Face Reading from Speech – Predicting Facial Action Units from Audio Cues

Fabien Ringeval[1,4], *Erik Marchi*[1], *Marc Mehu*[2], *Klaus Scherer*[3], *Björn Schuller*[3,4,5]

[1]Machine Intelligence & Signal Processing Group, MMK, Technische Universität München, Germany
[2]Department of Psychology, Webster Vienna Private University, Austria
[3] Swiss Center for Affective Sciences, University of Geneva, Switzerland
[4] Chair of Complex and Intelligent Systems, Universität Passau, Germany
[5] Department of Computing, Imperial College London, U.K.

fabien.ringeval@tum.de

Abstract

The automatic recognition of facial behaviours is usually achieved through the detection of particular FACS Action Unit (AU), which then makes it possible to analyse the affective behaviours expressed in the face. Despite the fact that advanced techniques have been proposed to extract relevant facial descriptors, the processing of real-life data, i. e., recorded in unconstrained environments, makes the automatic detection of FACS AU much more challenging compared to constrained recordings, such as posed faces, and even impossible when the corresponding parts of the face are masked or subject to low or no illumination. We present in this paper the very first attempt in using acoustic cues for the automatic detection of FACS AU, as an alternative way to obtain information from the face when such data are not available. Results show that features extracted from the voice can be effectively used to predict different types of FACS AU, and that the best performance are obtained for the prediction of the apex, in comparison to the prediction of onset, offset and occurrence.

Index Terms: computational paralinguistics, FACS action units

1. Introduction

The automatic analysis of affective and social behaviours from multimodal data has become a major field of research in the last decade. Recent work have shown that the complementarity of multimodal signals, such as speech, face and even physiology, can be successfully exploited to improve the automatic analysis of socio-affective behaviours, in comparison to mono-modal approaches [1, 2]. However, the processing of data captured in real-life interactions are challenging, since the variance of behaviours expressed is very high in relation to the data available (sparseness), and because state-of-the-art methods are largely affected by additive and convolutive background noise. Regarding the automatic analysis of facial expressions, most systems proposed in the literature focus on detecting facial action units (AU). These AU represent fundamental actions (e. g., contraction or relaxation) of individual muscles or group of muscles from the face, which are involved in the communication of facial expressions, such as smiling or frowning [3]. The automatic identification of these AU from video recordings is yet challenging, even when the data are captured under suitable conditions [4, 5]. Data captured in real-life environments are even more challenging, because the parts of the face related to the AU can be partially masked or not visible at all, or subject to very low or no illumination, which makes the automatic identification of the AU much more difficult or even impossible.

We therefore investigate in this paper the very first attempt at predicting facial AU from acoustic cues, as an alternative way to obtain information from the face when such data are not available or usable. Indeed, the activation of particular facial AU may impact the supra-glottal configuration of the speech production system in different ways, which in turn can provide some acoustic cues for the identification of the activated AU. Moreover, there exists some literature arguing for a close coupling of speech and facial expressions, all being governed by a coordinated planning at the cognitive level [6, 7]. Preliminary work has been conducted on the identification of the occurence of some facial expressions from speech data and has shown promising results [8]. However, this (only) study was not really focused on the prediction of AU as those defined in the Facial Action Coding Systems (FACS) [3]. In this paper, we focus on the time-continuous prediction of the onset, apex, offset and occurence of different FACS AU, from an updated version of the GEMEP database that was used for the first international challenge on facial expressions recognition [4].

This paper is structured as follows: we first introduce the system we designed for the prediction of FACS AU from speech data (Section 2), we then present the database (Section 3), the experiments and the results (Section 4) before concluding (Section 5).

2. System

We describe in the following sections the acoustic feature sets that were extracted from speech data and the machine learning algorithms that were used to predict FACS AU.

2.1. Acoustic feature sets

Acoustic low-level descriptors (LLD) were automatically extracted from the speech waveform using our open-source openSMILE feature extractor in its recent 2.1 release [9]. Two different feature sets were investigated: a large brute-forced feature set (ComParE) and a smaller, expert-knowledge based feature set (GeMAPS). The reader is referred to [10] for a detailed description of the implementation of these features sets.

2.1.1. *ComParE*

The ComParE feature set is the result of a continuous refinement of acoustic descriptors used for the analysis of paralinguistics in speech and language. It has been successfully employed for the automatic recognition of various paralinguistic traits and states, such as those investigated in the Interspeech Compu-

  September 6 – 10, 2015, Dresden, Germany

Table 1: COMPARE acoustic feature set: 65 low-level descriptors (LLD).

