Computing and Informatics, Vol. 42, 2023, 1233–1254, doi: 10.31577/cai_2023_5_1233

PREDICTION OF STRESS LEVEL FROM SPEECH – FROM DATABASE TO REGRESSOR

Marián Trnka, Sakhia Darjaa, Róbert Sabo, Milan Rusko*

Institute of Informatics, Slovak Academy of Sciences
Bratislava, Slovakia
e-mail: {trnka, utrrsach, robert.sabo, milan.rusko}@savba.sk

Meilin Schaper, Tim Stelkens-Kobsch

Institute of Flight Guidance, German Aerospace Center
Braunschweig, Germany
e-mail: {meilin.schaper, tim.stelkens-kobsch}@dlr.de

Abstract. The term stress can designate a number of situations and affective reactions. This work focuses on the immediate stress reaction caused by, for example, threat, danger, fear, or great concern. Could measuring stress from speech be a viable fast and non-invasive method? The article describes the development of a system predicting stress from voice – from the creation of the database, and preparation of the training data to the design and tests of the regressor. StressDat, an acted database of speech under stress in Slovak, was designed. After publishing the methodology during its development in [1], this work describes the final form, annotation, and basic acoustic analyses of the data. The utterances presenting various stress-inducing scenarios were acted at three intended stress levels. The annotators used a "stress thermometer" to rate the perceived stress in the utterance on a scale from 0 to 100. Thus, data with a resolution suitable for training the regressor was obtained. Several regressors were trained, tested and compared. On the test-set, the stress estimation works well (R square = 0.72, Concordance Correlation Coefficient = 0.83) but practical application will require much larger volumes of specific training data. StressDat was made publicly available.

Keywords: Acoustic correlates of stress, speech under stress, stress database, stress-inducing scenarios, stress measurement

* Corresponding author

1 INTRODUCTION

As soon as in 1983, Streeter et al. noted that: "Voice indications of psychological stress are perhaps the most commonly studied emotional phenomena in speech production." [2] The interest of scientists in this topic still continues, and the development of technologies based on deep neural networks makes it possible to project theoretical knowledge into applications. The level of stress in a human being corresponds to the levels of stress hormones, for which fast non-invasive online measurement is not yet available. Therefore, various bio-signals are used to estimate stress levels (for a comprehensive survey see [3]). It is assumed that acoustic symptoms contained in speech can be used to identify increased levels of momentary (acute) stress [4, 5].

One of the major obstacles to this research is the lack of representative databases of speech under stress, with reliable assessment of stress levels. Creating a database of recordings of real-life stress-inducing situations and reliably assessing the level of current stress is extremely challenging, as ethical and health reasons do not allow scientists to expose subjects to high levels of stress. However, several attempts have been made to create databases of speech under stress, either acted or induced (e.g., [4]).

Defining "stress" is a notoriously difficult problem. No single definition will satisfy all circumstances, or, if it does, it will be too vague to have any practical use. Our definition was chosen so that it is appropriate for certain security applications of speech technology but may be unsuited to other areas, such as medical research. The issues of stress definition and measurement were well summarized in detail in the work of Epel et al. [6] and a unified view of stress was proposed.

In contrast to chronic stress, our present work focuses on the acute, or momentary stress that is experienced as an immediate perceived threat, either physical, emotional, or psychological [7]. Epel notes that "Acute psychological stress responses are often measured by capturing specific emotional states. This is because negative emotional responses (fear, anxiety, sadness, anger) to an acute stressor are considered a core component of an acute stress response." [6]. In their previous work, the authors of this paper also dealt with the possibility of measuring stress via the identification of expressed emotions, as emotions with high arousal and negative valence can be triggered by immediate stress [8]. However, the problem with this approach was that they were not able to estimate valence reliably, although this applied primarily to emotional expressions with lower emotional arousal. Therefore, in this work, we do not identify emotions, but instead, we try to estimate stress (i.e., the annotators were instructed to rate the perceived level of stress). Our aim is to design an automatic system that monitors the level of stress from acoustic cues in speech. Their increased intensity could indicate a worsened, dangerous, or critical situation. Speech is a psychophysiological process, influenced by environmental and/or internal challenges and reflects the level of stress [9], which entitles us to believe that the acoustic properties of speech can be used to estimate stress. As an example of an attempt at practical application, we can mention the Stressometer

and Stress Tracker – a mobile application that provides the user with a graphic representation reflecting the measured stress [10].

2 MATERIALS AND METHODS

Due to the lack of databases of speech under stress, representing critical situations, thoroughly annotated, and large enough for training, we decided to record our own database. To have control over the content of recordings (and not being able to record in real critical situations) we have chosen to record an acted speech database. We believe that the manifestations of stress in various stressful situations played by actors are, in the first approximation, sufficiently like spontaneous ones. So, they can be used for research of acoustic cues of speech under stress and for training and testing automatic systems for stress measurement.

In our former research, we experimented with training a stress-predictor [11] on an acted database CRISIS [12], containing utterances with various levels of emotional arousal (calming, neutral, warning and highly insisting). This database was originally designed for expressive speech synthesis purposes. Using it for stress-detector training was only a make-do solution. However, the functionality of the stress prediction was acceptable, and the stress-detector was implemented as an add-on tool to increase the security of air-traffic control [13].

2.1 StressDat Database

A new, more specialized database of speech under stress had to be designed. As the authors are primarily engaged in research and development of applications for the Slovak language, the database is in Slovak. The methodology and details of the content of this database have already been published in [1] as a work under progress. The actual paper briefly summarizes its final form and properties and presents how it was used for the development of regressors for predicting the level of stress from speech.

2.1.1 Database Creation Method

The database creation method assumes that appropriately designed sets of sentences corresponding to stress-inducing scenarios can be played with high naturalness at different stress intensities. Twelve scenarios of the stress-inducing situations were drawn up. Each scenario consists of 10 to 13 sentences with semantic content and a form that can be played at three levels of stress (neutral, low, high). In addition, four emotionally neutral situations with sentences corresponding semantically to the particular neutral (non-stress) situation were prepared to reach "really neutral speech", uninfluenced by the semantic content of the stress-inducing scenarios.

