In the first step, each mined glottal pulse is amplitude and length normalized to maximum values of 1 due to bringing the global pulse size into accord. Then, each normalized glottal pulse is divided into a series of pulse segments from the peak to n-percentage amplitude level which is shifted step by step along the amplitude axis. The 0 % level is at the top of glottal pulse, and the 100 % value lies at its bottom (see Fig. 1). Due to this fact, the used glottal pulse segmentation is called Top-To-Bottom (TTB).

The selected n-percentage pulse segment is further analysed in the time domain taking into consideration the opening and closing phase. Kurtosis α, skewness β and area γ are calculated for each wave part corresponding to T_{o_p} and T_{c_p}. Then, the Closing-To-Opening phase ratio (*CTO*) is calculated for each obtained parameter

$$CTO_p(n) = \frac{T_{c\_p}(n)}{T_{o\_p}(n)}, \tag{1}$$

where n is the actual n-percentage level, p substitutes one of the analysed parameters (*i.e.* skewness, kurtosis, area, respectively), T_{c_p} is the current closing phase value and T_{o_p} is the current opening phase value. Figure 2 shows the most illustrative example of glottal pulses in /u/ vowels' beginning estimated by the DIF algorithm for both states of one speaker. Here, given values of *CTO*s were averaged from 39 pulses.

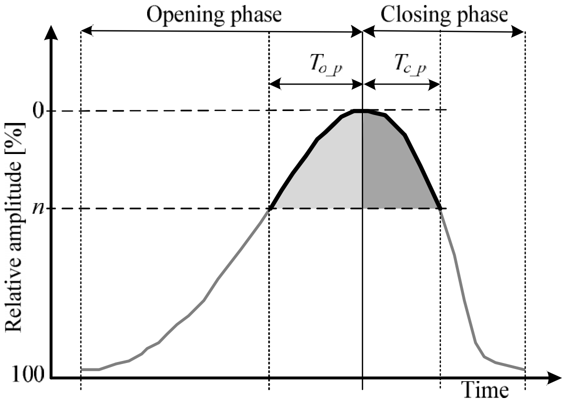

Fig. 1. An example of glottal pulse n-percentage division where the selected closing phase is illustrated by dark grey and the opening phase is represented by a light grey colour. The thick line marks the chosen curve part of the glottal pulse.

A similar *CTO* has been applied in previous research oriented on percentage segmentation of glottal pulses along the time axis [19]. In this case, Gaussian Mixture Models (GMM) were evaluated as the most appropriate feature processing approach under six different classifiers.

III. Experimental Results

This section describes realized experiments and achieved results. In the training process, 5 Czech vowels were spoken few times separately and automatically found [20] in recorded speech from 6 speakers in the ExamStress database to obtain reference *CTO* values in the vowels' beginning and center part by its averaging for all vowels. This fact means that for each speaker and each vowel there are 3 reference *CTO* values stored, leading to 90 reference values totally for each state of speaker and used method. These reference values are further used for training the GMM classifier. In the presented experiments, the GMM classifier standardly embedded in the Matlab environment is used and further is fitted on previously described reference values as a two-component Gaussian Mixture Model. Then, the binary decision (stress/no stress) is based on the higher probability reached for each state of fitted GMM to the investigated data. The flow chart of the algorithm used in our experiments is illustrated in Fig. 3.

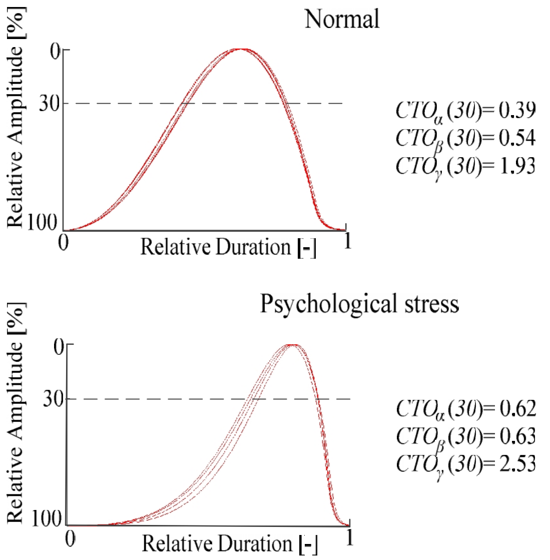

Fig. 2. An example of pulse differences varying on the speaker's state in glottal pulses estimated by DIF in /u/vowels' beginning for speaker 1 from the ExamStress database and 30 % selected interval with average *CTO*s.