4 energy related LLD	Group
RMS energy, zero-crossing rate	Prosodic
Sum of auditory spectrum (loudness)	Prosodic
Sum of RASTA-filtered auditory spectrum	Prosodic
55 spectral LLD	**Group**
MFCC 1–14	Cepstral
Psychoacoustic sharpness, harmonicity	Spectral
RASTA-filt. aud. spect. bds. 1–26 (0–8 kHz)	Spectral
Spectral energy 250–650 Hz, 1 k–4 kHz	Spectral
Spectral flux, centroid, entropy, slope	Spectral
Spectral Roll-Off Pt. 0.25, 0.5, 0.75, 0.9	Spectral
Spectral variance, skewness, kurtosis	Spectral
6 voicing related LLD	**Group**
F_0 (SHS and Viterbi smoothing)	Prosodic
Prob. of voicing	Voice qual.
log. HNR, jitter (local and δ), shimmer (local)	Voice qual.

tational Paralinguistic Challenge (COMPARE), e. g., personality [11], pathology [12], cognitive and physical load [13] and eating condition [14]. The latest version of the COMPARE feature set contains 65 LLD of speech – 130 LLD in total with their first order derivate. The LLD cover spectral, cepstral, prosodic and voice quality information, cf. Table 1. Voicing related LLD are extracted from 60 ms frames using a Gaussian window function ($\sigma = 0.4$), whereas all other LLD are extracted from 25 ms frames using a Hamming window function. All windows are overlapping and are sampled at a common rate of 100 Hz. A symmetric moving average window of 3 frames length (one previous, one current, and one future frame) is used to smooth the LLD, and first order delta regression coefficients are computed with a context window size of 2 frames for all LLD.

2.1.2. GeMAPS

In contrast to large scale brute-force feature sets, which have been successfully applied to many speech and music classification tasks, e. g., [1, 15], smaller, expert-knowledge based feature sets have also shown high robustness for the modelling of short-term paralinguistic states, such as emotion [16, 17]. Indeed, a minimalistic standard parameter set presents the advantage to reduce the danger of over-adaptation of classifiers to the training data in machine learning problems, in comparison to the use of a brute-forced features set. Some recommendations for the definition of such a minimalistic standard parameter set for the acoustic analysis of speech and other vocal sounds has been recently investigated, and has led to the definition of the Geneva Minimalistic Acoustic Parameter Set (GEMAPS) [18], cf. Table 2. Features were mainly selected based on their potential to index affective physiological changes in voice production, for their proven value in former studies, and for their theoretical definition. The implementation of this feature set has been conducted similarly to the COMPARE feature set, i. e., overlapping windows were used and sampled at a common rate of 100 Hz to extract the LLD, and a symmetric moving average window of 3 frames length was used for smoothing purpose.

2.2. Machine learning algorithms

We investigated two types of machine learning algorithms to perform time-continuous prediction of FACS AU from acoustic features: Support Vector Machines (SVM) [19] and Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN) [20, 21]. SVM were employed for their well known ability to generalise well over large features set, whereas LSTM-RNN were used for their ability to learn long-term contextual dependencies between features and labels.

Table 2: GEMAPS acoustic feature set: 28 low-level descriptors (LLD).

1 energy related LLD	Group
Sum of auditory spectrum (loudness)	Prosodic
11 spectral LLD	**Group**
Alpha ratio (50–1000 Hz / 1–5 kHz)	Spectral
Energy proportion (0–500 Hz, 0–1 kHz)	Spectral
Energy slope (0–500 Hz, 0–1 kHz)	Spectral
Hammarberg index	Spectral
MFCC 1–4	Cepstral
Spectral flux	Spectral
16 voicing related LLD	**Group**
F_0 (linear & semi-tone)	Prosodic
Formants 1, 2, 3 (freq., bandwidth, ampl.)	Voice qual.
Harmonic difference H1–H2, H1–A3	Voice qual.
log. HNR, jitter (local), shimmer (local)	Voice qual.

3. Database

We describe below the database that was used in our experiments, the procedure that was followed for the coding of the FACS AU, as well as those that were retained for our study.

