The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism

Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus R. Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, et al.

The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism*

Björn Schuller[1,2], Stefan Steidl[3], Anton Batliner[2], Alessandro Vinciarelli[4,5], Klaus Scherer[1]
Fabien Ringeval[6], Mohamed Chetouani[7], Felix Weninger[2], Florian Eyben[2], Erik Marchi[2],
Marcello Mortillaro[1], Hugues Salamin[4], Anna Polychroniou[4], Fabio Valente[5], Samuel Kim[5]

[1]Université de Genève, Swiss Center for Affective Sciences, Switzerland
[2]TU München, Machine Intelligence & Signal Processing group, MMK, Germany
[3]FAU Erlangen-Nuremberg, Pattern Recognition Lab, Germany
[4]University of Glasgow, School of Computing Science, Scotland
[5]IDIAP Research Institute, Martigny, Switzerland
[6]Université de Fribourg, Document Image and Voice Analysis group, Switzerland
[7]Université Pierre et Marie Curie, ISIR, Paris, France

schuller@tum.de

Abstract

The INTERSPEECH 2013 Computational Paralinguistics Challenge provides for the first time a unified test-bed for Social Signals such as laughter in speech. It further introduces conflict in group discussions as a new task and deals with autism and its manifestations in speech. Finally, emotion is revisited as task, albeit with a broader range of overall twelve enacted emotional states. In this paper, we describe these four Sub-Challenges, their conditions, baselines, and a new feature set by the openSMILE toolkit, provided to the participants.

Index Terms: Computational Paralinguistics, Challenge, Social Signals, Conflict, Emotion, Autism

1. Introduction

With the INTERSPEECH's 2009 Emotion Challenge [1], 2010 Paralinguistic Challenge [2], 2011 Speaker State Challenge [3], and the recent 2012 Speaker Trait Challenge [4] we organised challenges and official exchange fora as exist for many more 'traditional' speech tasks, comparable to the NIST evaluations (cf. e. g., [5]) or related audio and text processing disciplines such as the MIREX [6], CLEF, and TREC challenges in the field of Music and Text Information Retrieval [7]. The novel Challenge for INTERSPEECH 2013 broadens the scope and increases the number of tasks and new databases provided in response to the increased participation [8]: In line with the INTERSPEECH 2013's theme "Speech in Life Sciences and Human Societies", we address the novel tasks social signals [9] and conflict in communication [10] – extending to dyadic speech and speaker group analysis in realistic every-day conditions. As our previous tasks, these have never been addressed in such an open and strictly regulated comparison, but bear highest application potential. Further – due to the former popularity – we revisit speech emotion and speech pathology with new data. This includes as novelty the recognition of speakers with autism spectrum condition by their acoustics [11, 12, 13] and enacted emotion data taking into account potential differences to naturalistic data [14, 15]. We subsume these tasks under the umbrella of Computational Paralinguistics [16] and introduce the INTERSPEECH 2013 COMPUTATIONAL PARALINGUISTICS CHALLENGE (COMPARE). Due to space limitations, we cannot elaborate on tasks and methodologies but have to confine ourselves to the basics and 'technicalities' necessary for the participants. Four Sub-Challenges are addressed:

In the *Social Signal Sub-Challenge*, non-linguistic events such as laughter or sigh of a speaker have to be detected and localised based on acoustics.

In the *Conflict Sub-Challenge*, group discussions have to be automatically evaluated with the aim of recognising conflict.

In the *Emotion Sub-Challenge*, the emotion of a speaker within a closed set has to be determined by a suited learning algorithm and acoustic features.

In the *Autism Sub-Challenge*, the type of pathology of a speaker has to be determined by a suited classification algorithm and acoustic features.

The measures of competition will be Unweighted Average Recall (UAR) and temporal deviation depending on the Sub-Challenge. In addition, Area Under the receiver operating Curve (AUC) and Correlation Coefficient (CC) will partially be given. The transcription of the train and development sets will be known. Contextual knowledge may be used, as the sequence of chunks will be given. All Sub-Challenges allow contributors to find their own features and machine learning algorithm. However, a novel standard feature set will be provided per corpus that may be used. Participants will have to stick to the definition of training, development, and test sets. They may report on results obtained on the development set, but have only five trials to upload their results on the test sets, whose labels are unknown to them. Each participation will be accompanied by a paper presenting the results that 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.