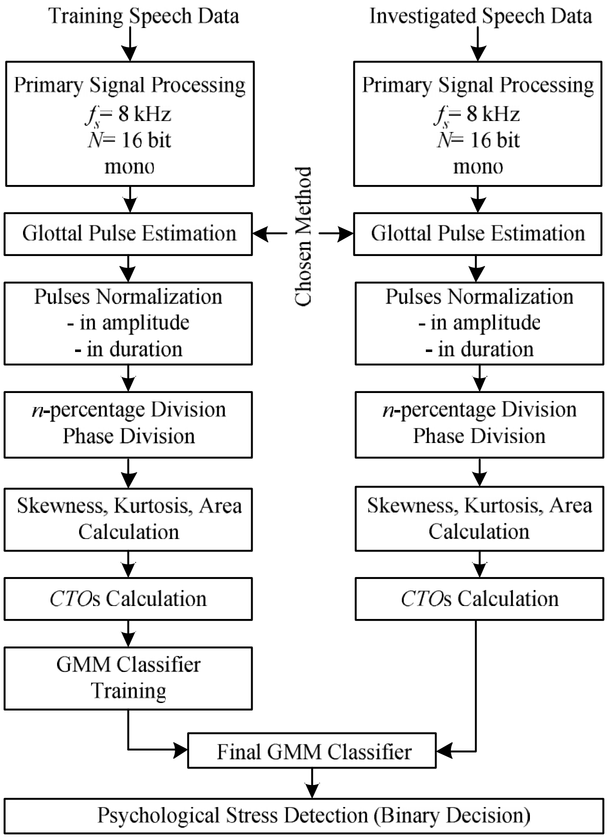

Fig. 3. The flow chart of the used psychological stress recognition algorithm.

In our experiments, the investigated data, *i.e.* observed glottal pulses, were automatically extracted from fluent speech, exactly over all spoken phonemes, by using the flowing rectangular window of duration 300 ms with 50 % overlapping. Then, the estimated glottal pulses were normalized in two dimensions and were filtered for removing the parasitic pulses.

Investigated speech data was automatically achieved for all voiced parts of speech. The second group of 6 ExamStress speakers was used for testing the designed classifier. Due to the records having high quality and their length, approximately 1500 glottal pulses were analysed and classified for each speaker.

For investigating the language-independency of the presented methods, the SUSAS database was used. Compared to the ExamStress database, the low quality (*e.g.* voice distortion, clipping, loud background noise, *etc.*) of these records rapidly decreases the total number of estimated glottal pulses. It has been observed experimentally that processing only short parts (50 ms) of SUSAS records leads to satisfactory glottal pulse mining. Other lengths of analysed speech signals lead to estimating glottal pulses which do not match the Liljencrants-Fant model [21]. All mined glottal pulses were filtered automatically, because incorrectly estimated glottal pulses occurred even for an analysed signal with short lengths. For each speaker in SUSAS, approximately 130 glottal pulses were received correctly and further used irrespectively of sound normalization and the vowels' parts for psychological stress detection.

The reached efficiency results using Method 1 and Method 2 are listed in Table I, where a few facts are evident. Sound normalization causes a decrease of psychological stress detection applied on the ExamStress database, but generally achieved efficiency is high and more than satisfactory. Contrary to previous statements, results achieved for the SUSAS database are high and more or less constant over the entire chosen *n*-percentage intervals for both methods which leads to much higher efficiency achieved than by using Method 2. These observations can lead to the statement that low quality records are less prone to sound normalization of testing sequences.

Table II shows the efficiency obtained by psychological stress detection based on the IAIF estimation algorithm and vowels' beginning (Method 3 and Method 4).

By comparing the results reached using the ExamStress database, the negative influence of sound normalization can be seen by a significant decrease of efficiency over all the observed *n*-percentage intervals. This effect is not that evident for the SUSAS database where almost all efficiencies are lower than its ExamStress equivalent (except the 80 % and 100 % level for Method 3 as well as 15 % and 55 % level for Method 4, achieving a stress detection efficiency of 95 %).

Obviously, psychological stress detection based on the IAIF estimation algorithm applied on the vowels' beginning is not appropriate on low quality records.

