The INTERSPEECH 2017 Computational Paralinguistics Challenge: Addressee, Cold & Snoring

Björn Schuller[1,2], *Stefan Steidl*[3], *Anton Batliner*[2,3], *Elika Bergelson*[4], *Jarek Krajewski*[5], *Christoph Janott*[6], *Andrei Amatuni*[4], *Marisa Casillas*[7], *Amanda Seidl*[8], *Melanie Soderstrom*[9], *Anne S. Warlaumont*[10], *Guillermo Hidalgo*[5], *Sebastian Schnieder*[5], *Clemens Heiser*[6], *Winfried Hohenhorst*[11], *Michael Herzog*[12], *Maximilian Schmitt*[2], *Kun Qian*[6], *Yue Zhang*[1,6], *George Trigeorgis*[1], *Panagiotis Tzirakis*[1], *Stefanos Zafeiriou*[1,13]

[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]Psychology and Neuroscience, Duke University, USA
[5]University of Wuppertal, Germany
[6]Technische Universität München, Germany
[7]Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
[8]Speech, Language, and Hearing Sciences, Purdue University, USA
[9]Psychology, University of Manitoba, Canada
[10]Cognitive and Information Sciences, University of California, Merced, USA
[11]Clinic for ENT Medicine, Head and Neck Surgery, Alfried Krupp Krankenhaus, Essen, Germany
[12]Clinic for ENT Medicine, Head and Neck Surgery, Carl-Thiem-Klinikum, Cottbus, Germany
[13]University of Oulu, Finland
schuller@IEEE.org

Abstract

The INTERSPEECH 2017 Computational Paralinguistics Challenge addresses three different problems for the first time in research competition under well-defined conditions: In the *Addressee* sub-challenge, it has to be determined whether speech produced by an adult is directed towards another adult or towards a child; in the *Cold* sub-challenge, speech under cold has to be told apart from 'healthy' speech; and in the *Snoring* sub-challenge, four different types of snoring have to be classified. In this paper, we describe these sub-challenges, their conditions, and the baseline feature extraction and classifiers, which include data-learnt feature representations by end-to-end learning with convolutional and recurrent neural networks, and bag-of-audio-words for the first time in the challenge series.

Index Terms: Computational Paralinguistics, Challenge, Addressee, Child Directed Speech, Speech under Cold, Snoring

1. Introduction

In this INTERSPEECH 2017 COMPUTATIONAL PARALINGUISTICS CHALLENGE COMPARE – the ninth since 2009 [1], we address three new problems within the field of Computational Paralinguistics [2] in a challenge setting:

In the *Addressee* ***(A)*** *Sub-Challenge*, speech produced by an adult has to be classified as directed to either adult or child. A possible application is the monitoring of adult-(parent-) child interaction – it is well-known that already babies should be exposed to a highly elaborated verbal interaction.

In the *Cold* ***(C)*** *Sub-Challenge*, speech under cold has to be told apart from speech under 'normal' health conditions. A possible application is the monitoring of call-centre and other telephone interactions in order to predict propagation of a cold.

In the *Snoring* ***(S)*** *Sub-Challenge*, a four-class classification of snoring sounds has to be performed. Identification of the type of snoring [3, 4] can be highly useful for a targeted and thus successful medical treatment [5].

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 that may be used. Participants have to use predefined training/development/test splits for each sub-challenge. They may report development results obtained from the training set (preferably with the supplied evaluation setups), but have only a limited number of five trials to upload their results on the test sets for the 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 in the Challenge. As evaluation measure, 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 the next section 2, we describe the challenge corpora. Section 3 describes the baselines experiments and metrics for each sub-challenge as well as the baseline results; a short conclusion is given in section 4.

