INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

The INTERSPEECH 2019 Computational Paralinguistics Challenge: Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity

Björn W. Schuller[1,2,3], *Anton Batliner*[2,4], *Christian Bergler*[4], *Florian B. Pokorny*[5], *Jarek Krajewski*[6,7], *Margaret Cychosz*[8], *Ralf Vollmann*[9], *Sonja-Dana Roelen*[7], *Sebastian Schnieder*[7], *Elika Bergelson*[10], *Alejandrina Cristia*[11], *Amanda Seidl*[12], *Anne S. Warlaumont*[13], *Lisa Yankowitz*[14], *Elmar Nöth*[4], *Shahin Amiriparian*[2], *Simone Hantke*[2,3], *Maximilian Schmitt*[2]

[1]GLAM – Group on Language, Audio & Music, Imperial College London, UK
[2] Chair of Embedded Intelligence for Health Care & Wellbeing, University of Augsburg, Germany
[3]audEERING GmbH, Gilching, Germany
[4]Pattern Recognition Lab, FAU Erlangen-Nuremberg, Germany
[5]Division of Phoniatrics, Medical University of Graz, Austria
[6]University of Wuppertal, Germany
[7]Rhenish University of Applied Science Cologne, Germany
[8]Department of Linguistics, University of California, Berkeley, USA
[9]Department of Linguistics, University of Graz, Austria
[10]Psychology and Neuroscience, Duke University, USA
[11]Laboratoire de Sciences Cognitives et Psycholinguistique, École Normale Supérieure, France
[12]Speech, Language, and Hearing Sciences, Purdue University, USA
[13]Department of Communication, University of California, Los Angeles, USA
[14]Department of Psychology, University of Pennsylvania, USA

schuller@IEEE.org

Abstract

The INTERSPEECH 2019 Computational Paralinguistics Challenge addresses four different problems for the first time in a research competition under well-defined conditions: In the *Styrian Dialects* Sub-Challenge, three types of Austrian-German dialects have to be classified; in the *Continuous Sleepiness* Sub-Challenge, the sleepiness of a speaker has to be assessed as regression problem; in the *Baby Sound* Sub-Challenge, five types of infant sounds have to be classified; and in the *Orca Activity* Sub-Challenge, orca sounds have to be detected. We describe the Sub-Challenges and baseline feature extraction and classifiers, which include data-learnt (supervised) feature representations by the 'usual' ComParE and BoAW features, and deep unsupervised representation learning using the AUDEEP toolkit.

Index Terms: Computational Paralinguistics, Challenge, Styrian Dialects, Sleepiness, Baby Sounds, Orca Activity

1. Introduction

In this INTERSPEECH 2019 COMPUTATIONAL PARALINGUISTICS CHALLENGE (COMPARE) – the eleventh since 2009 [1], we address four new problems within the field of Computational Paralinguistics [2] in a challenge setting: In the **Styrian Dialects** Sub-Challenge, three Austrian-German regional variants have to be told apart. Dialect classification is not only relevant for interaction optimisation in voice control applications or call centre services, but also for baseline estimation in speech diagnostics and speech-language therapy, and plays a valuable role as a tool in forensics [3, 4]. In the **Continuous Sleepiness** Sub-Challenge, the sleepiness of a speaker has to be assessed as regression problem. Monitoring and detecting sleepiness [5] is of high relevance for Advanced Driver Assistance Systems and other safety sensitive fields [6, 7]. In the **Baby Sound** Sub-Challenge, five types of infant speech sounds have to be classified. A possible application is diagnostics for developmental delay, based on the type and quantity of sounds that infants produce [8]. Finally, in the **Orca Activity** Sub-Challenge, orca sounds have to be detected. Collecting such bioacoustic data is an essential and practical tool to study and gain information about vocally active marine species [9, 10, 11, 12], providing invaluable information for ecosystem monitoring.

