The INTERSPEECH 2015 Computational Paralinguistics Challenge: Nativeness, Parkinson’s & Eating Condition*

Björn Schuller1,2, Stefan Steidl3, Anton Batliner3,4, Simone Hanke4, Florian Hö nig3, Juan Rafael Orozco-Arroyave3,5, Elmar Nö th3, Yue Zhang4, Felix Weninger4

1Department of Computing, Imperial College London, UK
2Chair of Complex & Intelligent Systems, University of Passau, Germany
3Pattern Recognition Lab, FAU Erlangen-Nuremberg, Germany
4Machine Intelligence & Signal Processing Group, TUM, Munich, Germany
5Faculty of Engineering, Universidad de Antioquia, Medellín, Colombia

Abstract

The INTERSPEECH 2015 Computational Paralinguistics Challenge addresses three different problems for the first time in research competition under well-defined conditions: the estimation of the degree of nativeness, the neurological state of patients with Parkinson’s condition, and the eating conditions of speakers, i.e., whether and which food type they are eating in a seven-class problem. In this paper, we describe these sub-challenges, their conditions, and the baseline feature extraction and classifiers, as provided to the participants.

Index Terms: Computational Paralinguistics, Challenge, Degree of Nativeness, Parkinson’s Condition, Eating Condition

1. Introduction

In this INTERSPEECH 2015 COMPUTATIONAL PARALINGUISTICS CHALLENGE (COM-PARE) – the seventh since 2009 [1], we address, for the first time within a challenge setting, four problems within the field of Computational Paralinguistics [2]:

In the Brave New Approach (BNA) Sub-Challenge, the pronunciation quality of non-native utterances has to be assessed, based on prosodic annotations, and using regression as measure. This is an ‘open’ challenge with known test labels; the task is not to obtain the highest performance for unknown test data but to come up with new ideas and interesting ‘alternative’ approaches spanning the spectrum from ‘good old phonetic/linguistic approaches’ to innovative ideas and paradigm shifts in paralinguistic methods for this rather new and difficult problem of addressing degree of nativeness within a speaker- and item-independent cross-corpus setting. Generally, it is well known that non-native pronunciations and prosody can be recognised automatically [3]; previous works targeting in particular the ‘degree of nativeness’ include, e.g., [4, 5, 6].

In the Degree of Nativeness (DN) Sub-Challenge, the training set from the BNA Sub-Challenge is used as training, and the test data from BNA as development set. In addition, a new test set with unknown labels is provided.

In the Parkinson’s Condition (PC) Sub-Challenge, the neurological state of Parkinson patients has to be estimated according to the Unified Parkinson’s Disease Rating Scale, motor subscale: UPDRS-III [7], within a regression task. PC is a neurological disorder affecting functions of the basal ganglia; it is characterised by the progressive loss of dopaminergic neurons in the substantia nigra of the midbrain [8]. PC leads to vocal impairment for approximately 90% of the patients [9].

Telemonitoring of the mostly elderly patients by vocal features has been shown to be feasible to some degree [10, 11, 12].

In the Eating Condition (EC) Sub-Challenge, the eating condition of a speaker has to be classified: whether s/he is eating or not, and if so, which type of food (six food types). So far, there have been only a few studies investigating speaking whilst speakers bite on a block [13] or considering muscle movements under speaking and eating [14]. In addition, chewing sounds (without speaking) have been recognised automatically in [15], and with special hardware in [16, 17].

Due to space limitations, we cannot elaborate in-depth on state-of-the-art and importance of the tasks: the assessment of non-native speech plays a pivotal role in language teaching, the same way as the assessment of the severity of Parkinson’s condition does in speech therapy; in both fields, automatic approaches are promising and worth any effort. Speech under eating is not yet an established field; however, we can imagine several promising applications such as adapting automatic speech recognition (for instance, for dictation under eating [18]) to EC, health (in-gestive behaviour) and security (when eating is not allowed) monitoring, forensics, or ethnography of communication [19] (analysing speaking and/under eating as essential communicative systems) [20].

