The INTERSPEECH 2014 Computational Paralinguistics Challenge: Cognitive & Physical Load*

Björn Schuller[1,2], *Stefan Steidl*[3], *Anton Batliner*[2,3], *Julien Epps*[4,5]
Florian Eyben[2], *Fabien Ringeval*[2], *Erik Marchi*[2], *Yue Zhang*[2]

[1]Imperial College London, Department of Computing, Machine Learning Group, UK
[2]TU München, Machine Intelligence & Signal Processing group, Germany
[3]FAU Erlangen-Nürnberg, Pattern Recognition Lab, Germany
[4]The University of New South Wales, Australia
[5]National ICT Australia, Australia
bjoern.schuller@imperial.ac.uk

Abstract

The INTERSPEECH 2014 Computational Paralinguistics Challenge provides for the first time a unified test-bed for the automatic recognition of speakers' cognitive and physical load in speech. In this paper, we describe these two Sub-Challenges, their conditions, baseline results and experimental procedures, as well as the COMPARE baseline features generated with the openSMILE toolkit and provided to the participants in the Challenge.

Index Terms: Computational Paralinguistics, Challenge, Cognitive Load, Physical Load

1. Introduction

So far, there have been five consecutive paralinguistic challenges at INTERSPEECH since 2009; cf. the challenge series' repository at http://www.compare.openaudio.eu, chapter 6.2 in [1], and [2]. They covered short-term states (emotion in 2009, interest in 2010), short-term events (laughter and conflict in 2013), medium-term states (intoxication and sleepiness in 2011), long-term traits (personality in 2012), atypical traits (autism in 2013), and biological trait primitives (age and gender in 2010). The Interspeech 2014 COMputational PARalinguistics ChallengE (COMPARE) is again an open challenge, dealing with – relatively – short-term states of speakers as manifested in the acoustic properties of the speech signal. The Cognitive-Load with Speech and EGG database (CLSE) and the Munich Bio-voice Corpus (MBC) covering different languages (Australian English and German) are provided by the organisers. CLSE features Australian speakers recorded during different cognitive load. MBC contains speech under physical exercising; heart rate and skin conductance were measured by sensors. Two Sub-Challenges are addressed:

In the *Cognitive Load Sub-Challenge*, three levels of cognitive load have to be classified automatically, based on acoustics (ternary classification).

In the *Physical Load Sub-Challenge*, the binary exercising state (running / resting) and by that the heart rate state (high pulse / low pulse) have to be classified automatically.

* The research leading to these results has received funding from the European Community's Seventh Framework Programme under grant agreements No. 338164 (ERC Starting Grant iHEARu) and No. 289021 (STREP ASC-Inclusion). The authors would further like to thank the sponsor of the Challenge, the Association for the Advancement of Affective Computing (AAAC). The responsibility lies with the authors.

The measure of competition will be Unweighted Average Recall (cf. below). The orthographic transcription of the train and development sets will be known. Both Sub-Challenges allow contributors to find their own features with their own machine learning algorithm. However, a standard feature set will be provided that may be used. Participants will have to abide by the definition of training, development, and test sets. They may report on results obtained on the development set, but have only five trials to upload their results on the test sets, whose labels are unknown to them. Each participation will be accompanied by a paper presenting the results that undergoes peer-review and has to be accepted for the conference in order to participate in the Challenge. The organisers preserve the right to re-evaluate the findings, but will not participate themselves in the Challenge.

In the following we introduce the Challenge corpora (Section 2) and describe the COMPARE baseline features (Section 3) and the baseline results (Section 4) before concluding (Section 5).

2. Challenge Corpora

2.1. Cognitive Load with Speech and EGG (CLSE)

In the *Cognitive Load Sub-Challenge*, the "Cognitive Load with Speech and EGG" (CSLE) database [3] serves to evaluate features and algorithms for determining cognitive load and working memory of speakers; it was recorded in Sydney/Australia using a close-talk microphone sampled at 16 kHz from 26 native Australian English speakers (20 male and 6 female).