For all tasks, a target value/class has to be predicted for each case. Contributors can employ their own features and machine learning algorithms; standard feature sets and procedures are provided. Participants have to use predefined training/development/test splits for each Sub-Challenge. They may report results obtained from the training/development set (preferably with the supplied evaluation setups), but have only five trials to upload their results on the test sets per Sub-Challenge, whose labels are unknown to them. Each participation must be accompanied by a paper presenting the results, which 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 in the Challenge. As evaluation measure, we employ: (1) **Unweighted Average Recall (UAR)** as used since the first Challenge from 2009 [1], especially because it is more adequate for (unbalanced) multi-class classifications than Weighted Average Recall (i. e., accuracy) [2, 13]; (2) **Spearman's Correlation Coefficient (ρ)** [14] as the more 'conservative' and robust alternative to Pearson's [15] or Concordance Correlation Coefficient [16]; and (3) **Area Under the Receiver Operating Characteristic Curve (AUC)**. Ethical approval for the studies has been obtained from the pertinent committees. In section 2, we describe the challenge corpora. Section 3 details baseline experiments, metrics, and baseline results; concluding remarks are given in section 4.

http://dx.doi.org/10.21437/Interspeech.2019-1122

2. The Four Sub-Challenges

2.1. The Styrian Dialect (SD) Sub-Challenge

Styria, the south-eastern province of Austria with the city of Graz as its provincial capital, is embedded in a linguistic dialectal continuum termed Middle Bavarian in the North and a transition zone towards South Bavarian over the other parts [17, 18]. Urban centres such as Graz differ from rural regions, amongst other factors due to a greater mobility – neither 'dialect' nor 'standard' is spoken; rather, these two 'ideals' can be regarded as orientational points for speakers mixing characteristics of both systems to a variable extent [19]. Most knowledge of Styrian dialects is still based on descriptions from the late 1960s; cf. [18]. Therefore, a dialect task force at the University of Graz recently carried out a comprehensive collection of data. The so-called 'STYRIALECTS' dataset comprises audio recordings of Austrian-German speakers representative for different dialect areas of Styria. All recordings were conducted with an Edirol R-09 wave recorder (single-channel/22.05 kHz/16 bits/PCM) in an interview setting usually consisting of three parts, namely (1) a questionnaire compiled with regard to expected dialectal features, (2) a picture naming task, and (3) a short free conversation about language attitudes towards standard language and dialect. For the **Styrian Dialects** Sub-Challenge, the recordings of 55 speakers (22 males, 33 females; mean age 48.0 ± 22.3 years) from 25 different Styrian places were automatically segmented into target utterances of length 0.3 s–1.5 s by means of a speaker diarisation algorithm. Subsequently, all segments were manually revised yielding 9 732 samples of three different Styrian dialects to be differentiated: Northern Styrian (NorthernS), Urban Styrian (UrbanS), and Eastern Styrian (EasternS). Partitioning was speaker- and interviewer-independent.

2.2. The Continuous Sleepiness (CS) Sub-Challenge

For the **Continuous Sleepiness** Sub-Challenge, the SLEEP (Düsseldorf Sleepy Language) Corpus was created at the Institute of Psychophysiology, Düsseldorf, and the Institute of Safety Technology, University of Wuppertal, Germany. The sub-set of the corpus used for this Sub-Challenge consists of 915 subjects (364 females, 551 males, age from 12 to 84 years, mean age 27.6 ± 11.0 years). The recordings were made in quiet rooms using a microphone/headset/hardware setup with the tasks presented on a computer in front of the participants. Audio files were recorded with 44.1 kHz and down-sampled to 16 kHz, with a quantisation of 16 bit. The speech material consists of different reading passages and speaking tasks. Furthermore, spontaneous narrative speech was elicited by asking subjects to briefly comment on, e. g., their last weekend, the best present they ever got, or to describe a picture. A session of one subject lasted from 15 minutes to 1 hour; recordings took place from 6 am to midnight. Each participant had to report sleepiness on the well-established Karolinska Sleepiness Scale (KSS) [20] with a range of 1 (extremely alert) to 9 (very sleepy). Two raters assigned post-hoc observer KSS ratings. The scores from self-assessment and observers are averaged to form the reference sleepiness values, cf. [21]. Note that speakers and problem (correlation instead of classification) differ from the task addressed in the Interspeech 2011 Speaker State Challenge [22].

