sciendo
DOI 10.2478/jazcas-2021-0053

# StressDat – DATABASE OF SPEECH UNDER STRESS IN SLOVAK

RÓBERT SABO[1] – ŠTEFAN BEŇUŠ[1,2] – MARIAN TRNKA[1] – MARIAN RITOMSKÝ[1] – MILAN RUSKO[1] – MEILIN SCHAPER[3] – JAKUB SZABO[4]
[1] Institute of Informatics, Slovak Academy of Sciences, Bratislava, Slovakia
[2] Constantine the Philosopher University, Nitra, Slovakia
[3] Institute of Flight Guidance, German Aerospace Center, Braunschweig, Germany
[4] Institute of Molecular Biomedicine, Faculty of Medicine, Comenius University, Bratislava, Slovakia

SABO, Róbert – BEŇUŠ, Štefan – TRNKA, Marian – RITOMSKÝ, Marian – RUSKO, Milan – SCHAPER, Meilin – SZABO, Jakub: StressDat – Database of speech under stress in Slovak. Journal of Linguistics, 2021, Vol. 72, No 2, pp. 579 – 589.

**Abstract:** The paper describes methodology for creating a Slovak database of speech under stress and pilot observations. While the relationship between stress and speech characteristics can be utilized in a wide domain of speech technology applications, its research suffers from the lack of suitable databases, particularly in conversational speech. We propose a novel procedure to record acted speech in the home of actors and using their own smartphones. We describe both the collection of speech material under three levels of stress and the subsequent annotation of stress levels in this material. First observations suggest a reasonable inter-annotator agreement, as well as interesting avenues for the relationship between the intended stress levels and those perceived in speech.

**Keywords:** speech database, speech under stress, stress annotation, inter-annotator agreement

## 1 INTRODUCTION

Areas of potential application for automatic speech technologies have been rapidly growing in the last decades. Despite great advances in statistical modelling and processing of speech, current state-of-the-art solutions still commonly use dedicated speech databases for specific domains of application. For example, in order to detect alcohol intoxication from speech, a system requires a specific database with speech under alcohol intoxication. In short, in order to be able to advance progress in the field of research and training of tools for automatic identification of speech expressiveness, it is necessary to build speech databases that will contain such speech expressions.

One such specific domain with great potential for making real world applications more effective and reliable is modelling stress based on speech characteristics. Understanding speech characteristics when people are under stress might help in mitigating stress-related problems and dangers in various situations,

such as air traffic control or crisis situations. This domain falls within a larger research area of modelling speech emotion and expressiveness ([1], [2], 3]). Recent years have witnessed immense progress in this modelling. This partly also results in the creation of and public access to several speech databases, which are specifically dedicated to research on emotions and expressiveness in speech ([4], [5]).

However, there is still a lack of suitable databases in speech under stress. The best-known database is SUSAS (Speech Under Simulated and Actual Stress [6]) that consists of four domains, encompassing a wide variety of stresses and emotions. It contains 32 speakers who have uttered more than 16,000 utterances. However, there are also limitations using SUSAS for training of tools for automatic identification of speech under stress and in performing acoustic analyses. Most sentences are one-word or two-word commands, there is often significant background noise, and the recordings have a low sampling frequency of 8 kHz.

Due to the lack of available corpora of naturalistic continuous speech containing stress level annotation, we decided to design and develop a dedicated Slovak database (StressDat) that would facilitate modelling speech under stress. The current paper describes our approach to the design, data collection, and processing. Specifically, section 2 describes the stimuli, section 3 speech recording, and section 4 the annotation of stress in the recordings. We also present first observations, particularly regarding the relationship between the intended and the perceived levels of stress in section 5. Section 6 presents further lines/directions of research and the conclusion.

## 2 StressDat STIMULI

### 2.1 Acted out vs. spontaneous speech

The first decision that had to be made was whether to use acted out or real speech. Naturally, spontaneous speech reflects the state of mind of the speaker in the best possible way. Additionally, some of the physical signs of stress are not possible to act out and, thus, their manifestation in speech is very difficult to imitate even for experienced actors.