* This is a preliminary version. Baselines may be updated during the ongoing Challenge.

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007–2013) under grant agreement No. 289021 (ASC-Inclusion). The authors would further like to thank the sponsors of the challenge, the HUMAINE Association and the Social Signal Processing Network (SSPNet). The responsibility lies with the authors.

2. Challenge Corpora

2.1. SSPNet Vocalisation Corpus (SVC)

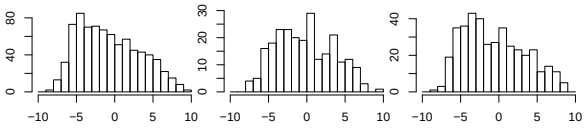
In the *Social Signals Sub-Challenge*, the "SSPNet Vocalization Corpus" (SVC) serves for analyses and comparison. It is composed of 2 763 audio clips (11 seconds length each) annotated in terms of laughter and fillers. Laughter [17, 18, 19] can indicate amusement, joy, scorn, or embarrassment. Fillers [20] are vocalisations like "uhm", "eh", "ah", etc.; they indicate attempts to hold the floor. The corpus was extracted from a collection of 60 phone calls involving 120 subjects (63 female, 57 male) [21]. The participants of each call were fully unacquainted and never met face-to-face before or during the experiment. The calls revolved around the Winter Survival Task: The two participants had to identify objects (out of a predefined list) that increase the chances of survival in a polar environment. The subjects were not given instructions on how to conduct the conversation, the only constraint was to discuss only one object at a time. The conversations were recorded on both phones (model Nokia N900) used during the call. The clips were extracted from the microphone recordings of the phones. Therefore they contain the voice of one speaker only. Each clip lasts for 11 seconds and was selected in such a way that it contains at least one laughter or filler event between $t = 1.5$ seconds and $t = 9.5$ seconds. Clips from the same speaker never overlap. In contrast, clips from two subjects participating in the same call may overlap (for example in the case of simultaneous laughter). However, they do not contain the same audio data because they are recorded with different microphones. Overall, the database contains 3.0 k filler events and 1.2 k laughter events. Both types of vocalisation can be considered fully spontaneous. The SVC will serve to evaluate features and algorithms for the determination and localisation of speakers' social signals in speech. By that, the *Social Signals Sub-Challenge* for the first time introduces a frame-wise detection and localisation task instead of supra-segmental classification as in the other Sub-Challenges and all previous Challenges. For the purpose of the Challenge, the data were divided into speaker disjoint subsets for training, development, and testing. For transparency, this was simply done by using calls 1–35 (70 speakers) for training, calls 36–45 (20 speakers) for development, and calls 46–60 for testing. The Challenge data are delivered with a manual segmentation of the training and development data into 'garbage', 'laughter', and 'filler' segments, in the 'master label file' (MLF) format used by the Hidden Markov Model Toolkit (HTK) [22]. Further meta data is not provided. The resulting partitioning by numbers of utterances, number of vocalisation segments (filler, laughter) as well as vocalisation and garbage frames (100 per second), is shown in Table 1.

