Affect recognition in real-life acoustic conditions – A new perspective on feature selection.

Florian Eyben [1], *Felix Weninger*[1], *Björn Schuller* [1,2]

[1]audEERING UG (haftungsbeschränkt), Gilching, GERMANY
[2]Institute for Sensor Systems, University of Passau, GERMANY
fe, fw, bs @ audeering.com

Abstract

Automatic emotion recognition and computational paralinguistics have matured to some robustness under controlled laboratory settings, however, the accuracies are degraded in real-life conditions such as the presence of noise and reverberation. In this paper we take a look at the relevance of acoustic features for expression of valence, arousal, and interest conveyed by a speaker's voice. Experiments are conducted on the GEMEP and TUM AVIC databases. To simulate realistically degraded conditions the audio is corrupted with real room impulse responses and real-life noise recordings. Features well correlated with the target (emotion) over a wide range of acoustic conditions are analysed and an interpretation is given. Classification results in matched and mismatched settings with multi-condition training are provided to validate the benefit of the feature selection method. Our proposed way of selecting features over a range of noise types considerably boosts the generalisation ability of the classifiers.

Index Terms: paralinguistics, affect, emotion, noise robustness, acoustic features

1. Introduction

In the light of rapidly growing interest in and market value of social signal and media analysis [1, 2, 3], interactive speech systems [4, 5], and multi-modal user profiling [6, 7, 8] or stress measurement [9], technologies for automatic affect recognition from speech gain increasing commerical attention. While good results are reported in research papers under laboratory conditions (cf. [10]) or with systems tailored towards specific databases, real-life applications still remain challenging [11, 12, 13] due to various factors. These factors can be roughly summaries by three categories: The large variability of affective expression across different speakers, languages, and cultures; contextual dependencies of the meaning and significance of affective expressions; and varying and degraded acoustic conditions caused by reverberation, background noise, and acoustic properties of the recording devices used.

In this paper we adress the challenge of degraded acoustic conditions through reverberation (convolutive noise) in conjunction with additive background noise. Previous work in this line has mainly focussed on additive noise (e.g., [14]) or reverberation in isolation ([15]). Techniques from Automatic Speech Recognition (ASR) for acoustic pre-processing and signal enhancement or multi-condition training have typically been applied to boost performances in these conditions (cf. [14, 16]). Now we take a deeper look at the acoustic features. Thereby the main idea is to find acoustic features which are least degraded by noise and most correlated to affective states of interest. We investigate different strategies for feature selection in combination with training on multi-condition noisy data as well as training on clean data.

The paper is structured as follows: In Section 2 we introduce the two affective speech corpora which are analysed in this study and we describe how these corpora were degraded in four different acoustic conditions to produce noisy test and training sets. Next, the set of acoustic parameters considered is described in Section 3 followed by a description of the feature selection method and analysis of selected features in Section 4. Finally we provide and discuss results for automatic regontion of affect under various conditions in Section 5 before concluding our study in Section 6.

2. Databases

In order to capture both acted emotions as well as natural, spontaneous affect, we consider two corpora for this study: The "Geneva Multimodal Emotion Portrayals" (GEMEP) and the TUM Audio-Visual Interest Corpus (TUM-AVIC).

2.1. GEMEP

The GEMEP database [17] contains 1.2 k instances of emotional speech enacted by ten professional actors (five female) in 18 emotional categories which cover all quadrants of the arousal / valence space. The categories comprise the well-known 'Big Six' emotions as well as more subtle differentiations of these (e. g., anxiety). Actors did not receive any instruction how to express the emotion. They worked with a professional director during the recording session and choose one personally relevant scenario to induce the emotion, either by recall or mental imagery. In contrast to earlier work (cf. [18]), for the present study a subset of 154 instances is used which comes with a continuous valued dimensional observer annotation. Continuous dimensional annotations are required in order to show the full potential of our feature selection approach which is based on correlations. To obtain the annotation, twenty participants (10 male) rated each of these expressions in terms of arousal and valence by using a continuous slider (range -1 to +1). The evaluator weighted estimator [19] of the twenty raters is computed as 'gold standard' per instance. Actors expressed each emotion by using three verbal contents (two pseudo-sentences and one sustained vowel) – in this study, only the pseudo-sentences are used. A fixed training and evaluation partitioning is used where actors 1, 2, 3, 4, 6, and 7 are assigned to the training set and actors 5, 8, 9, and 10 are assigned to the evaluation (test) set. Classification experiments are performed on discretised binary arousal and valence labels (high and low) with -0.1 as threshold.