'Working memory' refers to the limited temporary information store of the brain [4], and this is often investigated by 'span' tasks, which require participants to remember a number of concepts or objects in the presence of distractors [5]. The reading span task [6] employed in this database required participants to read a series of possibly illogical short sentences (e. g., "I like to walk in the sky"), indicate whether each was true or false, and then remember a single letter presented briefly between sentences [7]. The number of sentences presented in each set, after which participants were asked to recall all letters shown, varied from two to five. The working memory load level labels for this database were low (L1) after the first sentence, medium (L2) after the second sentence, and high (L3) after the third, fourth, and fifth sentence. This task contains 75 utterances per speaker, spread across 3 working memory load levels, each with an average utterance duration of 4 s.

'Cognitive load' refers to the load that a task places on

Table 1: *Partitioning of the Cognitive Load with Speech and EGG (CLSE) database into train(ing), dev(elopment), and test sets. Sum of number of utterances over the three tasks participants had to perform for each load level (L1–L3), excluding the reading span task letter recordings.*

#	train	dev	test	Σ
L1	297	189	216	702
L2	297	189	216	702
L3	429	273	312	1 014
Σ	1 023	651	744	2 418

the cognitive system [8], which may comprise both working memory and central executive components. The cognitive load tasks employed in this database – variants of the Stroop test [9] – required participants to name the font colour of words corresponding to different colour names. In the low level (L1), the font colours and the colour words were congruent, while in the medium and high levels (L2 and L3), they were incongruent. In the Stroop time pressure task (after [10, 11]), the high level required participants to name the font colour in a short period of time (0.8s), while in the Stroop test with dual task, participants were required to perform a tone-counting task in the high level, as well as naming the font colour. As part of the data collection, participants were also asked to perform a post-task subjective rating of the load they experienced, similarly to other studies of this kind [12]. Analysis showed that the ratings exhibited statistically significant differences across different load levels, using a paired t-test with a Bonferroni-adjusted significance level of 0.025 [3]. The Stroop test with time pressure and the Stroop test with dual task each contain three utterances for each of three cognitive load levels per speaker, with average utterance durations of 17 s and 21 s, respectively.

The CLSE database also includes a story reading task, in which participants read aloud a passage about smoke detectors. This produced recordings of neutral speech with an average duration of 80 s, which were intended for use as background speech data (after [10, 11]) – named UBM (Universal Background Model) in the data package. The database also contains electroglottograph (EGG) data for all tasks (not provided for the test data), recorded simultaneously with the speech at a sampling frequency of 48 kHz (downsampled to 16 kHz) and a resolution of 16 bits using a device from Laryngograph Ltd.

For the purpose of the Challenge, the data were divided into speaker disjoint subsets for training, development, and testing. For the reading span task, both the sentences and the letters read aloud were recorded. However, preliminary baseline results showed that the cognitive load level cannot be inferred well from the letters only. Thus, it was decided to remove the reading span task letter recordings from the evaluations, leaving only the sentence recordings. The letter recordings are provided in the Challenge sets, but are excluded in all evaluations. Participants may use the data at their own convenience. Note that during the evaluation, the Stroop test with time pressure, the Stroop test with dual task, and the reading span tasks should be treated separately! The best baseline results are obtained in the case of per task modelling (cf. Section 4). Table 1 summarises the partitioning of the database.