Table 1: *Databases: Number of instances per class in the train/devel/test splits used for the Challenge; CD: child-directed, AD: adult-directed, C: cold, NC: non-cold, V: velum, O: oropharyngeal lateral walls, T: tongue base, E: epiglottis.*

#	Train	Devel	Test	Σ
Homebank Child/Adult Addressee Corpus (HB-CHAAC)				
CD	2 302	2 182	blinded	during challenge
AD	1 440	1 368	blinded	during challenge
Σ	3 742	3 550	3 594	10 886
Upper Respiratory Tract Infection Corpus (URTIC)				
C	970	1 011	blinded	during challenge
NC	8 535	8 585	blinded	during challenge
Σ	9 505	9 596	9 551	28 652
Munich-Passau Snore Sound Corpus (MPSSC)				
V	168	161	blinded	during challenge
O	76	75	blinded	during challenge
T	8	15	blinded	during challenge
E	30	32	blinded	during challenge
Σ	282	283	263	828

2. Challenge Corpora

2.1. Addressee (A)

In this sub-challenge, we introduce the HOMEBANK CHILD/ADULT ADDRESSEE CORPUS (HB-CHAAC) (see Table 1). The task is to differentiate between speech produced by an adult that is directed to a child (child-directed speech, CDS) or directed to another adult (adult-directed speech, ADS). CDS is understood to have particular acoustic-phonetic and linguistic characteristics that distinguish it from ADS and is theorized to play a critical role in promoting language development (e. g., [6] and references therein). However, to date there have been few formal attempts to discriminate these forms of speech computationally (cf. [7, 8]). Furthermore, analyses of CDS vs ADS have been restricted to highly constrained contexts. The HB-CHAAC consists of a set of conversations (see below) selected from a much larger corpus of real-world child language recordings known as HomeBank [9] (*homebank.talkbank.org*). A set of 20 such conversations was selected from a subset of the available HomeBank recordings, from the following corpora: [10], [11], and [12]. The subset of recordings that were sampled (61 homes in total across four cities in North America) featured: North American English as the primary language being spoken, typically-developing children, participants who granted permission to share the audio with the research community, and a spread of ages sampled as uniformly as possible between 2 and 24 months and across the four contributing laboratory datasets, with each child only sampled once, cf. Table 1. Recordings were collected using the LENA recording device and software [13] that provides automated identification of 'conversational blocks' (bouts of audio identified as speech bounded by 5 s of non-speech on either end), which were used to select the 1 220 conversations, consisting of total 2 523 minutes of recording. Individual adult speaker audio clips within each conversation (as identified by the proprietary LENA algorithms speaker diarisation) were then subjected to hand-annotation for the challenge. Three trained research assistants judged whether each clip was directed to a child (CDS) or an adult (ADS) using both acoustic-phonetic information and context (see *https://osf.io/d9ac4/* for more detail). Clips deemed to be non-speech, not produced by an adult, or ambiguous between CDS/ADS were excluded using a "Junk" category. All CDS and ADS clips were additionally labeled by the research assistant as to whether the speaker was male or female. Annotators achieved high reliability in differentiating CDS/ADS (Fleiss' kappa $> .75$, $p < .001$).

Table 2: *Distribution of recordings and child age across the four contributing datasets; rec.: recordings, range: Age range in months (mean)*

Sub-Corpus	**# of rec.**	**range**
Bergelson Seedlings	44	6 -17 (11.75)
McDivitt	7	4 -19 (11.29)
VanDam2	3	5 -18 (12.33)
Warlaumont	7	2 -9 (3.71)

2.2. Cold (C)

For this sub-challenge, the UPPER RESPIRATORY TRACT INFECTION CORPUS (URTIC) is provided by the Institute of Safety Technology, University of Wuppertal, Germany (see also Table 1). The corpus consists of recordings of 630 subjects (382 m, 248 f, mean age 29.5 years, standard deviation 12.1 years, range 12-84 years), made in quiet rooms with a microphone/headset/hardware setup (sample rate 44.1 kHz, downsampled to 16 kHz, quantisation 16 bit).