2.2. TUM-AVIC

As a second and larger set, containing natural, spontaneous affect, the TUM-AVIC database was chosen [6]. TUM AVIC was used for the interest sub-challenge of the INTERSPEECH 2010 Paralinguistic Challenge [20]. For this challenge, continuous labels for the level of interest were provided, which have been estimated by averaging the four human ratings which have discrete labels from -2 – +2. For details please refer to [20] and [6]. The mean level of interest was discretized to three discrete LOI labels (loi1, loi2, loi3), as used in [10, 13], for example. Thereby LOI -2, -1, and 0 were combined to loi1 due to sparseness of LOI -2 and -1. LOI +1 was assigned loi2, and +2 was assigned loi3. The partitioning of the INTERSPEECH 2010 Challenge was kept and the training and development partitions were jointly used for training.

2.3. Noise and Reverberation

Realistic noise samples of three types as used in [21] serve as additive noise: Babble noise (*babble*), city street noise (*city*), and music (*music*). Babble noise recordings are samples from the *freesound.org* website out of the categories pub-noise, restaurant chatter, and crowd noise. Music recordings are instrumental and classical music from the *last.fm* website. The city recordings were recorded in Munich, Germany with smartphones while cycling and walking through the city similar to the task described in [22]. The noise samples for the training and test sets are fully disjunctive, i.e., no original sample occurrs in both sets. The length of the noise pool is 30 minutes for each type in the test set and 94 minutes for babble, 116 minutes for city, and 176 minutes for music noise in the training set.

Furthermore, room impulse responses (RIRs) from the Aachen Impulse Response Database [23] were used to add convolutive noise. To keep complexity low, yet simulate realistic conditions, we selected a few meaningful combinations of noise types and RIRs: babble noise and lecture room, babble noise and stairway, city noise and meeting room, and music noise and chapel (Aula Carolina). Conditions range from rather favorable reverberation conditions (meeting room) to heavily reverberated (chapel) and represent a wide range of non-stationary additive noises. Three different virtual microphone distances from the virtual sound sources in an azimuth angle of 90 ° (facing the sound source) in 'near', 'mid', and 'far' distance categories are employed to simulate various signal to convolutive noise ratios. The further the virtual microphone is away from the virtual sound source, the larger the amount of convolutive noise is in relation to the direct source signal.

In order to eliminate the influence of the average energy per utterance, all utterances in GEMEP and TUM-AVIC were normalised to -1 dB peak amplitude. From these normalised utterances, speech samples with degraded acoustic conditions were created for each corpus. An original utterance is thereby convolved with a RIR, then normalised to -6 dB peak amplitude, and finally mixed with an additive noise sample, which is scaled in order to achieve a given signal to noise ratio (SNR). The test set of each corpus is convolved with the 'near', 'mid', and 'far' impulse responses and noise at SNRs from 0 to 12 dB in steps of 3 dB is added resulting in 18 test sets (including three reverberated sets without noise) for each acoustic condition. The training set for each acoustic condition has three times the size of the original training set because each utterance is included once for the 3 RIR distances. Noise at random SNRs (uniformly distributed on the range 0–15 dB and with 10% probability of clean utterances) is added. SNRs are calculated after first order high pass filtering from the difference signal of speech and noise, approximating A-weighting to better match human perception. Additive noise samples matching the length of the speech samples are picked at random positions in the training and test noise pools. These are then convolved with the RIR of the current acoustic condition ('far' distance) and normalised to -6 dB peak amplitude before mixing with the speech sample.