2.3. The Baby Sound (BS) Sub-Challenge

In the **Baby Sounds** Sub-Challenge, five types of infant sounds have to be classified: (1) canonical babbling (with a consonant and vowel), (2) non-canonical babbling, (3) crying, (4) laughing, and (5) junk/other. Our current dataset contains 12 445 vocalisations from 46 healthy infants (2-36 months) without any known speech or developmental delays. The children were exposed to a range of languages: English, Spanish, Tsimane, Tseltal, and Quechua, and were recorded as part of several studies on child language development, cf. [23, 24, 25, 26, 27]. Recordings were made using the Language ENvironment Analysis (LENA) Digital Language Processor [28], a lightweight audio recording device. Children wore the recorder for extended periods, between 6 and 16 hours, inside a clothing pocket specially designed for the device. Recordings were then processed using the proprietary LENA analysis system which assigns utterances to speakers in the child's environment (e. g., female adult) or to the target child. We randomly sampled 100 of these child vocalisations from one recording per child. The only exception to this was for the Casillas corpus [25] where an Olympus audio recorder (WS-832 or WS-835) was used and the vocalisations were hand-segmented. The vocalisations could vary in length from 36 ms to 18.34 s; they were segmented into chunks of 36 ms to 500 ms length, with a modal value of 400 ms. Chunks were then categorised according to our 5-way annotation scheme at least three times on the citizen science platform iHEARu-PLAY [29]. Only those vocalisations were kept where the majority of the annotators (at least two) agreed on the label. All chunks < 70 ms were padded with silence to match 70 ms as ComParE features require a minimum length of 65 ms. The final dataset consisting of 11 304 chunks was partitioned in a baby-disjunct way, while keeping the ages balanced across partitions.

2.4. The Orca Activity (OA) Sub-Challenge

For the **Orca Activity** Sub-Challenge, we use parts of the DeepAL Fieldwork Data (DLFD), collected on a 15-meter research trimaran in 2017 and 2018 in Northern British Columbia. A custom-made high sensitivity and low noise towed-array was deployed, which has a flat frequency response of within +/-2.5 dB between 10 Hz and 80 kHz. Underwater sounds were digitised with a sound acquisition device (MOTU 24AI), sampling at 96 kHz, recorded by PAMGuard, and stored on hard drives as multichannel wav-files (4 hydrophones in 2017; 8 hydrophones in a towed array in 2018). The total amount of collected audio data comprises 157 hours (1 channel). The overall number of annotations comprises $\sim$5.66 h; pure orca annotations amount to $\sim$1.40 h, distributed over 3 197 audio clips (1 channel). Including the multiple channels, the whole dataset comprises $\sim$1 007 h total audio, $\sim$40.93 h overall annotations, and $\sim$9.88 h pure orca annotations. For this sub-challenge, we use a sub-sample that amounts to a total duration of 4.6 hours (sound files: range 0.3-5.0 s; mean duration 1.23 ± 0.96 s). The two classes to be told apart are noise vs orca sounds.

3. Experiments and Results

For all Sub-Challenges, the segmented and categorised audio was converted to single-channel 16 kHz, 16 bits PCM format, except for the Orca Activity Sub-Challenge, where it is provided with 44.1 kHz sampling rate and multi-channel audio is provided (4/8 channels) in addition to the usual single-channel audio files.

3.1. ComParE Acoustic Feature Set

The official baseline feature set is the same as has been used in the six previous editions of the INTERSPEECH ComParE challenges, starting from 2013 [30]. This feature set contains

Table 1: *Databases: Number of instances per class in the train/dev/test splits: Test split distributions are blinded during the ongoing challenge and will be given in the final version.*