2.2. SSPNet Conflict Corpus (SC^2)

In the *Conflict Sub-Challenge*, the "SSPNet Conflict Corpus" (SC^2) is used [23]. It contains 1 430 clips of 30 seconds extracted from the Canal9 Corpus – a collection of 45 Swiss political debates (in French) – including 138 subjects in total: 23 females (1 moderator and 22 participants) and 133 males (3 moderators and 120 participants). The clips have been annotated in terms of conflict level by roughly 550 assessors recruited via Amazon Mechanical Turk. Each clip is assigned a continuous conflict score in the range $[-10, +10]$, giving rise to a straightforward regression task ('Score' task). A binary classification task is created based on these labels, namely to classify into 'high' (≥ 0) or 'low' (< 0) level of conflict ('Class' task). As several subjects occur in debates with different moderators, a truly speaker independent partitioning is not possible on the data. Since all participants except the moderators do not occur more than a few times (most of them only once), the following strategy was followed to reduce speaker dependence to a minimum. All broadcasts with the female moderator (speaker # 50) were assigned to the training set. The development set consists of all broadcasts moderated by the (male) speaker # 153, and the test set comprises the rest (male moderators). This also ensures that the development and test sets are similar in case that the gender of the moderator should have an influence. The resulting partitioning is shown in Table 2 along with the distribution of binary class labels and continuous ratings (Figure 1) among the partitions. As meta data, manual speaker segmentation, as well as role (participant / moderator) and gender of the subjects are provided for the training and development sets. Participants are encouraged to use the manual speaker segmentation for development of features extraction, but for the test set an automatic speaker diarisation system has to be used. Freely available alternatives for this task comprise the LIUM [24][1] and Alize LIA-RAL[2] toolkits.

Table 1: *Partitioning of the SSPNet Vocalisation Corpus into train, dev(elopment), and test sets: Numbers of utterances, vocalisation segments (laughter, filler), and vocalisation / 'garbage' frames.* [1]*: 79 572 after training set balancing by re-sampling.*

#	train	dev	test	Σ
Utterances				
Σ	1 583	500	680	2 763
Segments				
laughter	649	225	284	1 158
filler	1 710	556	722	2 988
Frames				
laughter	59 294	25 750	23 994	109 038
filler	85 034	29 432	35 459	149 925
garbage	1 591 442[1]	492 607	684 937	2 768 986
Σ	1 735 770	547 789	744 390	3 027 949



(a) train (b) devel (c) test

Figure 1: *Level of conflict (*$\in [-10, +10]$*) histograms for the Challenge partitions of the SSPNet Conflict Corpus.*

Table 2: *Partitioning of the SSPNet Conflict Corpus into train, dev(elopment), and test sets for binary classification ('low'* $\equiv [-10, 0[$*, 'high'* $\equiv [0, +10]$*).*

#	train	dev	test	Σ
low	471	127	226	824
high	322	113	171	606
Σ	793	240	397	1 430

2.3. Geneva Multimodal Emotion Portrayals (GEMEP)

For the *Emotion Sub-Challenge* we selected the "Geneva Multimodal Emotion Portrayals" (GEMEP) [25]. It contains 1.2 k instances of emotional speech from ten professional actors (five

[1] http://lium3.univ-lemans.fr/diarization/doku.php/welcome
[2] http://mistral.univ-avignon.fr/index_en.html

Table 3: *Partitioning of the GEMEP database into train, dev(elopment), and test sets for 12-way classification by emotion category, and binary classification by pos(itive) / neg(ative) arousal (A) and valence (V).* $^{+}$*: Mapped to 'other' and excluded from evaluation in 12-class task.* **: Mapped to 'undefined' and excluded from evaluation in binary tasks.*

#	train	dev	test	A	V	Σ
admiration^{+}	20	2	8	pos	pos	30
amusement	40	20	30	pos	pos	90
anxiety	40	20	30	neg	neg	90
cold anger	42	12	36	neg	neg	90
contempt^{+}	20	6	4	neg	neg	30
despair	40	20	30	pos	neg	90
disgust^{+}	20	2	8	–*	–*	30
elation	40	12	38	pos	pos	90
hot anger	40	20	30	pos	neg	90
interest	40	20	30	neg	pos	90
panic fear	40	12	38	pos	neg	90
pleasure	40	20	30	neg	pos	90
pride	40	12	38	pos	pos	90
relief	40	12	38	neg	pos	90
sadness	40	12	38	neg	neg	90
shame^{+}	20	2	8	pos	neg	30
surprise^{+}	20	6	4	–*	–*	30
tenderness^{+}	20	6	4	neg	pos	30
Σ	602	216	442			1 260