For the purpose of training a regressor, it is necessary to have utterances with an assigned stress value available. The values should cover as large a range as possible and should be able to reach all the values of this interval. If analog representation

is not possible, the resolution of the digital representation should be as high as possible. However, from the previous recording of expressive speech databases, we had the experience that speakers were not able to consistently maintain more than three levels of expressiveness when increasing from neutral speech through expressive to extremely expressive. We therefore decided to instruct the actors to play three levels of stress – neutral speech, speech under stress and speech under extremely high stress. We will refer to these levels as three intended levels. We could accept a number representing the level of stress that the actor tried to express during the realization of the given utterance (i.e., 1, 2, or 3) as one way of evaluating the intensity of stress. But the resolution of three levels only is pitifully small. However, each person has a different range of stress manifestation intensity in speech, which means that many speakers will jointly cover a larger range than an individual speaker. On the other hand, the intensity of stress cues in utterances with different intended levels of different speakers (or even of the same speaker) can be perceived as similar by the listener. The boundaries of intended levels clusters overlap on the receiving part of the communication. Therefore, we decided to use the rating of stress, perceived in the utterance by annotators. The interface for collecting ratings, which uses a so-called stress-thermometer will be described later in Section 2.1.3.

2.1.2 Recorded Subjects

Our subjects were professional and non-professional actors recruited mostly from the pool of personal contacts of the authors. The database was created at a time of COVID-19 pandemic, so it was not possible to record in studio. The actors recorded their speech using their own phones in their own apartments. They received detailed instructions for positioning the smartphones and selecting a suitable acoustic environment to ensure as similar recording conditions across the speakers as possible [1]. Recordings were captured using the default voice-recorder application of a specific actor's phone with the highest available audio quality and stored in the .wav or .m4a format. All audio files were later down-sampled to 16 kHz, 16-bit, mono. The final database includes 30 speakers (16 females, 14 males). The subjects received a small payment for recording. The average age of the speakers was 29 years. The youngest was 21 and the oldest was 38. During the recording of the database, we came to the decision, that the length of the audio material from each recorded speaker was too small, and that the "neutral" level can be influenced by the semantic content of the sentences of the stress-inducing scenarios. 6 more scenarios were therefore added (3 semantically neutral situations to be presented at "really neutral" level, and 3 additional stress-inducing situations to be presented at three levels of stress). 20 speakers recorded the full battery of 16 situations and 10 speakers recorded 10 situations in 3 levels (situations No. 1, 2, 3, 5, 6, 8, 9, 10, 11, 12) and 2 neutral situations (situations 13, 14). The detailed description of the scenarios is presented in Table 1. The transcription of each recording was manually corrected. So, the text representation of each recorded utterance is exact even in case the actor made a mistake, pronounced a different word, or omitted a word from the scenario. In the

following, we will refer to the intended stress levels as a – neutral, b – medium, and c – high.

Category	**No.**	**Intended Stress Level**	**Description**
Threat of losing control over the situation	1	a, b, c	As an airline pilot you need to make an emergency landing.
	2	a, b, c	Navigating a plane at the airport during very bad weather.
	3	a, b, c	As a pilot you need an undisciplined passenger to comply with the ban on using laptops during take-off/landing.
	4	a, b, c	As a firefighter coordinator you organize firefighting in a burning building.
Psycho-social stress	5	a, b, c	As a parent, you must organize the morning routine for your kids before leaving for school.
	6	a, b, c	You are finishing last-minute changes for an important presentation with a colleague.
	7	a, b, c	As a passenger, you need information on train departures urgently.
The threat of life/health injury of self/close ones	8	a, b, c	You call an ambulance for your father who suffered a stroke.
	9	a, b, c	You are trying to pacify your drunk brother who is trying to forcefully enter your flat.
	10	a, b, c	You are calling the police to resolve the situation with your drunken brother above.
	11	a, b, c	As a pilot, you organize evacuation from a burning aircraft.
	12	a, b, c	You report an insured event after a car accident by phone.
Neutral	13	a	You talk about school with your son.
	14	a	You are buying shoes.
	15	a	You teach students at school.
	16	a	You are reading a text to a colleague.

Table 1. Description of the scenarios

2.1.3 Annotation of the Perceived Stress Level by Listeners

The perceived stress level in the recorded sentences was rated by five annotators. We have created a simple graphic interface for utterances evaluation. The annotators evaluated the utterance in sets of 100 with shuffled intended stress levels. They were advised to take breaks of at least 5 minutes between sets. As the task demands high attention, they were strongly advised not to evaluate more than 3 sets per day, which

corresponds to approximately 1 hour of work. In each set, there were 100 utterances, but each annotator had a different order of sentences in the corresponding set in order to minimize the influence of the previously heard sentences on the evaluation. To limit the speaker influence, the annotators evaluated sentences from several different speakers in each set. Annotators were instructed to assess "distress, anxiety or discomfort on a scale of 0 to 100". They were told: "Imagine you have a thermometer that measures the stress according to the indicated scale. Rate on that scale how you think the person felt when pronouncing the utterance."

The descriptions for the individual points of the scale were the following (translated from Slovak):

0: Totally relaxed;

10: Alert, well concentrated;

20: Minimal anxiety/discomfort;

30: Mild anxiety/discomfort, does not interfere with performance;

40: A little upset, manages the activity but does not feel well;

50: Moderate anxiety/discomfort, feels uncomfortable but can continue activities;

60: Moderate to severe feeling of anxiety/discomfort;

70: Quite anxious/significant discomfort, it disrupts activity;

80: Great anxiety/discomfort, unable to concentrate;

90: Extreme anxiety/discomfort;

100: The most intense anxiety/fear/discomfort he/she has ever felt.

A picture of a thermometer with a verbal description of the states corresponding to individual "temperature" degrees is often used as a good aid in the more detailed evaluation of affective phenomena on a wide scale. In general, the scale can be continuous, but the finite number of "degrees" leads to discretization (e.g. [14, 15, 16, 17, 18]). We used a "Stress Thermometer" tool (see Figure 1) based on the Subjective Units of Distress Scale [19].

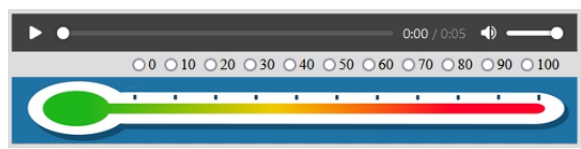


Figure 1. "Stress Thermometer" tool that allows the annotator to listen to the utterance and to assign a perceived stress level

2.1.4 Inter-Annotator Agreement

In [1] the authors assessed the agreement among the annotators using Fleiss' kappa reaching values around 0.3, which suggests a fair agreement. The interrater correlation coefficient (ICC), however, provides a better estimate of the agreement among raters [20] since the distance between the disagreements on the scale is taken into consideration (those between adjacent values are considered better than disagreements spanning multiple scale levels). We thus created a 2-way model based on a mean-rating ($k = 5$) and absolute agreement within the irr package [21] in R, which returned $\mathrm{ICC}(A, 5) = 0.935$ and 95 % confident intervals $(0.921, 0.945)$ and a more conservative model based on single rater gives $\mathrm{ICC}(A, 1) = 0.74$. These values correspond to excellent and moderate-good reliability, respectively.