For all tasks, the target value/class has to be predicted per speech file. Contributors can employ their own features and machine learning algorithm; however, a standard feature set is provided that may be used. Participants will have to stick to the
pre-defined training/development/test splits. They may report development results obtained from the training set (preferably with the supplied evaluation setups), but have only a limited number of trials to upload their results on the test sets for the DN, PC (ten, each) and EC (five) Sub-Challenges, 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 themselves in the Challenge. As evaluation measures, for the BNA, DN, and PC Sub-Challenges, we use Spearman’s Correlation Coefficient ($\rho$) as the more ‘conservative’ and robust alternative to Pearson’s correlation coefficient. For the EC task, we employ Unweighted Average Recall (UAR) as used since the first Challenge held in 2009 [1], especially because it is more adequate for (more or less unbalanced) multi-class classifications than Weighted Average Recall (i.e., accuracy).

In section 2, the challenge corpora, and in section 3, the baseline experiments are introduced. Novelties of this year’s challenge are in the BNA, DN, and PC Sub-Challenges the use of multiple databases in cross-corpus settings within a highly realistic mismatch of recording conditions between train (development) and test sets.

2. Challenge Corpora

2.1. Brave New Approach (BNA)

For the training set of the BNA Sub-Challenge, we employ data from the AUWL [21] and ISLE [22] corpora. In AUWL, learners of English as a second language practised pre-scripted dialogues. These data are more natural and contain less reading-related hesitations than read non-native speech. Microphones and recording hardware were heterogeneous and partly low-quality since learners were using their own equipment. The material used here comprises 31 speakers (13 f, 18 m; 36.5 ± 15.3 years; native languages: 16 German, 4 Italian, 3 Chinese, 3 Japanese, 5 other), 5.5 hours, and 3,732 speech files (423 distinct sentences/phrases).

Each speech file was annotated by five phoneticians with respect to its prosody (sentence melody and rhythm) on a five-point scale ranging from 1 (normal) to 5 (very unusual). With the (simplifying) assumption of an interval scale, we took the arithmetic average of the five labellers to obtain inter-subjective prosody scores [23], with an average of 1.7 and a standard deviation of 0.5 (range 1.0–3.8). From ISLE, we used material comprising 36 speakers (11 f, 25 m; native languages: 20 German, 16 Italian), 0.3 hours, 158 speech files (5 distinct sentences); prosody scores were collected in a similar manner (2.1 ± 0.5, range 1.3–3.4). These few sentences were included to take advantage of the fact that the speakers of the ISLE database are disjoint from the speakers of our databases. For the test set with known labels, we use a subset of the C-AuDiT database [24] which contains read non-native English (sentences from short stories; sentences containing different types of phenomena such as intonation or position of phrase accent, tongue twisters, etc.). Heterogeneous microphones and recording hardware were used for recording. The material is disjoint from the training set with respect to both speakers and sentences. It comprises 58 speakers (31 f, 27 m; native languages: 26 German, 10 French, 10 Spanish, 10 Italian, 2 Hindi), 2.7 hours, and 999 speech files (19 distinct sentences). Prosodic scores were collected similarly, except for using a 3-point scale from 0 for good to 2 for bad (0.5 ± 0.3, range 0.0–1.6). Additional material that may but need not be used comprises: (a) the word sequence the learners were supposed to produce, which can be used as a transcription since recordings with word errors were excluded; (b) a pronunciation dictionary with syllable boundaries and word accent positions; (c) an approximate phoneme segmentation automatically generated from (a) and (b); (d) speaker identities; and (e) the corpus each file came from. All recordings are given with a sampling rate of 16 kHz.

