Interspeech 2024
1-5 September 2024, Kos, Greece

BESST Dataset: A Multimodal Resource for Speech-based Stress Detection and Analysis

Jan Pešán[1], *Vojtěch Juřík*[2], *Martin Karafiát*[1], *Jan Černocký*[1]

[1]Faculty of Information Technology, Brno University of Technology, Brno, Czech Republic
[2]Faculty of Civil Engineering, Brno University of Technology, Brno, Czech Republic

ipesan@fit.vutbr.cz, jurik.vojtech@gmail.com, cernocky@fit.vutbr.cz, karafiat@fit.vutbr.cz

Abstract

The Brno Extended Stress and Speech Test (BESST) dataset is a new resource for the speech research community, offering multimodal audiovisual, physiological and psychological data that enable investigations into the interplay between stress and speech. In this paper, we introduce the BESST dataset and provide a details of its design, collection protocols, and technical aspects. The dataset comprises speech samples, physiological signals (including electrocardiogram, electrodermal activity, skin temperature, and acceleration data), and video recordings from 90 subjects performing stress-inducing tasks. It comprises 16.9 hours of clean Czech speech data, averaging 15 minutes of clean speech per participant. The data collection procedure involves the induction of cognitive and physical stress induced by Reading Span task (RSPAN) and Hand Immersion (HIT) task respectively. The BESST dataset was collected under stringent ethical standards and is accessible for research and development.

Index Terms: BESST dataset, stress recognition, multimodal data, speech research, physiological signals, cognitive load, speech production

1. Introduction

The estimation of cognitive and physical load through speech is an ongoing challenge in the field of speech research. As various applications emerge that require a deep understanding of human behavior through speech, the need for specialized and comprehensive datasets has become evident. In response to this need, we present the Brno Extended Stress and Speech Test (BESST), a dataset aimed at facilitating further research in this area.

Existing datasets such as Speech under Stress Conditions-0 (SUSC-0) and Speech under Stress Conditions-1 (SUSC-1)[1] are primarily military-centric and lack stress load labels, restricting their applicability in broader contexts. The Speech Under Simulated & Actual Stress Database (SUSAS) dataset[2],[3], while accessible, falls short in providing data that captures cognitive load in natural conversational settings. The Munich Biovoice Corpus (MBC) dataset [4], has concerns related to generalization due to its focus on specific fixed vocabulary. Cognitive Load with Speech and EGG (CLSE) [5], meanwhile, emphasizes cognitive load tasks but does not provide a framework for identifying physical stress. These limitations in generalization, replicability, and authenticity are significant challenges that BESST aims to address. Our dataset builds upon these foundational datasets while introducing a novel multimodal framework that facilitates a more comprehensive investigation of the stress phenomenon.

This paper intends to provide an overview of the BESST dataset. The specific focus on using speech for cognitive and physical load estimation is presented, followed by an overview of the dataset's structure, methodology, and potential applications.

2. BESST Protocol

2.1. Participants

The participant group comprised 90 volunteers (21F, 69M), primarily Caucasian young adults. Technical issues and data loss led to 79 participants (19F, 60M) forming the final dataset. Recruitment targeted individuals aged 19 to 26, facilitated through social networks, university platforms, posters, and emails. Participants met specific criteria, excluding those with epilepsy, heart conditions, acute health issues, or non-Czech native language speakers.

2.2. BESST Stress Induction Protocol

The BESST is a modified version of the Maastricht Acute Stress Test (MAST) [6]. Adjustments aim to maximize speaker's speech output. The Hand Immersion Task (HIT) and Mental Arithmetic Task (MAT) components of MAST were expanded.

2.2.1. Hand Immersion Task (HIT)

The HIT is a core component of the BESST experimental protocol designed to induce physical stress. During this task, participants are instructed to immerse their non-dominant hand, including the wrist joint, in a container of ice-cold water (approximately 4-8°C) while describing images displayed on a computer screen. The goal is to elicit a stress response by exposing participants to a discomforting physical stimulus. The images are presented for a variable duration, up to 90 seconds per image, and participants are required to provide detailed descriptions using at least three sentences. This task is inspired by the Cold Pressor Test (CPT) [7], a well-established method for inducing physical stress through cold exposure. In the BESST dataset, the HIT comprises a series of trials, each involving different sets of images and immersion times, to capture varying stress levels.