The participants had to complete different tasks, presented to them on a computer monitor. The subjects were asked to read out short stories, e. g., *The North Wind and the Sun* (widely used within the field of phonetics), and *Die Buttergeschichte* (a standard reading passage in German, used in speech and language pathology). Furthermore, the participants were producing voice commands as needed, e. g., for controlling driver assistance systems, and numbers from 1 to 40. Besides scripted speech, spontaneous narrative speech was recorded. Subjects were asked to briefly tell about, e. g., their last weekend, their best vacation, or to describe a picture. The whole session lasted from 15 minutes to 2 hours, while the number of tasks varied for each subject. The available recordings were split into 28 652 chunks with a duration between 3 s and 10 s.

To obtain the state of health, each participant reported a binary one-item measure based on the German version of the *Wisconsin Upper Respiratory Symptom Survey (WURSS-24)* [14], assessing the symptoms of common cold. The global illness severity item (on a scale of $0 =$ *not sick* to $7 =$ *severely sick*) was binarised using a threshold at 6. According to this binary label, the chunks were divided into speaker-independent partitions (balanced w. r. t. gender, age, and experimenter) with 210 speakers for each partition. In the training and development partitions, 37 participants were having a cold and 173 participants were not having a cold. The number of chunks per subject varies; the total duration is approximately 45 hours.

2.3. Snoring (S)

The MUNICH-PASSAU SNORE SOUND CORPUS (MPSSC) is introduced for classification of snore sounds by their excitation location within the upper airways. Snoring is generated by vibrating soft tissue in the upper airways during inspiration in sleep. Although simple snoring is not harmful for the snorer him- or herself, it can affect sleep quality of the bed partner, cause social disturbance, and is known to affect partnerships. There a numerous conservative and surgical methods attempting to improve or cure snoring, many of them showing only moderate success. Key to better clinical results is a treat-

ment exactly targeting the area in the upper airways where the snoring sound is generated in the individual patient. Basis material for the corpus are uncut recordings from Drug Induced Sleep Endoscopy (DISE) examinations from three medical centres recorded between 2006 and 2015. Recording equipment, microphone type, and location differ between the medical centers, so do the background noise characteristics.

During a DISE procedure, a flexible nasopharyngoscope is introduced into the upper airways while the patient is in a state of artificial sleep. Vibration mechanisms and locations can be observed while video and audio signals are recorded. DISE is an established diagnostic tool, which has a number of disadvantages: it is time consuming, puts the patient under strain, and cannot be performed during natural sleep. Therefore it is desirable to develop alternative methods for the classification of snore sounds, e. g., based on acoustic features.

More than 30 hours of DISE recordings have been automatically screened for audio events. The extracted events were manually selected, non-snore events and events disturbed by non-static background noise (such as speech or acoustic signals from medical equipment) were discarded. The remaining snore events have been classified by medical ENT (ear, nose, and throat) experts based on findings from the video recordings. Only events with a clearly identifiable, single site of vibration and without obstructive disposition were included in the database. Four classes are defined based on the VOTE scheme, a widely used scheme distinguishing four structures that can be involved in airway narrowing and obstruction [15, 16]:

V - Velum (palate), including soft palate, uvula, lateral velopharyngeal walls; **O** - Oropharyngeal lateral walls, including palatine tonsils; **T** - Tongue, including tongue base and airway posterior to the tongue base; **E** - Epiglottis.

The resulting database contains audio samples (raw PCM, sample rate 16 000 Hz, quantisation 16 bit) of 843 snore events from 224 subjects (see Table 1). The number of events per class in the database is strongly unbalanced, with 85 % of samples from the classes V and O, 11 % E-events and 5 % T-snores. This is in line with the likelihood of occurance during normal sleep [17, 18]. Subject number, centre and class are coded in the filename, the meta-information age and gender are available.