To account for the inter-annotator variability, z-score normalization was used and linearly projected on the 0–100 interval. All the following calculations were performed on this normalized annotation.

In accordance with the methodology, all annotators were Slovak native speakers. However, we also tried to evaluate the utterances by two annotators who do not speak, nor understand Slovak at all. One was from India (IND) and one was from the Netherlands (NL). Both live in Europe in an English-speaking environment. While NL's evaluations were in good agreement with Slovak annotators, IND used only a very narrow range of values. He said he was not able to perceive well the presence and intensity of stress in Slovak speech. Although it is possible that this was due to the individuality of the annotators, it may also indicate that the evaluation is strongly culturally dependent. Figure 2 shows how the three acted (intended) stress levels were subjectively rated on a scale from 0 to 100 by the annotators (perceived stress levels).

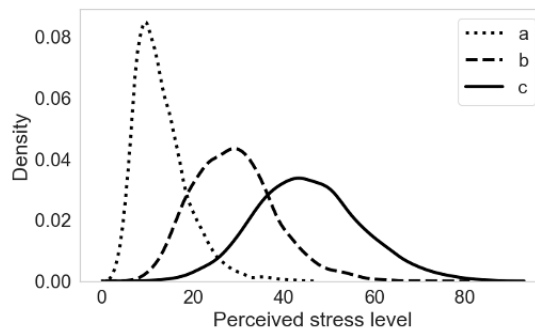


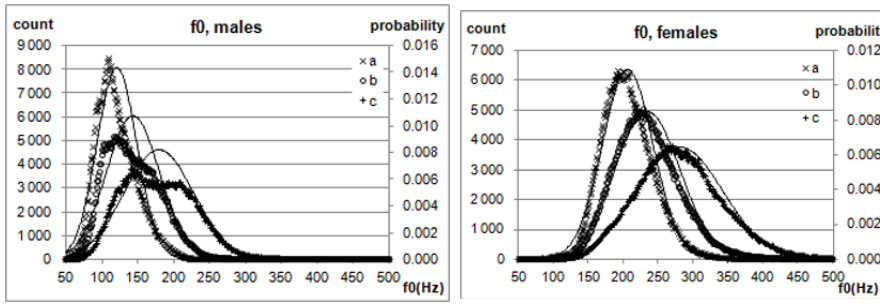
Figure 2. Distribution of perceived levels assigned by the annotators to the utterances with respect to the three intended stress levels

2.2 Acoustical Analyses

To get an idea of how the acoustic properties of speech differ at individual levels of stress in StressDat, analyzes of characteristics representing prosody and voice quality were performed, namely fundamental frequency of vocal fold vibrations, sound pressure level, long-term-average-spectrum, formant positions of vowels, speech rate and counts and durations of pauses.

2.2.1 Fundamental Frequency (F0)

F0 was measured on 25 ms frames through the entire corpus. The distribution of F0 values and their Gaussian approximation curves for the three intended levels of stress in male and female speakers are presented in Figure 3 a) and 3 b), respectively.



a) b)

Figure 3. Histograms of F0 values with Gaussian approximation curves for a) male and b) female speakers and three levels of intended stress (a – neutral, b – medium, and c – high)

The values of F0 increase with increasing stress level. While the distributions of female speakers are well approximated with Gaussians, the distribution of F0 in male utterances of levels b and c looks more like mixtures of two Gaussians. We can only speculate that some parts of these utterances are pronounced with moderate voice effort and the other parts are more expressive. This depends on the actor's way of realization (and thus also on the choice of the actors).

Mean F0 was computed per speaker and stress level. The average value of the F0 means and standard deviations with respect to the three levels of intended stress are presented in Table 2.

2.2.2 Sound Pressure Level

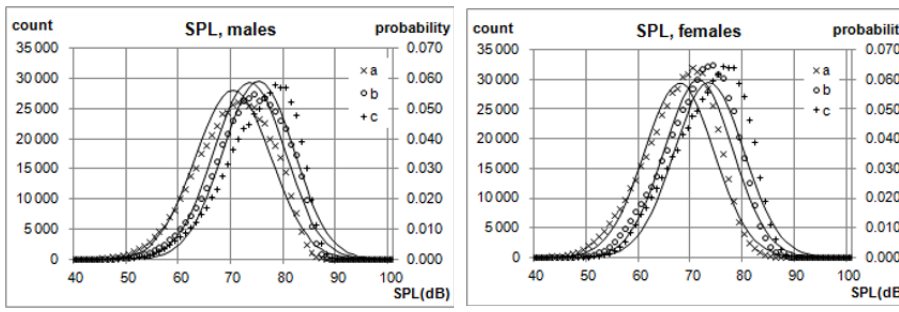
Sound pressure level of the speech signal is known to be highly correlated with manifestations of affect in speech. According to Praat documentation, "Intensity object represents an intensity contour at linearly spaced time points $t_i = t_1 + (i -$

Intended Stress Level	Average F0 Mean [Hz]		Average F0 stdev [Hz]	
	Males	Females	Males	Females
a	120.9	205.8	27.8	36.6
b	143.4	234.1	37.1	47.1
c	179.4	281.0	48.7	62.0

Table 2. Average F0 mean and standard deviation

1) dt, with values in dB SPL, i.e. dB relative to $2{\cdot}10^{-5}$ Pascal, which is the normative auditory threshold for a 1 000 Hz sine wave" [22].