2.2. Munich Bio-voice Corpus (MBC)

Physical load of users is of interest in manifold applications [13], but only few databases exist that contain speech under physical stress, e. g., the Dismounted Close Combat Database; see as well [14, 15]. In the *Phsyical Load Sub-Challenge*, the "Munich Bio-voice Corpus" (MBC) [16, 17] is used. The corpus was recorded at TUM in Munich/Germany. It contains speech from healthy subjects in two distinct physical load conditions. To produce such data, an experiment in which heart rate (HR) and skin conductivity (SC) were recorded simultaneously with vocal expressions was carried out. Several studies have shown that there exist significant correlations between speech features and heart rate [18, 19, 20], as well as with the level of sweating [21, 22, 23]. In the MBC corpus, HR and SC data were recorded with the Wild Divine Inc.'s "iom" – a lightweight hardware sensor device. Data were collected from three sensors attached to a subject's fingers. Audio was recorded with a Zoom Q3Hd camcorder equipped with an X-Y HD microphone placed on the table in front of the sitting subject. The audio has a sampling rate of 96 kHz in PCM-wave 24 bits format. Overall, 19 subjects (4 female, 15 male, 3 Chinese, 15 German, 1 Italian) gave their consent and participated in the experiment. All were free of temporary diseases, but the subjects include smokers and such with cardiac and neurological disorders. All subjects were recorded breathing, pronouncing the sustained vowel /a/, and reading a German or English text ("Der Nordwind und die Sonne" – "The Northwind and the Sun"), according to their mother tongue – both with low heart rate and with high heart rate under constant, pre-defined conditions. Subjects participated in the first recording session after a short introduction and a practice session, to ensure that their heart rate was low and they were in an 'idle' physical load state. After the first session, subjects had to perform a series of exercises, such as fast stair-climbing and running. A second session was recorded immediately after the exercise.

Table 2: *Partitioning of the Munich Bio-voice Corpus into train, dev(elopment), and test sets for binary classification ('low', 'high').*

#	train	dev	test	Σ
low	199	199	154	552
high	186	185	165	536
Σ	385	384	319	1 088

For the purpose of the Challenge, we only use the read text as main data since it conveys more speech variabilities than the sustained vowels. Consequently, two subjects were removed from the original database (1 German female, 1 Italian male), because this task was absent for one of them, and the high quality speech recording from the Zoom Q3Hd camcorder was not available for the other. Start and stop time-stamps of each session, i. e., before and after intense sport exercise, were manually segmented for each subject. Obtained wave files were then passed to a voice activity detector configured with a low energy threshold to allow the detection of breathing events in speech. In order to avoid having too short speech segments, we recursively merged those for which the duration was below 1 s. Data were finally divided into speaker disjoint subsets for training, development and testing, by stratifying (balancing) on gender, spoken language (English / German), age and body mass index. Table 2 summarises the partitioning of the database.

3. Challenge Features

For the baseline acoustic feature set used in this Challenge, we use the same acoustic feature set as in the INTERSPEECH 2013 Computational Paralinguistics ChallengE (COMPARE) [2] – the most effective set used in this series of Challenges so far. Again,

Table 3: COMPARE *acoustic feature set: 65 provided low-level descriptors (LLD).*

4 energy related LLD	**Group**
Sum of auditory spectrum (loudness)	prosodic
Sum of RASTA-filtered auditory spectrum	prosodic
RMS Energy, Zero-Crossing Rate	prosodic
55 spectral LLD	**Group**
RASTA-filt. aud. spect. bds. 1–26 (0–8 kHz)	spectral
MFCC 1–14	cepstral
Spectral energy 250–650 Hz, 1 k–4 kHz	spectral
Spectral Roll-Off Pt. 0.25, 0.5, 0.75, 0.9	spectral
Spectral Flux, Centroid, Entropy, Slope	spectral
Psychoacoustic Sharpness, Harmonicity	spectral
Spectral Variance, Skewness, Kurtosis	spectral
6 voicing related LLD	**Group**
F_0 (SHS & Viterbi smoothing)	prosodic
Prob. of voicing	voice qual.
log. HNR, Jitter (local & δ), Shimmer (local)	voice qual.

we use TUM's open-source openSMILE feature extractor [24] in its recent 2.0 release [25] and provide extracted feature sets on a per-chunk level. Configuration files for openSMILE with which the baseline features are reproducible will be provided soon together with the next openSMILE public release. The COMPARE feature set contains 6 373 static features – functionals of low-level descriptor (LLD) contours. For details on the feature set including an in-depth analysis of features for speech and music tasks the reader is referred to [26]. The LLD and functionals included in the set are summarised in Tables 3 and 4, respectively.