#	Train	Dev	Test	Σ
Styrian Dialects (STYRIALECTS)				
NorthernS	1 365	431	463	2 259
UrbanS	2 455	1 597	423	4 475
EasternS	1 407	542	1 049	2 998
Σ	5 227	2 570	1 935	9 732
Düsseldorf Sleepy Language Corpus (SLEEP)				
1-9 (KSS)	5 564	5 328	5 570	16 462
Baby Sounds (BS)				
Canonical	444	378	604	1426
Crying	243	163	263	669
Junk	1 826	1 357	1 392	4 575
Laughing	46	41	62	149
Non-canonical	1 437	1 678	1 370	4 485
Σ	3 996	3 617	3 691	11 304
DeepAL Fieldwork Data (DLFD)				
Noise	3 766	2 795	4 065	10 626
Orca	1 057	720	1 006	2 783
Σ	4 823	3 515	5 071	13 409

6 373 static features resulting from the computation of various functionals (statistics) over low-level descriptor (LLD) contours [30]. The configuration file is the ComParE_2016.conf, which is included in the 2.3 public release of openSMILE [31]. A full description of the feature set can be found in [32].

3.2. Bag-of-Audio-Words

In addition to the default ComParE feature set, where functionals are applied to the acoustic LLDs, we provide Bag-of-Audio-Words (BoAW) features. BoAW has already been applied successfully for, e. g., acoustic event detection [33] and speech-based emotion recognition [34]. Audio chunks are represented as histograms of acoustic LLDs, after quantisation based on a codebook. One codebook is learnt for the 65 LLDs from the ComParE feature set and another one for the 65 deltas of these LLDs. In Table 2, results are given for different codebook sizes. Codebook generation is done by *random sampling* from the LLDs/deltas in the training data. Each LLD/delta is assigned to the 10 audio words from the codebooks with the lowest Euclidean distance. Both BoAW representations, one from the LLDs and one from their deltas, are concatenated. Finally, a logarithmic term frequency weighting is applied to compress the numeric range of the histograms. LLDs are extracted with the openSMILE toolkit, BoAW are computed using openXBOW [35].

3.3. auDeep

Another feature set is obtained through unsupervised representation learning with recurrent sequence to sequence autoencoders, using the auDeep toolkit[1] [36, 37]. Representation learning commonly requires less human intervention than manually engineering a feature set such as the ComParE acoustic feature set. The recurrent sequence to sequence autoencoders which are employed by auDeep, in particular, explicitly model the inherently sequential nature of audio with RNNs within the encoder and decoder networks [36, 37]. In the auDeep approach, Mel-scale spectrograms are first extracted from the raw waveforms in a data set. In order to eliminate some background noise, power levels are clipped below four given thresholds in these spectrograms, which results in four separate sets of spectrograms per data set. Subsequently, a distinct recurrent sequence to sequence autoencoder is trained on each of these sets of spectrograms in an unsupervised way, i. e., without any label information. The learnt representations of a spectrogram are then extracted as feature vectors for the corresponding instance. Finally, these feature vectors are concatenated to obtain the final feature vector. For the results shown in Table 2, the autoencoders' hyperparameters were not optimised.

[1] https://github.com/auDeep/auDeep

Table 2: *Results for the four Sub-Challenges. The **official baselines** for test are highlighted (bold and greyscale). Dev: Development. C: Complexity parameter of the SVM/SVR. N: Codebook size for Bag-of-Audio-Words (BoAW) splitting the input into two codebooks (ComParE-LLDs/ComParE-LLD-Deltas) of the same given size, with 10 assignments per frame, and optimised complexity parameter of the SVM. $S2SAE$: Sequence to Sequence Autoencoder. X: Power levels which are clipped below four given thresholds. **UAR**: Unweighted Average Recall. ρ: Spearman's correlation coefficient. **AUC**: Area under ROC curve.*