On the other hand, a database of acted out speech in this domain has multiple advantages over spontaneous speech. First, using acted out speech has been a standard in the research of emotional and expressive speech for decades ([7], [8]). The main upshot is to have more control over the speech material and the level of stress. Spontaneous speech would have to be selected from various heterogeneous sources, situations, or recording conditions in a time-consuming effort. Acted out speech allows for recordings of more speakers in less time. Second, we can also control the situational context and limit the potential effects of a particular situation on the speech material in spontaneous speech. Rather, we can create specific stress situations that best reflect the intended use and coverage of the speech data. Third, the goal is to achieve balance in the amount of material for different levels of stress in speech, which would be immensely laborious to achieve with samples from speech under real stress. Fourth,

the quality of the acoustic signal is also important and it is essential to have comparable recordings regarding the same quality and the same amount of background noise. Finally, exposing participants to real stress might be ethically problematic while acting out a stressing situation does not pose these problems.

In addition to the above advantages of using acted out speech, we also tested the physiological indicators of stress when actors imitate a stressful situation. We asked an actor to read several sentences in neutral stress level and then imitate low, medium and high levels of stress. Our preliminary results suggest that even when an actor acts out a stressing scenario, physiological symptoms of real stress, such as increased skin conductivity and increased heart rate, can be observed. For each stress level, we measured average beats per minutes (BPM): neutral – 65.8, low – 65.4, medium – 68.05, high 72.7). These symptoms are, of course, weaker than when a person is actually exposed to a stressful situation. Figure 1 shows the temporal variability of skin conductivity G and heart rate HR during four acted out levels of stress (neutral, low, medium, high). For each of the three non-neutral stress levels, we can observe a typical step increase in both physiological indicators of stress, followed by their gradual decrease, indicating how the actor copes with this stress.

Hence, acted out speech is less natural than real speech, but a preferred option due to multiple practical considerations. Moreover, acted out speech is also linked to physiological indicators of stress in a similar way as real stress is expected to be linked to the same.

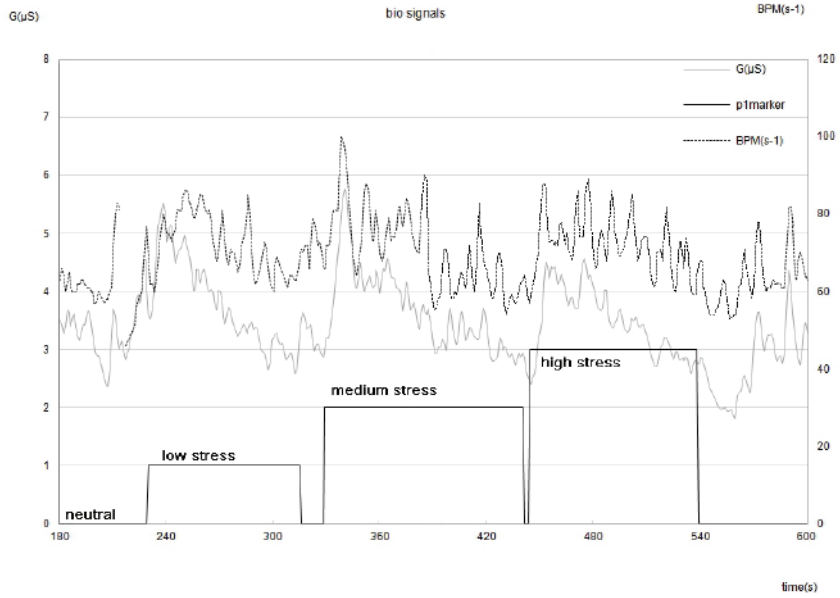


**Fig. 1.** Skin conductivity G (gray) and heart rate HR (dotted line) of the speaker acting out three levels of increased stress. Black solid line presents the stress level intended by the speaker

## 2.2 Stress levels

The decision to use acted out speech allows for some control over the levels of stress in the recording. Ideally, the database should cover as many levels of stress as people are able to produce systematically in a non-overlapping manner. We hypothesized that it would be two or three such levels.

This hypothesis is motivated in part by our prior research on developing a database of speech in crisis situations [9]. Naive subjects were able to produce three levels of activation corresponding to three levels of danger in a situation and the acoustic-prosodic characteristic in these three levels showed sufficient separation and reasonably acceptable overlap.