Correct interpretation of the correlation between intensity and stress is hampered by the fact that differences in SPL (which reflect vocal effort [23]), are highly non-specific. They are influenced not only by stress, but also by the distance of the speaker from the addressee, directional orientation, Lombard effect, emotion, mood, and personality of the speaker. In addition, a calibrated measuring system would be needed to measure SPL correctly [24]. Phone recording presents the problem of automatic gain control, which is specific to each phone model and its parameters are unknown to the user. The following analyses are therefore only informative, valid for this database, and cannot be freely generalized. The histograms of SPL values with Gaussian approximation curves for a) males, and b) females, at three levels of intended stress are presented in Figure 4.



a) b)

Figure 4. Histograms of SPL values with Gaussian approximation curves for a) males, b) females, and three levels of intended stress (a – neutral, b – medium, and c – high)

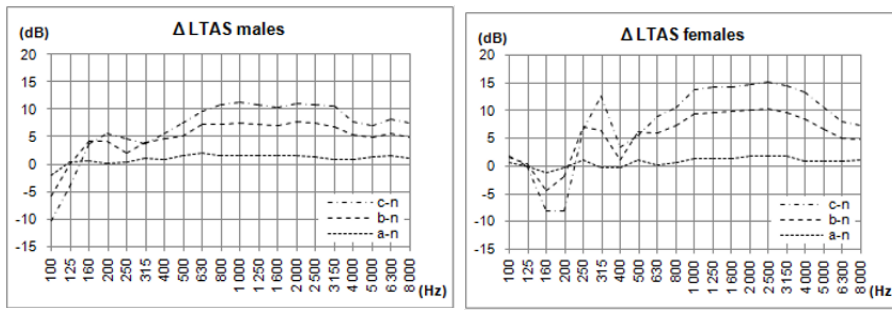
Female speakers produced speech with a slightly lower SPL, but otherwise the histograms of male and female subjects do not differ.

2.2.3 Long-Term-Average-Spectrum (LTAS)

Long-term-average-spectrum (LTAS) is known to contain information on the vocal effort [25], which is one of the means used by speakers to express various levels of expressivity.

Rather than measuring the overall amplitude, some studies have tended to determine the amplitude of different frequency bands. Although different techniques were employed to achieve these calculations, there is general agreement that amplitude measurements are greater in higher frequencies i.e., above 1 000 Hz in stress conditions compared to control conditions [26].

As it was already mentioned, before the recording of the three levels of stress (a, b, c), the speakers were asked to relax and record a set of emotionally neutral sentences to get an independent sample of "really neutral" speech (i.e., speech not influenced by stress). This level is referred to as "n" or "level n" in Figure 5. To study the behavior of LTAS at various stress levels, we took the LTAS of level n as a reference and plotted the deviations of the spectra of levels a, b, c, from this reference value to the graph (Figure 5). This difference spectrum is marked as ΔLTAS. It was analyzed in third-octave frequency bands.



a) b)

Figure 5. Comparison of ΔLTAS for a) male and b) female speakers for 3 intended levels

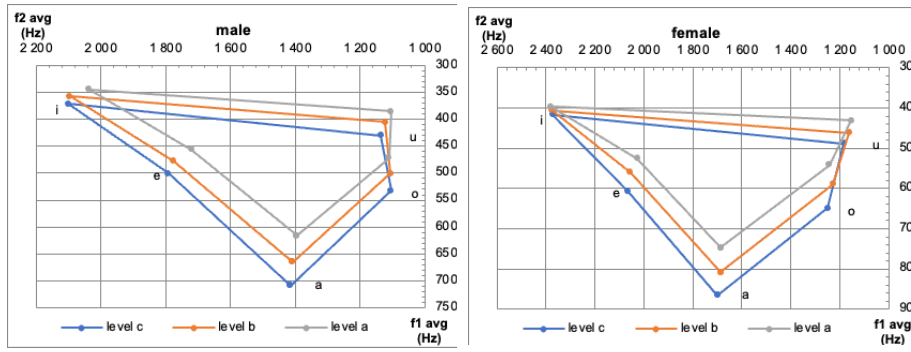
It can be seen from the Figure 5, that biggest differences of LTAS are in the range from 125 to 400 Hz, which is caused by energetically rich F0, shifting higher with increasing stress level. Much more notable are the higher values in the range from 1 000 to 3 150 (males) or to 4 000 Hz (females), which is caused by increasing energy of higher harmonics with increasing vocal effort.

2.2.4 Formant Positions of Vowels (F1, F2)

Ruiz et al. [27] analyzed spectral balance frequency in vowels and tentatively suggested that this might be a promising parameter that is sensitive to stress [26].

We measured the mean frequencies of the formants of the Slovak vowels a, e, i, o, u and presented them in the F1 versus F2 formant space in Figure 6.

While increasing the central frequency of F1 formant (caused mostly by the up-down movement of the jaw) is clearly observable in most of the vowels with increasing levels of stress, the movement of F2 (that depends mostly on the back-front movement of the tongue) is negligible.



a) b)

Figure 6. F1 versus F2 formant diagram of the five Slovak vowels a, e, i, o, u for the three intended levels of stress

2.2.5 Speech Rate

The speech rate of each utterance was measured in syllables per second and in words per second and the results are presented in Table 3.

Intended Stress Level	Syllables per Second		Words per Second	
	Males	Females	Males	Females
a	6.137	6.112	2.916	2.912
b	6.452	6.317	3.069	3.010
c	6.640	6.466	3.155	3.079

Table 3. Speech rate: syllables per second and words per second for 3 intended stress levels

As many of the utterances are short, syllables per second is the more suitable a measure to reflect the speech rate in StressDat than words per second. The syllable rate slightly increases with intended stress level and the gender differences are insignificant.

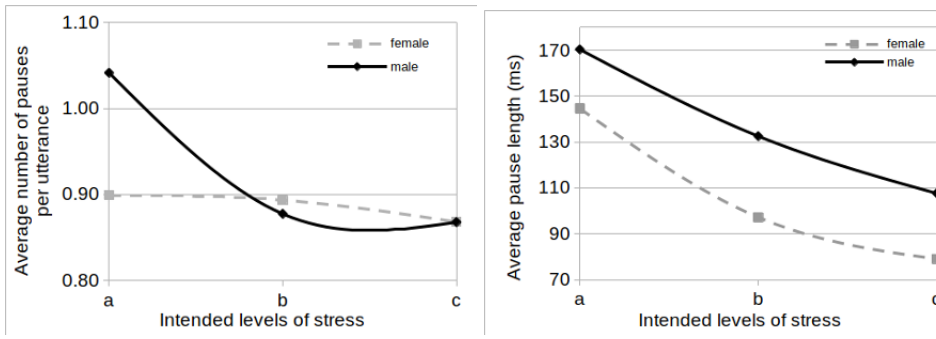
2.2.6 Pause Counts and Durations

Pauses can partly be influenced by respiration, which is the driving force of both stress and voice production, and Van Puyvelde et al. [28] hypothesize it to be the missing link in our understanding of the underlying mechanisms of the dynamic between speech and stress.

All recordings in StressDat were manually transcribed into text and then subjected to forced alignment. This algorithm determined the exact boundaries of phonemes and identified pauses. The average number of pauses per utterance (a)

and mean duration of pauses (b) with respect to three intended levels of stress are presented in Figure 7.

