Influence of Psychological Stress on Formant Structure of Vowels

M. Sigmund[1]
[1]*Brno University of Technology, Faculty of Electrical Engineering and Communication,*
Purkynova 118, 61200 Brno, Czech Republic, phone: +420 541 149 153
sigmund@feec.vutbr.cz

***Abstract*—This paper deals with speech signal under psychological stress. The investigation into the speaker's stress is based on statistical analysis of smooth vowel spectra. In total, fifteen features based on formants, antiformants and inflexion points of spectral envelope were analysed and evaluated. The most effective features for stress detection across all speakers were as follows: frequency of the second formant of /a/, frequency of the first formant of /i/, and lower part of bandwidth of the third formant of /u/ and /a/. The results indicate that speakers tend to follow an individual trend and speaker dependent stress recognition might be more appropriate.**

***Index Terms*—Speech processing, spectral analysis, frequency estimation, speech analysis.**

I. INTRODUCTION

Stress becomes the most spread diagnosis of the 21st century. This is characterizing especially for the today's modern "west society". Stress is generally defined as a psychological state that is a response to a perceived threat or task demand and is usually accompanied by some specific emotions (e.g., fear, anger, anxiety, etc.). A comprehensive reference source on stressors, the effects of activating stress response, and the disorders that may arise as a consequence of acute or chronic stress are provided, for example, in the four-volume Encyclopaedia of Stress [1]. How do we measure that a person is under stress? The most accurate estimations of the person's stress level can be obtained by measuring physiological parameters, such as electroencephalogram, heart rate or some biochemical markers in blood. However, many of these measurements are invasive. Non-invasive methods to assess the subjective stress level are mostly based on behavioural and psycho-physiological measurements [2].

In daily life, we often use the term "stress" to describe negative situations. However, there is a difference between eustress, which is the term for a positive stress, and distress, which refers to a negative stress. The positive stress motivates, focuses energy, feels exciting, and improves performance. In contrast, the negative stress causes anxiety, feels unpleasant, and decreases performance. Generally, it is difficult to define an optimal level of stress corresponding to the maximal speaker's performance as shown in Fig. 1, because individual persons have different reactions to particular stress situations. Besides the short-term effects of stress, the long-term occurrence of stress can have serious health consequences.

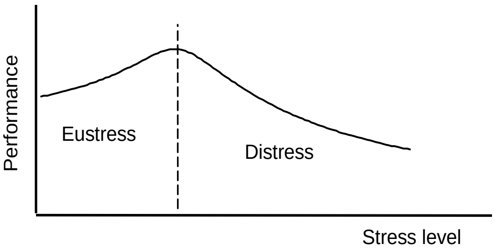


Fig. 1. General effect of positive and negative stress on speaker's performance.

A number of studies during the last decades have investigated the influence of emotion and stress on speech production and speech signal from different points of view. The most widely used methods usually start in time-domain from pitch, intensity and word/phoneme duration [3], [4]. Some spectral-domain parameters were used for characterization of stressed speech, for instance, in [5] and [6]. A specific feature derived from teager energy operator was proposed and found to be well responsive to speech under stress in [7]. As demonstrated in [8], some phonemes can differentiate emotions better than others.

In general, two types of information sources are available to determine the emotional status of a speaker from his/her speech, the linguistic content and the acoustic properties of the speech. In this study, we consider only the acoustic properties by using spectral features of vowels. While emotion researchers reported mostly on formant positions only, we investigated the influence of stress on the complete formant structure of vowel's spectrum.

II. DATABASES OF SPEECH UNDER STRESS

A typical speech corpus of extremely stressed speech from a real situation 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 [9]. The advantage of this

Manuscript received March 16, 2012; accepted May 12, 2012.
The research was performed in laboratories supported by the SIX project; the registration number CZ.1.05/2.1.00/03.0072, the operational program Research and Development for Innovation. The support of the project CZ.1.07/2.3.00/20.0007 WICOMT, financed from the operational program Education for Competitiveness, is also gratefully acknowledged.

