16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

STATISTICAL ANALYSIS OF GLOTTAL PULSES IN SPEECH UNDER PSYCHOLOGICAL STRESS

Milan Sigmund[+], Ales Prokes[+], and Zdenek Brabec[°]

[+]Institute of Radio Electronics, BUT Brno
Purkynova 118, CZ-61200 Brno, Czech Republic

[°]Department of Telecommunication Engineering, CTU Prague
Technicka 2, CZ-16627 Prague, Czech Republic

phone: + (420) 541149153, fax: + (420) 541149244, email: sigmund@feec.vutbr.cz
web: www.urel.feec.vutbr.cz

ABSTRACT

In this paper, the problem of speech signal under psychological stress is addressed. The investigation into the speaker's stress is based on statistical analysis of glottal pulse derivative extracted from the vowel signals. A pitch synchronous selection of segments from the glottal pulse waveform is used. Selected segments are fixed in their maxima and overlaid. The generated distribution matrix is analysed using special cuts. A new database of speech under stress is created for use in our experiments consisting of data collected during oral final examinations at our university. The database contains read and conversational speech of 31 male speakers, both in neutral and in stressed state. The stress recognition rate in the speaker dependent binomial classification (stress/no stress) reaches 88%.

1. INTRODUCTION

The emotional state of a speaker can be identified from the facial expression, speech, perhaps brainwaves, and other biological features of the speaker. A significant part of information contained in a speech signal refers to the speaker. These phonologically-linguistically irrelevant speaker characteristics make speech recognition less effective but can be used for speaker recognition and analysis of the speaker's emotional and health state. Such a speech cue would allow an analysis without the physical presence of the speaker.

In this paper, we focus on actual effects of stress on speech signal. Stress is a psycho-physiological state characterized by subjective strain, dysfunctional physiological activity and deterioration of performance. Psychological stress has a broad sense and a narrow sense effect. The broad sense reflects the underlying long-term stress and the narrow sense refers to the short-term excitation of the mind that prompts people to act. In automatic recognition of stress, a machine would not distinguish whether the emotional state is due to long-term or short-term effect so well as it is reflected in facial expression. Stress is more or less present in all professions in today's hectic and fast-moving society. The negative influence of stress on health, professional performance as well as interpersonal communication is well known. A comprehensive reference source on stressors, effects of activating the stress response mechanisms, and the disorders that may arise as a consequence of acute or chronic stress is provided, for example, in the Encyclopedia of Stress [1].

1.1 Speech under stress

Stress may be induced by external factors (noise, vibration, etc.) and by internal factors (emotion, fatigue, etc.). Physiological consequences of stress are, among other things, changes in the heart rate, respiration, muscular tension, etc. The muscular tension of vocal cords and vocal tract may, directly or indirectly, have an adverse effect on the quality of speech. The entire process is extremely complex, and is shown in a simplified model in Fig. 1. The accepted term for the speech signal carrying information on the speaker's physiological stress is "stressed speech".

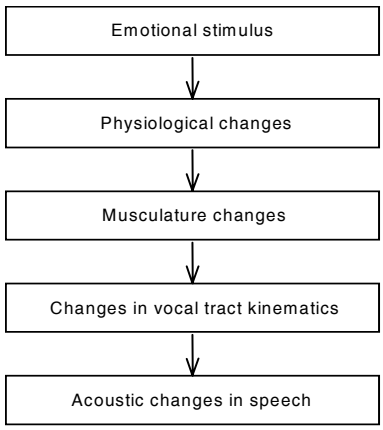

Figure 1 – Model of how emotion causes changes in speech.

The most widely used methods for investigating acoustic indicators of stress in speech usually start from pitch, duration, formant frequencies [2], [3], and spectral variation [4]. A specific feature derived from teager energy operator was proposed and found to be more responsive to speech under stress in [5]. However, use of these features with an HMM trained stressed speech classifier provides high error rates of 25% to 13% for some types of stress and neutral speech detection. In this study, a new technique of stress detection based on the evaluation of glottal pulse shape is described. While examining stress we are only concerned with the physically measurable characteristics of the speech signal. Besides these changes in the spoken language the content of the language, e.g. repetition of selected words, structure of the sentence, etc. is also very important for psychologists.