	Styrian		**Sleepiness**		**Baby**		**Orca**	
	Dev	Test	Dev	Test	Dev	Test	Dev	Test
	UAR [%]		ρ		UAR [%]		AUC	
C	**openSMILE: ComParE functionals + SVM**							
10^{-5}	37.8	33.9	<0	.007	50.2	52.2	.680	.759
10^{-4}	38.8	35.9	.074	.237	54.0	57.7	.767	.841
10^{-3}	37.3	33.3	.251	.314	51.1	54.2	.810	.866
10^{-2}	37.4	33.5	.206	.290	45.6	49.5	.795	.855
10^{-1}	37.2	35.7	.163	.227	40.8	45.3	.767	.826
10^{0}	38.0	36.0	.127	.172	39.1	43.6	.754	.806
N	**openXBOW: ComParE BoAW + SVM**							
125	38.2	31.9	.240	.291	51.5	52.7	.772	.815
250	38.2	32.4	.236	.268	51.0	54.3	.763	.822
500	38.2	31.2	.250	.304	51.2	53.7	.762	.831
1000	37.4	32.2	.265	.286	51.1	54.3	.770	.823
2000	38.0	32.0	.269	.260	51.0	54.9	.771	.836
X dB	**auDeep: S2SAE + SVM**							
-40	43.7	37.3	.128	.205	48.4	44.1	.714	.772
-50	44.4	**47.0**	.213	.301	49.0	43.8	.700	.781
-60	44.6	39.4	.243	.325	49.8	46.9	.730	.776
-70	46.7	34.0	.261	.310	49.6	47.8	.712	.774
fused	45.9	35.5	.257	.321	51.6	48.1	.740	.798
	Fusion (Majority Vote)							
3-best	–	40.0	–	**.343**	–	**58.7**	–	**.866**

3.4. Challenge Baselines

For the sake of transparency and reproducibility of the baseline computation and in line with the previous years, we use an open-source implementation of Support Vector Machines (SVM) with linear kernels. This year, for the first time, the provided scripts employ the scikit-learn toolkit with its classes LinearSVC and LinearSVR, respectively, for the classification based on functionals, BoAW, and auDeep features. All feature representations were scaled to zero mean and unit standard deviation (MinMaxScaler of scikit-learn), using the parameters from the respective training set (when training and development sets were fused for the final classifier, the parameters were calculated on this fusion). For all tasks, the complexity parameter

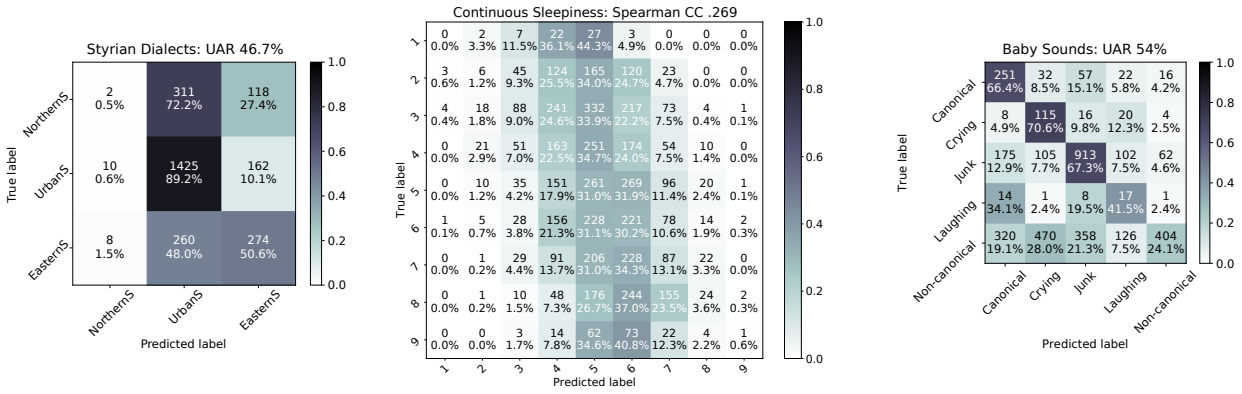

Figure 1: *Confusion matrices on the development set; overall number of instances per task given in Table 1. For each Sub-Challenge, the individual approach/hyperparameters performing best on the dev set were chosen, i. e.,* AUDEEP *with a clipping threshold of -70 dB and optimised complexity for the Styrian Dialects task, BoAW with a codebook size of 2 000 and optimised complexity for the Continuous Sleepiness task, and* COMPARE *with a complexity of 10^{-4} for the Baby Sounds task. In the cells, absolute number of cases is given, and percent of 'classified as' of the class displayed in the respective row; percentage also indicated by colour-scale: the darker, the higher.*