Further, the recognition efficiency was calculated for methods based on the vowels' centre part.

For DIF based methods (Method 5 and Method 6) and the ExamStress database, efficiency is more or less similar (over 90 %). However, for n percentage intervals higher than 50 %, efficiency slightly decreases to a value of 77 %. By applying Method 5 and Method 6 on the SUSAS database, similar efficiency is reached as for the ExamStress database and achieves high values almost over all the *n*-percentage intervals. Some exceptions can be found in the 45 %, 65 % and 80 % intervals, where both methods obtained poor and unsatisfactory efficiency.

TABLE I. THE EFFICIENCY OF PSYCHOLOGICAL STRESS DETECTION REACHED BY METHOD 1 AND METHOD 2.

***n* [%]**	***Efficiency* [%]**			
	Method 1		**Method 2**	
	ExamStress	**SUSAS**	**ExamStress**	**SUSAS**
5	94.8	95.0	85.8	95.0
10	95.0	95.0	27.0	95.0
15	94.1	95.0	95.0	95.0
20	95.0	95.0	49.9	95.0
25	94.8	94.8	34.6	94.6
30	94.3	86.7	85.9	94.1
35	94.0	94.6	84.6	95.0
40	85.8	95.0	58.3	94.8
45	92.5	95.0	82.8	95.0
50	79.6	94.7	82.4	94.7
55	84.5	95.0	49.5	94.7
60	90.3	94.9	82.9	95.0
65	83.9	90.2	70.8	94.5
70	83.4	95.0	83.2	95.0
75	83.1	94.9	95.0	95.0
80	82.8	90.6	82.5	92.2
85	82.4	85.6	80.9	72.2
90	81.6	95.0	80.1	90.6
95	82.8	82.8	80.1	51.1
100	82.4	94.7	79.4	82.8

According to the made observations, the DIF glottal pulse estimation algorithm has been found to also be appropriate for psychological detection. The effect of sound normalization on stress recognition can also be classified as minimal as well as the effect of low quality records and spoken language captured on analysed records.

TABLE II. STRESS DETECTION EFFICIENCY REACHED BY METHOD 3 AND METHOD 4.

***n* [%]**	***Efficiency* [%]**			
	Method 1		**Method 2**	
	ExamStress	**SUSAS**	**ExamStress**	**SUSAS**
5	75.0	63.9	94.8	64.6
10	92.5	60.5	94.7	60.5
15	95.0	63.3	89.4	95.0
20	93.9	59.2	94.7	51.0
25	93.1	65.9	74.6	68.7
30	91.9	67.3	74.2	46.3
35	91.4	52.4	74.0	63.9
40	40.7	66.7	73.6	59.2
45	90.6	65.9	73.4	68.0
50	90.2	68.7	72.8	68.7
55	89.3	69.4	27.3	95.0
60	88.4	58.5	71.6	68.0
65	40.9	70.7	70.1	59.2
70	86.1	74.1	94.5	72.1
75	94.8	76.9	69.1	75.5
80	95.0	95.0	94.6	74.8
85	95.0	79.6	67.6	80.9
90	81.9	53.6	50.9	78.9
95	81.6	82.3	81.3	52.4
100	94.7	95.0	83.9	83.7

Psychological stress detection efficiency results obtained by the IAIF estimation based on the vowels' centre part are described in the following text and are equivalent to the previously mentioned Table II.

As in previous cases (see Table I and Table II) for the ExamStress database, the IAIF estimation algorithm reaches generally lower recognition efficiency than the DIF algorithm, but it still gives satisfactory results almost on all *n*-percentage intervals. Applying Method 7 and Method 8 on the SUSAS database, the recognition efficiency generally sharply decreases by 20 % on average, except for 4 *n*-percentage intervals where it reaches much higher values than for the ExamStress database.

Apparently, as in the previous case, the IAIF algorithm is sensitive to analysed records quality, and in some cases also on spoken language. Methods based on the IAIF algorithm applied on the vowels' beginning are more suitable for psychological stress detection than Method 7 and Method 8.

IV. EFFICIENCY EVALUATION

To summarize the results listed in the previous section, it is necessary to make a final evaluation of the used methods and *n*-percentage intervals. Firstly, evaluating all investigated *n*-percentage intervals is appropriate for finding the most consecutive glottal pulse parts where the highest differences between normal and stressed speech occur. Table III lists average efficiency values ε reached for each *n*-percentage interval for all used methods and databases.