2. DATABASES OF STRESSED SPEECH

There are not many corpora designed to allow the study of speech under stress. It is extraordinarily difficult to obtain realistic voice samples of speakers in various states of stress, recorded in realistic situations. "Normal" people (as well as professional actors) cannot simulate real case stress perfectly with their voices. However, there are some approaches how to simulate stressful events, for instance by using vocal noises, quick question-answer-quizzes with the possibility of winning a prize, negotiation concerning an important contract, etc.

2.1 Existing databases

A typical corpus of extremely stressed speech from a real case is extracted from the cockpit voice recorder of a crashed aircraft. Such speech signals together with other corresponding biological factors are collected for example in the NATO corpus SUSC-0 [6]. The advantage of this database is that an objective measure of workload was obtained, and that physiological stress measures (heart rate, blood pressure, respiration, and transcutaneous pCO_2) were recorded simultaneously with the speech signal. However, such extreme situations as crashed aircraft occur seldom in everyday life. The most frequently mentioned corpus in the literature is the SUSAS (Speech Under Simulated and Actual Stress) database of stressed American English described in [7] and distributed by Linguistic Data Consortium at the University of Pennsylvania. For the French speech, the Geneva Emotion Research Group at the University of Geneva conducts research into many aspects of emotions including stress, and it also collected emotion databases. Their website provides access to a number of databases and research materials [8]. The German database of emotional utterances including panic was recorded at the Technical University of Berlin. A complete description of the database called Berlin Database of Emotional Speech can be found in [9]. A list of existing emotional speech data collections including all available information about the databases such as the kinds of emotions, the language, etc. was provided in [10]. Most of the studies reported in the literature concern English. For Slavic languages, no research in stressed speech is known and no appropriate database is available.

2.2 Creating the ExamStress database

For our studies conducted within research into speech signals we created and used our own database. The most suitable event with realistic stress took place during the final state examinations at Brno University of Technology held in oral form in front of a board of examiners. The test persons were 31 male pre-graduate and post-graduate students, mostly Czech native speakers. The created database called ExamStress consists of two kinds of stressed speech material: Defence and Pre-Defence. The speech data in the Defence part were collected from spontaneously spoken utterances during the state exams. The speech data in the Pre-Defence part are read speech data recorded ca. 10 minutes before the state exam started. All speakers were asked to read the same text of an approximate length of 1 minute. A few days later, the same speakers read this text again and the speech signal recorded in speaker's neutral state of mind was added to the database. Finally, the recorded utterances were manually preprocessed and evaluated. The main features of the database are summarized in Table 1. A complete description of the ExamStress database can be found in [11].

Table 1. Description of the ExamStress database.

Database part	Defence	Pre-Defence
Stressor	Exam nerves, fatigue	Exam nerves, fatigue
Language	Czech	Czech
Speech type	Spontaneous, read sentences	Read sentences
Total length	180 minutes	30 minutes
Speakers	31	19
Gender	Male	Male
Occupation	Student	Student
Microphone	C 417	AKG
Quantization	16-bit linear	16-bit linear
Sampling rate	22 kHz	22 kHz
Quality	Good	Very good

In some cases the heart rate *HR* of students was measured simultaneously with the speech recordings in both stressed and neutral state. A comparison of these measured data proves the influence of exam nerves on the speaker's emotional state. The oral examination seems to be a reliable stressor. Figure 2 shows a typical run of the *HR* curve in the first 900 seconds after the start of the examination. As expected, the *HR* is the highest at the beginning and then slowly decreases. On average, the *HR* values obtained for stressed state were almost doubled compared to the neutral state (such values usually occur if a person is under medium physical activity). Surprising is the rapid short-time increase at the beginning of recordings in neutral state. Probably, this represents an initial effect of "stress due to the attached measuring equipment", which is similar to the well-known "stress from physicians".