The results for the average number of pauses for men and women are not quite consistent. This characteristic is highly dependent on the individuality of the speaker, and due to the small number of speakers, the statistical "smoothing" of the results was not effective enough. On the other hand, the shortening of average pause lengths with increasing stress levels has the same tendency in men and women. However, pauses for women are approximately 30 ms shorter.



a) b)

Figure 7. Average number of pauses per utterance a) and mean duration of pauses b) in StressDat utterances with respect to three intended levels of stress

2.2.7 Sentence Lengths

StressDat contains sentences of various lengths. The graph of distribution of sentence lengths is shown in Figure 8. The minimum duration of the utterance is 0.6 s, and the maximum is 31 s. The average is 4.57 s. 98.9 % of sentences have a length in the range of 1–10 s.

2.3 Representative Features and Regressors

From a machine learning perspective, the most significant difference between regression versus classification is that while regression helps predict a continuous quantity, classification predicts discrete class labels. However, a regression algorithm can predict a discrete value which is in the form of an integer quantity.

Trnka et al. have already published experiments on the use of three class classificators trained on StressDat [29]. In this work we will focus on the approach using regression.

Experiments with stress level regression focused on comparing 3 diverse types of parameters (GeMAPS, X-vector and TRILL) using 5 diverse types of regres-

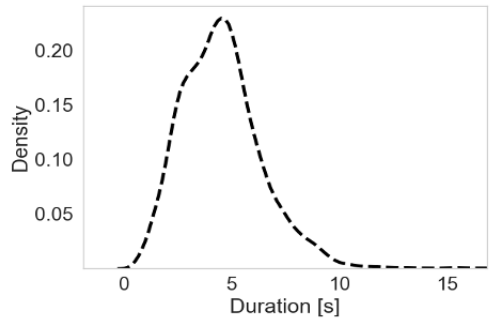


Figure 8. Distribution of the time lengths of the audio files in seconds

sors: Random Forest regressor, Gradient Boosting regressor, Decision Tree regressor, Multi-layer Perceptron regressor and Support Vector Regressor. The default hyperparameters of the regressors were used in training. The Scikit-learn machine learning library [30] was used for training. The training and test sets for regression are organized in pairs of feature vectors (X-vector, TRILL or GeMAPS), representing particular utterances, and the corresponding value of the annotated perceived stress level.

In the data preparation phase, representative characteristics are calculated from the digitized audio signal. We used three different sets of features. Acoustic features were extracted using the OpenSMILE toolkit [31] and the GeMAPSv01b subset of the OpenSMILE features was used. This subset, containing 62 features, was designed especially for affective speech recognition [32]. Therefore, we assumed it will work well for stress prediction.

For comparison we also used modern auditory non-semantic speech representation, X-vectors [33] and TRILLs (TRIpLet Loss network vectors) [34]. We used Kaldi toolkit [35] to train X-vectors extractor to compute the 512-dimensional X-vectors. The procedure was like that presented in [33]. 39 MFCC features (13 MFCC + delta MFCC + delta delta MFCC) were used as the input spectral information for the X-vector extractor. The frame length was 25 ms and frame shift 10 ms. The X-vector extractor was trained on VoxCeleb [36] and VoxCeleb2 [37] speech databases. The energy-based Voice Activity Detector (VAD) was used to filter out silence frames.

To compute TRILL embedding we used publicly available pretrained network [38]. The size of TRILL embedding was 512 – same as X-vector. For TRILL computing we did not use VAD.

3 DATABASE EVALUATION

To evaluate the potential of the created database, it was used to train and test the regressors. We first divided StressDat into independent training set and test set.

As one of the speakers had to be temporarily excluded from the database due to incomplete annotation, the number of speakers was decreased to 29. The train set had 24 speakers (10 female, 12 male) which is 10 591 audio files (utterances) in total. The test set had 5 speakers (2 male, 3 female), which is 2 172 utterances.

Figure 9 shows density of perceived stress levels in the training set and test set.

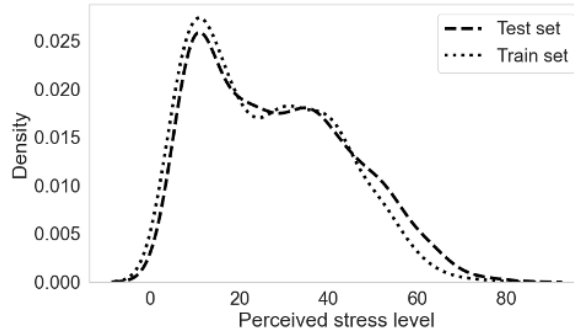


Figure 9. Distribution of perceived stress levels in training set and test set

It is clear from Figure 9, that the largest part of utterances was evaluated as expressing lower levels of stress. Elevated levels are still quite common, but there are few manifestations of extremely high stress. However, the distribution of stress levels in the test set replicates well in the training set.

3.1 Regression Results

3.1.1 Evaluation Metrics

In different works, different measures were used for evaluation of the quality of regression. For better comparison we used 4 metrics to evaluate the quality of the regression: R2 (Coefficient of determination), MAE (Mean Absolute Error), CC (Pearson correlation coefficient) a CCC (Congruence Correlation Coefficient). CCC is a correlation measure, which is widely used in affective speech research. It was used for instance in OMG – Emotion Challenge at IEEE World Congress on Computational Intelligence in 2018 [39].

3.1.2 Comparison of Features and Regressors

Five machine learning algorithms were compared when using three different sets of representative features. Regressors were trained on StressDat training set and tested on StressDat test set. Results are presented in Table 4.

It can be seen that the overall best result is achieved by the Gradient Boosting Regressor model trained on X-vector features (CCC 0.83). However, we can conclude

Regressor	Feature	R^2	MAE	CC	CCC
Support Vector	X-vector	0.71	7.2	**0.86**	0.8
Random Forest	X-vector	0.64	8.23	0.83	0.75
Gradient Boosting	X-vector	**0.72**	**7.1**	**0.86**	**0.83**
Decision Tree	X-vector	0.27	11.16	0.61	0.6
ML Perceptron	X-vector	0.55	9.03	0.77	0.74
Support Vector	GeMAPS	**0.7**	**7.11**	**0.85**	**0.8**
Random Forest	GeMAPS	0.66	7.77	0.83	0.79
Gradient Boosting	GeMAPS	0.62	8.24	0.82	0.79
Decision Tree	GeMAPS	0.26	10.86	0.64	0.63
ML Perceptron	GeMAPS	0.51	9.29	0.78	0.76
Support Vector	TRILL	**0.64**	7.9	**0.83**	0.74
Random Forest	TRILL	0.57	8.87	0.78	0.68
Gradient Boosting	TRILL	**0.64**	**7.73**	0.81	**0.75**
Decision Tree	TRILL	0.05	12.49	0.48	0.48
ML Perceptron	TRILL	0.23	11.6	0.66	0.66

Table 4. Performance of regression of different acoustic features on various machine learning algorithms

that the most stable results were obtained by the Support Vector Regressor for all three types of parameters. In the CC metric it was always the best (for the X-vector it had the same result as Gradient Boosting).

