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

Björn W. Schuller1,2,3, Anton Batliner2,4, Christian Berger4, Florian B. Pokorny5, Jarek Krajewski6,7, Margaret Cychosz8, Ralf Vollmann9, Sonja-Dana Roelen7, Sebastian Schnieder7, Erika Bergelson10, Alejandrina Cristia11, Amanda Seidl12, Anne S. Warlaumont13, Lisa Yankowitz13, Elmar Nöth14, Shahin Amiriparian9, Simone Hantke2,3, Maximilian Schmitt2

1GLAM – Group on Language, Audio & Music, Imperial College London, UK
2Chair of Embedded Intelligence for Health Care & Wellbeing, University of Augsburg, Germany
3audEERING GmbH, Gilching, Germany
4Pattern Recognition Lab, FAU Erlangen-Nuremberg, Germany
5Division of Phoniatrics, Medical University of Graz, Austria
6University of Wuppertal, Germany
7Rhenish University of Applied Science Cologne, Germany
8Department of Linguistics, University of California, Berkeley, USA
9Department of Linguistics, University of Graz, Austria
10 Psychology and Neuroscience, Duke University, USA
11Laboratoire de Sciences Cognitives et Psycholinguistique, École Normale Supérieure, France
12Speech, Language, and Hearing Sciences, Purdue University, USA
13Department of Communication, University of California, Los Angeles, USA
14Department 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 a 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-learned (supervised) feature representations by the ‘usual’ ComParE and BoAW features, and deep unsupervised representation learning using the AU/DEEP toolkit.

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

1. Introduction

In this INTERSPEECH 2019 COMPUTATIONAL PARALINGUISTICS CHALLENGE (COMPAR) – 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 a 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 (ρ) 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.
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 dia
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;
and vowel), (2) non-canonical babbling, (3) crying, (4) laughing,
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 fea
tures, (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 9732 samples of three different Styrian dialects
to be differentiated: Northern Styrian (NorthernS), Urban Sty-
rian (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 In-
istitute 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 multi-channel audio is provided
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.

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,
Audio-Words (BoAW) features. BoAW has already been applied to the recurrent sequence to sequence autoencoders which are engineering a feature set such as the C feature set and another one for the 65 deltas of the C acoustic feature set. All feature representations were scaled to zero mean and unit standard deviation AU functionals, BoAW, and representation learning with recurrent sequence to sequence autoencoders, we provide Bag-of-Audio-Words (BoAW) representations, one assigned to the 10 audio words from the codebooks with the lowest Euclidean distance. Both BoAW representations, one 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 OPENSMILE [31].

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.

<table>
<thead>
<tr>
<th></th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
<th>Σ</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Styrian Dialects (STYRIALECTS)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NorthernS</td>
<td>1 365</td>
<td>431</td>
<td>463</td>
<td>2 259</td>
</tr>
<tr>
<td>UrbanS</td>
<td>2 455</td>
<td>1 597</td>
<td>423</td>
<td>4 475</td>
</tr>
<tr>
<td>EasternS</td>
<td>1 407</td>
<td>542</td>
<td>1 049</td>
<td>2 998</td>
</tr>
<tr>
<td><strong>Σ</strong></td>
<td>5 227</td>
<td>2 570</td>
<td>1 935</td>
<td>9 732</td>
</tr>
<tr>
<td><strong>Düsseldorf Sleepy Language Corpus (SLEEP)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I-9 (KSS)</td>
<td>5 564</td>
<td>5 328</td>
<td>5 570</td>
<td>16 462</td>
</tr>
</tbody>
</table>

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 OPENSMILE [35].

### 3.3. auDeep

Another feature set is obtained through unsupervised representation learning with recurrent sequence to sequence autoencoders, using the auDeep toolkit

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/SMV. 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.

|                          | Dev  | Test | Dev  | Test | Dev  | Test | Dev  | Test |έ |
|--------------------------|------|------|------|------|------|------|------|------|   |
| **Styrian**              |      |      |      |      |      |      |      |      |   |
| Sleepiness               |      |      |      |      |      |      |      |      |   |
| Baby                      | 37.8 | 33.9 | 0.243 | 0.325 | 49.8 | 46.9 | 0.730 | 0.776 |   |
| Orca                      |      |      |      |      |      |      |      |      |   |
| **Σ**                    | 4 823 | 3 515 | 5 071 | 13 409 |   |

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.

### 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

https://github.com/auDeep/auDeep

https://github.com/auDeep/auDeep
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 ρ = .343, for the Baby Sounds Sub-Challenge, it is UAR = 58.7%, and for the Orca Activity Sub-Challenge, it is AUC = .860. 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, Urbanan 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 CompParE 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