2.2.2. Reading Span Task (RSPAN)

RSPAN is a component of the BESST experimental protocol designed to induce cognitive load. In this task, participants are presented with short paragraphs containing emoticons that replace certain words. Participants are required to read the paragraphs out loud, including the replaced words, and memorize the last word of each paragraph. Subsequently, they reproduce the memorized words in the correct order at the end of each segment. The complexity of the task gradually increases as par-

10.21437/Interspeech.2024-42

ticipants progress through different segments, involving paragraphs with varying numbers of emoticons and sentences.

2.3. Procedure

The BESST session consists of segments shown in Table 1. Each run lasts about 43 minutes, led by two instructors.

Table 1: *Session run.*

Action	Duration [mm:ss]	Materials
Instrumentation	7:00	Consent signature Headset Empatica E4 Faros 180
Introduction	2:00	PowerPoint Introduction Questions
Evaluation 1	2:00	
Relax 1	3:00	
Evaluation 2	2:00	PSS14 STAI-Y1
Experiment run	13:00	See details in Table 2
Evaluation 3	3:00	
Relax 2	5:00	STAI-Y1 NASA-TLX
De-briefing	5:00	Exit interview
De-instrumentation	1:00	
Total	43:00	

- **Instrumentation and Introduction:** Participants sign consent forms and are fitted with sensors. General info is provided, questionnaires filled, and relaxation follows.
- **Experiment run:** Participants practice without ice water (HIT and Reading Span task (RSPAN) dry-runs). They then complete 5 HIT and 4 RSPAN tasks alternately, lasting 1-1.5 minutes each. State-Trait Anxiety Inventory-Y1 (STAI-Y1) and relaxation follow the last HIT task. Detailed breakdown of the experiment run can be seen in Table 2
- **Debriefing and de-instrumentation** After Experiment run, participants are debriefed, where they are explained the purposes of the study, and any questions or concerns are addressed.

2.4. Psychological Measurements

Psychological assessments in the BESST protocol include three questionnaires: the Perceived Stress Scale 14 (PSS14) [8], State-Trait Anxiety Inventory-Y2 (STAI-Y2) [9], and NASA Task Load Experience (NASA-TLX) [10]. These tools were used to gauge the mental states and psychological traits of the participants about stress coping.

2.5. Data Streams

The dataset captures multiple data streams from different modalities. Table 3 provides an overview of the available data streams, including details about the devices used, sampling rates, and file formats.

2.6. Language specifics

The dataset was captured at the Faculty of Information Technology - Brno University of Technology (FIT-BUT). To ensure realism, native Czech participants were chosen. Consequently, the entire dataset is in Czech. However, we anticipate no influence on our intended use case, as cognitive load and physical stress effects are expected to be language agnostic.

Table 2: *Experiment run.*

Action	Duration [mm:ss]	Material
HIT 0 - Dry-run	1:30	
RSPAN 0- Dry-run	1:00	Rebus with 2 paragraphs
HIT 1	1:30	Ice cold bucket Pictures
RSPAN 1	1:00	Rebus with 3 paragraphs
HIT 2	1:00	Ice cold bucket Pictures
RSPAN 2	1:15	Rebus with 4 paragraphs
HIT 3	1:00	Ice cold bucket Pictures
RSPAN 3	1:30	Rebus with 4 paragraphs
HIT 4	1:30	Ice cold bucket Pictures
RSPAN 4	1:45	Rebus with 5 paragraphs
HIT 5	1:00	Ice cold bucket Pictures

3. BESST Dataset

Dataset is organized into subfolders for each modality. In there, another subfolder contains data per each participant. Detailed breakdown of the dataset content is in Table 4. We define two types of the data-streams: **Native** streams are as close as possible to the data source e.g. direct recordings from the audio recorder. **Derived** are created using the Native streams and postprocessed in non-trivial way, creating separate datastream. Different numbers of valid participants per datastream are caused mostly by protocol error and/or recording device malfunction.

For ease of use, the dataset is split into 5 different archives: audio data with Voice Activity Detection (VAD) segmentations (see Section 4.2); each type of video streams; biological, psychological and semantic segments. This way, it will be easier for the user to download only relevant parts of the database (since video data is especially voluminous).

3.1. Data Subsets

We extracted two distinct subsets from the experimental data: one focusing on cognitive load and the other on physical stress.