C was optimised during the development phase. For the Baby Sounds Sub-Challenge, we upsampled the minority classes in order to balance the five classes in the training (and development) sets; for the Styrian Dialects Sub-Challenge, training and development sets were not fused as the development set is considerably smaller in size owing to the sparseness of the data. Apart from this, the pipelines are basically the same, except from the fact that a linear Support Vector *Regression* is used for the Continuous Sleepiness Sub-Challenge and confidences are computed for the Orca Activity Sub-Challenge, as the *area under the ROC curve* is used as a metric.

Each Sub-Challenge package includes scripts that allow participants to reproduce the baselines and perform the testing in a reproducible and automatic way (including pre-processing, model training, model evaluation on the development set, and scoring by the competition and further measures).

This year, we provide the three above outlined approaches to Computational Paralinguistics: besides the usual COMPARE features plus SVM, we employ for the third time BoAW plus SVM, and for the second time sequence-to-sequence autoencoder (AUDEEP) learnt acoustic features, classified with an SVM, leaving, however, end-to-end learning from the raw time signal out. The same way as in the last two years, we chose the highest results on test for defining the baselines, irrespective of the corresponding results on development, in order to prevent participants from surpassing the official baseline by simply repeating or slightly modifying other constellations that can be found in Table 2. A fusion of the three models has been made by *Majority Voting* for the Styrian Dialects and Baby Speech Sub-Challenges and by taking the mean of the outputs for the Continuous Sleepiness and Orca Activity Sub-Challenges.

As can be seen in Table 2, for the Styrian Dialects Sub-Challenge, the baseline is UAR $= 47.0\,\%$, for the Continuous Sleepiness Sub-Challenge, it is Spearman's $\rho = .343$, for the Baby Sounds Sub-Challenge, it is UAR $= 58.7\,\%$, and for the Orca Activity Sub-Challenge, it is AUC $= .866$. Note that there is no 'official' baseline for Dev!

Figure 1 displays a 'good' confusion for the Baby Sounds Sub-Challenge (high frequencies in most of the diagonal cells) and the difficulty of the other tasks (low frequencies in some of the diagonal cells, high frequencies in some of the off-diagonal cells). Laughing is very sparse and it might therefore not be possible to model it robustly enough; in contrast, Non-canonical might show too much variability and similarity to the other classes. For the Styrian Dialects Sub-Challenge, UrbanS as majority class with mixed characteristics, by that possibly displaying greatest variance, is classified best; NorthernS seems to be least distinct from the other two varieties. In the SLEEP corpus, the extreme labels 1, 2, 8, 9, and to a slightly lesser extent, 3 and 7, are underrepresented – maybe less distinct – and cannot be modelled robustly; the classes in the middle – 4, 5, and 6 – are more frequent and confused with each other, and they attract the sparse, extreme labels, cf. the uniformly colour-scaled columns for 4, 5, and 6.

4. Concluding Remarks

This year's challenge is new by four new tasks (Styrian Dialects, Continuous Sleepiness, Baby Sounds, and Orca Activity, all of them highly relevant for applications). We further featured sequence-to-sequence autoencoder-based audio features by the AUDEEP toolkit using deep learning for audio classification for the second time as baselines and the popular OPENXBOW toolkit. For all computation steps, scripts are provided that can, but need not be used by the participants. We expect participants to obtain better performance measures by employing novel (combinations of) procedures and features including such tailored to the particular tasks.

5. Acknowledgements

We acknowledge funding from the EU's HORIZON 2020 Grants No. 115902 (RADAR CNS), the Bavarian State Ministry of Education, Science and the Arts in the framework of the Centre Digitisation.Bavaria (ZD.B). We thank the sponsor of the Challenge, audEERING GmbH. We also thank and remember Stefan Steidl of FAU Erlangen-Nuremberg in Germany who organised the first ten consecutive Interspeech Challenges with us since the first edition in 2009. He unexpectedly passed away at a young age in October 2018. In his honour and memory, the ComParE awards shall be named the STEFAN STEIDL COMPUTATIONAL PARALINGUISTICS AWARDS from 2019 onwards. Stefan, we will truly miss you – as a friend, colleague, and scientist.