3. Experiments and Results

3.1. End-to-end Learning

For the first time in the COMPARE challenge, we provide results using end-to-end learning (e2e) models. These deep models have had huge success in the vision community and even more recently in speech applications such as emotion recognition [19], speaker verification [20], speech recognition [21], and further audio analysis tasks (e. g., [22]). An attractive characteristic of these models is that the optimal features for a given task can be learnt purely form the data at hand, i. e., we aim to learn simultaneously the optimal features and the classifier in a single optimisation problem. Similar to [19] we use a convolutional network to extract features from the raw time representation and then a subsequent recurrent network (LSTM) which performs the final classification. For training the network, we split the raw waveform into chunks of 40 ms each, which are fed into a convolutional network. The convolutional network is comprised by a series of alternating convolution and pooling operations which try to find a robust representation of the original signal. The extracted features are then subsequently fed to M LSTM modules (cf. Table 3) which compress the temporal signal to a single final hidden state of the recurrent network which is then used to perform the final classification [1]. As these models rely purely on the statistics of the available data to learn the optimal features, we assume the available data to contain a large amount of variation. We expect that the performance of the e2e models can be improved by using smart data augmentation techniques modelling the data distribution properly.

3.2. COMPARE Acoustic Feature Set

The official baseline feature set is the same as has been used in the four previous editions of the INTERSPEECH COMPARE challenges [23, 24, 25, 26]. This feature set contains 6 373 static features resulting from the computation of various functionals over low-level descriptor (LLD) contours. The configuration file is the IS13_ComParE.conf, which is included in the 2.1 public release of openSMILE [27, 28]. A full description of the feature set can be found in [29].

3.3. Bag-of-Audio-Words

In addition to the default ComParE feature set, where functionals (statistics) are applied to the acoustic LLDs, we provide bag-of-audio-words (BoAW) features. In the BoAW method, the 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 one for the 65 deltas of these LLDs. In Table 3, results are given for different codebook sizes. Codebook generation is done by *random sampling* from the LLDs in the training data. When fusing training and development data for the final model, the codebook is learnt again from the fused data.

BoAW has already been applied successfully for, e. g., acoustic event detection [30], speech-based emotion recognition [31], and classification of snore sounds [32]. The LLDs have been extracted with the openSMILE toolkit [28], BoAW have been computed using openXBOW [33].

3.4. Basics for the Challenge Baselines

The primary evaluation measure for the sub-challenges (all being classification tasks) is Unweighted Average Recall (UAR). The motivation to consider *unweighted* rather than weighted average recall ('conventional' accuracy) is that it is also meaningful for highly unbalanced distributions of instances among classes (as is the case for the **S** sub-challenge).

For the sake of transparency and reproducibility of the baseline computation, we use open-source implementations from the data mining algorithms (WEKA 3, revision 3.8.1; [34]) for functionals and BoAW; in line with previous years, the machine learning paradigm chosen is Support Vector Machines (SVM), in particular WEKA's SVM implementation with linear kernels. In all tasks the Sequential Minimal Optimisation (SMO; [35]) as implemented in WEKA was used as training algorithm.

Features were scaled to zero mean and unit standard deviation (option `-N 1` for Weka's SMO), using the parameters from the training set (when multiple folds where used for development, the parameters were calculated on the training set of each fold). For all tasks, the complexity parameter C was optimised during the development phase.

The results for late fusion are also reported. We fused the predictions of the 3-layer e2e model and the COMPARE functionals and BoAW models that performed best on the respective

[1] A detailed implementation of these models can be found at `https://github.com/trigeorgis/ComParE2017`

Table 3: *Results for the three sub-challenges. The* **official baselines** *for test are highlighted (bold and greyscale). Dev: Development. M: Number of LSTM layers in end-to-end (e2e) learning. C: Complexity parameter of SVM. N: Codebook size of Bag-of-Audio-Words (BoAW) splitting the input into two codebooks (ComParE-LLDs/ComParE-LLD-Deltas), with 10 assignments per frame, optimised complexity parameter of SVM.* ***UAR****: Unweighted Average Recall.*