2.2. Degree of Natenessness (DN)

The training set is the same as for the BNA Sub-Challenge, and the development set is the test set of the BNA Sub-Challenge. The DN test set was created at TUM. The recordings were made in a quiet office room with a single microphone/hardware setup. The participants were asked to read aloud sentences of two short stories in English: “The North Wind and the Sun” (widely used within phonetics, speech pathology, and alike), and “The Rainbow” (standard reading passage used in speech/language pathology). The speech material comprises 54 speakers (28 f, 26 m; 31.3 ± 8.9 years; native languages: 23 German, 12 Chinese, 19 other; 1.4 hours, 594 speech files, 11 distinct sentences). Prosodic scores were collected in the same manner as for AUWL, using 16–23 annotators. Labels range from 1.1 to 5.0, with an average of 2.9 and a standard deviation of 0.7. Additional information that may but not need be used comprises the target texts (can be used as transcription since recordings with word errors were excluded) and the respective entries in the pronunciation dictionary. The sampling rate was 16 kHz. The material is disjoint from the training and development sets with respect to both speakers and sentences.

2.3. Parkinson’s Condition (PC)

Recordings of the training and development sets were done at UdA [25] in a sound-proof booth (dynamic omnidirectional microphone, professional audio card, sampling at 44.1 kHz) with a total of 50 patients with Parkinson’s disease (25 f, 25 m). 35 of the patients are included in the training set, and the remaining 15 comprise the development set. Each speaker performed a total of 42 speech tasks including 24 isolated words, 10 sentences, one reading text, one monologue, and the rapid repetition of the syllables /pa-ta-ka/, /pa-ka-ta/, and /pe-ta-ka/. The test set consists of the same 42 tasks produced by 11 patients (5 f, 6 m), recorded with the same microphone, sound card, resolution bits, and sampling frequency as the training and development sets – yet not in a sound-proof booth but in quiet office environments. The total duration of recordings included in the training, development, and test sets are 81, 33, and 43 minutes. Reading texts comprise a total of 36 words. The average duration of monologues per speaker in the training, development, and test sets are 48 ± 26, 42 ± 19, and 112 ± 21 seconds. The mean age of the participants included in the train, development, and test sets are 61.3 ± 10, 62 ± 6.5, and 63 ± 7. All of the patients were diagnosed and labelled by a neurologist according to the UPDRS-III scale, with a mean of 38.5 and a standard deviation of 19.1 (range 5 to 92). The speech samples were recorded with the patients in ON-state, i.e., no more than 3 hours after the morning medication. All speakers were evaluated by a phoniatrician; if they showed any speech atypicality different from those due to PC, they were excluded from the database. For training and development, we provide additional material that may but need not be used: (a) speaker identity; (b) task type; and (c) the target sentences, where applicable (not necessarily usable as transcription due to reading errors).
ComParE.conf, which is included in the 2.1 public release of openSMILE [29, 30]. A pre-release version of openSMILE 2.1 was used, resulting in slightly different baseline features for some descriptors in comparison to the features extracted with the latest 2.1 version. As evaluation measure for the EC Sub-Challenge, we use UAR; given the ordinal-scaled annotations of the BNA, DN, and PC Sub-Challenges, we use ρ as the official competition measure for these sub-challenges as outlined above. For transparency and reproducibility, we use open-source implementations from the Weka 3 data mining toolkit [31]. We apply linear kernel Support Vector Machines (SVM) / linear Support Vector Regression (SVR) with epsilon-insensitive loss, which are known to be robust against overfitting.

### 3. Challenge Baselines

For the baseline feature set, we use the same COMPARE set of supra-segmental (utterance-level) acoustic features as in the previous two editions of Interspeech ComParE [27, 28]. None of the additional material supplied for BNA, DN, and PC is used. The COMPARE feature set contains 6 373 static features as functionals of low-level descriptor (LLD) contours. The configuration file is the IS13_ComParE.conf, which is included in the 2.1 public release of openSMILE [29, 30].