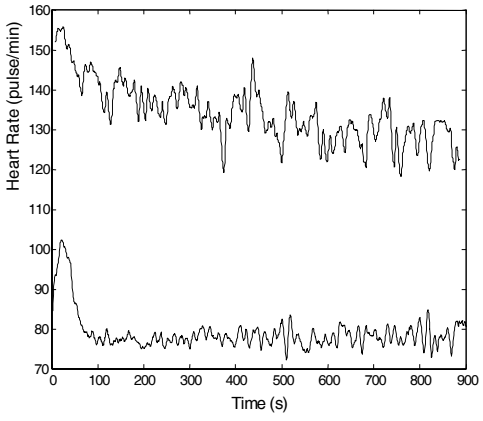

Figure 2 - Typical heart rate measured during the exam (upper curve) and in neutral state (lower curve).

3. GLOTTAL EXCITATION

Glottal source estimation has a great potential for use in identifying emotional states, non-invasive diagnosis of voice disorders, etc. Voiced speech is typically modeled as the output of a linear and time-invariant filtering process. A quasi-periodic glottal flow signal at the glottis, denoted $g(n)$, acts as an acoustic source and excites the vocal tract filter of impulse response $h(n)$. The output speech pressure waveform measured in front of the lips can be expressed in the time domain by

$$s(n) = g(n) * h(n), \qquad (1)$$

where * denotes convolution.

3.1 Estimation of glottal pulses

In our experiments, glottal pulses were obtained from speech by applying the IAIF (Iterative Adaptive Inverse Filtering) algorithm, which is one of the most effective techniques for extracting excitation from a speech signal [12]. Other techniques for obtaining glottal pulses can be found, for example, in [13]. The block diagram of the IAIF is shown in Fig. 3. This method operates in two repetitions, hence the word iterative in the name of the method. The first phase (blocks LPC 1st order, filter $H_1^{-1}(z)$, LPC 12th order, filter $H_2^{-1}(z)$) generates an estimate of glottal excitation, which is subsequently used as input of the second phase (blocks LPC 4th order, filter $H_3^{-1}(z)$, LPC 12th order, filter $H_4^{-1}(z)$) to achieve a more accurate estimate. The steps of the method are described in detail below. Firstly, the input speech signal is analyzed by first-order LPC predictor. This step gives an initial estimate of the effect of glottal flow on the speech spectrum. Using the obtained filter $H_1^{-1}(z)$ of 1st order, the input signal is inversely filtered. This step effectively removes the spectral tilt caused by the spectrum of the excitation signal. The output of the previous step is analyzed by the LPC predictor of 12th order to obtain a model of the vocal tract transfer function. The order of the LPC analysis is related to the number of formants to be modelled. The input signal is then inversely filtered by filter $H_2^{-1}(z)$ using the inverse of the 12th order model from the previous step. This yields the first estimate of the glottal pulse derivative and completes the first repetition. The second repetition runs analogously.

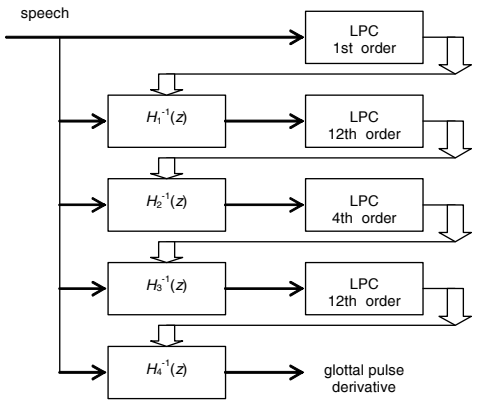

Figure 3 - Block diagram of IAIF.

Figure 4 shows a typical waveform $s(n)$ of the vowel "a" and its corresponding glottal pulse derivative $v(n)$ estimated using IAIF. In order to minimize the influence of voice intensity (i.e. loud vs. soft voice), the amplitude normalization was used before applying the IAIF-procedure. For the analysis, a pitch synchronous selection of segments from the obtained glottal pulse waveform was used. A position determining the special phase of the glottis (circles in Fig. 4) such as the maximum and the minimum of the glottal pulse derivative waveform was marked for every segment. The waveform was multiplied by rectangular window of one fundamental period in length. Selected segments were fixed in one of the two phases and overlaid.