UAR [%]	**A**ddressee		**C**old		**S**noring	
	Dev	Test	Dev	Test	Dev	Test
M	**e2e: CNN + LSTM**					
2	59.8	60.1	59.1	60.0	37.0	37.9
3	60.9	59.1	58.6	59.6	40.3	40.3
C	**ComParE functionals + SVM**					
10^{-6}	55.8	65.8	62.9	63.9	29.3	48.4
10^{-5}	60.5	67.7	64.0	70.2	31.1	51.4
10^{-4}	61.8	67.6	61.7	66.5	40.6	**58.5**
10^{-3}	59.4	64.6	58.1	61.9	39.2	55.6
10^{-2}	57.4	60.9	58.8	59.5	39.2	55.6
10^{-1}	57.4	59.6	60.0	58.4	39.2	55.6
N	**ComParE BoAW + SVM**					
$125/125$	63.2	67.5	55.9	62.8	43.8	48.7
$250/250$	61.4	66.6	62.8	66.5	46.6	49.9
$500/500$	62.4	68.2	63.9	66.7	44.2	51.2
$1000/1000$	62.2	67.2	64.2	67.3	42.8	50.0
$2000/2000$	63.4	67.7	64.1	67.3	41.0	48.3
$4000/4000$	63.4	68.2	63.8	67.2	39.8	48.2
$8000/8000$	63.3	68.3	64.0	69.7	36.6	47.8
Models	**Late fusion**					
e2e+func	66.3	69.0	62.6	64.8	38.9	55.8
e2e+BoAW	67.8	68.4	62.7	62.5	45.1	46.0
func+BoAW	62.8	68.7	64.2	70.1	42.1	52.4
All (conf.)	66.4	**70.2**	66.1	70.7	43.5	53.0
All (maj.)	64.0	68.0	65.2	**71.0**	43.4	55.6

development partitions. For the fusion of two models, for each instance, the label with the highest confidence was chosen. For the fusion of all three models, we selected the final prediction after two different rules: the label with the highest sum of confidence (conf.) and a majority vote (maj.).

Each sub-challenge package includes scripts that allows participants to reproduce the baselines and perform the testing in a reproducible and automatic way (including pre-processing, model training, model evaluation, and scoring by the competition and further measures).

3.5. Baselines

This year, we introduced several new approaches: Besides the usual ComParE features plus SVM, we employ e2e and BoAW plus SVM; additionally, we present different late fusion results. By that, we – as organisers – have more than 5 trials available; doing that, at the same time, we open new avenues of research for the participants. This comes at a cost – the results given in Table 3 show a dilemma: If we followed "the rules of the game" and take those results for test that correspond to the best results for Dev(elopment), we would end up with challenge baselines that are markedly below other results for test depicted in the table. By that, participants could surpass the official baseline by just repeating and/or slightly modifying the procedures leading to the better results in Table 3. To avoid that, we thus decided simply to choose the best test results for each sub-challenge as official baseline. These results are still obtained by employing non-optimised standard procedures; thus, there is ample space for surpassing these figures.

As can be seen in Table 3, for the Adressee sub-challenge, the baseline is UAR $= 70.2\,\%$; for the Cold sub-challenge, it is UAR $= 71.0\,\%$, and for the Snoring sub-challenge, it is UAR $= 58.5\,\%$. All but two of the 60 classifications (20 for each of the three tasks) show the expected gain for Test in comparison to Dev, due to the increased training set (Train plus Dev). This is most pronounced for Snoring when functionals are employed (up to >20% absolute difference). Note that the four classes in this task display a highly un-balanced distribution, and that we use UAR – which means that mis-classifying a few cases (due to different acoustic properties of a few subjects) or modelling such cases better in a sparse class, influences UAR to a larger extent than weighted average recall. The two factors responsible for the mismatch of best Dev vs. best Test might be: (1) in the 20 classifications for each task, this might happen once simply by chance, cf. the Addressee and Cold tasks. (2) the marked difference that we only observe for Snoring might be due as well to the unbalanced distribution between classes.