4 energy related LLD	**Group**
Sum of auditory spectrum (loudness),	pros.
Sum of RASTA-style filtered auditory spectrum,	pros.
RMS Energy, Zero-Crossing Rate.	pros.
55 spectral LLD	**Group**
RASTA-style auditory spectrum, bands 1–26 (0–8 kHz)	spec.
MFCC 1–14.	ceps.
Spectral energy 250–650 Hz, 1 k–4 kHz.	spec.
Spectral Roll Off Point 0.25, 0.50, 0.75, 0.90.	spec.
Spectral Flux, Centroid, Entropy, Slope,	spec.
Psychoacoustic Sharpness, Harmonicity,	spec.
Variance, Skewness, Kurtosis.	spec.
6 voicing related LLD	**Group**
F_0 (SHS & Viterbi smoothing),	pros.
Prob. of voice,	voice qual.
log. HNR, Jitter (local, delta), Shimmer (local).	voice qual.

Table 1: 64 ComParE low-level descriptors (LLD).

Functionals applied to LLD / Δ LLD	**Group**
quartiles 1–3, 3 inter-quartile ranges	percentiles
1 % percentile (≈ min), 99 % percentile (≈ max)	percentiles
percentile range 1 %–99 %	percentiles
position of min / max, range (max – min)	temporal
arithmetic mean[1], root quadratic mean	moments
contour centroid, flatness	temporal
standard deviation, skewness, kurtosis	moments
rel. duration LLD is above 25 / 50 / 75 / 90% range	temporal
rel. duration LLD is rising	temporal
rel. duration LLD has positive curvature	temporal
gain of linear prediction (LP), LP Coefficients 1–5	modulation
mean, max, min, std. dev. of segment length[2]	temporal
Functionals applied to LLD only	**Group**
mean value of peaks	peaks
mean value of peaks – arithmetic mean	peaks
mean / std.dev. of inter peak distances	peaks
amplitude mean of peaks, of minima	peaks
amplitude range of peaks	peaks
mean / std.dev. of rising / falling slopes	peaks
linear regression slope, offset, quadratic error	regression
quadratic regression a, b, offset, quadratic error	regression
percentage of non-zero frames[3]	temporal

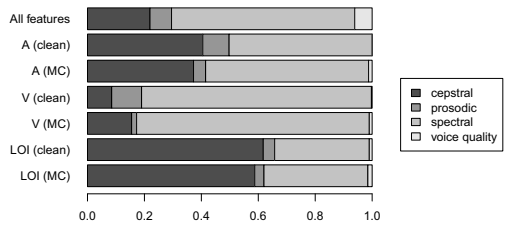
Table 2: Applied functionals. [1]: arithmetic mean of LLD / positive Δ LLD. [2]: not applied to voice related LLD except F_0. [3]: only applied to F0.

3. ComParE feature set

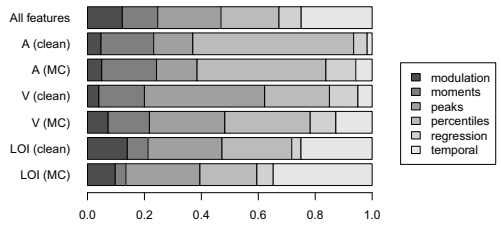
In order to cover an exhaustive set of acoustic features, we decided for the ComParE feature set – the baseline feature set for the INTERSPEECH 2013 Computational Paralinguistics Evaluation (ComParE) Challenge [18]. The set contains a total of 6 373 acoustic features, which are brute-force combinations of an extensive set of acoustic low-level descriptors (LLD) and their delta coefficients with a large set of functionals (cf. Tables 1 and 2, respectively). The acoustic features were extracted with the openSMILE feature extraction toolkit [24], which is actively maintained by the authors.

