SEMI-AUTONOMOUS DATA ENRICHMENT BASED ON CROSS-TASK LABELLING OF MISSING TARGETS FOR HOLISTIC SPEECH ANALYSIS

Yue Zhang[1], *Yuxiang Zhou*[1], *Jie Shen*[1], *Björn Schuller*[1]

[1]Department of Computing, Imperial College London, London, Unites Kingdom
{yzhang9,yuxiang.zhou10,jie.shen07,bjoern.schuller}@imperial.ac.uk

ABSTRACT

In this work, we propose a novel approach for large-scale data enrichment, with the aim to address a major shortcoming of current research in computational paralinguistics, namely, looking at speaker attributes in isolation although strong interdependencies between them exist. The scarcity of multi-target databases, in which instances are labelled for different kinds of speaker characteristics, compounds this problem. The core idea of our work is to join existing data resources into one single holistic database with a multi-dimensional label space by using semi-supervised learning techniques to predict missing labels. In the proposed new Cross-Task Labelling (CTL) method, a model is first trained on the labelled training set of the selected databases for each individual task. Then, the trained classifiers are used for the crosslabelling of databases among each other.To exemplify the effectiveness of the 'CTL' method, we evaluated it for likability, personality, and emotion recognition as representative tasks from the INTERSPEECH Computational Paralinguistics ChallengE (ComParE) series. The results show that 'CTL' lays the foundation for holistic speech analysis by semi-autonomously annotating the existing databases, and expanding the multi-target label space at the same time, while achieving higher accuracy as the baseline performance of the challenges.

Index Terms— Data enrichment, semi-supervised learning, missing labels, multi-target learning, holistic speech analysis

1. INTRODUCTION

With recent technology advances, the automatic analysis and understanding of speaker characteristics has received major attention due to the multiplicity of possible applications in information and communication systems, including not only natural human-machine interaction (e.g., speech-based assistants adapting to the personality and age of the user), but also multimedia information retrieval (e.g., video tagging), and monitoring of safety and security critical environments (e.g., detection of intoxication or sleepiness).

Today, analysis of speaker characteristics and non-linguistic information has emerged to a major research field with a somewhat standardised methodology that provides 'reasonable' results for a variety of single tasks, which consider speaker characteristics in isolation, i.e., single or only few speaker attributes are considered at once. Consequently, there is very little exploitation of the interplay and synergies between different speaker traits, states, and speaking styles, yet in reality, strong interdependencies between bits of paralinguistic information exist. On a 'superficial' level, biological primitives such as height, gender, and ethnicity are coupled to some degree. For example, acoustic models for gender classification (male vs female) should be different by age, since arguably the most important feature, pitch, is also influenced by age. A seminal study has recently shown that considering interdependencies between age, height, gender and ethnicity can significantly improve accuracy of speaker trait prediction [1]. In a more subtle manner, personality influences the way that emotional states manifest in speech, and especially also the subjective likability (i. e., agreeableness) of a speaker's voice by others. Still, before these interdependencies can be exploited on a larger scale, richly annotated data sets will have to be created: At present, typically encountered databases are labelled for single tasks only. Following the belief 'there is no data like more data', speech data covering as many speaker characteristics as possible should be collected to help generalisation of systems to real-world use cases. Unfortunately, one of the major barriers of today's research is the costly consequences of obtaining human annotations, which are time-consuming and expensive to obtain.

In this work, we propose a Cross-Task Labelling (CTL) method that enables large-scale data enrichment to overcome the aforementioned shortcomings. The idea is to join existing data resources into one holistic database annotated in multi-faceted paralinguistic dimensions. To this end, we consider Semi-Supervised Learning (SSL) techniques to successively complete the missing labels in each single database without the need for human annotation [2, 3]. Particularly, we use *self-training* that trains a classifier on a small labelled data set and re-trains the model iteratively with the most confident machine predictions for a unlabelled data pool [4]. A great advantage of this technique is that it does not require any intervention of human oracles for the annotation of unlabelled data. Considering actual scenarios 'in the wild', for which we assume that the reference labels in form of 'ground truth' or 'gold standard' are not available, we evaluate the correctness of the predicted labels by reproducing the baseline with the created holistic database as the new training set and the original test set for each task. In this way, the annotation quality can be accessed by comparing the new and previous baseline accuracy.

Through data fusion and multi-target learning, we aim to bridge the major gap between today's systems and humans analysing speech in a holistic fashion, learning how speaker characteristics influence each other, and continuously improving their skills from interactions with others.