Additionally, in the pilot data shown in Figure 1, it can be observed that despite the typical step increase and gradual lowering of the two physiological indicators present for all three (low, medium, high) stress levels, the actual difference between the medium and high levels is non-existent in this particular speaker. Hence, while two levels of stress seem to be easily induced also in terms of physical indicators, three levels might be questionable and four rather unlikely.

Hence, given these considerations, and the fact that we planned to involve professional actors used to acting out various states of mind, we opted for three levels of stress (neutral, low, high).

## 2.3 Linguistic material

In order to obtain phonetically comparable recordings enabling measurements focused on acoustic and phonetic signs of stress level in speech, we created situations and associated textual material for acting out all three levels of stress (neutral, low stress and high stress). To capture the heterogeneity of stress and its manifestations in speech, we modeled 12 stressful situations. These could be grouped into three broad categories as follows: a) Threat of losing control over the situation, b) Psycho-social stress c) Threat to life/health or of an injury of self or the close ones.

Table 1 shows brief descriptions of these situations organized along three categories.

| Category | Nr. | Stress level | Description |
|---|---|---|---|
| Threat of losing control over the situation | 1 | neutral, medium, high | As an airline pilot you need to make an emergency landing. |
| | 2 | neutral, medium, high | Navigating a plane at the airport during very bad weather. |
| | 3 | neutral, medium, high | As a pilot, you need an undisciplined passenger to comply with the ban on using laptops during takeoff/landing. |
| | 4 | neutral, medium, high | As a firefighter coordinator, you organize firefighting in a burning building. |

| Category | Nr. | Stress level | Description |
|---|---|---|---|
| Psycho-social stress | 5 | neutral, medium, high | As a parent, you have to organize the morning routine for your kids before leaving for school. |
| | 6 | neutral, medium, high | You and your colleague are making last-minute changes to an important presentation with a colleague. |
| | 7 | neutral, medium, high | As a passenger, you need information on train departures urgently. |
| Threat of to life/ health injury of self/close ones | 8 | neutral, medium, high | You are calling an ambulance for your father who has suffered a stroke. |
| | 9 | neutral, medium, high | You are trying to pacify your drunk brother who is trying to forcefully enter your flat. |
| | 10 | neutral, medium, high | You are calling the police to resolve the situation with your drunken brother above. |
| | 11 | neutral, medium, high | As a pilot, you organize evacuation from a burning aircraft. |
| | 12 | neutral, medium, high | You are reporting an insurance event after a car accident by phone. |
| Neutral | 13 | neutral | You are talking about school with your son. |
| | 14 | neutral | You are buying shoes. |
| | 15 | neutral | You are teaching students at school. |
| | 16 | neutral | You are reading a text to a colleague. |

**Tab. 1.** Description of situations in the database

The goal was to achieve balance among different factors that might cause stress and different linguistic material for these factors. Each situation included between 10 to 13 sentences naturally expected in the given context and the sentences were created in a way that makes them appropriate for each stress level: neutral, under low stress, and under high stress. For example, a sample of sentences in situation #2 from Table 1 is shown in the following bulleted list.

- Gama 2305, it is important to quickly finish fueling the aircraft.
- Please speed up the loading of luggage, it is necessary to finish the loading of luggage as soon as possible.
- Runway number seven is not cleared from snow. Runway seven needs to be cleared. I repeat, runway seven needs to be cleared.
- The weather is getting worse, you need to take off immediately.
- ...

In addition to the 12 emotionally charged situations, we also included four 4 emotionally neutral situations with sentences corresponding to the neutral level of

stress only. These sentences were included since we cannot rule out the effect of text expressiveness on the neutral level of stress in the 12 expressive situations. These are shown at the bottom of Table 1. This inclusion will allow for testing the effect of linguistic material on the acoustic-prosodic rendering under the intended neutral level of stress (sentences corresponding to situations 1–12 vs. 13–16).

## 3 StressDat RECORDING

### 3.1 Speaker selection

To maintain the highest possible naturalness of elicited speech in the database, professional actors were recruited. The current pandemic situation facilitated recruiting of the actors since many of them experienced decreased demands on their time.

Currently the database includes 30 speakers (16 females, 14 males) who provided their recordings in exchange for payment. 20 speakers recorded the full battery of 16 situations in Table 1 and 10 speakers recorded 10 situations in 3 levels and 2 neutral situations.