female) in 18 categories. In the Challenge task, these will need to be classified into 12 categories (multi-class task). We further provide results for the two dimensions arousal and valence (binary tasks) training on these binary targets and mapping from the 12 categories after classification. The GEMEP database contains prompted speech comprising sustained vowel phonations, as well as two 'nonsensical' phrases with two different intended sentence modalities, each expressed by each actor in various degrees of regulation (emotional intensity) ranging from 'high' to 'masked' (hiding the true emotion). Given this layout, a partitioning that is both text and speaker independent is not feasible. Hence, the following strategy was followed: Vowels and phrase #2 are used for training and development, subdividing by speaker ID, and phrase #1 is used for testing. Masked regulation utterances are only contained in the test set in order to alleviate potential model distortions. By the above partitioning, we obtain text independence. Since six of the 18 emotional categories are extremely sparse (≤ 30 instances in the entire GEMEP database), we restrict the evaluation to the 12 most frequent ones in the multi-class classification task. For the arousal / valence tasks, mappings are only defined for selected categories such as to obtain a balanced distribution of positive / negative arousal and valence among the categories. Nevertheless, the remaining data is given to the participants (with labels in 18 categories for the training and development sets); it can be used, e. g., to train 'background' or 'garbage' models. The resulting partitioning is shown in Table 3. As meta data, actor IDs, prompts, and intended regulation are provided for the training and development sets.

2.4. Child Pathological Speech Database (CPSD)

The *Autism Sub-Challenge* is based upon the "Child Pathological Speech Database" (CPSD) [26]. It provides speech as recorded in two university departments of child and adolescent psychiatry, located in Paris, France (Université Pierre et Marie Curie/Pitié-Salpêtière Hospital and Université René Descartes/Necker Hospital). The dataset used in the Sub-Challenge contains 2.5 k instances of speech recordings from 99 children aged 6 to 18 years. 35 of these children show Pervasive Development Disorders either of autism spectrum condition (PDD, 10 male, 2 female), specific language impairment such as dysphasia (DYS, 10 male, 3 female) or PDD Non-Otherwise Specified (NOS, 9 male, 1 female) according to the DSM-IV criteria. A monolingual control group consists of 64 further children (TYP, 52 male, 12 female). The French speech includes prompted sentence imitation of 26 sentences representing different modalities (declarative, exclamatory, interrogative, and imperative) and four types of intonations (descending, falling, floating, and rising). Two evaluation tasks have been defined: a binary 'Typicality' task (typically vs. atypically developing children), and a four-way 'Diagnosis' task (classifying into the above named categories). Partitioning into training, development and test data is done by order of speaker ID, stratified by age and gender of the children, and speaker-independent; class distribution is given in Table 4. As speaker meta data, age and gender of the children are given.

Table 4: *Partitioning of the Child Pathological Speech Database into train, dev(elopment), and test sets for four-way classification by diagnosis, and binary classification by typical / atypical development. Diagnosis classes: TYPically developing, Pervasive Developmental Disorders (PDD), pervasive developmental disorders Non-Otherwise Specified (NOS), and specific language impairment such as DYSphasia.*

#	train	dev	test	Σ
Typically developing				
TYP	566	543	542	1651
Atypically developing				
PDD	104	104	99	307
NOS	104	68	75	247
DYS	129	104	104	337
Σ	903	819	820	2542

3. Challenge Features

For the baseline acoustic feature set used in this Challenge, we slightly modified the acoustic feature set used for the INTERSPEECH 2012 Speaker Trait Challenge [4] – the most effective used in the series so far. Again, we use TUM's open-source openSMILE feature extractor [27] and provide extracted feature sets on a per-chunk level (except for SVC). Configuration files for openSMILE will be provided together with the next openSMILE public release. Voice quality features (jitter and shimmer) were slightly improved, Viterbi smoothing for F_0 was added, and the exceptions which functionals are applied to which LLD were simplified. The set includes energy, spectral, cepstral (MFCC) and voicing related low-level descriptors (LLDs) as well as a few LLDs including logarithmic harmonic-to-noise ratio (HNR), spectral harmonicity, and psychoacoustic spectral sharpness. Altogether, the 2013 COMPARE feature set contains 6 373 features. For the *Social Signals Sub-Challenge* that requires localisation, a frame-wise feature set is derived from the above. Taking into account space and memory requirements, only a small set of descriptors are calculated per frame, following a sliding window scheme to combine frame-wise LLDs and functionals. In particular, frame-wise MFCCs 1–12 and logarithmic energy are computed along with their first and second order delta (Δ) regression coefficients as typically used in speech recognition. They are augmented by voicing probability, HNR, F0 and zero-crossing rate, as well as their first order Δs. Then, for each frame-wise LLD the arithmetic mean and standard deviation across the frame itself and eight of its neighbouring frames (four

before and four after) are calculated. This results in $47 \times 3 = 141$ descriptors per frame.