3.1.1. Cognitive Load Subset

The structure of the Cognitive Load subset was intentionally aligned with that of the CLSE dataset to facilitate possible comparative analyses. Given the sensitivity of paralinguistic tasks to lexical content overlap between data splits, we employed carefully designed train/validation/test splits to mitigate this effect. While this approach introduces some degree of speaker leakage across the splits, we contend that this does not significantly impact the target use case, as speaker identity should not significantly influence cognitive load levels. For each list, 5-fold random splits are created. Due to varying availability of modalities across participants, we created three list variants: **audio**, **audio+video** and **audio+video+bio**. Each variant includes the respective available data types. A detailed breakdown is provided in Table 5.

3.1.2. Physical Stress Subset

The Physical Stress subset adopted the labeling methodology employed in the MBC [4], thereby ensuring a consistent approach to capturing speech patterns induced by physical load. Similar to the Cognitive Load Subset, we carefully addressed

Table 3: *Available data streams*

Modality	Details	Count	Device	Format
Audio	Main stream, 48000 Hz, 2 channels, 24 bit, Pulse Code Modulation (PCM)	2	Zoom H4n	WAV
Audio	Auxiliary stream 48000 Hz, 2 channels, 16 bit, AAC	4	Panasonic HC-VX9805	WAV
Video	(face, left posture, right posture, back), 1080p, 25 fps, h.264	4	Panasonic HC-VX9805	MP4
Electro-Dermal Activity (EDA)	Auxiliary skin conductance, 4 Hz	1	Empatica E4	CSV
Temperature	Auxiliary skin temperature, 4 Hz	1	Empatica E4	CSV
Acceleration	Acceleration (x,y,z), 32 Hz	3	Empatica E4	CSV
Heart activity	RR (Interbeat) intervals	1	Empatica E4	CSV
Acceleration	Acceleration (x,y,z), 250 Hz	3	Faros 180	CSV
Heart activity	Electrocardiogram (ECG), 1000 Hz, 24 bit	1	Faros 180	CSV

Table 4: *Valid data - Kind is either N - Native, or D - Derived*

Data	# Part.	Kind	File name
Audio - tabletop	79	N	tabletop.wav
Audio - close-talk	79	N	close_talk.wav
Video - face	79	N	face.mp4
Video - left posture	77	N	left.mp4
Video - right posture	75	N	right.mp4
Video - back	72	N	bucket.mp4
EDA - wrist	76	N	EDA.csv
Skin Temperature - wrist	76	N	TEMP.csv
Acceleration - wrist	75	N	ACC2.csv
Acceleration - chest	76	N	ACC.csv
Heart activity - ECG	76	N	ECG.csv
Heart activity - RR	76	D	RR.csv
Voice Activity Detection	79	D	vad-segments.csv
Semantic segmentation	72	D	segments.csv
Questionaire - self-reported	90	N	psychoload.csv
Rebus performance	90	N	results.csv

lexical overlap and speaker leakage concerns, as discussed earlier. Dataset structure is similar to the Cognitive load subset. Detailed breakdown is provided in Table 6.

4. Validation and processing

4.1. Validation Process for Each Modality

To ensure accurate data representation, each modality underwent a thorough validation process:

- **Audio:** Our careful setup of microphones and recording devices maintained the fidelity of audio signals. Regular quality checks were performed to address any potential issues.
- **Video:** Cameras were manually time synchronized with the precision of ± 1 second as our camera models does not have external time signal input. This synchronization facilitated rudimentary alignment with other modalities.
- **Physiological Signals:** Rigorous validation protocols, including calibration and testing, were applied to physiological sensors (ECG, EDA, accelerometers) to guarantee reliable recordings.
- **Psychological Assessments:** Standardized procedures were followed for psychological questionnaires, yielding consistent self-reported stress and cognitive load scores.

4.2. Preprocessing

Data quality was upheld through comprehensive preprocessing techniques:

- **Audio:** We conducted speaker-level volume normalization to maintain uniform audio levels within speaker sessions. This ensured clarity without compromising natural speech variations.
- **Video:** Timestamp alignment synchronized video streams with other modalities, enabling accurate association of visual cues with physiological and psychological responses.
- **Physiological Signals:** Noise and artifact removal enhanced physiological data quality. Outliers were identified and removed, providing trustworthy continuous data.
- **ECG Signal Preprocessing:** A robust QRS complex detection algorithm was applied, employing three independent detectors based on different principles [11]. Manual verification by an ECG expert ensured precise QRS complex identification. Detected positions were carefully corrected for accuracy.
- **Psychological Assessments:** Thorough quality checks were performed on psychological questionnaire responses, capturing participants' self-reported psychological states.