### 3.2 Recording procedure

We needed to create a database of speech under stress at a time when people's face-to-face interactions were limited by the corona virus pandemic. For this reason, a novel procedure of database creation was developed. This allowed to not only achieve the required speech-under-stress recordings, but also to limit physical contact normally required in traditional speech elicitation protocols.

The goal was to utilize, and adjust if needed, the actors´ home environment and their own smartphones. We instructed the actors to select a room with the smallest possible reverberations, for example having as few bare walls and surfaces as possible, and make adjustments to further improve the acoustic environment, such as spreading the curtains, opening wardrobes, or covering sharp furniture edges. Additionally, instructions for positioning their smartphones during the recording were also given to ensure as comparable a recording environment across the speakers as possible.

The core of the instruction was to describe the three stress levels and facilitate actors' getting into the character. This was achieved in two ways. First, there were instructions regarding the three stress levels generally. For the neutral level, we asked them to imagine that they are completely calm, they navigate the situation with sufficient perspective, and that the situation does not affect their mental state in any adverse way. For the medium stress level, we asked them to imagine that they are under stress, that the situation is serious and its resolution should be done with care but assertively. For the high level of stress, we explained that they are under very high stress, that the situation is extremely critical and almost impossible to manage, and that they have to resolve it immediately.

Second, inspired by [10] for each situation, we specifically described a) the character (e.g., a parent of two schoolchildren that are difficult to manage), b) the situation (e.g., a Monday morning, an important meeting with grave consequences at work and parental duties involving the morning routine), c) the stressing factor (the family overslept and the kids are not cooperative), d) the goal (e.g., to manage to send the kids off to school and come to work on time), and e) the approach corresponding to three levels of stress (e.g., calmness usually works best with the kids (neutral), radio announces traffic jams and you need to be very efficient and effective with the kids (medium), you are very late, kid still doesn't behave, the situation is critical (high)).

At first, each actor recorded their first attempt at the two situations. We assessed both the acoustic quality of the recording and the differentiation of speech under the three stress levels. If adjustments were deemed necessary, they were communicated to the actor. The actors then proceeded with recording the full set of the situations.

## 4 StressDat ANNOTATION

After speech elicitation and processing, the annotation of the perceived level of stress in the recorded sentences was organized. A web-based speech stress assessment tool "Stress Thermometer" was designed based on the Subjective Units of Distress Scale [11], which allows the annotator to listen to the utterance and to assign a perceived stress level according to the instructions (see Figure 2). The visual representation of the thermometer was adopted from [12].

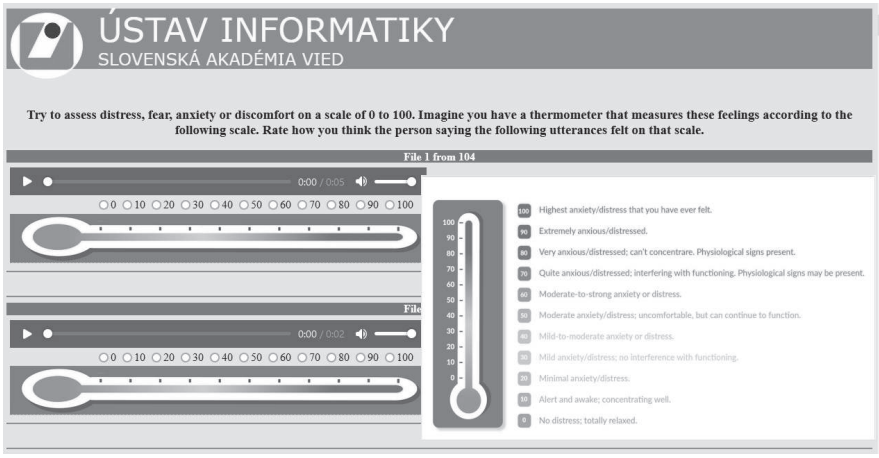


**Fig. 2.** The graphical user interface of the "Stress Thermometer" tool that allows the annotator to listen to the utterance and to assign a perceived stress level according to the verbal descriptions in the rightward panel [13]

Five annotators listened to each utterance and rated it on a discrete eleven-point scale according to the following instruction: “Try to assess distress, fear, anxiety, or discomfort on a scale from 0 to 100. Imagine you have a thermometer that measures these feelings on such scale. Rate how you think the person saying the following utterances felt on that scale.” Each utterance can thus be characterized with the mean of the values of the perceived stress level assigned by the annotators. The ratings of perceived stress can, therefore, reach real number values in the interval 0 to 100 in steps of 10. This allows regression to be used in stress assessment instead of classification. To limit the influence of the speaker, the annotators evaluated sentences from different speakers in random order. During the evaluation, each annotator had a different order of sentences in order to minimize the influence of the previously heard sentences on the evaluation.