In what follows, we describe the related work in Section 2. In Section 3, we explain our proposed 'CTL' method. Then, in Section 4 and Section 5, we describe the database and feature set, respectively. The experimental setup and the results are presented in Section 6. In Section 7, we discuss our findings and explore possible extensions of this work.

The research leading to these results has received funding from the European Unions Framework Programme HORIZON 2020 under the Grant No. 645378 (ARIA-VALUSPA).

2. RELATED WORK

In contrast to traditional *single-target* learning that deals with learning from data, where each example is associated with a single label l from a set of disjoint labels $L, |L| > 1$, *multi-target learning* is concerned with learning from data, where each training example is associated with multiple labels taken form a finite set of labels $Y \subseteq L$ [5]. For simplicity, we do not distinguish between the terms 'multi-label', 'multi-task', and 'multi-target' in this work. For details on the discrimination between the concepts the reader is referred to [6]. There is currently a wealth of multi-target learning methods, which can be categorised into two main categories [7]: a) problem transformation methods, and b) algorithm adaption methods. Besides these two categories of methods for multi-target learning, Madjarov et al. [8] have introduced a third category: ensemble methods. In our previous work [6], a few initial experiments were carried out on the INTERSPEECH 2012 Speaker Trait Challenge dataset (Likability Subchallenge) by using state-of-the-art multi-target learning methods of the MEKA toolkit, which is an extension to the WEKA machine learning framework by adding support for multi-target classification [9]. More specifically, we compared the performance of the Ensembles of Classifier Chains (ECC) [10], the Ensembles of Class Relevance (ECR) method, and the 'Oracle' Multi-Target (OMT) learning method, which trains single-task classifiers, while including the correct labels of the other tasks as features. As the obtained benchmark results show clear signs of overfitting, poor generation performance can be attributed to very limited amounts of training data. Yet, there have been a few promising impulses, which will be the starting point of our work. In this paper, we address the scarcity of multi-target databases and alleviate overfitting to enable holistic processing of paralinguistic information.

3. METHODOLOGY

In this section, we explain the principle of the 'CTL' method based on self-training and confidence measure. The data fusion process aims to overcome the scarcity of multi-target databases by joining existing corpora that are annotated in a related context (e. g., emotion, likability, personality). The main issue here is that typically encountered databases are labelled for single or a few tasks only. Thus, the (partially) missing labels in each single database need to be completed in order to create a 'universal' data collection pool that is labelled in a common multi-dimensional label space shared by all databases. For crosslabelling, we resort to self-training which is a well-known SSL method based on the principle of confidence measure, in such a way that the predicted classes with higher certainty levels are automatically labelled and added to the training set.

3.1. Confidence Measure

For the confidence measure, we use Support Vector Machines (SVMs) as the classification model. SVMs are supervised learning models that construct decision hyperplanes to separate instances of different classes by using the decision function $f(\mathbf{x})$, while maximising the functional margin. For each instance, the output distances to the decision boundaries are then transformed into probability values through a parametric method of logistic regression [11]. As there can be more than two classes when it comes to multi-target data, the following consideration needs to be made. In case of binary classification, the confidence value for the predicted class is obtained by forming the difference of the posterior probabilities $P_0(\mathbf{x}), P_1(\mathbf{x})$ for classes '0' and '1', respectively.

$$C(\mathbf{x}) = |P_1(\mathbf{x}) - P_0(\mathbf{x})| \tag{1}$$

For the classification of more than two classes, the highest two probabilities are deducted from each other.

Formally, the query function for self-training is defined as:

$$\mathbf{x} = \arg\max_{\mathbf{x}} |C(\mathbf{x})|, \tag{2}$$

where $C(\mathbf{x})$ denotes the confidence value assigned to the predicted label of a given instance $\mathbf{x}$.

3.2. Cross-Task Labelling

The algorithm of the 'CTL' method based on self-training is depicted in Figure 1. We define the following notations: $\mathcal{L}$ denotes the labelled training set of a specific task. $\mathcal{U}$ comprises all the training data of the other tasks, where the missing target labels are indicated with the question mark '?'. The self-learning process starts by training a model on the labelled data and subsequently using this model to classify the instances from the unlabelled data pool $\mathcal{U}$. The confidence values are ranked and stored in a queue (in descending order). Accordingly, a subset $\mathcal{N}_{st} \subset \mathcal{U}$ that is classified with the highest confidence is selected and added to the training set, together with their predicted labels. Then, the classifier is retrained and the iterative process is repeated until all unlabelled instances are annotated. In the outer loop, the self-training process is performed for every task, resulting in a holistic database containing the completely filled label space regarding all tasks.