3.1. The GEMEP corpus

The Geneva Multimodal Emotion Portrayals (GEMEP) corpus has been proposed with the goal to provide audiovisual material for the analysis of the mechanisms involved in the perception and the expression of emotions [22]. The GEMEP corpus includes more than 7.000 audiovisual emotion portrayals, representing 18 emotions portrayed by 10 professional French-speaking theater actors (5 male, 5 females). Pseudo-linguistic phoneme sequences (‘*nekal ibam soud molen!*’ and ‘*koun se mina lod belam?*’) and a sustained vowel (/aaa/) were uttered in ongoing interactions between the actors and a professional director. The expressions were recorded with three digital cameras at a frame rate of 25 Hz, including one camera to zoom in on the facial expressions of the actors. Speech was recorded with separate microphones located at each of the three cameras, and with an additional microphone located over the left ear of the actor (44.1 kHz frame rate). Recordings were manually segmented and synchronised using a dedicated software [22].

3.2. FACS AU coding and selection

A subset of the GEMEP corpus was designed, using only the pseudo speech sentences for which the emotional expressions were best recognised by humans, while keeping a balance in the number of instances obtained for each actor. 158 portrayals were retained in total and annotated in terms of facial expression following the guidelines provided in the FACS manual [3]. All AU were annotated by a certified FACS coder [23], using the annotation software Anvil 4.5 [24]. Start and end times for the onset, apex, and offset of each AU were recorded. The onset begins on the frame where the first appearance change associated with the AU was observed. Although periods of little

Table 3: *Description, number of instances (in thousands) and inter-rater agreement (% of agreement and Cohen's κ) for each AU selected from the* GEMEP *corpus; r.: raiser, p.: puller.*

AU	Description	On.	Apex	Off.	% ag.	κ
1	Inner brow r.	2.73	5.23	4.91	90.9	0.60
2	Outer brow r.	2.34	5.29	4.76	89.9	0.72
4	Brow lowerer	3.29	7.41	2.79	95.2	0.63
6	Cheek raiser	2.94	11.0	1.73	67.5	0.53
7	Lid tightener	3.50	10.1	3.91	55.3	0.30
10	Upper lip r.	3.31	7.12	5.05	47.7	0.05
12	Lip corner p.	6.86	10.8	5.63	57.8	0.37
17	Chin raiser	2.39	3.32	3.23	75.5	0.19

change could be observed during the onset phase, they were still considered part of the onset provided a subsequent increase was observed before the AU reached a plateau, i. e., the apex. The apex begins on the frame that follows the last increase in intensity observed for the particular AU. Finally, the first sign of decrease in the AU's intensity determined the end of the apex and the start of the offset. The decrease should be continuous (i. e., last at least two frames) and terminated with the disappearance of the AU, or with a new increase in intensity, which marked a new onset phase. A new AU was recorded when it appeared out of a neutral position or when it was seen increasing after a decrease in intensity.

In order to provide a sufficient amount of data to perform machine learning in suitable conditions, we selected a subset of the coded AU based on a minimum of 5% of occurence for onset, apex and offset. Over a total of 39.1 k instances, a subset of 8 different AU was retained for our experiments[1], cf. Table 3. Frame-based inter-rater reliability was calculated for 10 videos (total of 824 frames) coded by two certified FACS coders. The percentage of agreement and the Cohen's κ coefficient between these two coders are shown in Table 3. The analysis shows that raters agree more than chance alone would predict (all p-values for $\kappa < .001$). Low values for κ can be partly explained by the low prevalence of certain AU in the coded portrayals, therefore percentage of agreement should be taken into account. A revision of the coding used in [4] was performed and problematic annotations corrected.

4. Experiments and results

We describe in the followings the experimental setup and the results obtained in the time-continuous prediction of FACS AU from acoustic cues.

4.1. Experimental setup

The speech recordings captured by the head-mounted microphone were post-processed in order to optimise the synchronisation with the video data, using the same procedure as in [25]. Obtained wave files were normalised to 0dB peak amplitude to compensate for various levels of loudness due to the expression of emotions. We did not use any voice activity detection algorithm because the portrayals were already segmented and contained very few pauses or only short ones. LLD were extracted from the normalised speech wave form, using either the ComParE or the GeMAPS feature set, and binary labels (e. g., '*unactive*' vs. '*active*') were defined according to the timecode provided by the FACS coding of the selected AU. Variability of the acoustic features over the actors was compensated by performing a speaker dependent z-score normalisation.