#### 3.1. Brave New Approach and Degree of Nativeness

The two sub-challenges on degree of nativeness are cross-corpus tasks, with different text material and different recording conditions. This should be accounted for during development, otherwise the system might overfit to matched data during development and perform poorly on the mismatched data in test. We can account for the text mismatch in a straightforward way by training and evaluating on disjoint text material during development. Therefore, we use a double nested loop (K=2) over speakers and texts for the cross-validation on train. As we have a sufficient number of recordings in the training set, we limit computation time by using just N=2 speaker and text folds, resulting in a total number of N^K=4 folds. Thus, per fold about \( \frac{2N}{K} \) = 25% of the data is used for training, and similarly about \( \frac{1}{\sqrt{K}} \) = 25% of the data for testing (see Table 2).

For BNA, the best result in CV is obtained for the complexity \( C=10^{-5} \) with \( p_m=0.403 \). This complexity is optimal on test, too, and yields \( p_m=0.415 \) here. Given the nature of this Sub-Challenge, this is, however, merely a rough guide line rather than a real baseline – the spirit here is to generally compare interesting brave new approaches rather than optimise to beat a number.

For DN, things are a bit more complicated: while optimal complexity is \( C=10^{-6} \) for both provided development schemes (CV on train; train vs development), with \( p_m=0.403 \) and \( p_m=0.415 \),
were recorded in non-controlled noise conditions, so a lower ρ with (1) training just with train, or (2) merging train and development.

This difference can likely be explained because the test data differs very similar within each of the development schemes, differing respect to speakers. The two development schemes provided lead to different optimal values for C: for CV on train, C=10^{-2} is optimal, with ρ=.434, while for train vs development, C=10^{-3} is best, with ρ=.492. However, the results for C=10^{-2} and C=10^{-3} are very similar within each of the development schemes, differing only after the fifth and fourth decimal, (when not rounded). For this sub-challenge there are two options to build the final system: (1) training just with train, or (2) merging train and development for training. In Table 3 we consider only the first of these two options. The second option led to a downgrade rather than an upgrade, likely due to a too large mismatch between the development and test partitions: The highest result was obtained with C=10^{-5} as ρ=.390 when training only on train, but drops to ρ=.354 if using train and development data for training with the same C. However, participants are free to decide what option they choose for the final system, e.g., by considering suited domain adaptation or data and/or feature transfer learning methods to reduce the differences between the partitions [32, 33]. In fact, also the optimal value C=10^{-5} considering the test data is unexpected, as it is quite lower than for the development scheme. This difference can likely be explained because the test data were recorded in non-controlled noise conditions, so a lower complexity can prevent the system from overfitting to the very clean acoustic conditions in the train and development data.

### 3.2. Parkinson’s Condition

For PC, we provide two development schemes: CV on train, and train vs development. To reflect the fact that the system is going to be applied on unknown speakers, each fold is constructed to partition the training set with respect to the speakers (single nested CV: K=1). Since we use N=4 folds, within each fold, about 75% of the data is used for training, and 25% for testing.

The development set is disjunct from the training set with respect to speakers. The two development schemes provided lead to different optimal values: for train, C=10^{-2} is optimal, with ρ=.434 when training only on train, but drops to ρ=.390 when training and test swap roles in folds 1/4 and 2/3.

Table 2: Double nested speaker-and text-independent cross-validation for BNA/DN. Speakers are partitioned into to sets S₁ and S₂, text material into T₁ and T₂. Note that with our N=2, train and test swap roles in folds 1/4 and 2/3.

<table>
<thead>
<tr>
<th>Fold</th>
<th>Speaker Fold</th>
<th>Text Fold</th>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>S₁∧T₁</td>
<td>S₂∧T₂</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>S₁∧T₂</td>
<td>S₂∧T₁</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>1</td>
<td>S₂∧T₁</td>
<td>S₁∧T₂</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>2</td>
<td>S₂∧T₂</td>
<td>S₁∧T₁</td>
</tr>
</tbody>
</table>

respectively, it is better to use a higher complexity for train vs test: With C=10^{-4}, we get ρ=.425 on test. This can be explained by the fact that unlike DN-train and DN-development, DN-test has been recorded under homogeneous recording conditions. Thus, allowing a model with some more complexity pays off. Note that we use only train for building the final system, since train+development cannot simply be combined due to the different scales used for annotation. However, participants are allowed to combine both sets with suitable measures for handling the different scales.