4. Challenge Baselines

As primary evaluation measure, we retain the choice of Unweighted Average Recall (UAR) as used since the first Challenge held in 2009 [27]: the unweighted (by number of instances in each class) mean of the percentage correctly classified in the diagonal of the confusion matrix. The motivation to consider *unweighted* rather than weighted average recall ('conventional' accuracy) is that it is also meaningful for highly unbalanced distributions of instances among classes, as is given in the *Cognitive Load Sub-Challenge*. For transparency and reproducibility, we use open-source classifier implementations of Support Vector Machines (SVM) from the WEKA data mining toolkit [28]. To this end, linear kernel SVM are used, which are known to be robust against overfitting. As training algorithm, we use Sequential Minimal Optimisation (SMO).

Reproducible balancing of the training set for the *Cognitive Load Sub-Challenge* is implemented by integer upsampling of the classes L1, L2, and L3, by the factors 2, 2, and 3, respectively. For evaluation on the test set, we re-train the models using the training and development set. For CLSE, in this case, the upsampling is applied to the training and development set. No balancing was performed for the MBC set, because the data are well balanced by the experimental setting, cf. Table 2.

In the *Cognitive Load Sub-Challenge*, the way that cognitive load is expressed in speech produced by the participants depends strongly on the task they had to perform. Thus, for the baseline we compare modelling each task individually to modelling all three tasks together. In the former case, the data for each partition (train/dev/test) are split into three sub-partitions by the task ID (READINGSPAN, TIMEPRESSURE, DUALTASK) and independet evaluations on each task sub-partition are performed (resulting in one model for each task); the predictions from the three evaluations are concatenated and scored for UAR normally. In global modelling, no sub-partitioning is performed and a single model is trained. Table 5 shows detailed results (UAR for each task individually) for CLSE obtained with the best baseline setting. Global modelling is compared to per task modelling. For the *Physical Load Sub-Challenge*, only global modelling was used since a single task was performed by all subjects, i. e., reading the text before and after exercising.

In the baseline systems, the SVM complexity C was optimised on the development set, by investigating C values from 0.00001 to 0.5 in roughly double increments, i. e., 0.00001, 0.00002, 0.00005, 0.0001, ... 0.5. Figure 1 shows the results (UAR) obtained with different complexities as well as different feature normalisation methods for both Sub-Challenges. The following six normalisation methods were investigated on both the CLSE and MBC development set: max/min normalisation (range -1 to +1) (**n-**) and mean 0 and variance 1 normalisation (**z-**), both individually for each speaker (**-spk**), individually for the respective training/test partitions (**-part**), and for training/test

Table 4: COMPARE *acoustic feature set: Functionals applied to LLD contours (Table 3).* [1]*: arithmetic mean of LLD / positive* Δ *LLD.* [2]*: not applied to voicing related LLD except* F_0. [3]*: only applied to F0.*

Functionals applied to LLD / Δ LLD	**Group**
quartiles 1–3, 3 inter-quartile ranges	percentiles
1 % percentile ($\approx$ min), 99 % pctl. ($\approx$ max)	percentiles
percentile range 1 %–99 %	percentiles
position of min / max, range (max – min)	temporal
arithmetic mean[1], root quadratic mean	moments
contour centroid, flatness	temporal
standard deviation, skewness, kurtosis	moments
rel. dur. LLD is above 25 / 50 / 75 / 90 % range	temporal
relative duration LLD is rising	temporal
rel. duration LLD has positive curvature	temporal
gain of linear prediction (LP), LP Coeff. 1–5	modulation
mean, max, min, std. dev. of segment length[2]	temporal
Functionals applied to LLD only	**Group**
mean value of peaks	peaks
mean value of peaks – arithmetic mean	peaks
mean / std.dev. of inter peak distances	peaks
amplitude mean of peaks, of minima	peaks
amplitude range of peaks	peaks
mean / std. dev. of rising / falling slopes	peaks
linear regression slope, offset, quadratic error	regression
quadratic regression a, b, offset, quadratic err.	regression
percentage of non-zero frames[3]	temporal