Of the material material recorded by 30 actors, two thirds have been fully annotated and the rest is currently approaching completion.

## 5 PILOT OBSERVATIONS

### 5.1 Annotation normalization

It is common in annotating tasks using a scale that annotators use the scale in different ranges and variances. To normalize for this variability, we use z-score normalization [14] by annotator.

Figure 3 shows that normalizing annotations makes sense. Consider the neutral (Level 1) stress for raters a1–2 vs. a3–4. It is clear that the ratings of a1–2 are shifted lower compared to a3–4 in all three levels. Hence, the stress level was perceived similarly, only the first group used the lower range of the scale compared to the second group. This similarity among raters is reflected in the right panel after normalization. The figure also shows consistent and robust separation among the three stress levels in the annotations.

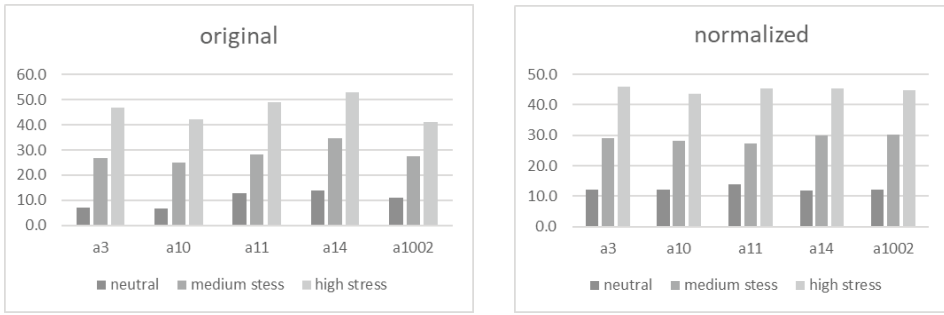


**Fig. 3.** Mean stress assessments for the three stress levels (neutral, medium, high) for five annotators (a1–a5) before (left) and after (right) normalization

### 5.2 Inter-annotator agreement

To find out the degree of agreement among all annotators using the 11-point scale, we calculated Fleiss' kappa [15] for the original and the normalized assessments, see the mean values for all utterances in Table 2. The values between 0.2 and 0.4 are considered a fair agreement. Given the subjective nature of stress perception, and as many as 11 discrete points, we consider this agreement reasonably good for this task.

It should be kept in mind, however, that Fleiss' kappa considers the discrete rating as independent of each other and penalizes a one-step difference between two annotators (e.g. 2–3) in the same way as a seven-point difference (2–9). To capture the distance among annotators, we also calculated variance and standard deviation for each sentence; average of values for all sentences is shown in Table 2.

| Original annotations | | | Normalized annotations | | |
|---|---|---|---|---|---|
| Fleiss | Variance | stdev | Fleiss | Variance | stdev |
| 0.31 | 1.22 | 9.71 | 0.35 | 0.63 | 7.84 |

**Tab. 2.** Evaluation of inter-annotator agreement

### 5.3 Intended vs. perceived level of stress

Figure 4 shows how the intended levels of stress produced by the actors corresponds to the levels of stress perceived by the annotators. We plotted average stress ratings for three levels of stress and three categories of stressful situations from Table 1 in section 2.3. The figure provides several initial observations. First, the ratings show that the actors were consistently successful in separating the three levels of stress. Second, there is a difference between sentences in completely neutral situations and the fourth bar of Level 1 and the other three bars, i.e., those acted out in a neutral way but including stress semantically. This difference may stem either from the effect of text semantics on the actors, the annotators, or both. Third, the situations grouped under psycho-social stress are perceived/produced as less stressful than the situations in the other two groups consistently at all three stress levels. We may speculate that the nature of these situations (at home with kids or at work with a colleague) elicits lower stress levels either due to the less severe stressors, or certain amount of control over the situation compared to the other two groups involving less control and greater severity.