TABLE III. AVERAGE EFFICIENCY VALUE FOR ALL 8 METHODS AND BOTH DATABASES.

***n* [%]**	**5**	**10**	**15**	**20**	**25**	**30**	**35**
***ε* [%]**	83.2	81.8	84.4	77.5	78.4	81.9	82.6
***n* [%]**	**40**	**45**	**50**	**55**	**60**	**65**	**70**
***ε* [%]**	77.6	74.2	80.9	80.8	83.2	69.1	82.5
***n* [%]**	**75**	**80**	**85**	**90**	**95**	**100**	
***ε* [%]**	85.1	78.4	79.1	73.5	74.4	80.9	

Obviously, the band of the best *n*-percentage TTB amplitude intervals lies between 5 % and 40 % where average efficiency ε reaches consequently higher values than 77.5 %. Table IV lists average efficiency ε values for each used method for both databases and all *n*-percentage intervals in the range from 5 to 40 % with a step of 5 %. As can be seen, the efficiency value depends on used methods, exactly on different vowel parts performed for training the classifier.

TABLE IV. AVERAGE EFFICIENCY VALUE FOR ALL 8 METHODS AND BOTH DATABASES.

	Method				**Σ DIF**
	1	**2**	**5**	**6**	
***ε* [%]**	93.7	80.0	91.9	88.6	88.5
	86.8		90.2		
	Method				**Σ IAIF**
	3	**4**	**7**	**8**	
***ε* [%]**	73.3	73.7	69.4	77.0	73.3
	73.5		73.2		

Results listed in Table IV show that not so significant positive impact exists on reached ε over *n*-percentage intervals in the case of sound normalization. Obviously, similar ε results are achieved by similar glottal pulse estimation methods trained only on a varying vowel part. Finally, the highest average efficiency on the observed *n*-percentage intervals are reached by using the DIF estimation method (88.5 %) which achieved higher ε by a significant 15.2 % compared to the IAIF estimation algorithm (73.3 %).

V. Conclusions

According to all achieved results, it can be concluded that the DIF based methods give better stress detection, glottal pulse normalization is sensitive to the sound quality, and the vowel's part used for classifier training does not have a significant effect on recognition efficiency. The usage of the presented algorithms of glottal pulse processing estimated by DIF and applied on TTB *n*-percentage intervals from 5 % to 40 % can lead to high efficient psychological stress recognition in speech. Obviously by achieved high values of recognition efficiency (in some cases approaching 95 %), the presented technique could be classified as possibly text and language independent which can lead to further analysis of glottal flow in more detail to deploy it into real applications.

Nevertheless, in future work, it is necessary to verify the achieved results on other languages and expand the speaker database.

References

[1] S. E. Bou-Ghazale, “A comparative study of traditional and newly proposed features for recognition of speech under stress”, *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 4, pp. 429–442, 2000. [Online]. Available: http://dx.doi.org/ 10.1109/89.848224

[2] M. Sigmund, “Statistical analysis of fundamental frequency based features in speech under stress”, *Information Technology and Control*, vol. 42, no. 3, pp. 286–291, 2013. [Online]. Available: http://dx.doi.org/10.5755/j01.itc.42.3.3895

[3] D. Gharavian, M. Sheikhan, F. Ahoftedel, “Emotion recognition improvement using normalized formant supplementary features by hybrid of DTW-MLP-GMM model”, *Neural Computing & Applications*, vol. 22, no. 6, pp. 1181–1191, 2013. [Online]. Available: http://dx.doi.org/10.1007/s00521-012-0884-7

[4] H. Lu, D. Frauendorfer, M. Rabbi, M. Schmid Mast, G. T. Chittaranjan, A. T. Campbell, D. Gatica-Perez, T. Choudhury, “StressSense: detecting stress in constrained acoustic environments using smartphones”, in *Proc. UbiComp ‘12 –ACM Conf. Ubiquitous Computing.* New York, 2012, pp. 351–360. [Online]. Available: http://dx.doi.org/10.1145/2370216.2370270

[5] M. Stanek, M. Sigmund, “Finding the most uniform changes in vowel polygon caused by psychological stress”, *Radioengineering*, vol. 24, no. 2, pp. 604–609, 2015. [Online]. Available: http://dx.doi.org/ 10.13164/re.2015.0604