4.3. Synchronization of Data Streams

All data streams are time synchronized. The time t_0 is set at the moment when the audio recording from the Zoom recorder started. The Zoom recorder and video cameras have no explicit way to programmatically set the correct time from an external source (i.e., Network Time Protocol (NTP)). Instead, we manually set each device's time with an acceptable margin of error of 1 second. The Empatica E4 device implicitly synchronizes its internal Real-Time Clock (RTC) each time data is downloaded. The Faros 180 allows explicit synchronization of its RTC when connected to its software. In our pilot trial, the Empatica E4 bracelet served as the reference time source. Although we did not use its event marker function for experiment start, accidental markers visible on video were utilized to calculate relative shifts in camera clocks for video alignment. Audio-to-video synchronization was achieved using the Audalign[1] project, extracting relative lag from cross-correlation and fingerprinting of the audio streams. The Faros 180's RTC exhibited time drift issues, leading to incorrect timestamps. We corrected these timestamps by cross-correlating RR-intervals between Faros ECG and Empatica E4 RR-intervals per participant.

4.4. Speech recognition system

Voice Activity detection (VAD) was based on a simple feed-forward NN with two outputs and two layers with 400 neurons. Standard Mel-filter bank features (15 coefficients) with F0 estimates (3 coefficients) [12] were taken as the input after cepstral mean and variance normalization.

Following Automatic Speech Recognition (ASR) system was based on hybrid Deep Neural Network-Hidden Markov Model (DNN-HMM) and trained with Kaldi toolkit [13]. The architecture followed the one published in [14] but trained on

[1] `https://github.com/benfmiller/audalign`

Table 5: *Cognitive load dataset duration and label distribution*

Modality	# Part.	# Segments	Total (s)	Low [L1] (s)	Med [L2] (s)	High [L3] (s)
Audio	79	1021	10543.64	3162.93	3221.74	4158.98
Audio-Video	79	1021	10543.64	3162.93	3221.74	4158.98
Audio-Video-Bio	72	932	9572.39	2874.74	2924.86	3772.79

Table 6: *Physical load dataset duration and label distribution*

Modality	# Part.	# Segments	Total (s)	No load[L0] (s)	Load [L1] (s)
Audio	79	2915	40078.86	17071.91	23006.95
Audio-Video	79	2915	40078.86	17071.91	23006.95
Audio-Video-Bio	72	2642	36166.08	15287.60	20878.48

Czech data (∼2300 hours).

We utilized the 1-best word transcriptions obtained from the Automatic Speech Recognition (ASR) output and merged them with permissible gaps of up to 2 seconds. Segments shorter than 3 seconds were discarded. This process resulted in 4753 segments, with a median duration of 11.6 seconds, contributing to a total of 16.9 hours of clean speech data.

4.5. Semantic Labeling

The dataset includes manual semantic labels that annotate distinct parts of the whole seesion (see Table 1 and 2 for semantic segments details). These labels offer context for interpreting recorded signals and identifying stress-related events. Manual semantic labeling of segments was carried out using our BESS-TiANNO tool.

5. Availability, Ethical Considerations and Potential Applications

The BESST dataset, along with relevant code and tools, is accessible for research and development purposes for free at https://speech.fit.vutbr.cz/besst. To access the dataset, researcheres needs to fill in a request form and sign the provided license agreement.

Ethical considerations played a pivotal role in the dataset collection. This dataset collection was assessed and approved by the Ethical Committee Faculty of Electrical Engineering and Communication For Biomedical Research, Brno University of Technology, Brno, Czech Republic under number EK:02b/2022.

The BESST dataset's unique multimodal characteristics present an array of possibilities for advancing speech research:

- **Speech and Physical Stress:** By aligning physiological and psychological states with speech responses, investigations into stress-induced speech variations become feasible.
- **Speech and Cognitive Load:** The dataset enables the examination of how speech attributes are influenced by cognitive load.
- **Multimodal Interaction:** Researchers can explore the interplay between different modalities, analyzing how speech interacts with physiological and psychological markers.
- **Real-world Applications:** BESST's multimodal insights have the potential to be translated into real-world applications, such as stress detection systems, human-computer interaction, and virtual communication environments.