4. Feature analysis

We select features relevant for the target task by computing the Pearson correlation coefficients (CC) of each feature with the continuous target label; this method is referred to as CC-FS in the ongoing. In this section, we summarize the top 400 features in the CC-FS feature set, for each of the arousal, valence, and LOI tasks, by LLD and functional type. We contrast the features obtained on the clean training set with those obtained on the multi-condition training sets (training sets of all four noisy conditions combined). For both arousal and valence, it can be seen that the relative importance of prosodic features decreases when introducing noise and reverberation. For example, enery change features seem to be much less reliable in the reverberated and noisy cases, such as the root quadratic mean of loudness delta coefficients (CC = .651 with arousal on clean, CC = -.052 on multi-condition). Regarding the functional types, we observe that the relative importance of the temporal functionals (e. g., up-level and rise times) increases in the reverberant and noisy case. This might indicate that in the presence of noise the overall distribution of a signal over time remains more intact than other parameters like means and moments. Overall, however, there is more change in feature group relevance for between the individual tasks than between selecting features on the clean or noisy multi-condition sets.



(a) LLD



(b) Functionals

Figure 1: Full ComParE feature set vs. 400 top features selected by CC-FS on clean and multi-condition (MC) sets for arousal (A), valence (V), and level of interest (LOI) classification: Percentage of LLD and functional types.

5. Noise robust classification

We show how the reduction of the feature set size to 400 features affects the automatic classification performance and how the data set on which the CC based feature selection is performed on influences the results. We contrast clean and multi-condition training. In multi-condition training (MCT) we join the training sets of three noisy acoustic conditions (e.g, A, B, C) and evaluate for all test sets of the other condition (e.g., D) in the case of mismatched evaluations and on the test sets for each of the three training conditions (e.g., A, B, C) for matched evaluations. We repeat the evaluations four times, in order to have each condition once as a mismatched test set. Note, that the training set in multi-condition training is 9 times the size of the original (clean) training set (three acoustic conditions and three RIR distances).

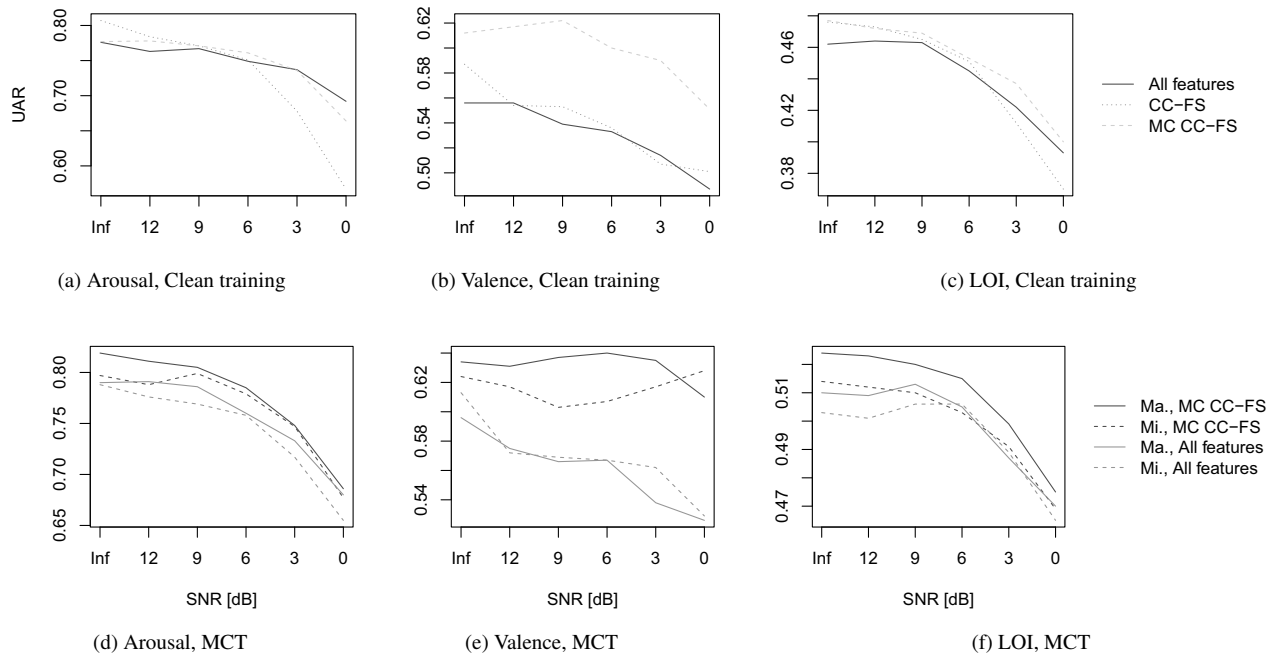
[%] UAR	A	B	C	D
Binary Arousal/Valence				
clean T	73.4/52.0	73.7/54.9	73.4/56.6	70.2/52.4
MCT mi	74.0/59.2	78.8/65.0	75.3/63.2	77.7/58.9
MCT ma	76.4/63.5	79.1/61.8	77.4/66.1	77.3/61.1
Ternary Level of Interest				
clean T	42.9	44.0	48.7	40.8
MCT mi	50.2	50.6	53.0	46.1
MCT ma	50.6	51.6	52.2	49.3