For all the experiments, we used a leave-one-speaker-out (LOSO) evaluation framework, to ensure speaker independency in the prediction of the AU. Each combination of AU and activation, i. e., onset, apex, offset or occurence, was processed separately. To cope with imbalanced class distribution, up-sampling of the underrepresented class was performed prior to the training of SVM. This procedure was not used for the training of LSTM-RNN, as it will have broken the time continuum. SVM were trained with the Sequential Minimal Optimisation (SMO) algorithm implemented in the Weka toolkit [26] and a linear kernel was used. The complexity parameter was fixed to a low value ($C = 10^{-3}$) to ensure generalisation abilities.

The LSTM networks have a layout composed of 3 hidden layers with 156, 256 and 156 units respectively. The learning rate was set to 10^{-5} for all the experiments and a cross-entropy cost function was used for the training of the networks. To improve generalisation and prevent overfitting, Gaussian noise with a standard deviation 0.1 was added to all input features. A logistic function was used for both the activation function and the output layer. Binary classification of the AU was achieved by thresholding the output with the observed median value. We used our CURRENNT toolkit for the implementation of LSTM-RNN [27].

4.2. Results

Performance is measured as the unweighted average recall (UAR) of the classes, which represents the accuracy in a dataset with equal class priors. This is especially important when the class distribution is imbalanced and high accuracy could be achieved by picking the majority class. It is calculated by the sum of recall-values (class-wise accuracy) for all classes divided by the number of classes. This is the standard measure of the INTERSPEECH ComParE Challenge series [28].

Results obtained in the prediction of the different types of AU with the SVM are given in Table 4. Except for the onset of AU06 (Cheek raiser) and the offset of AU07 (Lid tightener), all values are above the chance level ($UAR = 50\%$), which thus show that speech contains relevant cues for the prediction of various FACS AU. The GeMAPS minimalistic feature set shows an higher robustness than the ComParE brute-forced features set, as two third of the best results are obtained with the former set. The best recognised AU in terms of occurrence are those related to the raising of the inner and outer eyebrows, i. e., AU01 and AU02. This result might be explained by the fact that eyebrow movements have been shown to correlate well with the fundamental frequency [7]. However, the performance obtained on the detection of AU04 (Brow lowerer) is slightly lower than for AU01 and AU02, which suggests an higher degree of correlation between eyebrows raising and speech than for eyebrows lowering. The apex of the AU is generally better recognised than their transitions, i. e., onset or offset, or their occurence. Indeed, the apex of the AU may have generated less variability in the features as well as more prominent cues.

Results obtained in the prediction of the different types of AU with the LSTM-RNN are given in Table 5. All values are significantly above the chance level, and almost all better than the results obtained with the SVM ($p < .05$). This result thus shows the importance of using contextual information in the modelling of AU from speech data, as it was also found for others paralinguistic tasks [1, 29]. Even if the performance is

[1]An exception was made for the AU06 since it was below the 5% threshold only for its onset.

Table 4: *Results (%UAR) for different AU prediction from* ComParE *and* GeMAPS *acoustic feature sets with SVM; LOSO evaluation. Values given in bold style correspond to the best performance obtained for each AU, i. e., for either onset, apex, offset or occurence, with either* ComParE *or* GeMAPS *acoustic features set.*

	ComParE				**GeMAPS**			
AU	**Onset**	**Apex**	**Offset**	**Occu.**	**Onset**	**Apex**	**Offset**	**Occu.**
1	58.34	63.00	58.26	62.28	53.98	**64.83**	56.38	63.30
2	55.39	60.42	51.34	61.72	55.29	**62.09**	52.31	60.77
4	57.82	53.89	55.58	55.48	57.18	55.40	58.32	**58.99**
6	48.02	53.31	51.21	52.16	46.01	53.83	52.22	**54.48**
7	**55.02**	53.00	54.10	48.67	53.14	54.82	51.99	52.60
10	53.19	51.98	**54.12**	51.68	53.71	53.39	53.01	46.25
12	54.05	54.40	52.76	53.43	52.52	51.54	**54.97**	54.22
17	56.05	62.52	60.29	59.29	57.96	**62.57**	56.67	59.75
Avg.	54.74	56.57	54.71	55.59	53.72	**57.31**	54.48	56.30

