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

Björn Schuller[1,2], *Stefan Steidl*[3], *Anton Batliner*[3,4], *Simone Hantke*[4], *Florian Hönig*[3], *Juan Rafael Orozco-Arroyave*[3,5], *Elmar Nöth*[3], *Yue Zhang*[4], *Felix Weninger*[4]

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

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 (COMPARE) – 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 (ingestive 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

*The research leading to these results has received funding from the European Community's Seventh Framework Programme through ERC Starting Grant No. 338164 (iHEARu), the German Federal Ministry of Education, Science, Research and Technology (BMBF) under grant 01IS07014B (C-AuDiT), and the German Ministry of Economics (BMWi) under grant KF2027104ED0 (AUWL). Juan Rafael Orozco-Arroyave is under grants of "Convocatoria 528 para estudios de doctorado en Colombia, generación del bicentenario 2011" funded by COLCIENCIAS. This work was also financed by COLCIENCIAS, project № 111556933858. We want to thank *digital publishing (dp)* for permitting the use of the C-AuDiT and AUWL databases, as well as ELRA for using parts of the ISLE corpus (ELRA catalogue (http://catalog.elra.info), ISLE Speech Corpus, catalogue reference: ELRA-S0083), and Eduardo Coutinho for help with the test data. The authors would further like to thank the sponsors of the Challenge: audEERING UG (limited) and the Association for the Advancement of Affective Computing (AAAC). The responsibility lies with the authors.

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 (ρ) 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 mis-match 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, 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) for normal to (5) for 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 disjunct 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 Nativeness (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 annotaters. 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 need not 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 disjunct 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 UdeA [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).

Table 1: *The iHEARu-EAT database: Number of instances per class in the CV-train/test split used for the Challenge.*

#	Train	Test	Σ
No Food	140	70	210
Apple	140	56	196
Nectarine	133	63	196
Banana	140	70	210
Crisp	140	70	210
Biscuit	133	70	203
Gummi bear	119	70	189
Σ	945	469	1 414

2.4. Eating Condition (EC)

For the EC Sub-Challenge, the audio tracks of the audio-visual iHEARu-EAT database are used [20]. 30 subjects (15 f, 15 m; 26.1±2.7 years) were recorded in a quiet, low reverberant office room at TUM (27 German; 1 Chinese, 1 Indian, 1 Tunisian origin, all of them having a close-to-native competence in German; no speaker displayed significant speech impediments.). Prior to the actual recording, subjects performed practice trials to familiarise themselves with the procedure. Food classes were chosen with partly similar consistency (for instance, crisps and biscuits) and partly dissimilar consistency (for instance, nectarine vs crisps). These food classes represent snacks which are likely to be encountered in practical scenarios and enable the subjects to speak while eating. In order to control for the amount of food being consumed, and in particular to encourage subjects to actually eat while speaking, an assistant provided the subjects with a serving of fixed size prior to the recording of each utterance. The serving size was chosen such as to enable a significant effect on the subjects' speech. For read speech, the German version of the phonetically balanced standard story (cf. also the DN test partition) "The North Wind and the Sun" ("Der Nordwind und die Sonne") was chosen (71 word types with 108 tokens, 172 syllables [26]). The subjects had to read the whole text with each sort of food. Spontaneous narrative speech was elicited by prompting subjects to briefly comment on, e. g., their favourite travel destination, genre of music, or sports activity. A typical session of one subject lasted about one hour. The narratives were segmented into units whose length roughly equals the length of the six pre-defined units in the read story. The speech files were segmented manually, in order to remove non-speech parts at the beginning and the end with only 'eating noise', which could make the classification task too easy. All in all, 1 414 turns and 2.9 hours of speech (sampled at 16 kHz) were recorded. By construction, 1/7 of the speech files contain spontaneous speech. Note that there is a slight difference in the amount of utterances per class, because some subjects chose not to eat all types of food.

For the Challenge, the data were split speaker-independently into a training set (20 speakers) and test set (10 speakers), stratified by age and gender. The resulting numbers of instances per class and set are shown in Table 1.

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]. 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. As training algorithm, we use Sequential Minimal Optimisation (SMO). We scale all features to a standard deviation of 1 (option `-N 1` for Weka's SMO/SMOreg). For SVR, a fixed ϵ of 1.0 is used. As a novelty, we introduce CV in all sub-challenges: 4-fold speaker-independent CV for BNA, DN, and PC, and leave-one-speaker-out cross-validation (LOSO-CV) for EC. By that, it is hoped that the results obtained on the training set are more representative and hence the benefits of CV outweigh the increased computational cost. Performance is computed as a single ρ or UAR value over the combined results of all CV folds (i. e., not averaged over the results in the individual folds). For DN and PC, an alternative evaluation scheme for development is given by train vs BNA test (DN) and train vs development (PC). The complexity parameter C was optimised up to a power of ten through CV on train; however, this did not always result in optimal values for test (see below). For all sub-challenges, a baseline recipe is provided to the participants that performs CV on the training set in a reproducible and automatic way, including preprocessing, model training, model evaluation, and scoring by the competition and further measures. A novelty of this year's BNA Sub-Challenge is the provision of a recipe for re-producing the baseline regression results on the test set with known labels. In the following, we will briefly summarise the baseline results as displayed in Table 3.