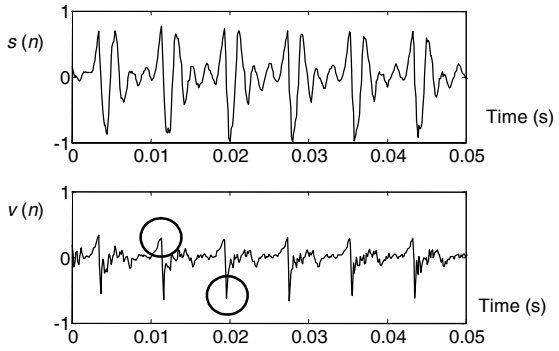

Figure 4 - Example of a speech signal (upper graph) and the corresponding glottal pulse derivative (lower graph).

3.2 Statistical description of glottal pulses

Based on the graphical interpretation, a two-dimensional distribution matrix was generated as shown in Fig. 5. The amplitude-time space is divided into small elements via horizontal and vertical lines (180 intervals on the time axis, 100 intervals on the amplitude axis). The number of glottal pulse waveform lines going through a cell is equal to the numerical value of the corresponding element of the distribution matrix 100x180.

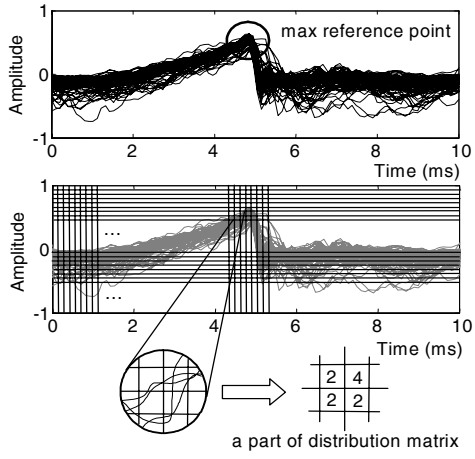

Figure 5 - Generating the distribution matrix of glottal pulses derivative.

The distribution matrix obtained was displayed as a grey scale image where the maximum and minimum values of the matrix are black and white. An example of such an image created from about 4000 segments can be seen in Fig. 6. In this case, the fixation point for all segments was in each period the maximum of the glottal pulse derivative waveform (upper circle in Fig. 4).

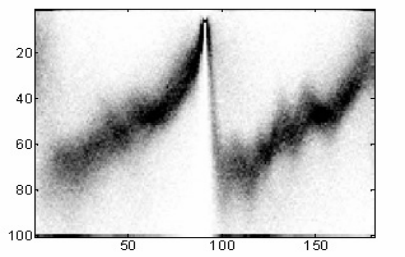

Figure 6 - Illustration of a distribution matrix of glottal pulses derivative waveform interpreted in grey scale.

4. EXPERIMENTS AND RESULTS

The ultimate goal in our experiments was to find common speech characteristics of stressed speech based on distribution matrices of glottal pulses. In order to compare the distribution matrices automatically with each other, it is inevitable to find a useful description (a few significant features) of the matrices. An effective criterion seems to be straight cuts made at a reference position. Figure 8 shows the positions of applied cuts and the form of the intersection for two speakers in both neutral and stressed state. For the stressed state, the distribution matrix seems to be "blacker" than for the neutral state; it means that if the speaker is under stress, the derivative waveforms of the glottal pulses produced are more concentrated about the average waveform and the distribution form in the cuts is more asymmetric; in the cut the mass of the distribution is concentrated on the right of the figure. These effects are obvious in almost any speaker. Another type of cuts provides less useful information [14].