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 by researchers is the SUSAS (Speech under Simulated and Actual Stress) database of stressed American English presented in [10] and distributed by Linguistic Data Consortium at the University of Pennsylvania. A German database of emotional utterances including also panic was recorded at the Technical University of Berlin. A complete description of this database called Berlin Database of Emotional Speech can be found in [11]. A list of existing emotional speech data collections including all available information about the databases such as the type of emotions, the language, etc. was provided in [12]. Generally, most of the corpora and studies reported in scientific works concern English. For Czech as well as other Slavic languages, no appropriate database was available.

For our studies conducted within research into speech signals under stress 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 created database called ExamStress consists of neutral speech and 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. A complete description of the ExamStress database can be found in [13]. The database currently contains 34 male speakers, mostly Czech natives.

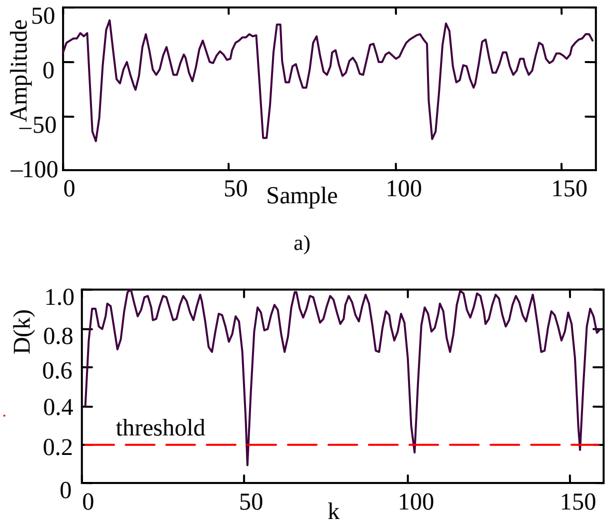
III. Estimation of Smoothed Spectrum of Vowels

Vowels generally are produced by a fixed vocal tract configuration. Their short-time spectra are characterized by local maxima called formants and denoted as F1, F2, F3, F4, etc. Formants are counted from the lowest frequency upwards and usually only the first three (F1, F2 and F3) contribute significantly to the intelligibility of speech. Physically, formants represent the primary resonances of the vocal tract caused by different geometric configurations of the articulating organs. In order to eliminate the effects of transitions from vowels into different consonants in continuous speech [14], only the stationary short-time segments of vowels were considered in this study using a two-stage segment selection. First, we apply automatic speech segmentation tool based on mel-frequency cepstral coefficients [15] to detect the individual vowels in speech signal with subsequent manual correction aimed to remove the wrong detected vowels. Then, well periodic single or multiple segments of constant size N were selected from the remaining vowel signals by means of difference function. This function is based upon the idea that for a truly periodic signal $s(n)$ of period P, the sequence

$$D(k) = \sum_{n=1}^{N} \left| s(n) - s(n+k) \right|, \qquad (1)$$

where for k=0,1…N-1 would be zero for k=0, P, 2P,… For short segments of vowels, it is reasonable to expect that $D(k)$ will be small at multiples of the period, but not identically zero. The non-zero value effectively represents non-periodicity in the waveform.

Figure 2 shows the difference function $D(k)$ normalized to 1 and a threshold value used as criterion in selection of well periodic signal segments. If the two lowest significant minima of $D(k)$ fall below the threshold (as the case in Fig. 2), the vowel segment is selected for further processing. The decision threshold was set empirically to 20%.



a)

b)

Fig. 2. Example of short-time speech segment of vowel /a/ in the word “jedna” (English “one”) (upper graph) and corresponding difference function (lower graph).