Algorithm: *Cross-Task Labelling*
Repeat for each task:
Repeat until $\mathcal{U} \in \{\}$:

1. (Optional) Upsample training set $\mathcal{L}$ to even class distribution $\mathcal{L}_D$
2. Use $\mathcal{L}/\mathcal{L}_D$ to train classifier $\mathcal{H}$, then classify $\mathcal{U}$
3. Select a subset $\mathcal{N}_{st}$ that contains those instances predicted with the highest confidence values
4. Remove $\mathcal{N}_{st}$ from the unlabelled set $\mathcal{U}, \mathcal{U} = \mathcal{U} \setminus \mathcal{N}_{st}$
5. Add $\mathcal{N}_{st}$ to the labelled set $\mathcal{L}$, $\mathcal{L} = \mathcal{L} \cup \mathcal{N}_{st}$

Fig. 1. Pseudocode description of the Cross-Task Labelling (CTL) algorithm

4. DATABASE

4.1. Speaker Likability Database

The "Speaker Likability Database" (SLD) [12] is a subset of the German Agender database [13] (800 speakers, age and gender balanced). Likability ratings for each recording used were obtained from 32 annotators on a 7-point Likert scale. To establish a consensus from the individual likability ratings, the evaluator weighted estimator (EWE) [14] was used. The EWE is a weighted mean, with weights corresponding to the 'reliability' of each rater, which is the cross-correlation of her/his rating with the mean rating (over all raters). The EWE rating was discretised into the 'likable' (L) and 'non-likable' (NL) classes based on the median EWE rating of

Table 1. *Partitioning into training and test sets. Binary classification: Speaker Likability Database by (L: likable / NL: non-likable); Speaker Personality Corpus by (X: high on trait X / NX: low on trait X, X ∈ { O, C, E, A, N }); GEMAP by (pos(itive) / neg(ative); arousal (A) and valence (V), where 'undefined' is excluded from evaluation in binary tasks)*

Sub-Task	#	Train	Test	Σ
LIKABILITY	L	281	119	400
	NL	291	109	400
OPENNESS	O	167	80	247
	NO	272	121	393
CONSCIENTIOUS.	C	191	99	290
	NC	248	102	350
EXTRAVERSION	E	213	107	320
	NE	226	94	320
AGREEABLENESS	A	178	105	323
	NA	221	96	317
NEUROTICISM	N	228	90	318
	NN	211	111	322
AROUSAL	pos	380	220	600
	neg	390	210	600
	–*	48	12	60
VALENCE	pos	384	216	600
	neg	386	214	600
	–*	48	12	60
Σ		1829	871	2700

all stimuli in the SLD. The data were partitioned into training and test based on the subdivision in the INTERSPEECH 2012 Speaker Trait Challenge (IS12 STC) (Age and Gender Sub-Challenges). The resulting partitioning is shown in Table 1.

4.2. Speaker Personality Corpus

The "Speaker Personality Corpus" (SPC) bases on 640 clips extracted from the French speaking Swiss national broadcasting service. The number of speakers is 322. Eleven raters annotated each recording in terms of the perceived personality of the speakers using BFI-10, a personality assessment questionnaire. The personality traits assessed correspond to the Big-Five personality dimensions: OPENNESS to new experiences; CONSCIENTIOUSNESS; EXTRAVERSION; AGREEABLENESS; and NEUROTICISM. Each clip is labelled to be above average (X) for a given trait X ∈ { O, C, E, A, N } if at least six judges (the majority) assign to it a score higher than their average for the same trait; otherwise, it is labelled NX. Training and test set are defined by speaker independent subdivision of the SPC, stratifying by speaker gender (cf. Table 1).

4.3. Geneva Multimodal Emotion Portrayals

The "Geneva Multimodal Emotion Portrayals" (GEMEP) contains 1.2 k instances of emotional speech from ten professional actors in 18 categories [15]. Applying the same heuristic approach as in the INTERSPEECH 2013 ComParE Emotion sub-challenge, these classes are mapped to the two dimensions arousal and valence (binary tasks). The category 'undefined' is considered in the cross-labelling process, but excluded from evaluation in binary tasks to ensure a fair comparison between the challenge baseline and the 'CTL' method. The resulting partitioning is shown in Table 1.