[6] M. Jessen, *Einfluss von Stress auf Sprache und Stimme*. Schulz-Kirchner: Idstein, 2006. (in German)

[7] M. El Ayadi, M. S. Kamel, F. Karray, “Survey on speech emotion recognition: Features, classification schemes, and databases”, *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011. [Online]. Available: http://dx.doi.org/10.1016/j.patcog.2010.09.020

[8] D. Ververidis, C. Kotropoulos, “Emotional speech recognition: Resources, features, and methods”, *Speech Communication*, vol. 48, no. 9, pp. 1162–1181, 2006. [Online]. Available: http://dx.doi.org/10.1016/j.specom.2006.04.003

[9] C. Kirchhubel, D. M. Howard, A. W. Stedmon, “Acoustic correlates of speech when under stress: Research, methods and future directions”, *Int. Journal of Speech, Language and the Law*, vol. 18, no. 1, pp. 75–98, 2011. [Online]. Available: http://dx.doi.org/10.1558/ijsll.v18i1.75

[10] A. I. Iliev, M. S. Scordilis, J. P. Papa, A. X. Falcao, “Spoken emotion recognition through optimum-path forest classification using glottal features”, *Computer Speech & Language*, vol. 24, no. 3, pp. 445–460, 2010. [Online]. Available: http://dx.doi.org/10.1016/j.csl.2009.02.005

[11] H. Muthusamy, K. Polat, S. Yaacob, “Improved emotion recognition using Gaussian Mixture Model and extreme learning machine in speech and glottal signals”, *Mathematical Problems in Engineering*, 2015. [Online]. Available: http://dx.doi.org/10.1155/2015/394083

[12] P. K. Mongia, R. K. Sharma, “Estimation and statistical analysis of human voice parameters to investigate the influence of psychological stress and to determine the vocal tract transfer function of an individual”, *Journal of Computer Networks and Communications*, 2014. [Online]. Available: http://dx.doi.org/ 10.1155/2014/290147

[13] D. Fisher, K. Chang, J. Canny, “A speech analysis library for analyzing affect, stress, and mental health on mobile phones”, in *Proc. PhoneSense*, 2011.

[14] M. Sigmund, A. Prokes, Z. Brabec, “Statistical analysis of glottal pulses in speech under psychological stress”, in *Proc. EUROSIPCO – European Signal Proc. Conf.*, Lausanne, 2008, pp. 1–5.

[15] M. Airas, “TKK Aparat: an environment for voice inverse filtering and parameterization”, *Logopedics, phoniatrics, vocology*, vol. 33, no. 1, pp. 49–68, 2008. [Online]. Available: http://dx.doi.org/ 10.1080/14015430701855333

[16] P. Alku, “Glottal wave analysis with pitch synchronous Iterative Adaptive Inverse Filtering”, *Speech Communication*, vol. 11, no. 2–3, pp. 109–118, 1992. [Online]. Available: http://dx.doi.org/ 10.1016/0167-6393(92)90005-R

[17] M. Sigmund, “Introducing the database ExamStress for speech under stress”, in *Proc. 7th Nordic Signal Processing Symposium,* Reykjavik, 2006, pp. 290–293. [Online]. Available: http://dx.doi.org/ 10.1109/NORSIG.2006.275258

[18] J. H. L. Hansen, S. E. Bou-Gazale, “Getting started with SUSAS: A speech under simulated and actual stress database”, in *Proc. EUROSPEECH ’97–European Conf. Speech Communication Technology*, Rhodes, 1997, pp. 1743–1746.

[19] M. Stanek, M. Sigmund, “Psychological stress detection in speech using return-to-opening-phase ratios in glottis”, *Elektronika ir Elektrotechnika*, vol. 21, no. 5, pp. 59–63, 2015. [Online]. Available: http://dx.doi.org/ 10.5755/ j01.eee.21.5.13336

[20] M. Stanek, L. Polak, “Algorithms for vowel recognition in fluent speech based on formant positions”, in *Proc. 36th Int. Conf. Telecommunications and Signal Processing*, Rome, 2013, pp. 521–525. [Online]. Available: http://dx.doi.org/10.1109/ tsp.2013.6613987

[21] G. Fant, J. Liljencrants, Q. Lin, “A four-parameter model of glottal flow”, *Speech Transmission Laboratory Quarterly Progress Report 4/85*, 1985, pp. 1–3.