Table 5: *Results (%UAR) for different AU prediction from* ComParE *and* GeMAPS *acoustic feature sets with LSTM-RNN; LOSO evaluation. Values given in bold style correspond to the best performance obtained for each AU, i. e., for either onset, apex, offset or occurence, with either* ComParE *or* GeMAPS *acoustic features set.*

	ComParE				**GeMAPS**			
AU	**Onset**	**Apex**	**Offset**	**Occu.**	**Onset**	**Apex**	**Offset**	**Occu.**
1	61.88	64.54	59.48	67.03	62.08	**67.96**	61.18	67.62
2	64.38	67.43	62.27	69.33	64.60	**70.99**	62.98	70.90
4	61.35	66.21	63.33	64.66	61.90	66.67	63.19	**67.05**
6	64.01	65.17	63.35	63.78	64.25	**67.71**	63.17	63.11
7	62.71	56.62	58.29	54.47	**62.85**	59.74	61.46	52.72
10	61.82	60.53	56.39	60.39	**61.92**	61.03	58.02	60.34
12	59.34	59.85	56.71	58.67	58.04	**60.92**	57.77	58.72
17	60.47	64.78	62.34	64.53	61.62	64.94	63.13	**65.88**
Avg.	62.00	63.14	60.27	62.86	62.16	**65.00**	61.36	63.29

significantly higher when using LSTM-RNN instead of SVM, the same observations hold: (i) the GeMAPS minimalistic features set is more robust than the brute-force ComParE features set, (ii) the apex of the AU is better recognised than their transitions (i. e., onset, offset) or occurence and (iii) the inner / outer brow raiser are the best recognised AU. An additional interesting result is that the AU12, which is involved in smiles, is the one that achieved the worst performance in average, which can be related with the low performance usually obtained on the prediction of emotional valence from speech data [1]. One may further notice that the AUs involving muscles located around the mouth, e. g., AU10 and AU12, are usually not considered for the automatic prediction of FACS AU when speech is present [4, 5], because it is difficult to distinguish physical actions that are necessary to produce speech from those that are performed for conveying non-verbal information.

5. Conclusions

We investigated in this paper the very first attempt in using acoustic features to predict facial action units, as an alternative way to obtain information from the face when such data are not available. We used as acoustic features a brute-forced set (ComParE) and a minimalistic set (GeMAPS). Machine learning algorithms, such as SVM and LSTM-RNN were trained to predict FACS coded AU from the acoustic features at a 10 ms frame rate and in a leave-one-speaker-out evaluation framework. Results obtained on the GEMEP database have shown that speech contains relevant cues for the prediction of various facial action units, especially for the apex, and that the use of contextual information, by using LSTM-RNN instead of SVM, helps to improve the performance. However, these promising results need to be confirmed on other databases, such as those used in [5], because the GEMEP corpus contains few instances of speech with a constrained production, which makes the recognition of the FACS AU from acoustic data easier.

Future work will thus consist in exploiting larger databases with unconstrained speech production to perform the FACS AU detection from acoustic features. Models of speech rhythm, such as those introduced in [30, 31], will be used as additional acoustic features since the temporal dimension may convey some relevant information for the detection of FACS AU. The automatic prediction of FACS AU's intensity will be also investigated, as well as the fusion with state-of-the-art system, to quantify the complementarity of acoustic descriptors with facial descriptors for the automatic recognition of FACS AU. Emotion recognition experiments will also be performed with automatically detected FACS AU from acoustic data to evaluate the interest of such approach for affective computing.

6. Acknowledgements

The research leading to these results has received funding from the EC Seventh Framework Programme under grant agreements no. 338164 (ERC iHEARu), no. 230331 (ERC PROPEREMO), no. 645094 (IA SEWA) and no. 645378 (RIA ARIAVALUSPA).

7. References

[1] F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J.-P. Thiran, T. Ebrahimi, D. Lalanne, and B. Schuller, "Prediction of asynchronous dimensional emotion ratings from audiovisual and phys-

iological data," *Pattern Recognition Letters*, November 2014, in press.

[2] F. Schwenker, S. Scherer, and L.-P. Morency, *Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction*. Springer, 2015.

[3] P. Ekman, W. V. Friesen, and J. Hager, *Facial action coding system*. Salt Lake City, UT: Research Nexus, 2002.

[4] M. Valstar, B. Jiand, M. Mehu, M. Pantic, and K. Scherer, "The first facial expression recognition and analysis challenge," in *Proc. of the IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, Santa Barbara (CA), USA, March 2011, pp. 921–926.