Table 5: *UAR* ***for each task individually*** *for CLSE and best baseline configuration (cf. Table 7). SVM complexity* $C = 0.0001$*, per task modelling vs. global model, normalisation: z-train (see text).*

UAR [%]	**Per task model**		**Global model**	
Task	Devel	Test	Devel	Test
Reading sentence	61.2	61.5	61.3	61.7
Stroop time pressure	74.6	66.7	54.0	44.4
Stroop dual task	63.5	56.9	44.4	37.5

Figure 1: *UAR vs. SVM complexity (C) on the CLSE (left) and MBC (right) development sets for four feature normalisation methods: n/z-train/part (see text for details). Per task modelling for CLSE.*

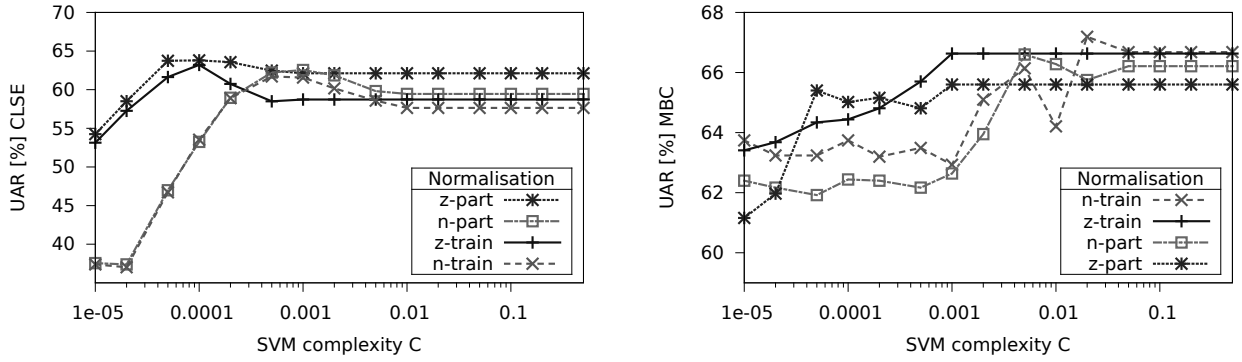

Table 6: *Effect of feature normalisation (Normal.) methods (see text for details). UAR in % (development set) for best SVM complexity parameter C for each normalisation method; per task modelling on CLSE.*

	CLSE		MBC	
Normal.	C	**UAR [%]**	C	**UAR [%]**
n-train	0.0005	61.7	0.02	67.2
z-train	0.0001	63.2	0.001	66.6
n-part	0.001	62.5	0.005	66.6
z-part	0.001	63.8	0.001	65.6
n-spk	0.0002	69.4	0.005	64.5
z-spk	0.0001	68.9	0.0005	66.7

partitions with parameters computed only from the training partition (**-train**). The per speaker normalisations (*n/z-spk*) cannot be applied on the test set because no speaker information is provided there. Thus, they are excluded from the valid official baseline configurations. Participants are free to employ an unsupervised speaker ID method, though.

In previous Challenges, the normalisation parameters have been computed on the training set and have been applied to the test set. This method must be used for on-line evaluations, i. e., where only one test instance is known at a time. As in this Challenge, however, all test instances are available and can be used in a single batch during evaluation, we also consider this case. Thus, normalisation of the test set is additionally performed as a whole, independent of the training set (*n/z-part*). For consistency with previous Challenges, however, we only consider the methods *n/z-train* for the official baseline results. Table 6 compares the results obtained with different feature normalisation strategies for both corpora. Based on Table 6, the official baseline results are shown in Table 7 in bold-face. Due to the different nature of the tasks and evaluation measures, we also present chance level baselines, which, for UAR, are defined by assuming a classifier which predicts only a single class label for all instances. Let us briefly summarise the baseline results here: With the best configurations (chosen on the development set for each Sub-Challenge), in the *Cognitive Load Sub-Challenge*, 63.2 % UAR and 61.6 % UAR are achieved on development and test sets, for the ternary classification; in the *Physical Load Sub-Challenge*, 67.2 % UAR and 71.9 % UAR are obtained, respectively, for the binary classification. For the challenge, only the best result on test out of the up to five uploads per site will be considered.