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 of N^K=4 folds. Thus, per fold about $(\frac{N-1}{N})^K = 25\%$ of the data is used for training, and similarly about $(\frac{1}{N})^K = 25\%$ of the data for testing (see Table 2).

For BNA, the best result in CV is obtained for the complexity C=10^{-5} with ρ=.403. This complexity is optimal on test, too, and yields ρ=.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^{-5} for both provided development schemes (CV on train; train vs development), with ρ=.403 and ρ=.415,

Table 2: *Double nested speaker-and text-independent cross-validation for BNA/DN. Speakers are partitioned into to sets S_1 and S_2, text material into T_1 and T_2. Note that with our N=2, train and test swap roles in folds 1/4 and 2/3.*

Fold	Speaker Fold	Text Fold	Train	Test
1	1	1	$S_1 \wedge T_1$	$S_2 \wedge T_2$
2	1	2	$S_1 \wedge T_2$	$S_2 \wedge T_1$
3	2	1	$S_2 \wedge T_1$	$S_1 \wedge T_2$
4	2	2	$S_2 \wedge T_2$	$S_1 \wedge T_1$

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.

3.2. 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 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.

3.3. 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.

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.*

	(a)	**(b)**	**(c)**
C	CV train	train/dev	train/test
Degree of Nativeness (ρ)			
10^{-6}	.333	.311	.223
10^{-5}	**.403**	**.415**	.359
10^{-4}	.399	.411	$\boxed{\textbf{.425}}$
10^{-3}	.368	.347	.354
10^{-2}	.368	.338	.355
Parkinson's Condition (ρ)			
10^{-5}	.238	.368	$\boxed{\textbf{.390}}$
10^{-4}	.433	.467	.300
10^{-3}	.434	**.492**	.236
10^{-2}	**.434**	.491	.237
Eating Condition (UAR [%])			
10^{-5}	51.1	–	48.0
10^{-4}	59.7	–	60.6
10^{-3}	**61.3**	–	$\boxed{\textbf{65.9}}$
10^{-2}	60.9	–	65.9
10^{-1}	60.9	–	65.9

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.

5. 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, Special Issue on Sensing Emotion and Affect - Facing Realism in Speech Processing*, 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] C. Teixeira and I. Trancoso, "Continuous and semi-continuous hmm for recognising non-native pronunciations," in *Proc. IEEE Workshop ASR*, 1993, pp. 26–27.

[4] C. Teixeira, H. Franco, E. Shriberg, K. Precoda, and K. Somnez, "Prosodic features for automatic text-independent evaluation of degree of nativeness for language learners," in *Proc. ICSLP*, Beijing, 2000, 4 pages.

[5] J. Tepperman, T. Stanley, K. Hacioglu, and B. Pellom, "Testing suprasegmental english through parroting," in *Proc. of Speech Prosody 2010, May 11-14, 2010, Chicago IL, USA,*, 2010.

[6] F. Hönig, T. Bocklet, K. Riedhammer, A. Batliner, and E. Nöth, "The automatic assessment of non-native prosody: Combining classical prosodic analysis with acoustic modelling," in *Proc. Interspeech, Portland Oregon, USA*, 2012, pp. 823–826.

[7] G. Stebbing and C. Goetz, "Factor structure of the unified parkinsons disease rating scale: Motor examination section," *Movement Disorders*, vol. 13, pp. 633–636, 1998.

[8] O. Hornykiewicz, "Biochemical aspects of parkinson's disease," *Neurology*, vol. 51, no. 2, pp. S2–S9, 1998.

[9] L. Ramig, C. Fox, and S. Sapir, "Speech treatment for parkinson's disease," *Expert Review Neurotherapeutics*, vol. 8, no. 2, pp. 297–309, 2008.

[10] M. Little, P. McSharry, E. Hunter, J. Spielman, and L. Ramig, "Suitability of dysphonia measurements for telemonitoring of parkinson's disease," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 4, pp. 1015–1022, 2009.

[11] A. Tsanas, M. Little, P. McSharry, J. Spielman, and L. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of Parkinsons disease," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 5, pp. 1264–1271, 2012.

[12] J. C. Vásquez-Correa, J. R. Orozco-Arroyave, J. D. Arias-Arias-Londoño, J. F. Vargas-Bonilla, and E. Nöth, "New computer aided device for real time analysis of speech of people with Parkinson's disease," *Fac. Ing. Univ. Antioquia*, vol. Sept., no. 72, pp. 87–103, 2014.

[13] J. E. Flege, S. G. Fletcher, and A. Homiedan, "Compensating for a bite block in /s/ and /t/ production: Palatographic, acosutic, and perceptual data," *J. Acoust. Soc. Am*, vol. 83, pp. 212–228, 1988.