### 3.3. Eating condition

For EC, we again employ CV (single nested CV: K=1); since we use N=20 folds, within each fold, about 95% of the data is used for training, and 5% for testing. The optimal complexity for CV on train is C=10^{-3} with 61.3% UAR. The same complexity is optimal for test, with an UAR of 65.9%. Note that these results cannot be directly compared to those in [20] because of different evaluation and classifier setups. We cannot provide meaningful estimates of mean / standard deviation of accuracy or UAR in LOSO-CV, since not all classes are present for all speakers.

### 4. Conclusion

The tasks in this year’s challenge are new in several ways: with EC, we introduce a new field of research; for BNA, DN, and PC – all being representative for established fields, namely assessment of non-native and pathological speech – we have to face a sometimes severe acoustic mismatch due to different recording conditions between training/development and test sets. Moreover, for BNA and DN, the task is speaker- and item-independent, as well as cross-corpus. The acoustic mismatches caused in turn mismatches in performance between optimal complexity settings for CV train and/or development on the one hand, and the optimal complexity settings for the test sets. As baselines, we established the results with the optimal complexity parameters obtained for the test set. We report five attempts on test used for their determination – the participants have either five (EC) or ten (DN, PC) attempts per Sub-Challenge. Ten attempts are allowed in the cross-corpus Sub-Challenges given the higher complexity due to acoustic and further mismatches. Yet, feature sets and learning procedures are standard – competitive but not optimised and kept generic across the tasks, despite their obvious differences. We hope that the meta-information that was not used in the baselines by intention for the sake of simplicity will make it possible for the participants to come up with both competitive and interesting new approaches towards the general challenge we all will face when ‘going real-life’, by that shifting from well-designed lab constellations to more realism.

Table 3: Challenge Baselines. C: Complexity parameter of SVM/SVR. Column (a): results of cross-validation on train. Column (b): results of train vs development. Column (c): results of train vs test. The official challenge baselines are highlighted by frames.

<table>
<thead>
<tr>
<th>Degree of Nativeness (ρ)</th>
<th>C</th>
<th>CV train</th>
<th>train/dev</th>
<th>train/test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parkinson’s Condition</td>
<td>10^{-6}</td>
<td>.333</td>
<td>.311</td>
<td>.223</td>
</tr>
<tr>
<td></td>
<td>10^{-5}</td>
<td>.403</td>
<td>.415</td>
<td>.359</td>
</tr>
<tr>
<td></td>
<td>10^{-4}</td>
<td>.399</td>
<td>.411</td>
<td>.425</td>
</tr>
<tr>
<td></td>
<td>10^{-3}</td>
<td>.368</td>
<td>.347</td>
<td>.354</td>
</tr>
<tr>
<td></td>
<td>10^{-2}</td>
<td>.368</td>
<td>.338</td>
<td>.355</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Eating Condition (UAR [%])</th>
<th>C</th>
<th>CV train</th>
<th>train/dev</th>
<th>train/test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10^{-6}</td>
<td>51.1</td>
<td>–</td>
<td>48.0</td>
</tr>
<tr>
<td></td>
<td>10^{-5}</td>
<td>59.7</td>
<td>–</td>
<td>60.6</td>
</tr>
<tr>
<td></td>
<td>10^{-3}</td>
<td>61.3</td>
<td>–</td>
<td>65.9</td>
</tr>
<tr>
<td></td>
<td>10^{-2}</td>
<td>60.9</td>
<td>–</td>
<td>65.9</td>
</tr>
<tr>
<td></td>
<td>10^{-1}</td>
<td>60.9</td>
<td>–</td>
<td>65.9</td>
</tr>
</tbody>
</table>
5. References