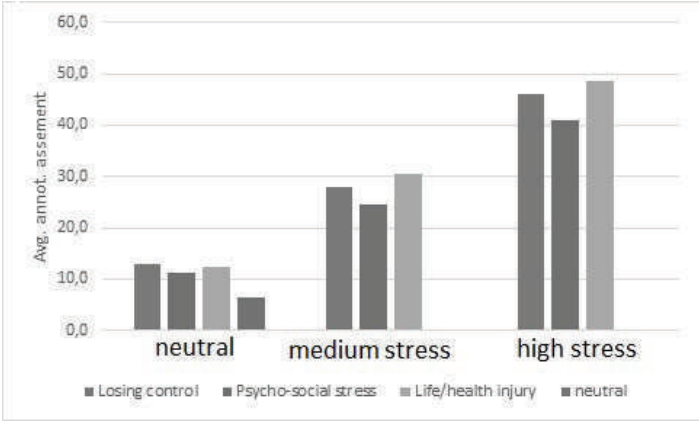


**Fig. 4.** Comparison of the average evaluation of annotators in relation to the played level of stress and the classification of situations into groups

## 6 DISCUSSION AND FUTURE WORK

The sampling of both the actors and the annotators provides richness and variability in that each utterance from the corpus is produced by multiple speakers and its stress level is assessed by multiple annotators. Thus, the information about the intended level of stress in speech production and the associated perceived level of stress for each utterance of StressDat provide the basis for developing the statistical models predicting the level of stress in speech.

The complete database will contain 30 speakers, and will be accessible for research purposes.

## ACKNOWLEDGMENTS

This work is supported by European Union's Horizon 2020 research and innovation programme under grant agreement number 832969, project SATIE "Security of Air Transport Infrastructure of Europe". This output reflects the views only of the author(s), 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 SATIE project see: http://satie-h2020.eu/. This work was also funded by the Slovak Scientific Grant Agency VEGA, grant number 2/0161/18.

## References

[1] Yogesh, C. K. et al. (2017). Bispectral features and mean shift clustering for stress and emotion recognition from natural speech. In Computers & Electrical Engineering, 62, pages 676–691.

[2] Robinson, C., and Nicolas, R. A. (2019). Sequence-to-sequence modelling of f0 for speech emotion conversion. In ICASSP 2019, pages 6830–6834.
[3] Rusko, M., Trnka, M., Darjaa, S., Stelkens-Kobsch, T., and Finke, M., (2018). Weaknesses of voice biometrics – sensitivity of speaker verification to emotional arousal. In ICSV25: 25th International Congress on Sound and Vibration. Hiroshima, Japan, pages 1–8.
[4] Alimuradov, A. K. et al. (2020). Development of Natural Emotional Speech Database for Training Automatic Recognition Systems of Stressful Emotions in Human-Robot Interaction. In 4th Scientific School on Dynamics of Complex Networks and their Application in Intellectual Robotics (DCNAIR), pages 11–16.
[5] Busso, C. et al. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. In Language resources and evaluation, pages 335–359.
[6] Hansen, J. H. et al. (1997). Getting started with SUSAS: a speech under simulated and actual stress database. In Eurospeech, 97(1), pages 1743–1746.
[7] Burkhardt, F. et al. (2005). A database of German emotional speech. In Ninth European Conference on Speech Communication and Technology.
[8] Campbell, N. (2000). Databases of emotional speech. In ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion.
[9] Rusko, M., Darjaa, S., Trnka, M., Sabo, R., and Ritomský, M. (2015). Expressive Speech Synthesis for Critical Situations. COMPUTING AND INFORMATICS, 33(6), pages 1312–1332.
[10] Enos F., and Hirschberg J. (2006). A framework for eliciting emotional speech: Capitalizing on the actors process. In First International Workshop on Emotion: Corpora for Research on Emotion and Affect LREC 2006, Genoa, Italy, pages 6–10.
[11] Wolpe, J. (1969). The Practice of Behavior Therapy. Pergamon Press, 314 p.
[12] Accessible at: https://ccp.net.au/suds-thermometer/.
[13] SATIE project: "D4.2 – Traffic Management Intrusion and Compliance System", Status: submitted.
[14] Accessible at: https://en.wikipedia.org/wiki/Standard_score.
[15] Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. Psychological Bulletin, 76(5), pages 378–382.