6. Conclusion and Future Work

The BESST dataset bridges the gap between speech and physiological responses, offering a valuable resource for multimodal speech research. By facilitating investigations into the intricate connections between stress, cognitive load, and speech, the dataset contributes to various domains, from affective computing to healthcare applications. As research in multimodal communication continues to evolve, the BESST dataset enables researchers to unravel the complex interplay of physiological, psychological, and speech components.

In future work, we aim to explore the potential of automatic stress detection within the context of large pre-trained models, such as HuBERT [15], Wav2Vec2 [16], and others.

7. Acknowledgements

The work was partly supported by BUT IGA project no. FIT-S-23-8278. The creation of the BESST dataset was made possible through the collaborative efforts of numerous individuals and institutions. We would like to express our gratitude to the Grey Lab research infrastructure held at theDepartment of Psychology, Faculty of Arts, Masaryk University, Brno, for providing essential support during the data collection process. We extend our appreciation to Tom Smeets, the author of the MAST protocol, for generously providing his materials and guidelines for executing the protocol, which served as a foundational basis for the BESST protocol. We also acknowledge the valuable contributions of the Aeroworks group at FIT-BUT for their exploratory work on pilot state estimation from speech applicable to ground-based synthetic flight training platforms.

We are thankful to all the participants who volunteered their time and effort to take part in the data collection sessions.

8. References

[1] J. H. L. Hansen and M. A. Clements, "Evaluation of speech under stress and emotional conditions," *The Journal of the Acoustical Society of America*, vol. 82, no. S1, pp. S17–S18, Nov. 1987. [Online]. Available: https://doi.org/10.1121/1.2024686

[2] J. H. L. Hansen, "Susas ldc99s78," https://catalog.ldc.upenn.edu/LDC99S78, 1999, last accessed on 2022-04-15.

[3] ——, "Susas transcripts ldc99t33," https://catalog.ldc.upenn.edu/LDC99T33, 1999, last accessed on 2022-04-15.

[4] B. Schuller, F. Friedmann, and F. Eyben, "The munich biovoice corpus: Effects of physical exercising, heart rate, and skin conductance on human speech production," in *International Conference on Language Resources and Evaluation*, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:1500634

[5] T. F. Yap, "Speech production under cognitive load: Effects and classification," Ph.D. dissertation, The University of New South Wales, 2012.

[6] T. Smeets, S. Cornelisse, C. W. Quaedflieg, T. Meyer, M. Jelicic, and H. Merckelbach, "Introducing the maastricht acute stress test (mast): A quick and non-invasive approach to elicit robust autonomic and glucocorticoid stress responses," *Psychoneuroendocrinology*, vol. 37, no. 12, pp. 1998–2008, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S030645301200162X

[7] E. A. Hines, "A standard stimulus for measuring vasomotor reactions: its application in the study of hypertension," in *Mayo Clin Proc*, vol. 7, 1932, pp. 332–335.

[8] S. Cohen, T. Kamarck, and R. Mermelstein, "A global measure of perceived stress," *Journal of Health and Social Behavior*, vol. 24, no. 4, p. 385, Dec. 1983. [Online]. Available: https://doi.org/10.2307/2136404

[9] A. G. Hedberg, "Review of state-trait anxiety inventory." *Professional Psychology*, vol. 3, no. 4, pp. 389–390, 1972. [Online]. Available: https://doi.org/10.1037/h0020743

[10] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (task load index): Results of empirical and theoretical research," in *Advances in Psychology*. Elsevier, 1988, pp. 139–183. [Online]. Available: https://doi.org/10.1016/s0166-4115(08)62386-9

[11] L. Smital, L. Marsanova, R. Smisek, A. Nemcova, and M. Vitek, "Robust qrs detection using combination of three independent methods," in *2020 Computing in Cardiology*, 2020, pp. 1–4.

[12] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. Florence, Italy: IEEE, May 2014.

[13] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[14] M. Kocour, J. Umesh, M. Karafiát, J. Švec, F. Lopez, K. Beneš, M. S. Diez, I. Szőke, J. Luque, K. Veselý, L. Burget, and J. Černocký, "Bcn2brno: Asr system fusion for albayzin 2022 speech to text challenge," in *Proceedings of IberSpeech 2022*. International Speech Communication Association, 2022, pp. 276–280. [Online]. Available: https://www.fit.vut.cz/research/publication/12859

[15] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," 2021.

[16] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020.