5. ACOUSTIC FEATURES

5.1. ComParE Acoustic Feature Set

The COMPARE set of supra-segmental (utterance-level) acoustic features is used, as for the baselines in the previous instalments of the challenge series [16, 17, 18]. The COMPARE feature set contains 6 373 static features, which are obtained as functionals of low-level descriptor (LLD) contours. We use TUM's open-source openSMILE feature extractor in its 2.1 release [19].

5.2. Extended Geneva Minimalistic Acoustic Parameter Set

For alleviating overfitting, we further applied the "extended Geneva Minimalistic Acoustic Parameter Set" (eGEMAPS) [20], which was selected based on a) their potential to index affective physiological changes in voice production, b) their proven value in former studies as well as their automatic extractability, and c) their theoretical significance. The minimalistic acoustic parameter set contains a compact set of 18 Low-level descriptors (LLD), containing 62 parameters. It does not contain any cepstral parameters and only very few dynamic parameters (i. e., it contains no delta regression coefficients and no difference features; only the slopes of rising and falling F_0 and loudness segments encapsulate some dynamic information). Further, especially cepstral parameters have proven highly successful in modelling of affective states, e. g., by [21], [22], [23]. Therefore, an *extension* set to the minimalistic set is used which contains the following 7 LLD in addition to the 18 LLD in the minimalistic set:

Spectral (balance/shape/dynamics) parameters:

- **MFCC 1–4** Mel-Frequency Cepstral Coefficients 1–4.
- **Spectral flux** difference of the spectra of two consecutive frames.

Frequency related parameters:

- **Formant 2–3 bandwidth** added for completeness of Formant 1–3 parameters.

As ***functionals***, the ***arithmetic mean*** and the ***coefficient of variation*** are applied to all of these 7 additional LLD to all segments (voiced and unvoiced together), except for the formant bandwidths to which the functionals are applied only in voiced regions. This adds 14 extra descriptors. Additionally, the arithmetic mean of the spectral flux in unvoiced regions only, the arithmetic mean and coefficient of variation of the spectral flux and MFCC 1–4 in voiced regions only is included. This results in another 11 descriptors. Additionally the **equivalent sound level** is included. This results in 26 extra parameters. In total, when combined with the Minimalistic Set, the eGeMAPS set contains 88 parameters.

6. EXPERIMENTS AND RESULTS

As performance measure, we retain the choice of unweighted average recall (UAR) in accordance with the IS challenges [24]. In the given case of two classes ('X' and 'NX'), it is calculated as (Recall(X)+Recall(NX))/2, i. e., the number of instances per class is ignored by intention to compensate imbalances. For transparency and reproducibility, we used open-source classifier implementations of SVMs from the WEKA data mining toolkit [9]. As classifiers, we chose linear kernel SVMs trained with Sequential Minimal Optimization (SMO), as they are robust against over-fitting in high dimensional feature spaces. For each task, we chose the complexity parameter $C \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ for the SMO algorithm

Table 2. Personality, Likability, *and* Emotion Sub-Challenge *performance measures by unweighted average recall (UAR) in percent; C: complexity parameter (by 5-fold cross-validation (CV); Test: Results on test set, by training on the original training set for baseline, and on the joint multi-target training set for CTL; Acoustic feature sets: ComParE vs eGeMAPS.*

UAR [%]	*ComParE*						*eGeMAPS*					
	Baseline			**CTL**			**Baseline**			**CTL**		
	C	**CV**	**Test**	C	**CV**	**Test**	C	**CV**	**Test**	C	**CV**	**Test**
IS12 STC Personality Sub-Challenge												
(N)O	10^{-2}	60.3	**60.8**	10^{-1}	76.4	55.0	10^{-3}	59.0	**56.0**	10^{-3}	39.3	**56.0**
(N)C	10^{-2}	72.5	**78.1**	10^{-2}	77.5	**78.1**	10^{-1}	70.8	**78.1**	10^{-2}	80.0	69.5
(N)E	10^{-1}	79.1	71.7	10^{-2}	86.6	**72.8**	10^{-1}	75.8	**71.4**	10^{-3}	81.7	68.0
(N)A	10^{-3}	65.9	58.2	10^{-2}	88.0	**61.0**	10^{-1}	65.1	58.6	10^{-1}	76.8	**63.1**
(N)N	10^{-2}	71.9	63.3	10^{-2}	90.1	**66.3**	10^{-1}	72.3	**65.5**	10^{-1}	89.5	65.0
Mean		69.9	66.4		83.7	**66.6**		68.6	**65.9**		73.5	64.3
IS12 STC Likability Sub-Challenge												
(N)L	10^{-4}	58.3	57.2	10^{-1}	82.3	**60.3**	10^{-3}	57.2	**58.6**	10^{-1}	82.2	53.5
IS13 ComParE Emotion Sub-Challenge												
Arousal	10^{-1}	97.0	68.9	10^{-1}	97.7	**69.0**	10^{-1}	85.0	72.9	10^{-1}	86.5	**73.2**
Valence	10^{-1}	96.7	**61.6**	10^{-1}	97.6	59.3	10^{-1}	68.3	**60.0**	10^{-1}	80.9	58.5
Mean		96.9	**65.3**		97.7	64.2		76.7	**66.5**		83.7	65.9