4. Conclusion

This year's challenge is new in several respects; besides three new tasks – Addressee, Cold, and Snoring, all of them being highly relevant for applications – we introduced several new procedures: Both e2e and BoAW are less knowledge-based because they either do not need the usual extraction of features (e2e) but only the time signal, or they do not need a lexicon but generate a quasi-word representation themselves (BoAW). The learning procedures employed for functionals and BoAW are standard - competitive, but not optimised and kept generic for all tasks by intention to provide transparent and easily redoable processing steps. For all computation steps, scripts are provided that can but need not be used by the participants.

We expect participants to obtain considerably better performance measures by employing novel (combinations of) procedures and features including such tailored to the particular tasks.

Beyond the tasks featured in this challenge series, there remains a broad variety of further information that is conveyed in the acoustics of speech and the spoken words themselves that have not been dealt with either at all or in a well-defined competition framework. Many of these bear, however, great application potential, and remain to be investigated more closely.

5. Acknowledgements

This research has received funding from the EU's Framework Programme HORIZON 2020 Grant No. 115902 (RADAR CNS), the EU's 7^{th} Framework Programme ERC Starting Grant No. 338164 (iHEARu), and the EU Horizon 2020 Research & Innovation Action Grant No. 645378 (ARIA-VALUSPA), as well as from SSHRC Insight Grant (#435-2015-0628), ERC Advanced Grant INTERACT (269484), NIH DP5-OD019812, and NSF SBE-1539129 and NSF BCS-1529127. Further, the support of the EPSRC Centre for Doctoral Training in High Performance Embedded and Distributed Systems (HiPEDS, Grant Reference EP/L016796/1) is gratefully acknowledged. The authors thank the research assistants who provided HB-CHAAC labels and Kelsey Dyck for help developing the labelling protocol, and the sponsors of the Challenge: audEERING GmbH, and the Association for the Advancement of Affective Computing (AAAC). The responsibility lies with the authors.

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, 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] T. Mikami, Y. Kojima, M. Yamamoto, and M. Furukawa, "Recognition of breathing route during snoring for simple monitoring of sleep apnea," in *Proceedings of SICE Annual Conference*. IEEE, 2010, pp. 3433–3434.

[4] D. L. Herath, U. R. Abeyratne, and C. Hukins, "Hmm-based snorer group recognition for sleep apnea diagnosis," in *Proceedings EMBC*. IEEE, 2013, pp. 3961–3964.

[5] C. Janott, B. Schuller, and C. Heiser, "Acoustic information in snoring noises," *HNO*, vol. 65, no. 2, pp. 107–116, February 2017.

[6] M. Soderstrom, "Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants," *Developmental Review*, vol. 27, pp. 501–532, 2007.

[7] A. Batliner, B. Schuller, S. Schaeffler, and S. Steidl, "Mothers, Adults, Children, Pets – Towards the Acoustics of Intimacy," in *Proceedings of ICASSP*, Las Vegas, NV, 2008, pp. 4497–4500.

[8] S. Schuster, S. Pancoast, M. Ganjoo, M. C. Frank, and D. Jurafsky, "Speaker-independent detection of child-directed speech," in *Spoken Language Technology Workshop (SLT)*, South Lake Tahoe, CA, 2014, pp. 366–371.

[9] M. VanDam, A. Warlaumont, E. Bergelson, A. Cristia, M. Soderstrom, D. P. P., and B. MacWhinney, "An online repository of daylong child-centered audio recordings," *Seminars in Speech and Language*, vol. 37, no. 2, pp. 128–42, 2016.

[10] E. Bergelson, "Bergelson Seedlings HomeBank corpus," 2016, doi:10.21415/T5PK6D.