Table 7: *Challenge Baselines for* ComParE 2014. *C: Complexity parameter in SVM training (tuned on development set). Devel: Result on development set, by training on training set. Test: Result on test set, by training on the training and development sets. Chance: Expected measure by chance (cf. text). Settings: max/min normalisation (n), mean/variance normalisation (z) on all instances with parameters computed only from the training set (-train); Cognitive Load Sub-Challenge: one model* per task *or one* global *model. Official baseline results are marked in bold-face.*

Setting	C	UAR [%]		
		Devel	**Test**	Chance
Cognitive Load Sub-Challenge				
Per task, z-train	0.0001	**63.2**	**61.6**	33.3
Global, z-train	0.0001	59.1	58.2	33.3
Physical Load Sub-Challenge				
Global, n-train	0.02	**67.2**	**71.9**	50.0

5. Conclusion

We introduced the INTERSPEECH 2014 Computational Paralinguistics Challenge. This year, we focused on cognitive and physical load – tasks that have not yet been addressed that often as, e. g., emotion or personality; yet, they are interesting in themselves and promising for potential applications such as monitoring of subjects who perform physically or cognitively demanding tasks. The baseline results show both the feasibility and the difficulty to model these states automatically. We have provided baselines using a standard feature set and classification approach for the sake of consistency across the Sub-Challenges. We tried to make the baselines, on the one hand, competitive, by using a very comprehensive feature vector; on the other hand, they should be beatable to provide room for improvement. Thus, we did not optimise the feature vector by feature selection or reduction, and we used standard classification algorithms with only basic tuning. Hence, it will be of interest to see the performance of methods that are more tailored to peculiarities of the presented tasks.

6. References

[1] B. Schuller and A. Batliner, *Computational Paralinguistics – Emotion, Affect, and Personality in Speech and Language Processing*. Chichester, UK: Wiley, 2014.

[2] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proc. of Interspeech*, Lyon, France, 2013, pp. 148–152.

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

[4] A. Baddeley, "Working memory," *Science*, vol. 255, no. 5044, pp. 556–559, 1992.

[5] A. R. Conway, M. J. Kane, M. F. Bunting, D. Z. Hambrick, O. Wilhelm, and R. W. Engle, "Working memory span tasks: A methodological review and users guide," *Psychonomic Bulletin & Review*, vol. 12, no. 5, pp. 769–786, 2005.

[6] M. Daneman and P. A. Carpenter, "Individual differences in working memory and reading," *Journal of Verbal Learning and Verbal Behavior*, vol. 19, no. 4, pp. 450–466, 1980.

[7] N. Unsworth, R. P. Heitz, J. C. Schrock, and R. W. Engle, "An automated version of the operation span task," *Behavior Research Methods*, vol. 37, no. 3, pp. 498–505, 2005.

[8] F. G. Paas and J. J. Van Merriënboer, "Instructional control of cognitive load in the training of complex cognitive tasks," *Educational Psychology Review*, vol. 6, no. 4, pp. 351–371, 1994.

[9] J. R. Stroop, "Studies of interference in serial verbal reactions." *Journal of Experimental Psychology*, vol. 18, no. 6, p. 643, 1935.

[10] B. Yin, F. Chen, N. Ruiz, and E. Ambikairajah, "Speech-based cognitive load monitoring system," in *Proc. of ICASSP*, 2008, pp. 2041–2044.