[%] UAR	A	V	LOI	A	V	LOI
	all features			CC-FS		
clean	74.7	53.1	44.2	72.6	54.0	44.1
clean MC-FS	-	-	-	74.8	59.9	45.1
MCT mi.	74.4	56.9	49.5	**76.5**	**61.6**	**50.0**
MCT ma.	75.7	56.1	49.9	**77.6**	**63.1**	**50.9**

Table 3: Average classification results (UAR) for each acoustic condition averaged over distances and SNRs (top) and further averaged over all acoustic conditions (bottom). Training (T) on clean, and multi-condition training (MCT) on other conditions than the test condition (mismatched, *mi.*) and conditions including the test condition (matched, *ma.*). Top: all features. Bottom: Comparison of all features and top 400 features after CC-FS. Feature selection for clean training on clean set (clean) and on joint training set of all four conditions (MC FS); best two results per target in bold.

As classifier we use Support-Vector Machines (SVM) with a linear kernel, trained with the Sequential Minimal Optimization Algorithm using the WEKA toolkit [25]. The complexity parameter was set to $C = 0.1$ for the GEMEP corpus (arousal and valence) and $C = 0.01$ for the larger TUM-AVIC corpus. Before training the classifier the features were standardised to have mean zero and unit variance on the training set. The means and variances estimated from the training set are then used to standardize the features in the test sets. To avoid overfitting to the majority class on the TUM-AVIC corpus (1.88k instances of loi2 vs. 349 instances of loi1), the training set was balanced by random subsampling so that each class had the same number of instances as the minority class. As an upper benchmark of performance the results in terms of Unweighted Average Recall (UAR) for training on the clean training set and testing on the clean test set are as follows: 77.2% for binary arousal, 61.7% for binary valence, and 56.0% for three levels of interest. Please note, that the result for interest is below results in other publications (e.g., [10, 6]). However, this is due to two factors: a) the full set of instances as in the Paralinguistic Challenge is used (other studies used a set where instances with low labeller agreement were removed) and b) each utterance is normalised to peak amplitude -1 dB, eliminating mean energy as cue.

Table 3 (top) shows the performance for the four test set acoustic conditions (A–D) averaged over all 18 sub-conditions (SNRs and RIR distances). Chapel RIR and music noise is clearly the most challenging condition, both for clean and MCT cases. Overall, MCT boosts the performance over clean training, which is in line with [14]. Matched conditions is thereby



(a) Arousal, Clean training (b) Valence, Clean training (c) LOI, Clean training

(d) Arousal, MCT (e) Valence, MCT (f) LOI, MCT

Figure 2: Unweighted average recall (UAR) of binary arousal, valence and ternary LOI classification: Influence of SNR across all acoustic conditions. Clean or multi-condition training (MCT) including (matched) or not including (mismatched) the test condition. Full ComParE feature set (All) vs. CC-FS on clean set or multi-condition sets.

only marginally better than mismatched conditions training. Table 3 (bottom) shows the performances averaged over all acoustic conditions in order to contrast CC based feature selection (CC-FS) with the full feature set. While for clean training there is only little difference between CC-FS and all features, the best result is achieved by combining MCT and CC-FS performed on the MCT training set. Even the performance of clean training can be significantly boosted by selecting features on the MCT sets instead of the clean set.