To get the short-time smoothed spectrum of speech signal, the approach based on the well-know linear prediction was applied. The importance of this method lies both in its ability to provide extremely accurate estimates of the spectral parameters and in its relative high speed of computation. Linear prediction theory is presented in many books; for instance in [16]. Denoting the sampling rate of speech signal by f_s, the relative amplitude spectrum can be estimated in the z-transform domain as

$$S(f) = \left| \frac{1}{1 - \sum_{m} a_m z^{-m}} \right|^2 =$$

$$= \frac{1}{\left| 1 - \left(a_1 e^{-j2\pi f/f_s} + a_2 e^{-j4\pi f/f_s} + \ldots + a_M e^{-j2M\pi f/f_s} \right) \right|^2}, \qquad (2)$$

where a_m are the predictive coefficients and M stands for the total number of the coefficients (i.e., predictor order) used to compute the speech spectrum. In general, the spectrum can be computed for frequency with a sweep from f=0 Hz up to the Nyquist frequency. To illustrate the nature of the spectral modelling capability of linear predictive spectra, Fig. 3

shows a comparison between logarithmic spectra estimated from the same speech segment once by Fourier transform (FT) and once by linear prediction (LP). In case of the LP-spectrum, the order *M* can effectively control the degree of spectral smoothness.

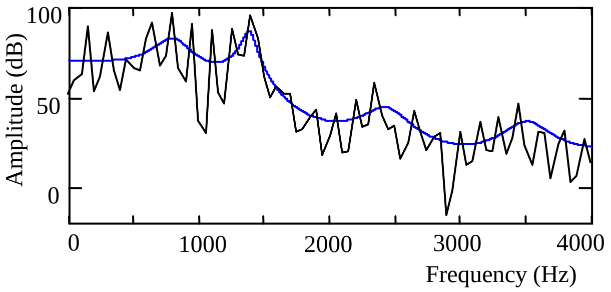


Fig. 3. Comparison between the log spectra obtained by FT and LP algorithm (f_s=8 kHz, M=10) for vowel /a/ in the frequency range 0–4 kHz for the same speech segment as in Fig. 2.

Using the LP-approach, the basic formant parameters can be computed directly from the set of predictive coefficients a_m solving (3) for complex roots z_i

$$1-\sum_{m=1}^{M} a_m z^{-m} = 0. \tag{3}$$

The frequency *F* and the bandwidth *B* of the *i*-th formant are related to the associated root z_i as follows:

$$F = \frac{f_s}{2\pi}\left|\arg(z_i)\right|, \tag{4}$$

$$B = \frac{f_s}{\pi}\ln\left|z_i\right|. \tag{5}$$

Table I presents formant parameters estimated from the LP-spectrum shown in Fig. 3 by roots using (4) and (5) and by spectral contour. Moreover, formant frequencies of the same speech segment were computed using the phonetic software Praat and added in the table for comparison. A low frequency false formant (with wide bandwidth) was detected by roots in this case. Generally, formant estimation by roots seems to be not precise and robust enough to analyse spectral changes for our purpose. Therefore, the vowel's formant structure was analysed by spectral contour using the first and second derivatives.

TABLE I. A COMPARISON OF COMPUTED FORMANT FREQUENCIES.

Formant	By roots		By contour	By Praat
no.	F	B	F	F
?	244	835	-	-
1	795	249	787	781
2	1382	94	1378	1377
3	2434	302	2415	2402
4	3508	311	3506	3490

IV. EXPERIMENTAL RESULTS

In our experiments, the short-time spectra of three cardinal Czech vowels /a/, /i/ and /u/ were analysed in both neutral and stressed speech in ten speakers. These vowels determine a basic vowel triangle in the F1-F2 space [16]. The speech signal from our database ExamStress (22 kHz, 16 bit) was resampled at 8 kHz and segmented by rectangular window (20 ms). Thus, the analysed frequency range is 0-4000 Hz which corresponds to spectrographic methods usually used in the forensic speaker recognition [17]. Two hundred segments from each vowel were extracted from the speech data of each speaker, 100 segments for neutral speech and 100 segments for stressed speech from the Defence part. A series of significant spectral points including local maxima (formants), local minima (antiformants) and points of inflexion was estimated. Formant bandwidths were defined separately for increasing and decreasing frequencies by means of inflexion points as can be seen in Fig. 4. Thus, the following features were investigated:

1) *Formant frequencies F1 to F4;*
2) *Antiformant frequencies A1 and A3;*
3) *Formant bandwidths Bd1 to Bd4;*
4) *Formant bandwidths Bu1 to Bu4.*

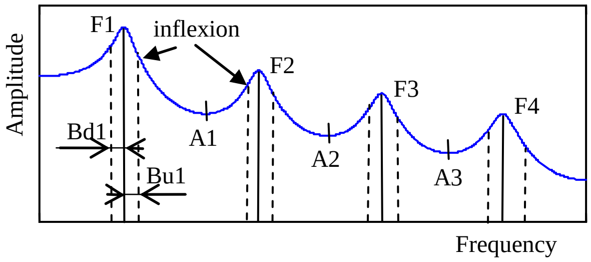


Fig. 4. Plot of a LP-spectrum with corresponding formant and antiformant positions, formant bandwidths are drawn dashed.

All 15 features were computed and evaluated separately for the vowels /a/, /i/ and /u/ from each speaker. Table II shows an example of the statistical parameters (mean μ and standard deviation σ) for all features of vowel /a/ obtained from speaker De in both neutral and stressed speech. To select the best feature candidates for long-term stress detection, two evaluation criteria were applied: the discrimination ratio for the *i*-th feature

$$DR(i) = \frac{\left(\mu_{N,i}-\mu_{S,i}\right)^2}{\sigma_{N,i}^2+\sigma_{S,i}^2} \tag{6}$$

and the difference of means for the *i*-th feature

$$\Delta\mu(i) = \left|\mu_{N,i}-\mu_{S,i}\right|. \tag{7}$$

TABLE II. STATISTICAL PARAMETERS OF ALL FEATURES OBTAINED FROM VOWEL /A/ SPOKEN BY SPEAKER DE.

Feature	**Neutral**		**Under stress**		**Evaluation**	
	μ_N	**σ_N**	**μ_S**	**σ_S**	**Δμ**	**DR**
*F*1	649	33	625	21	24	0.376
*F*2	1195	27	1151	14	44	2.093
*F*3	2507	58	2521	9	14	0.047
*F*4	3474	82	3560	37	86	0.914
*A*1	980	29	942	13	38	1.430
*A*2	2049	80	1963	13	86	1.126
*A*3	3097	47	3065	11	32	0.439
*Bd*1	54	10	68	12	14	0.803
*Bd*2	78	19	64	9	14	0.443
*Bd*3	171	18	132	15	39	2.770
*Bd*4	139	16	173	17	34	2.121
*Bu*1	72	15	82	11	10	0.289

Feature	Neutral		Under stress		Evaluation	
	μ_N	σ_N	μ_S	σ_S	$\Delta\mu$	DR
*Bu*2	94	17	91	14	3	0.019
*Bu*3	180	23	149	10	31	1.528
*Bu*4	161	9	166	13	5	0.100

TABLE III. FOUR MOST EFFECTIVE SPECTRAL FEATURES ACROSS ALL SPEAKERS ACCORDING TO THE DR-CRITERION.

Feature	Discrimination ratio		
	average	min	max
*F*2 /a/	2.560	1.784	2.903
*Bd*3 /u/	2.331	1.829	2.756
*Bd*3 /a/	2.284	1.455	2.770
*F*1 /i/	2.067	1.563	2.369

V. CONCLUSIONS

This paper presents speech signal analysis for stress detection in speakers. Fifteen spectral features based on linear prediction spectral model were investigated statistical in short-time spectra of three Czech vowels /a/, /i/ and /u/. The features summarized in Table III have been shown to be most effective features for stress detection among the fifteen features evaluated. However, obtained results indicate that speakers tend to follow an individual trend rather a global trend valid for all speakers. Therefore, speaker dependent recognition of stressed speech might be more appropriate.

In future work we plan to expand our database adding other types of stress (including simulated stress). Generally, speech features depend not only on the emotions but also may depend on the language. Thus, comparative measurements on vowels from stressed speech in more languages are needed. A natural goal is the development of algorithms for automatic detection and quantification of stress.