The comparison of different features turned out best for X-vector embedding. However, the Opensmile GeMAPS parameters performed only slightly worse, even though the vector size was only 62 compared to the X-vector size of 512.

On the contrary, Decision Trees clearly gave the worst results for our experiment setup.

4 DISCUSSION

We proposed a methodology for creating an acted database of speech under stress. We have successfully recorded 30 speakers, which gives in total nearly 16 hours of recordings. We proposed an approach of subjective assessment of stress by the annotators using a so-called stress thermometer. This allowed us to obtain a rating on a continuous scale and within a reasonable range of the intensity of acted manifestations of stress.

The speakers had to produce sentences from scenarios at three intended levels of stress. However, the affective setting is individual for each speaker and the levels are not consistent between speakers. When evaluating the perceived level of stress, the annotators did not have information about the intended level of stress and were therefore not influenced by this information. Their assessment was not limited to three levels, but they could use a scale from 0 to 10, which is suitable for creating a regressor assessing stress on a continuous scale.

We made basic acoustic measurements on the database to provide information about the impact of stress on the acoustic parameters of speech. Separate analyses for male and female speakers make it possible to study gender differences. Through experiments with three types of representative characteristics and with five types of regressors, we tried to design the best combination suitable for automatic system for stress level estimation from speech.

Training was done on the training-set and the tests were done on the test-set of the StressDat database. The speakers in the training and testing sets do not overlap. However, it should be noted that the scenario texts in the training and test sets are the same. Ideally, the quality of stress level prediction should be verified on a completely independent, reliably annotated test set, but this was not available.

Due to the pandemic situation, it was not possible to upload the database in the flight simulator. So, it was replaced by an acted database. This results in less naturalness of speech expression. Moreover, the database is too small to representatively cover all aspects of stress in speech. It is an ad-hoc solution that needs to be improved in the future by obtaining authentic speech data from real stressful situations. In the future, different time windows of stress assessment, both longer and shorter than one sentence, can be considered.

We hope the StressDat database partly fills the lack of databases of speech under stress with a reliable assessment of the stress level. Moreover, this Slovak database opens possibilities for cross-lingual and cross-cultural research.

Stress is an overly broad concept, which covers a large number of different situations and reactions to them. It is therefore highly likely that for the creation of practical applications, it will be necessary to build specific stress databases covering adequately the very manifestations of stress that the system is supposed to detect.

Funding: This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 832969. This output reflects the views only of the authors, and the European Union cannot be held responsible for any use which may be made of the information contained therein. For more information on the project, see: `http://satie-h2020.eu/`. The work was also funded by the Slovak Scientific Grant Agency VEGA, project No. 2/0165/21. The work was also funded by the The Slovak Research and Development Agency, project No. APVV-21-0373.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of the Slovak Academy of Sciences (protocol code 83296 from September 25, 2019).

Informed Consent Statement: Written informed consent for publication was obtained from all subjects involved in the study.

Data Availability Statement: StressDat database is available for non-commercial purposes on request from the corresponding author.

Conflict of Interest: The authors declare no conflict of interest.

REFERENCES

[1] Streeter, L. A.—Macdonald, N. H.—Apple, W.—Krauss, R. M.—Galotti, K. M.: Acoustic and Perceptual Indicators of Emotional Stress. The Journal of the Acoustical Society of America, 1983, Vol. 73, 1997, No. 4, pp. 1354–1360, doi: 10.1121/1.389239.

[2] Sharma, N.—Gedeon, T.: Objective Measures, Sensors and Computational Techniques for Stress Recognition and Classification: A Survey. Computer Methods and Programs in Biomedicine, Vol. 108, 2012, No. 3, pp. 1287–1301, doi: 10.1016/j.cmpb.2012.07.003.

[3] Hansen, J. H. L.—Bou-Ghazale, S. E.: Getting Started with SUSAS: A Speech Under Simulated and Actual Stress Database. Proceedings of 5th European Conference on Speech Communication and Technology (EUROSPEECH 1997), 1997, doi: 10.21437/Eurospeech.1997-494.

[4] Baird, A.—Triantafyllopoulos, A.—Zänkert, S.—Ottl, S.—Christ, L.—Stappen, L.—Konzok, J.—Sturmbauer, S.—Messner, E.-M.—Kudielka, B. M.—Rohleder, N.—Baumeister, H.—Schuller, B. W.: An Evaluation of Speech-Based Recognition of Emotional and Physiological Markers of Stress. Frontiers in Computer Science, Vol. 3, 2021, Art. No. 750284, doi: 10.1016/j.compind.2017.04.005.

[5] Epel, E. S.—Crosswell, A. D.—Mayer, S. E.—Prather, A. A.—Slavich, G. M.—Puterman, E.—Mendes, W. B.: More Than a Feeling: A Unified View of Stress Measurement for Population Science. Frontiers in Neuroendocrinology, Vol. 49, 2018, pp. 146–169, doi: 10.1016/j.yfrne.2018.03.001.

[6] Plarre, K.—Raij, A.—Hossain, S. M.—Ali, A. A.—Nakajima, M.—Al'Absi, M.—Ertin, E.—Kamarck, T.—Kumar, S.—Scott, M.—Siewiorek, D.—Smailagic, A.—Wittmers, L. E.: Continuous Inference of Psychological Stress from Sensory Measurements Collected in the Natural Environment. Proceedings of the 10th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN), 2011, pp. 97–108.

[7] Trnka, M.—Darjaa, S.—Ritomský, M.—Sabo, R.—Rusko, M.—Schaper, M.—Stelkens-Kobsch, T.: Mapping Discrete Emotions in the Dimensional Space: An Acoustic Approach. Electronics, Vol. 10, 2021, No. 23, Art. No. 2950, doi: 10.3390/electronics10232950.

[8] Hansen, J. H. L.—Patil, S.: Speech Under Stress: Analysis, Modeling and Recognition. In: Müller, C. (Ed.): Speaker Classification I. Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, Vol. 4343, 2007, pp. 108–137, doi: 10.1007/978-3-540-74200-5_6.