that achieves the best UAR value in the 5-fold crossvalidation on the training set. An argument in favour of the crossvalidation instead of the original training vs development partition as used in the challenges is that the data fusion process leads to increasing data volume of the training data while the development set would retain the same size. At each learning iteration of the self-training process, we selected a subset $\mathcal{N}_{st}$ comprising 30% of the unlabelled data to be merged into the labelled training set. This iterative process is repeated until the target label is completed in all databases. The values for C turned by CV as well as the performance measures are depicted in Table 2. According to the results, the following observations can be made. First, it can be seen that the CV results obtained through 'CTL' are markedly better than the corresponding baseline results for each feature set. The performance increase can be explained by the fusion of the three databases. Simply put, the joint training set contains only around $1/3$ of the samples originating from human annotation. The rest of the instances are estimates obtained through 'CTL'. Hence, in each iteration of the crossvalidation, the validation fold also contains $2/3$ of its samples, which are labelled by 'CTL', causing heavily biased results. Second, we found that the ComParE set leads to increased tendency of overfitting. This is expected because the dimensionality of this feature set is much higher than the total number of instances. In comparison, the CV results are closer to the 'test' results when using the eGeMAPS feature set. Third, there is a slight (although statistically insignificant) tendency that 'CTL' leads to improved performance on the test set by using the ComParE set. Finally, we computed the Student's t-test to statistically compare the test performances of baseline vs CTL for each feature set. The analysis of the significance levels ($p = 0.87 >> .05$) confirms the correctness of the crosslabelling process by showing that there is no significant difference between the test results. This fact is of paramount importance for our data enrichment approach because it demonstrates the effectiveness of the 'CTL' method to fuse existing databases into one multi-target database that can be expanded in a continuously increasing label space, thus enabling large-scale data collection with truly multi-dimensional ('universal') labels.

7. CONCLUSION

In this paper, we introduced a cross-task labelling method that overcomes the scarcity of multi-target databases by semi-autonomously completing the target labels which are partially or completely missing. Through data fusion, the 'CTL' method generates a multi-target database, while achieving the comparable test results as the original baseline performance of the ComParE challenges. In this way, a 'universal' database with a continuously expanding label space can be created, enabling large-scale data enrichment.

For future research, we aim to exploit the label correlations between different tasks and benefit from the created multi-target database. Methods such as Conditional Random Fields and Dynamic Bayesian Networks, are also promising candidates, which will be investigated for their suitability for multi-target learning. Dynamic Bayesian Networks allow for modelling of interdependencies between all desirable labels and the input data. Combined with nonparametric learning methods and hybrid/tandem approaches, they are one powerful method for holistic speaker characteristics analysis.

Through multi-target data collection and learning from them, we aim to achieve a holistic understanding of all the paralinguistic facets of human speech in tomorrows real-life information, communication and entertainment systems.

8. REFERENCES

[1] B. Schuller, M. Wöllmer, F. Eyben, G. Rigoll, and D. Arsić, "Semantic Speech Tagging: Towards Combined Analysis of Speaker Traits," in *Proc. AES 42nd International Conference*, Ilmenau, Germany, July 2011, pp. 89–97, Audio Engineering Society.

[2] X. Zhu, "Semi-supervised learning literature survey," Tech. Rep. TR 1530, Department of Computer Sciences, University of Wisconsin at Madison, Madison, WI, 2006.

[3] O. Chapelle, B. Schölkopf, A. Zien, et al., *Semi-Supervised Learning*, MIT Press, Cambridge, MA, 2006.