[11] K. McDivitt and M. Soderstrom, "McDivitt HomeBank corpus," 2016, doi: 10.21415/T5KK6G.

[12] A. S. Warlaumont and G. M. Pretzer, "Warlaumont HomeBank corpus," 2016, doi:10.21415/T54S3C.

[13] C. R. Greenwood, K. Thiemann-Bourque, D. Walker, J. Buzhardt, and J. Gilkerson, "Assessing childrens home language environments using automatic speech recognition technology," *Communication Disorders Quarterly*, vol. 32, pp. 83–92, 2011.

[14] B. Barrett, R. L. Brown, M. P. Mundt, G. R. Thomas, S. K. Barlow, A. D. Highstrom, and M. Bahrainian, "Validation of a short form wisconsin upper respiratory symptom survey (wurss-21)," *Health and Quality of Life Outcomes*, vol. 7, no. 1, p. 76, 2009.

[15] N. Charakorn and E. Kezirian, "Drug-Induced Sleep Endoscopy," *Otolaryngol Clin North Am.*, vol. 49, pp. 1359–1372, 2016.

[16] E. Kezirian, W. Hohenhorst, and N. de Vries, "Drug-induced sleep endoscopy: the VOTE classification," *Eur Arch Otorhinolaryngol.*, vol. 268, pp. 1233–1236, 2011.

[17] N. Hessel and N. de Vries, "Diagnostic work-up of socially unacceptable snoring. II. Sleep endoscopy," *Eur Arch Otorhinolaryngol.*, vol. 259, pp. 158–161, 2002.

[18] J. A. Fiz and R. Jane, "Snoring Analysis. A Complex Question," *J Sleep Disor: Treat Care*, vol. 1, pp. 1–3, 2012.

[19] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proceedings of ICASSP*. IEEE, 2016, pp. 5200–5204.

[20] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *Proceedings of ICASSP*. IEEE, 2016, pp. 5115–5119.

[21] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," 2016.

[22] P. Smaragdis, "End-to-end music transcription using a neural network," *Journal of the Acoustical Society of America*, vol. 140, no. 4, pp. 3038–3038, 2016.

[23] 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 *Proceedings of INTERSPEECH*, Lyon, France, 2013, pp. 148–152.

[24] 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 *Proceedings of INTERSPEECH*, Singapore, 2014, pp. 427–431.

[25] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang, and F. Weninger, "The INTERSPEECH 2015 Computational Paralinguistics Challenge: Degree of Nativeness, Parkinson's & Eating Condition," in *Proceedings of INTERSPEECH*, Dresden, Germany, 2015, pp. 478–482.

[26] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language," in *Proceedings of INTERSPEECH*, 2016, pp. 2001–2005.

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

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

[29] F. Eyben, *Real-time Speech and Music Classification by Large Audio Feature Space Extraction*, ser. Springer Theses. Switzerland: Springer International Publishing, 2015.

[30] H. Lim, M. J. Kim, and H. Kim, "Robust sound event classification using LBP-HOG based bag-of-audio-words feature representation," in *Proceedings of INTERSPEECH*. Dresden, Germany: ISCA, 2015, pp. 3325–3329.

[31] 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 *Proceedings of INTERSPEECH*. San Francisco, USA: ISCA, 2016, pp. 495–499.

[32] M. Schmitt, C. Janott, V. Pandit, K. Qian, C. Heiser, W. Hemmert, and B. Schuller, "A bag-of-audio-words approach for snore sounds' excitation localisation," in *Proceedings of ITG Speech Communication*, VDE. Paderborn, Germany: IEEE, 2016, pp. 230–234.

[33] M. Schmitt and B. W. Schuller, "openXBOW-introducing the passau open-source crossmodal bag-of-words toolkit," *preprint arXiv:1605.06778*, 2016.

[34] 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.

[35] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in large margin classifiers*, A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, Eds. Cambridge, MA: MIT Press, 1999, pp. 61–74.