An experiment with mathematical description of applied cuts resulted in the use of two effective parameters: □ and *k*. The first parameter, □, is defined by

$$\alpha = \frac{S_1}{S_2 + S_3}, \tag{2}$$

where S_1, S_2, and S_3 are the sub-areas of the cut located symmetrically to the maximum of the cut and bounded graphically by lines in 20 %, 40 %, 60 %, and 80 % of the total width of the cut, as illustrated in Fig. 7. The second parameter, *k,* is defined as

$$k = \frac{\mu_4}{\sigma^4} - 3, \tag{3}$$

where □4 is the fourth central moment and □ is the standard deviation. For our experiment, 31 male speakers from the newly created ExamStress database were used (defence part recordings only due to higher expected stress). Approximately 2000 voiced segments of 5 vowels were extracted from the speech data of each speaker for each state. The IAIF algorithm was applied to those segments to estimate

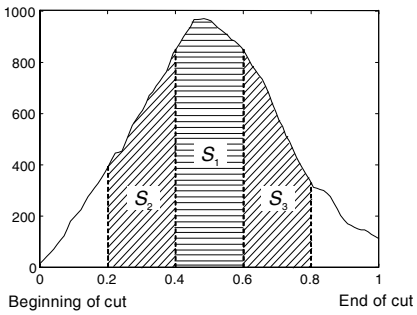

Figure 7 - Definitions of the sub-areas S_1, S_2, and S_3 in a distribution matrix cut.

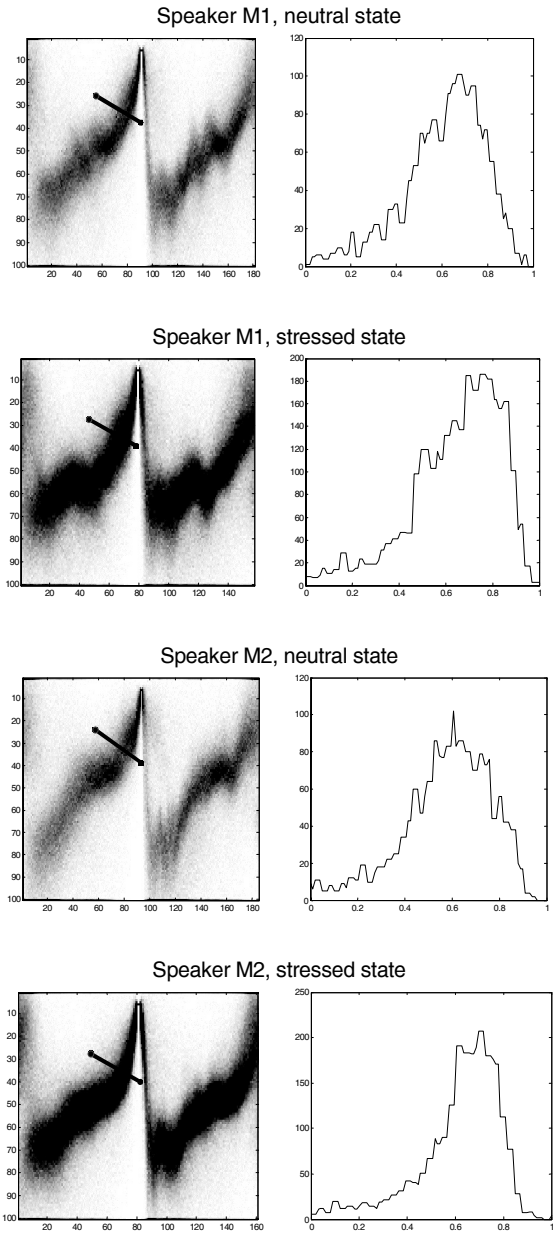

Figure 8 - Graphical samples of distribution matrices and their comparative cuts estimated for the vowel "a".

the glottal pulse waveforms. Two distribution matrices for both neutral and stressed state were calculated for each speaker and the speaker's state was estimated using the distribution parameters of cuts in a binomial classification (stress/no stress). The classification was performed by using the Mahalanobis distance measure. Table 2 shows samples of parameters ◘ and *k* computed from the speech signals of vowel phonemes for a group of ten speakers in neutral state (denoted N) and stressed state (denoted S). In the stressed state, slightly higher values of both applied parameters are indicated in most cases. The stress recognition rate in the speaker dependent recognition achieved 88%. In the speaker independent experiments without neutral reference speech data the recognition rate decreased to 72%.

Table 2. Values of applied parameters for neutral (N) and stressed (S) speech.