4. Challenge Baselines

As primary evaluation measure, we retain the choice of unweighted average recall as used since the first Challenge held in 2009 [1]. 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 given in the *Autism Sub-Challenge*. Given the nature of the *Social Signals Sub-Challenge's* detection task, we also consider the Area Under the Curve measure [28] for the laughter and filler classes on frame level (100 frames per second); the unweighted average (UAAUC) is the official competition measure of this Sub-Challenge. For this reason, participants are for the first time required to also submit posterior class probabilities ('confidences') per frame in this Sub-Challenge. Besides, in the *Conflict Sub-Challenge*, we additionally consider the Pearson correlation coefficient as evaluation criterion for regression on the 'continuous valued' original labels, following the 2010 Challenge which also featured a regression task [2]. A novelty of this year's Challenge is the provision of a 'recipe' for re-producing the baseline classification and regression results on the development set in an automated fashion, including pre-processing, model training, model evaluation, and scoring by the competition and further measures. For transparency and reproducibility, we use open-source classifier implementations from the WEKA data mining toolkit [29]. To provide a (somewhat) unified scheme for tackling the various tasks in this Challenge, we restrict ourselves to static classification (regression) for all tasks. To this end, linear kernel Support Vector Machines (SVM) / Support Vector Regression (SVR) are used, which are known to be robust against overfitting. As training algorithm, we use Sequential Minimal Optimisation (SMO). For each task, we choose the SVM complexity parameter $C \in \{10^{-3}, 10^{-2}, 10^{-1}, 1\}$ that achieves best UAR on the development set. To obtain (pseudo) class posteriors, logistic models are fitted to the SVM hyperplane distance based on the training set. To cope with imbalanced class distribution in the *Autism Sub-Challenge*, up-sampling is applied. The under-represented categories (PDD, PDD-NOS, SLI) in the four-way 'Diagnosis' task are upsampled by using a factor of five; in the binary 'Typicality' task a factor of two is applied. Conversely, for the *Social Signals Sub-Challenge*, down-sampling is used, where only 5 % of the 'garbage' frames are kept. No re-sampling of the training set is done for the other Sub-Challenges. The baseline recipe provided to the participants performs training set re-sampling in a reproducible way. For evaluation on the test set, we re-train the models using the training and development set, applying re-sampling as above.

Let us briefly summarise the baseline results as displayed in Table 5. Due to the different nature of the tasks and evaluation measures, we also present chance level baselines. For AUC and CC, these are obtained as mean and standard deviation over 25 random trials on the development set, using random class posteriors (AUC) or prediction of Gaussian random numbers with mean and standard deviation of the training set labels (CC). For UAR, they are defined as an equal class distribution (50 % for 2, 25 % for 4, and 8.33 % for 12 classes). In the *Social Signals Sub-Challenge*, detection of fillers seems slightly 'easier' than detection of laughter, and for both a somewhat acceptable performance in terms of AUC (83.3 % baseline UAAUC on test) is achieved – yet, showing the challenge of vocalisation localisation in naturalistic recordings of spontaneous speech. In the *Conflict Sub-Challenge*, it turns out that the simple baseline already delivers a remarkable performance (80.8 % baseline UAR on test) on the binary classification task – the baseline features and classification do not respect the multi-party conversation scenario (e. g., mean F0 is calculated on average across all participants). In the *Emotion Sub-Challenge*, the baseline shows Arousal to be easier to classify than Valence – this is a well known phenomenon when using acoustic features only. On the test set, a performance drop is observed for the binary tasks. In the 12-way Category task there is a large room for improvement (40.9 % baseline UAR on test), indicating the challenge of classifying subtle emotional differences even in enacted emotional speech. If the 12-way task is mapped to the binary tasks after classification, results differ (not shown in Table 5) for Arousal (74.4 % UAR on test) and Valence (64.0 % UAR on test). Finally, in the *Autism Sub-Challenge*, the binary Typicality task can again alternatively be solved by mapping from the 4-way task leading to 92.6 % UAR on test (not shown in Table 5). Better algorithms are clearly sought after for the Diagnosis task (67.1 % baseline UAR on test). For all four challenges, the official competition measure is UAAUC and UAR, respectively; these are given in boldface in Table 5.