[11] P. N. Le, E. Ambikairajah, J. Epps, V. Sethu, and E. H. Choi, "Investigation of spectral centroid features for cognitive load classification," *Speech Communication*, vol. 53, no. 4, pp. 540–551, 2011.

[12] F. Chen, N. Ruiz, E. Choi, J. Epps, M. A. Khawaja, R. Taib, B. Yin, and Y. Wang, "Multimodal behavior and interaction as indicators of cognitive load," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 2, no. 4, p. 22, 2012.

[13] S. Harada, J. Lester, K. Patel, T. Saponas, J. Fogarty, J. Landay, and J. Wobbrock, "Voicelabel: Using speech to label mobile sensor data," in *Proc. of the ACM International Conference on Multimodal Interfaces, ICMI*, Chania, Greece, 2008.

[14] J. H. L. Hansen and S. E. Bou-Ghazale, "Getting Started With SUSAS: A Speech Under Simulated And Actual Stress Database," in *Proc. of Eurospeech*, Rhodes, Greece, 1997, pp. 1743–1746.

[15] A. Stupakov, E. Hanusa, J. Bilmes, and D. Fox, "COSINE - A corpus of multi-party COnversational Speech In Noisy Environments ," in *Proc. of ICASSP*, Taipei, 2009, pp. 4153–4156.

[16] 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 *Proc. of LREC*, Reykjavik, Iceland, 2014, to appear.

[17] ——, "Automatic recognition of physiological parameters in the human voice: Heart rate and skin conductance," in *Proc. of ICASSP*, Vancouver, Canada, 2013, pp. 7219–7223.

[18] R. Orlikoff and R. Baken, "The effect of the heartbeat on vocal fundamental frequency perturbation," *Journal of Speech, Language, and Hearing Research (JSLHR)*, vol. 32, pp. 576–582, 1989.

[19] D. Skopin and S. Baglikov, "Heartbeat feature extraction from vowel speech signal using 2D spectrum representation," in *Proc. of the 4th International Conference on Information Technology (ICIT)*, Amman, Jordan, 2009, p. 6 pages.

[20] A. Mesleh, D. Skopin, S. Baglikov, and A. Quteishat, "Heart rate extraction from vowel speech signals," *Journal of Computer Science and Technology*, vol. 27, no. 6, pp. 1243–1251, 2012.

[21] R. Croft, C. Gonsalvez, J. Gander, L. Lechem, and R. Barry, "Differential relations between heart rate and skin conductance, and public speaking anxiety," *Journal of Behavior Therapy and Experimental Psychiatry*, vol. 35, no. 3, pp. 259–271, September 2004.

[22] J. Zhang, J. Kalinowski, T. Saltuklaroglu, and D. Hudock, "Stuttered and fluent speakers' heart rate and skin conductance in response to fluent and stuttered speech," *International Journal of Language & Communication Disorders*, vol. 45, no. 6, pp. 670–680, 2010.

[23] H. Kurniawan, A. Maslov, and M. Pechenizkiy, "Stress detection from speech and galvanic skin response signals," in *Proc. of the 26th IEEE International Symposium on Computer-Based Medical Systems (CBMS)*, Porto, Portugal, 2013, pp. 209–214.

[24] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE – the Munich versatile and fast open-source audio feature extractor," in *Proc. of the 18th ACM International Conference on Multimedia, MM 2010*, Florence, Italy, 2010, pp. 1459–1462.

[25] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent developments in opensmile, the Munich open-source multimedia feature extractor," in *Proc. of the 21st ACM International Conference on Multimedia, MM 2013*, Barcelona, Spain, 2013, pp. 835–838.

[26] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer, "On the acoustics of emotion in audio: What speech, music and sound have in common," *Frontiers in Psychology, Emotion Science, Special Issue on Expression of emotion in music and vocal communication*, vol. 4, no. Article ID 292, pp. 1–12, 2013.

[27] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 9/10, pp. 1062–1087, 2011.

[28] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The WEKA data mining software: An update," *SIGKDD Explorations*, vol. 11, 2009.