Speaker	Parameter α		Parameter *k*	
	N	S	N	S
M1	1.09	1.20	2.91	3.21
M2	0.85	1.92	2.66	3.50
M3	1.32	1.17	3.04	3.31
M4	0.93	1.18	2.88	2.96
M5	0.78	0.88	2.12	2.86
M6	0.76	1.40	2.45	3.30
M7	0.55	0.83	2.30	2.63
M8	1.03	1.40	2.68	3.19
M9	0.91	1.05	2.82	2.90
M10	0.87	1.32	2.71	3.25

5. CONCLUSIONS

There are broad categories of human performance which can be affected by stress, for instance psychomotor skills, memory function, situation awareness, attention, etc. This paper presents the speech signal as a possible indicator of stress. To analyse speakers, psychologically suitable databases with specific realistic speech signals are necessary. The newly developed Czech database ExamStress of speech under realistic stressed conditions was introduced. The presence of stress is proved objectively by the speakers' increased heart rate, which was measured simultaneously with the speech. Using the collected corpus some factors were studied in trying to characterize the changes in stressed speech. The results of the experiments show that an average accuracy of 88% can be achieved in speaker-dependent stress recognition. The results are better than the 75% accuracy achieved by human assessment. The high recognition rate indicates that glottal pulses, when analysed statistically, seem to be an effective means of expressing stress. It should be noted, however, that the reliability of a recognizer of speaker's state is strongly dependent on the speech data used (stressor effects, real situation during acquiring speech data, age of speakers, etc.).

The encouraging initial results should be taken as a starting point for future research. A natural next step is the development of algorithms for automatic detection and quantification of stress. To clarify the general significance of the results, it will be necessary to test more speakers (including female speakers) in connection with various stressful stimuli. The motivation for research into methods for detecting and measuring stress in speech signals is an application for monitoring highly emotional calls in an emergency telephone system. Such a system could also be used for monitoring aircraft voice communication.

REFERENCES

[1] G. Fink, *Encyclopedia of Stress*. London, New York: Academic Press, 2007.

[2] J. H. Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech Communication*, vol. 20(2), pp. 151–170, November 1996.

[3] M. Sigmund and T. Dostal, "Analysis of emotional stress in speech," in *Proc. Internat. Conf. on Artificial Intelligence and Applications – AIA 2004*, Innsbruck, Austria, February 16-18. 2004, pp. 317–322.

[4] R. Sarikaya, and J. N. Gowdy, "Subband based classification of speech under stress," in *Proc. ICASSP'98*, Seattle, USA, May 12-15. 1998, vol, I pp. 569–572.

[5] G. Zhou, J. H. Hansen, and J. F. Kaiser, "Nonlinear feature based classification of speech under stress," *IEEE Transactions on Speech & Audio Processing*, vol. 9(2), pp. 201–216, March 2001.

[6] D. Haddad, S. Walter, R. Ratley, and M. Smith, *Investigation and Evaluation of Voice Stress Analysis Technology*. Project Report: Rome Laboratory, NY, USA, 2002.

[7] J. H. Hansen and S. E. Ghazale, "Getting started with SUSAS," in *Proc. Eurospeech'97*, Rhodes, Greece, September 22-25. 1997, pp. 1743–1746.

[8] http://www.unige.ch/fapse/emotion

[9] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. Eurospeech'05*, Lisbon, Portugal, September 4-8. 2005, pp. 1517–1520.

[10] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Communication,* vol. 48(9), pp. 1162–1181, September 2006.

[11] M. Sigmund, "Introducing the database ExamStress for speech under stress," in *Proc. 7th IEEE Nordic Signal Processing Symposium - NORSIG 2006*, Reykjavik, Iceland, June 7-9. 2006, pp. 290–293.

[12] P. Alku, "An automatic method to estimate the time-based parameters of the glottal pulseform," in *Proc. ICASSP'92*, San Francisco, USA, March 23-26. 1992, pp. 29–32.

[13] M. Bostik and M. Sigmund, "Methods for estimation of glottal pulses waveforms exciting voiced speech," in *Proc. Eurospeech'03*, Geneva, Switzerland, September 1-4. 2003, pp. 2389–2392.

[14] M. Bostik, *Voice Analysis for Stress Recognition*. Doctoral thesis, BUT Brno, 2005.