[4] D. Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods," in *Proc. of the 33rd annual meeting on Association for Computational Linguistics*, Stroudsburg, PA, 1995, Association for Computational Linguistics, pp. 189–196.

[5] M. Zhang and Z. Zhou, "A Review On Multi-Label Learning Algorithms," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 26, no. 8, pp. 1819–1837, 2014.

[6] B. Schuller, Y. Zhang, F. Eyben, and F. Weninger, "Intelligent systems' Holistic Evolving Analysis of Real-life Universal speaker characteristics," in *Proc. of the 5th International Workshop on Emotion Social Signals, Sentiment & Linked Open Data (ES^{3}LOD 2014), satellite of the 9th Language Resources and Evaluation Conference (LREC 2014)*, B. Schuller, P. Buitelaar, L. Devillers, C.Pelachaud, T. Declerck, A. Batliner, P. Rosso, and S. Gaines, Eds., Reykjavik, Iceland, May 2014, pp. 14–20, ELRA.

[7] G. Tsoumakas and I. Katakis, "Multi-label Classification: An Overview," *International Journal of Data Warehousing and Mining (IJDWM)*, vol. 3, no. 3, pp. 1–13, 2007.

[8] G. Madjarov, D. Kocev, D. Gjorgjevikj, and S. Džeroski, "An Extensive Experimental Comparison of Methods for Multi-label Learning," *Pattern Recognition*, vol. 45, no. 9, pp. 3084–3104, 2012.

[9] 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.

[10] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier Chains for Multi-label Classification," *Machine learning*, vol. 85, no. 3, pp. 333–359, 2011.

[11] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in large margin classifiers*, pp. 61–74. MIT Press, Cambridge, MA, 1999.

[12] F. Burkhardt, B. Schuller, B. Weiss, and F. Weninger, "'Would You Buy A Car From Me?' – On the Likability of Telephone Voices," in *Proc. Interspeech*, Florence, Italy, 2011, pp. 1557–1560.

[13] F. Burkhardt, M. Eckert, W. Johannsen, and J. Stegmann, "A Database of Age and Gender Annotated Telephone Speech.," in *Proc. LREC*, 2010.

[14] M. Grimm and K. Kroschel, "Evaluation of natural emotions using self assessment manikins," in *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding*, Cancun, Mexico, 2005, pp. 381–385.

[15] T. Bänziger, M. Mortillaro, and K.R. Scherer, "Introducing the Geneva Multimodal expression corpus for experimental research on emotion perception," *Emotion*, vol. 12, pp. 1161–1179, 2012.

[16] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism," in *Proc. INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association*, Lyon, France, August 2013, pp. 148–152, ISCA.

[17] 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. of INTERSPEECH*, Singapore, Singapore, September 2014, ISCA.

[18] B. Schuller, S. Steidl, A. Batliner, S.Hantke, F. Hönig, J. Orozco-Arroyave, E. Nöth, Y. Zhang, and F. Weninger, "The INTERSPEECH 2015 Computational Paralinguistics Challenge: Degree of Nativeness, Parkinson's & Eating Condition," in *Proceedings INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association*, Dresden, Germany, September 2015, ISCA, 5 pages.

[19] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proc. of ACM MM*, Barcelona, Spain, 2013, pp. 835–838.

[20] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. André, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Transactions on Affective Computing*, 2015, 14 pages.

[21] B. Schuller, A. Batliner, D.Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson, "The Relevance of Feature Type for the Automatic Classification of Emotional User States: Low Level Descriptors and Functionals," in *Proc. INTERSPEECH 2007, 8th Annual Conference of the International Speech Communication Association*, Antwerp, Belgium, August 2007, pp. 2253–2256, ISCA.

[22] B. Schuller and G.Rigoll, "Recognising Interest in Conversational Speech – Comparing Bag of Frames and Supra-segmental Features," in *Proceedings INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association*, Brighton, UK, September 2009, pp. 1999–2002, ISCA.

[23] E. Marchi, A. Batliner, B. Schuller, S. Fridenzon, S. Tal, and O. Golan, "Speech, Emotion, Age, Language, Task, and Typicality: Trying to Disentangle Performance and Feature Relevance," in *Proc. First International Workshop on Wide Spectrum Social Signal Processing (WS^{3}P 2012), held in conjunction with the ASE/IEEE International Conference on Social Computing (SocialCom 2012)*, Amsterdam, The Netherlands, September 2012, IEEE.

[24] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 9, pp. 1062–1087, 2011.