6. References

[1] 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, pp. 1062–1087, 2011.

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

[3] O. Köster, R. Kehrein, K. Masthoff, and Y. H. Boubaker, "The telltale accent: Identification of regionally marked speech in German telephone conversations by forensic phoneticians," *International Journal of Speech, Language & the Law*, vol. 19, 2012.

[4] G. Brown, "Y-ACCDIST: An Automatic Accent Recognition System for Forensic Applications," Ph.D. dissertation, University of York, York, UK, 2014.

[5] J. Krajewski, A. Batliner, and M. Golz, "Acoustic sleepiness detection – Framework and validation of a speech adapted pattern recognition approach," *Behavior Research Methods*, vol. 41, pp. 795–804, 2009.

[6] S. Melamed, "Excessive Daytime Sleepiness and Risk of Occupational Injuries in Non-Shift Daytime Workers," *Sleep*, vol. 25, pp. 315–322, 2002.

[7] A. MacLean, "Sleepiness and Driving," *Sleep Medicine Reviews*, vol. 7, pp. 507–521, 2003.

[8] D. K. Oller, R. E. Eilers, A. R. Neal, and A. B. Cobo-Lewis, "Late onset canonical babbling: A possible early marker of abnormal development," *American Journal on Mental Retardation*, vol. 103, pp. 249–263, 1998.

[9] W. C. Cummings and D. V. Holliday, "Passive acoustic location of bowhead whales in a population census off Point Barrow, Alaska," *Journal of the Acoustical Society of America*, vol. 78, pp. 1163–1169, 1985.

[10] K. M. Stafford, C. G. Fox, and D. S. Clark, "Long-range acoustic detection and localization of blue whale calls in the northeast Pacific Ocean," *Journal of the Acoustical Society of America*, vol. 104, pp. 3616–3625, 1998.

[11] T. F. Norris, M. McDonald, and J. Barlow, "Acoustic detections of singing humpback whales (Megaptera novaeangliae) in the eastern North Pacific during their northbound migration," *Journal of the Acoustical Society of America*, vol. 106, pp. 506–514, 1999.

[12] A. B. Morton and H. K. Symonds, "Displacement of Orcinus orca (L.) by high amplitude sound in British Columbia, Canada," *ICES Journal of Marine Science*, vol. 59, pp. 71–80, 2002.

[13] A. Rosenberg, "Classifying skewed data: Importance weighting to optimize average recall," in *Proc. INTERSPEECH*, Portland, OR, 2012, pp. 2242–2245.

[14] C. S. Spearman, "The Proof and Measurement of Association Between Two Things," *The American Journal of Psychology*, vol. 15, pp. 72–101, 1904.

[15] K. Pearson, "Note on Regression and Inheritance in the Case of Two Parents," *Proceedings of the Royal Society of London*, vol. 58, pp. 240–242, 1895.

[16] L. I.-K. Lin, "A Concordance Correlation Coefficient to Evaluate Reproducibility," *Biometrics*, vol. 45, pp. 255–268, 1989.

[17] E. Kranzmayer, *Historische Lautgeographie des gesamtbairischen Dialektraumes*. Graz–Köln: Böhlau, 1956.

[18] P. Wiesinger, *Mundart und Geschichte in der Steiermark: Ein Beitrag zur Dialektgeographie eines österreichischen Bundeslandes*. Marburg an der Lahn: Elwert, 1967.

[19] J. Taeldeman, "The influence of urban centres on the spatial diffusion of dialect phenomena," in *Dialect Change. Convergence and Divergence in European Languages*, P. Auer, F. Hinskens, and P. Kerswill, Eds. Cambridge: Cambridge University Press, 2005, ch. 10, pp. 263–284.