In Figure 2, we show the UARs on the binary arousal and valence, as well as the ternary LOI tasks obtained with clean training or multi-condition training, and with optional feature selection. For arousal, it turns out that there is only a slight performance difference between clean and MCT if all features are used; however, a drop in performance at lower SNRs is observed (below 60 % UAR) for clean training and feature selection on the clean data. If features are selected from the MC set, results are similar to training on the whole MC set. This shows that the feature selected on the clean data do not generalize well to unfavourable acoustic conditions, yet MC feature selection can remedy this. For valence, unfortunately, we do not obtain robust results even in the noise free case (only up to 63 % UAR) – this can be attributed to the general difficulty of determining valence from pure acoustics. It is encouraging, though, that through MCT a result largely independent of the SNR can be obtained. Finally, in the task of LOI determination, we observe a hugh influence of MCT, yet a small influence of feature selection. Notably, the worst result (at 0 dB SNR) with MCT is similar to the best result (at SNR = ∞) without MCT. This points at a deterioration caused by reverberation. Overall, the LOI results indicate that there are little features that stand out across multiple acoustic conditions – for this task one apparently has to learn how the feature behavior changes in unfavorable acoustic conditions.

Regarding feature selection, we can conclude that in most cases, similar or better results with respect to the full feature set can be obtained with only a fraction (6.3 %) of the features. The combination of MCT and feature selection seems especially powerful to cope with challenging acoustic conditions, even under mismatched settings.

6. Conclusions and Outlook

We have investigated the influence of convolutive and additive noise on the performance of automatic affect recognition systems in realistic conditions. Considerable performance degradations are observed for both convolutive noise (10% UAR absolute on TUM-AVIC for three levels of interest) and additive noise (an additional 8% absolute loss in UAR for 0 dB SNR). As known from previous work, the performance can be improved by multi-condition training. In this study we have discovered that an additional gain in performance – and thus more robust systems – can be achieved by automatic selection of acoustic parameters on a multi noisy condition data-set. The increase in performance is also present when training only on clean data; it is thus complimentary to the gain by multi-condition training.

In future work a detailed study of the relevance of individual features for the specific SNR conditions could reveal more insight towards which features are robust noise to what extent. Human perception tests could be combined with the automatic analysis to verify at which SNRs human affect recognition performance suffers and to which acoustic features this relates. This will pave the way to automated methods for a better understanding of how humans perceive acoustic affect and how technical system be made ready for reliable use in real-world commercial applications.

7. References

[1] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[2] F. Eyben, F. Weninger, N. Lehment, G. Rigoll, and B. Schuller, "Violent scenes detection with large, brute-forced acoustic and visual feature sets," in *Proc. MediaEval 2012 Workshop, Pisa, Italy*, 2012.

[3] 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.

[4] J. Pittermann, A. Pittermann, and W. Minker, "Emotion recognition and adaptation in spoken dialogue systems," *International Journal of Speech Technology*, vol. 13, pp. 49–60, 2010.

[5] M. Schröder, E. Bevacqua, R. Cowie, F. Eyben, H. Gunes, D. Heylen, M. ter Maat, G. McKeown, S. Pammi, M. Pantic, C. Pelachaud, B. Schuller, E. de Sevin, M. Valstar, and M. Wöllmer, "Building Autonomous Sensitive Artificial Listeners," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 165–183, April–June 2012.

[6] B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker, and H. Konosu, "Being Bored? Recognising Natural Interest by Extensive Audiovisual Integration for Real-Life Application," *Image and Vision Computing, Special Issue on Visual and Multimodal Analysis of Human Spontaneous Behavior*, vol. 27, no. 12, pp. 1760–1774, November 2009.