REFERENCES

[1] G. Fink, *Encyclopedia of Stress*. London: Academic Press, 2007, p. 3000.

[2] B. Hilburn, P. G. Jorna, *Workload and Air Traffic Control. Stress, Workload & Fatigue*. Mahwah: Lawrence Erlbaum Assoc., 2001, pp. 384–394.

[3] M. Lugger, B. Yang, “Cascaded Emotion Classification via Psychological Emotion Dimensions Using a Large Set of Voice Quality Parameters”, in *Proc. of the International Conference on Acoustics, Speech and Signal Processing,* Las Vegas, vol. 4, 2008, pp. 4945–4948.

[4] S. E. Bou-Ghazale, J. H. Hansen, “A Comparative Study of Traditional and Newly Proposed Features for Recognition of Speech Under Stress”, *Speech and Audio Processing,* IEEE, vol. 4, no. 8, pp. 429–442, 2000. [Online]. Available: http://dx.doi.org/10.1109/89.848224

[5] S. Ramamohan, S. Dandapat, “Sinusoidal Model-Based Analysis and Classification of Stressed Speech”, *Audio, Speech and Language Processing,* IEEE, vol. 3, no. 14, pp. 737–746, 2006. [Online]. Available: http://dx.doi.org/10.1109/TSA.2005.858071

[6] S. Wu, T. H. Falk, W. Y. Chan, “Automatic Speech Emotion Recognition Using Modulation Spectral Features”, *Speech Communication*, Elsevier, vol. 5, no. 53, pp. 768–785, 2011. [Online]. Available: http://dx.doi.org/10.1016/j.specom.2010.08.013

[7] G. Zhou, J. H. Hansen, J. F. Kaiser, “Nonlinear Feature Based Classification of Speech Under Stress”, *Speech and Audio Processing,* IEEE, vol. 2, no. 9, pp. 201–216, 2011.

[8] V. Sethu, E. Ambikairaja, J. Epps, “Phonetic and Speaker Variations in Automatic Emotion Classification”, in *Proc. of the Interspeech*, Brisbane, 2008, pp. 617–620.

[9] D. Haddad, S. Walter, R. Ratley, M. Smith, “Investigation and Evaluation of Voice Stress Analysis Technology”, Project Report, New York, Air Force Research Lab Rome, 2002, p. 115.

[10] J. H. Hansen, S. E. Ghazale, “Getting Started with SUSAS”, in *Proc. of the Eurospeech,* Rhodes, 1997, pp. 1743–1746.

[11] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, B. Weiss, “A Database of German Emotional Speech”, in *Proc. of the Eurospeech,* Lisbon, 2005, pp. 1517–1520.

[12] D. Ververidis, C. Kotropoulos, “Emotional Speech Recognition: Resources, Features, and Methods”, *Speech Communication,* Elsevier, vol. 9, no. 48, pp. 1162–1181, 2006. [Online]. Available: http://dx.doi.org/10.1016/j.specom.2006.04.003

[13] M. Sigmund, “Introducing the Database ExamStress for Speech Under Stress”, in *Proc. of the Nordic Signal Processing Symposium,* Reykjavik, 2006, pp. 290–293. [Online]. Available: http://dx.doi.org/10.1109/NORSIG.2006.275258

[14] D. Balbonas, G. Daunys, “Movement of Formants of Sound *a* in Lithuanian Language”, *Elektronika ir Elektrotechnika (Electronics and Electrical Engineering),* no. 8, pp. 75–78, 2006.

[15] M. Sigmund, P. Matejka, “An Environment for Automatic Speech Signal Labeling” in *Proc. of the Artificial Intelligence and Applications,* Innsbruck, 2002, pp. 298–301.

[16] L. R. Rabiner, R. W. Schafer, *Digital Speech Processing*. London: Prentice Hall, 2011, p. 1042.

[17] A. Braun, J.-P. Köster, *Studies in Forensic Phonetics.* Trier: Wissenschaftlicher Verlag, 1995, p. 167.