[20] A. Shahid and K. Wilkinson, "Karolinska Sleepiness Scale (KSS)," in *STOP, THAT and One Hundred Other Sleep Scales*, A. Shahid, K. Wilkinson, S. Marcu, and C. M. Shapiro, Eds. Springer, 2012, pp. 209–210.

[21] F. Hönig, A. Batliner, E. Nöth, S. Schnieder, and J. Krajewski, "Acoustic-Prosodic Characteristics of Sleepy Speech – between Performance and Interpretation," in *Proc. of Speech Prosody*, Dublin, Ireland, 2014, pp. 864–868.

[22] B. Schuller, A. Batliner, S. Steidl, F. Schiel, and J. Krajewski, "The INTERSPEECH 2011 Speaker State Challenge," in *Proc. INTERSPEECH*, Florence, Italy, 2011, pp. 3201–3204.

[23] A. S. Warlaumont, G. M. Pretzer, S. Mendoza, and E. A. Walle, "Warlaumont HomeBank Corpus," 2016, doi:10.21415/T54S3C.

[24] E. Bergelson, "Bergelson Seedlings HomeBank Corpus," 2017, doi:10.21415/T5PK6D.

[25] M. Casillas, P. Brown, and S. C. Levinson, "Casillas HomeBank Corpus," 2017, doi:10.21415/T51X12.

[26] M. Cychosz, "Cychosz HomeBank Corpus," 2018, doi:10.21415/YFYW-HE74.

[27] C. Scaff, J. Stieglitz, and A. Cristia, "Daylong recordings from young children learning Tsimane in Bolivia," 2018, https://nyu.databrary.org/volume/445.

[28] C. R. Greenwood, K. Thiemann-Bourque, D. Walker, J. Buzhardt, and J. Gilkerson, "Assessing Children's Home Language Environments Using Automatic Speech Recognition Technology," *Communication Disorders Quarterly*, vol. 32, pp. 83–92, 2011.

[29] S. Hantke, F. Eyben, T. Appel, and B. Schuller, "iHEARu-PLAY: Introducing a game for crowdsourced data collection for affective computing," in *Proc. of 1st Workshop on Automatic Sentiment Analysis in the Wild (WASA) held in conjunction with ACII*. Xi'an, P. R. China: IEEE, 2015, pp. 891–897.

[30] 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. INTERSPEECH*. Lyon, France: ISCA, 2013, pp. 148–152.

[31] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor," in *Proc. ACM Multimedia*. Barcelona, Spain: ACM, 2013, pp. 835–838.

[32] F. Weninger, F. Eyben, B. Schuller, M. Mortillaro, and K. R. Scherer, "On the Acoustics of Emotion in Audio: What Speech, Music and Sound have in Common," *Frontiers in Emotion Science*, vol. 4, pp. 1–12, 2013.

[33] H. Lim, M. J. Kim, and H. Kim, "Robust Sound Event Classification Using LBP-HOG Based Bag-of-Audio-Words Feature Representation," in *Proc. INTERSPEECH*. Dresden, Germany: ISCA, 2015, pp. 3325–3329.

[34] M. Schmitt, F. Ringeval, and B. Schuller, "At the Border of Acoustics and Linguistics: Bag-of-Audio-Words for the Recognition of Emotions in Speech," in *Proc. INTERSPEECH*. San Francisco, CA: ISCA, 2016, pp. 495–499.

[35] M. Schmitt and B. W. Schuller, "openXBOW – Introducing the Passau Open-Source Crossmodal Bag-of-Words Toolkit," *Journal of Machine Learning Research*, vol. 18, pp. 1–5, 2017.

[36] S. Amiriparian, M. Freitag, N. Cummins, and B. Schuller, "Sequence to Sequence Autoencoders for Unsupervised Representation Learning from Audio," in *Proc. of 2nd Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE 2017)*. Munich, Germany: IEEE, 2017, pp. 17–21.

[37] M. Freitag, S. Amiriparian, S. Pugachevskiy, N. Cummins, and B. Schuller, "auDeep: Unsupervised Learning of Representations from Audio with Deep Recurrent Neural Networks," *Journal of Machine Learning Research*, vol. 18, pp. 1–5, 2018.