[7] C.-C. Lee, M. Black, A. Katsamanis, A. Lammert, B. Baucom, A. Christensen, P. G. Georgiou, and S. S. Narayanan, "Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples," in *Proc. of InterSpeech*, Makuhari, Japan, Sep. 2010.

[8] B. Schuller, S. Steidl, A. Batliner, E. Nth, 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. of INTERSPEECH, Portland, Oregon, USA*. ISCA, 2012.

[9] H. Lu, M. Rabbi, G. Chittaranjan, D. Frauendorfer, M. Schmid Mast, A. T. Campbell, D. Gatica-Perez, and T. Choudhury, "Stresssense: Detecting stress in unconstrained acoustic environments using smartphones," in *Proc. Ubicomp'12*, Sep. 2012.

[10] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth, "Acoustic Emotion Recognition: A Benchmark Comparison of Performances," in *Proceedings 11th Biannual IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2009*, IEEE. Merano, Italy: IEEE, December 2009, pp. 552–557.

[11] 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, no. 9/10, pp. 1062–1087, November/December 2011.

[12] E. Mower, M. J. Mataric, and S. S. Narayanan, "A framework for automatic human emotion classification using emotional profiles," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 5, pp. 1057–1070, 2011.

[13] B. Schuller, M. Valstar, R. Cowie, and M. Pantic, "AVEC 2012 – The Continuous Audio/Visual Emotion Challenge," in *Proc. 2nd International Audio/Visual Emotion Challenge and Workshop, AVEC 2012, Grand Challenge and Satellite of ACM ICMI 2012*. Santa Monica, CA: ACM, October 2012.

[14] F. Eyben, B. Schuller, and G. Rigoll, "Improving generalisation and robustness of acoustic affect recognition," in *Proc. Second International Audio/Visual Emotion Challenge and Workshop (AVEC 2012), Grand Challenge and Satellite of ACM ICMI 2012, Santa Monica*. New York NY, USA: ACM, 2012, pp. 517–522.

[15] B. Schuller, "Affective Speaker State Analysis in the Presence of Reverberation," *International Journal of Speech Technology*, vol. 14, no. 2, pp. 77–87, 2011.

[16] F. Weninger, B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognition of nonprototypical emotions in reverberated and noisy speech by nonnegative matrix factorization," *EURASIP Journal on Advances in Signal Processing*, Jan. 2011.

[17] T. Bänziger, M. Mortillaro, and K. R. Scherer, "Introducing the Geneva Multimodal Expression corpus for experimental research on emotion perception," *Emotion*, vol. 12, no. 5, pp. 1161–1179, 2012.

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

[19] M. Grimm and K. Kroschel, "Evaluation of natural emotions using self assessment manikins," in *Proc. ASRU*. IEEE, 2005, pp. 381–385.

[20] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Muller, and S. Narayanan, "The INTERSPEECH 2010 paralinguistic challenge," in *Proc. of INTERSPEECH, Makuhari, Japan*. ISCA, 2010.

[21] F. Eyben, F. Weninger, S. Squartini, and B. Schuller, "Real-life voice activity detection with lstm recurrent neural networks and an application to hollywood movies," in *Proc. of ICASSP, Vancouver, Canada*. IEEE, 2013, to appear.

[22] B. Schuller, F. Pokorny, S. Ladstätter, M. Fellner, F. Graf, and L. Paletta, "Acoustic geo-sensing: Recognising cyclists route, route direction, and route progress from cell-phone audio," in *Proc. of ICASSP, Vancouver, Canada*. IEEE, 2013, to appear.

[23] M. Jeub, M. Schäfer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proceedings of International Conference on Digital Signal Processing (DSP)*. Santorini, Greece: IEEE, July 2009, pp. 1–4.

[24] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE – the munich versatile and fast open-source audio feature extractor," in *Proc. of ACM Multimedia, Florence, Italy*. ACM, 2010, pp. 1459–1462.

[25] 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.