Table 5: *Challenge Baselines. C: Complexity parameter in SVM/SVR training (tuned on development set). Devel: Result on development set, by training on training set. Test: Result on test set, by training on the training and development sets. Chance: Expected measure by chance (cf. text). UAAUC: Unweighted average of AUC for detection of the laughter and filler classes. Official Challenge competition measures are highlighted.*

[%]	C	Devel	Test	Chance
Social Signals Sub-Challenge				
AUC [Laughter]	0.1	86.2	82.9	50.0 ± 0.18
AUC [Filler]	0.1	89.0	83.6	50.0 ± 0.21
UAAUC		87.6	**83.3**	50.0 ± 0.13
Conflict Sub-Challenge				
CC [Score]	0.001	81.6	82.6	-0.8 ± 2.3
UAR [Class]	0.1	79.1	**80.8**	50.0
Emotion Sub-Challenge				
UAR [Arousal]	0.01	82.4	75.0	50.0
UAR [Valence]	0.1	77.9	61.6	50.0
UAR [Category]	1.0	40.1	**40.9**	8.33
Autism Sub-Challenge				
UAR [Typicality]	0.01	92.8	90.7	50.0
UAR [Diagnosis]	0.001	52.4	**67.1**	25.0

5. Conclusion

We introduced the INTERSPEECH 2013 Computational Paralinguistics Challenge. As for previous Challenges, we focused on realistic settings including radio broadcast, and genuine pathologic speech – the baseline results show the difficulty of the investigated automatic recognition tasks. In contrast, however, we also included a task based on enacting in studio conditions for the first time. We have provided a baseline using a rather 'brute force' feature extraction and classification approach for the sake of consistency across the Sub-Challenges; particularly, for the *Conflict Sub-Challenge*, no information on speaker segmentation is used or assessed in the baseline. Hence, it will be of interest to see the performance of methods that are more tailored to peculiarities of the presented tasks.

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*, vol. 53, no. 9/10, pp. 1062–1087, 2011.

[2] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "Paralinguistics in Speech and Language – State-of-the-Art and the Challenge," *Computer Speech and Language, Special Issue on Paralinguistics in Naturalistic Speech and Language*, vol. 27, no. 1, pp. 4–39, January 2013.

[3] B. Schuller, S. Steidl, A. Batliner, F. Schiel, J. Krajewski, F. Weninger, and F. Eyben, "Medium-Term Speaker States – A Review on Intoxication, Sleepiness and the First Challenge," *Computer Speech and Language, Special Issue on Broadening the View on Speaker Analysis*, 2013, 14 pages.

[4] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, "The INTERSPEECH 2012 Speaker Trait Challenge," in *Proc. Interspeech 2012*. Portland, OR: ISCA, 2012, 4 pages.

[5] M. Przybocki and A. Martin, "NIST Speaker Recognition Evaluation Chronicles," in *Proc. Odyssey 2004, The Speaker and Language Recognition Workshop*, Toledo, Spain, 2004, pp. 12–22.

[6] J. Downie, A. Ehmann, M. Bay, and M. Jones, "The Music Information Retrieval Evaluation eXchange: Some Observations and Insights," in *Advances in Music Information Retrieval*, W. Zbigniew and A. Wieczorkowska, Eds. Springer, 2010, pp. 93–115.

[7] T. Mandl, "Recent developments in the evaluation of information retrieval systems: Moving towards diversity and practical relevance," *Informatica*, vol. 32, pp. 27–38, 2008.