[9] Gaggioli, A.—Pioggia, G.—Tartarisco, G.—Baldus, G.—Ferro, M.—Cipresso, P.—Serino, S.—Popleteev, A.—Gabrielli, S.—Maimone, R.—Riva, G.: A System for Automatic Detection of Momentary Stress in Naturalistic Settings. In: Wiederhold, B. K., Riva, G. (Eds.): Annual Review of Cybertherapy and Telemedicine 2012. IOS Press, Studies in Health Technology and Informatics, Vol. 181, 2012, pp. 182–186, doi: 10.3233/978-1-61499-121-2-182.

[10] Sabo, R.—Rusko, M.—Ridzik, A.—Rajčáni, J.: Stress, Arousal and Stress Detector Trained on Acted Speech Database. In: Ronzhin, A., Potapova, R., Németh, G. (Eds.): Speech and Computer (SPECOM 2016). Springer, Cham, Lecture Notes in Computer Science, Vol. 9811, 2016, pp. 675–682, doi: 10.1007/978-3-319-43958-7_82.

[11] Rusko, M.—Darjaa, S.—Trnka, M.—Cerňak, M.: Expressive Speech Synthesis Database for Emergent Messages and Warnings Generation in Critical Situations. Proceedings of the Language Resources for Public Security Applications Workshop, LREC 2012, Istanbul, Turkey, 2012, pp. 50–53.

[12] Rusko, M.—Trnka, M.—Darjaa, S.—Rajčáni, J.—Finke M.—Stelkens-Kobsch, T.: Enhancing Air Traffic Management Security by Means of Conformance Monitoring and Speech Analysis. In: Klempous, R., Nikodem, J., Baranyi, P. (Eds.): Cognitive Infocommunications, Theory and Applications. Springer, Cham, Topics in Intelligent Engineering and Informatics, Vol. 13, 2019, pp. 177–199, doi: 10.1007/978-3-319-95996-2_9.

[13] Sabo, R.—Beňuš, Š.—Trnka, M.—Ritomský, M.—Rusko, M.—Schaper, M.—Szabo, J.: StressDat – Database of Speech Under Stress in Slovak. Journal of Linguistics/Jazykovedný časopis, Vol. 72, 2021, No. 2, pp. 579–589, doi: 10.2478/jazcas-2021-0053.

[14] Feelings Thermometer. Wisconsin Office of Children's Mental Health. Available online: `https://children.wi.gov/Pages/FeelingsThermometer.aspx`. [Accessed 2022].

[15] Stress Thermometer. Ohio Center for Autism and Low Incidence. Available online: `https://www.ocali.org/project/resource_gallery_of_interventions/page/StressThermometer`. [Accessed 2022].

[16] Gorski, T.: The Stress Thermometer. Available online: `https://i.redd.it/jutk7qv4nkb41.jpg`.

[17] Stress Level Thermometer. Available online: `https://www.etsy.com/listing/1110274677/stress-level-thermometer-pdf`. [Accessed 2022].

[18] SUDS Thermometer. Centre for Clinical Psychology. Available online: `https://ccp.net.au/suds-thermometer/`.

[19] Wolpe, J.: The Practice of Behavior Therapy. Pergamon Press, 1969, pp. 314.

[20] Koo, T. K.—Li, M. Y.: A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. Journal of Chiropractic Medicine, Vol. 15, 2016, No. 2, pp. 155–163, doi: 10.1016/j.jcm.2016.02.012.

[21] Gamer, M.—Lemon, J.—Fellows, I.—Singh, P.: irr: Various Coefficients of Interrater Reliability and Agreement. R package version 0.84.1. Available online: `https://CRAN.R-project.org/package=irr`. [Accessed 2022].

[22] Boersma, P.—Weenink, D.: Praat: Doing Phonetics by Computer [Computer

program; Version 6.2.06]. 1992–2022. Available online: https://www.praat.org. [Accessed January 2022].

[23] Traunmüller, H.—Eriksson, A.: Acoustic Effects of Variation in Vocal Effort by Men, Women, and Children. The Journal of the Acoustical Society of America, Vol. 107, 2000, No. 6, pp. 3438–3451, doi: 10.1121/1.429414.

[24] Švec, J. G.—Granqvist, S.: Tutorial and Guidelines on Measurement of Sound Pressure Level in Voice and Speech. Journal of Speech, Language, and Hearing Research, Vol. 61, 2018, No. 3, pp. 441–461, doi: 10.1044/2017_JSLHR-S-17-0095.

[25] Liénard, J. S.: Quantifying Vocal Effort from the Shape of the One-Third Octave Long-Term-Average Spectrum of Speech. The Journal of the Acoustical Society of America, Vol. 146, 2019, No. 4, pp. EL369-EL375, doi: 10.1121/1.5129677.

[26] Kirchhübel, C.—Howard, D. M.—Stedmon, A. W.: Acoustic Correlates of Speech when Under Stress: Research, Methods and Future Directions. International Journal of Speech Language and the Law, Vol. 18, 2011, No. 1, pp. 75–98, doi: 10.1558/ijsll.v18i1.75.

[27] Ruiz, R.—Absil, E.—Harmegnies, B.—Legros C.—Poch, D.: Time- and Spectrum-Related Variabilities in Stressed Speech Under Laboratory and Real Conditions. Speech Communication, Vol. 20, 1996, No. 1-2, pp. 111–129, doi: 10.1016/S0167-6393(96)00048-9.

[28] Van Puyvelde, M.—Neyt, X.—McGlone, F.—Pattyn, N.: Voice Stress Analysis: A New Framework for Voice and Effort in Human Performance. Frontiers in Psychology, Vol. 9, 2018, Art. No. 1994, doi: 10.3389/fpsyg.2018.01994.

[29] Kejriwal, J.—Beňuš, Š.—Trnka, M.: Stress Detection Using Non-Semantic Speech Representation. Proceedings of the 2022 32nd International Conference Radioelektronika, Košice, Slovakia, 2022, pp. 1–5, doi: 10.1109/RADIOELEKTRONIKA54537.2022.9764916.

[30] Pedregosa, F.—Varoquaux, G.—Gramfort, A.—Michel, V.—Thirion, B.—Grisel, O. et al.: Scikit-Learn: Machine Learning in Python. Journal of Machine Learning Research, Vol. 12, 2011, pp. 2825–2830, https://jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.

[31] Eyben, F.—Wöllmer, M.—Schuller, B.: Opensmile: The Munich Versatile and Fast Open-Source Audio Feature Extractor. Proceedings of the 18th ACM International Conference on Multimedia (MM '10), 2010, pp. 1459–1462, doi: 10.1145/1873951.1874246.