[14] K. M. Hiiemae, J. B. Palmer, S. W. Medicis, J. Hegener, B. S. Jackson, and D. E. Lieberman, "Hyoid and tongue surface movements in speaking and eating," *Archives of Oral Biology*, vol. 47, pp. 11–27, 2002.

[15] O. Amft, "Automatic dietary monitoring using on-body sensors: Detection of eating and drinking behaviour in healthy individuals," Ph.D. dissertation, ETH Zurich, 2008.

[16] K. Yatani and K. N. Truong, "BodyScope: A Wearable Acoustic Sensor for Activity Recognition," in *UbiComp '12: Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. ACM, 2012, pp. 341–350.

[17] T. Rahman, A. T. Adams, M. Zhang, E. Cherry, B. Zhou, H. Peng, and T. Choudhury, "BodyBeat: A Mobile System for Sensing Non-Speech Body Sounds," in *MobiSys '14: Proceedings of the 12th annual international conference on Mobile systems, applications, and services*. ACM, 2014, pp. 2–13.

[18] G. C. David, A. C. Garcia, A. W. Rawls, and D. Chand, "Listening to what is said - transcribing what is heard: the impact of speech recognition technology (SRT) on the practice of medical transcription (MT)," *Sociology of Health & Illness*, vol. 31, no. 6, pp. 924–938, 2009.

[19] D. Hymes, "Introduction: Toward Ethnographies of Communication," *American Anthropologist*, vol. 66, pp. 1–34, 1964.

[20] S. Hantke, F. Weninger, R. Kurle, A. Batliner, and B. Schuller, "I hear you eat and speak: automatic recognition of eating condition and food type," ms., to appear, has been distributed to the participants.

[21] F. Hönig, A. Batliner, and E. Nöth, "Automatic Assessment of Non-Native Prosody – Annotation, Modelling and Evaluation," in *Proceedings of the International Symposium on Automatic Detection of Errors in Pronunciation Training (IS ADEPT)*, Stockholm, 2012, pp. 21–30, available at http://www5.informatik.uni-erlangen.de/Forschung/Publikationen/2012/Hoenig12-AAO.pdf.

[22] W. Menzel, E. Atwell, P. Bonaventura, D. Herron, P. Howarth, R. Morton, and C. Souter, "The ISLE corpus of non-native spoken English," in *Proc. LREC*, Athens, 2000, pp. 957–964.

[23] F. Hönig, A. Batliner, and E. Nöth, "How many labellers revisited – naives, experts and real experts," in *Proceedings of L2WS/SLaTE, Workshop on Second Language Studies: Acquisition, Learning, Education and Technology/Speech and Language Technology in Education, September 22-24, 2010, Tokyo, Japan*, 2010, pp. 137–140.

[24] F. Hönig, A. Batliner, K. Weilhammer, and E. Nöth, "Islands of failure: Employing word accent information for pronunciation quality assessment of english L2 learners," in *Proc. of SLATE*, Wroxall Abbey, 2009.

[25] J. Orozco-Arroyave, J. Arias-Londoño, J. Vargas-Bonilla, M. González-Rátiva, and E. Nöth, "New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease," in *Proc. LREC*, 2014, pp. 342–347, available at https://www5.informatik.uni-erlangen.de/Forschung/Publikationen/2014/Orozco14-NSS.pdf.

[26] T. Haderlein, C. Moers, B. Möbius, F. Rosanowski, and E. Nöth, "Intelligibility rating with automatic speech recognition, prosodic, and cepstral evaluation," in *Proceedings of Text, Speech and Dialogue (TSD)*, ser. Lecture Notes in Artificial Intelligence, vol. 6836. Berlin, Heidelberg: Springer, 2011, pp. 195–202.

[27] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani *et al.*, "The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism," in *Proc. Interspeech*, Lyon, France, 2013, pp. 148–152.

[28] B. Schuller, S. Steidl, A. Batliner, J. Epps, F. Eyben, F. Ringeval, E. Marchi, and Y. Zhang, "The INTERSPEECH 2014 Computational Paralinguistics Challenge: Cognitive & physical load," in *Proc. Interspeech*, Singapore, September 2014, pp. 427–431.

[29] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proc. ACM Multimedia*. Florence, Italy: ACM, 2010, pp. 1459–1462.

[30] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in *Proc. ACM MM*. Barcelona, Spain: ACM, October 2013, pp. 835–838.

[31] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[32] E. Coutinho, J. Deng, and B. Schuller, "Transfer Learning Emotion Manifestation Across Music and Speech," in *Proceedings of the 2014 International Joint Conference on Neural Networks (IJCNN) as part of the IEEE World Congress on Computational Intelligence (IEEE WCCI)*, Beijing, China, July 2014, pp. 3592–3598.

[33] J. Deng, Z. Zhang, and B. Schuller, "Linked Source and Target Domain Subspace Feature Transfer Learning – Exemplified by Speech Emotion Recognition," in *Proc. ICPR 2014*, Stockholm, Sweden, August 2014, pp. 761–766.