[8] B. Schuller, "The Computational Paralinguistics Challenge," *IEEE Signal Processing Magazine*, vol. 29, no. 4, pp. 97–101, July 2012.

[9] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing*, vol. 27, pp. 1743–1759, 2009.

[10] W.-M. Roth and K. Tobin, "Solidarity and conflict: aligned and misaligned prosody as a transactional resource in intra- and intercultural communication involving power differences," *Cultural Studies of Science Education*, vol. 5, p. 807, December 2010.

[11] J. McCann and S. Peppe, "Prosody in autism spectrum disorders: a critical review," *International Journal of Language and Communication Disorder*, vol. 38, pp. 325–350, October–December 2003.

[12] J. van Santen, E. Prudhommeaux, L. Black, and M. Mitchell, "Computational Prosodic Markers for Autism," *Autism*, vol. 14, pp. 215–236, 2010.

[13] D. Bone, M. P. Black, C.-C. Lee, M. E. Williams, P. Levitt, S. Lee, and S. S. Narayanan, "Spontaneous-Speech Acoustic-Prosodic Features of Children with Autism and the Interacting Psychologist," in *Proc. Interspeech*, Sep. 2012.

[14] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth, "Desperately Seeking Emotions: Actors, Wizards, and Human Beings," in *Proc. ISCA Workshop on Speech and Emotion*, Newcastle, Northern Ireland, 2000, pp. 195–200.

[15] T. Vogt and E. André, "Comparing Feature Sets for Acted and Spontaneous Speech in View of Automatic Emotion Recognition," in *Proc. International Conference on Multimedia and Expo (ICME)*, Amsterdam, The Netherlands, 2005, pp. 474–477.

[16] B. Schuller and A. Batliner, *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. Wiley, 2013, to appear.

[17] J. Bachorowski, M. Smoski, and M. Owren, "The acoustic features of human laughter," *Journal of the Acoustical Society of America*, vol. 110, pp. 1581–1597, 2001.

[18] J. Vettin and D. Todt, "Laughter in Conversation: Features of Occurrence and Acoustic Structure," *Journal of Nonverbal Behavior*, vol. 28, no. 2, pp. 93–115, January 2004.

[19] H. Tanaka and N. Campbell, "Acoustic Features of Four Types of Laughter in Natural Conversational Speech," in *Proc. 17th International Congress of Phonetic Sciences (ICPhS)*, Hong Kong, 2011, pp. 1958–1961.

[20] H. Clark and J. Fox Tree, "Using "uh" and "um" in spontaneous speaking," *Cognition*, vol. 84, no. 1, pp. 73–111, 2002.

[21] A. Vinciarelli, H. Salamin, A. Polychroniou, G. Mohammadi, and A. Origlia, "From nonverbal cues to perception: Personality and social attractiveness," *Cognitive Behavioural Systems*, pp. 60–72, 2012.

[22] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK book (v3.4)*. Cambridge University Press, 2006.

[23] S. Kim, M. Filippone, F. Valente, and A. Vinciarelli, "Predicting the conflict level in television political debates: an approach based on crowdsourcing, nonverbal communication and gaussian processes," in *Proc. of ACM International Conference on Multimedia*. Nara, Japan: ACM, 2012, pp. 793–796.

[24] S. Meignier and T. Merlin, "LIUM SpkDiarization: An open source toolkit for diarization," in *Proc. CMU SPUD Workshop*, Dallas, TX, USA, 2010.

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

[26] F. Ringeval, J. Demouy, G. Szaszák, M. Chetouani, L. Robel, J. Xavier, D. Cohen, and M. Plaza, "Automatic intonation recognition for the prosodic assessment of language impaired children," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 19, pp. 1328–1342, 2011.

[27] 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.

[28] I. H. Witten and E. Frank, *Data mining: Practical machine learning tools and techniques, 2nd Edition*. San Francisco: Morgan Kaufmann, 2005.

[29] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The WEKA Data Mining Software: An Update," *SIGKDD Explorations*, vol. 11, 2009.