[32] Eyben, F.—Scherer, K. R.—Schuller, B. W.—Sundberg, J.—André, E.—Busso, C.—Devillers, Y. L.—Epps, J.—Laukka, P.—Narayanan, S. S.—Truong, K. P.: The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. IEEE Transactions on Affective Computing, Vol. 7, 2016, No. 2, pp. 190–202, doi: 10.1109/TAFFC.2015.2457417.

[33] Snyder, D.—Garcia-Romero, D.—Sell, G.—Povey, D.—Khudanpur, S.: X-Vectors: Robust DNN Embeddings for Speaker Recognition. Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 5329–5333, doi: 10.1109/ICASSP.2018.8461375.

[34] Shor, J.—Jansen, A.—Maor, R.—Lang, O.—Tuval, O.—de Chaumont

Quitry, F.—Tagliasacchi, M.—Shavitt, I.—Emanuel, D.—Haviv, Y.: Towards Learning a Universal Non-Semantic Representation of Speech. Proceedings of INTERSPEECH 2020, 2020, pp. 140–144, doi: 10.21437/Interspeech.2020-1242.

[35] Povey, D.—Ghoshal, A.—Boulianne, G.—Burget, L.—Glembek, O.—Goel, N.—Hannemann, M.—Motlícek, P.—Qian, Y.—Schwarz, P.—Silovský, J.—Stemmer, G.—Veselý, K.: The Kaldi Speech Recognition Toolkit. Proceedings of the IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, 2011.

[36] Nagrani, A.—Chung, J. S.—Zisserman, A.: VoxCeleb: A Large-Scale Speaker Identification Dataset. Proceedings of INTERSPEECH 2017, 2017, pp. 2616–2620, doi: 10.21437/Interspeech.2017-950.

[37] Chung, J. S.—Nagrani, A.—Zisserman, A.: VoxCeleb2: Deep Speaker Recognition. Proceedings of INTERSPEECH 2018, 2018, pp. 1086–1090, doi: 10.21437/Interspeech.2018-1929.

[38] Shor, J.—Jansen, A.—Maor, R.—Lang, O.—Tuval, O.—de Chaumont Quitry, F.—Tagliasacchi, M.—Shavitt, I.—Emanuel, D.—Haviv, Y.: Nonsemantic-Speech-Benchmark/Trill. Available online: `https://aihub.cloud.google.com/p/products%2F41239b97-c960-479a-be50-ae7a23ae1561`. [Accessed 2022].

[39] Barros, P.—Churamani, N.—Lakomkin, E.—Siqueira, H.—Sutherland A.—Wermter, S.: The OMG-Emotion Behavior Dataset. CoRR, 2018, doi: 10.48550/arXiv.1803.05434.

[40] Rusko, M.—Darjaa, S.—Trnka, M.—Ritomský, M.—Sabo, R.: Alert! ... Calm Down, There Is Nothing to Worry About. Warning and Soothing Speech Synthesis. Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014), Reykjavík, 2014, pp. 1182–1187, `http://www.lrec-conf.org/proceedings/lrec2014/pdf/722_Paper.pdf`.

[41] Guoth, I.—Rusko, M.—Ritomský, M.—Trnka, M.—Darjaa, S.: Identifying Tense Arousal in Speech Using Phase Based Features. Proceedings of Meetings on Acoustics, Vol. 30, 2017, No. 1, Art. No. 06005, doi: 10.1121/2.0000659.

Marián Trnka is a researcher and coder at the Institute of Informatics, Slovak Academy of Sciences. His expertise lies in programming, artificial intelligence, machine learning, natural language processing, speech synthesis, speech, speaker, and emotion recognition. He earned his Master's degree in cybernetics from the Slovak Technical University in Bratislava in 1994. He has authored more than 90 scientific papers and has contributed to applications such as the Automatic Dictation System for Judicial Domain (APD), automatic broadcast news subtitling and interactive fairy-tale reading application Readmio.

Sakhia Darjaa is a researcher and coder at the Institute of Informatics, Slovak Academy of Sciences. He specializes in programming, artificial intelligence, machine learning, natural language processing, speech synthesis, speech signal analysis, speaker, and emotion recognition. He holds his Master's degree in automatic systems of control from the Technical University in Košice, awarded in 1980, and obtained his Doctoral degree in telecommunications from the same institution in 2013. He is the author of more than 90 scientific papers and has been involved in several projects aimed at designing applications.

Róbert Sabo is a researcher and linguist affiliated with the Institute of Informatics, Slovak Academy of Sciences. His expertise encompasses linguistic analysis, phonetics, prosody research, speech corpus development, and expressive speech analysis. He earned his Master's degree in Slovak language and literature from the Comenius University in Bratislava in 2007. He completed his Doctoral degree in linguistics at the Ľudovít Štúr Institute of Linguistics of the Slovak Academy of Sciences in 2013. He is the author and co-author of over 40 scientific papers in the field of speech processing and computational linguistics.

Milan Rusko serves as the Head of the Department of Speech Analysis and Synthesis and is a project manager and researcher at the Institute of Informatics, Slovak Academy of Sciences. His expertise is in acoustics, artificial intelligence, speech synthesis, speech analysis, speaker recognition, and emotion recognition. He obtained his Master's degree in radio-electronics from the Technical University in Bratislava in 1984 and later earned his Doctoral degree in telecommunications from the Technical University in Košice in 2013. He has authored more than 120 scientific papers in the field of speech processing and has held positions on the Board of the Slovak Acoustic Society.

Meilin Schaper joined DLR's Institute of Flight Guidance in 1997 as an applications programmer and received her diploma in computer science from the University of Hagen in 2003. Since 2003 she worked on the concept and tool development of the DLR/Eurocontrol DMAN research prototype supporting tower air traffic controllers in departure sequence optimization. She was a project manager of the continued work, enhancing the tool (now called CADEO), and integrating and validating it in shadow mode trials. Further research topics include airport collaborative decision making, total airport management and combining planning systems to support operators and controllers seamlessly. Currently, her main research interest is ATM security.

Tim Stelkens-Kobsch received his diploma in aeronautical engineering from the University of Braunschweig in 2001, where he continued scientific activities as a researcher and teaching assistant for Air Traffic Management. He has extensive experience in the fields of guidance and control, controller assistance systems, and simulation and validation. In 2010 he joined the DLR's Institute of Flight Guidance in Braunschweig. Within the DLR he works on ATM-Simulation and Aviation Security and is responsible for the Generic Cockpit Simulator (GECO). He headed several validations regarding Air Traffic Management and has extensive experience in the management of international projects.