[5] M. Valstar, G. McKeown, M. Mehu, L. Yin, M. Pantic, and J. Cohn, "FERA 2015 - Second Facial Expression Recognition and Analysis Challenge," in *Proc. of the IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, Ljubljana, Slovenia, May 2015.

[6] D. L. Bolinger, *Intonation and its parts*. New York: London, Edward Arnold, 1985.

[7] C. Cave, I. Guaitella, R. Bertrand, S. Santi, F. Harlay, and R. Espesser, "About the relationship between eyebrow movements and f0 variations," in *Proc. of the 4th International Conference on Spoken Language*, Philadelphia (PA), USA, 1996, pp. 2175–2178.

[8] G. Lejan, N. Souviraa-Labastie, and F. Bimbot, "Facial expression recognition from speech," in *Research Report, RR-8337*, 2013.

[9] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in *Proc. of the 21st ACM International Conference on Multimedia (ACM MM)*, Barcelona, Spain, October 2013, pp. 835–838.

[10] F. Eyben, "Real-time speech and music classification by large audio feature space extraction," Ph.D. dissertation, Technische Universität München, 2014, submitted, to appear.

[11] B. Schuller, S. Steidl, A. Batliner, E. Noth, 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 2012, 13th Annual Conference of the International Speech Communication Association (ISCA)*, Portland (OR), USA, September 2012, pp. 254–257.

[12] 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. of INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association (ISCA)*, Lyon, France, August 2013, pp. 148–152.

[13] 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 2014, 15th Annual Conference of the International Speech Communication Association (ISCA)*, Singapore, Republic of Singapore, September 2014, pp. 427–431.

[14] 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: Nativeness, Parkinson's & Eating Condition," in *Proc. of INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association (ISCA)*, Dresden, Germany, September 2015.

[15] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer, "On the acoustics of emotion in audio: What speech, music and sound have in common," *Frontiers in Psychology, Emotion Science, Special Issue on Expression of emotion in music and vocal communication*, vol. 4, no. Article ID 292, pp. 1–12, May 2013.

[16] D. Bone, C.-C. Lee, and S. S. Narayanan, "Robust Unsupervised Arousal Rating: A Rule-Based Framework with Knowledge-Inspired Vocal Features," *IEEE Transactions on Affective Computing (TAC)*, vol. 5, no. 2, pp. 201–213, April-June 2014.

[17] F. Ringeval, S. Amiriparian, F. Eyben, K. Scherer, and B. Schuller, "Emotion recognition in the wild: Incorporating voice and lip activity in multimodal decision-level fusion," in *2nd Emotion Recognition In The Wild Challenge and Workshop (EmotiW 2014), held in conjunction with the 16th International Conference on Multimodal Interaction (ICMI 2014)*, Istanbul, Turkey, September 2014, pp. 473–480.

[18] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. Andr, 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, to appear.

[19] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, September 1995.

[20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[21] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks, Special Issue, 18th International Joint Conference on Neural Networks (IJCNN)*, vol. 18, no. 5-6, pp. 602–610, July-August 2005.

[22] T. Bänziger, M. Mortillaro, and K. R. Scherer, "Introducing the Geneva Multimodal Expression Corpus for Experimental Research on Emotion Perception," *Emotion*, vol. 12, no. 2, pp. 1161–1179, 2012.

[23] M. Mortillaro, M. Mehu, and K. Scherer, "Subtly different positive emotions can be distinguished by their facial expressions," *Social Psychological and Personality Science*, vol. 2, no. 3, pp. 262–271, May 2011.

[24] M. Kipp, *Gesture generation by imitation – from human behavior to computer character animation*. Boca Raton, FL: Dissertation.com, 2004.

[25] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *Proc. of Face & Gestures 2013, 2nd IEEE International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE)*, Shanghai, China, April 2013.

[26] 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, June 2009.

[27] F. Weninger, J. Bergmann, and B. Schuller, "Introducing CURRENNT – the Munich Open-Source CUDA RecurREnt Neural Network Toolkit," *Journal of Machine Learning Research*, vol. 15, 2014, in press.

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

[29] M. Wöllmer, F. Wenninger, F. Eyben, B. Schuller, and G. Rigoll, "Acoustic-Linguistic Recognition of Interest in Speech with Bottleneck-BLSTM Nets," in *Proc. of INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association (ISCA)*, Florence, Italy, August 2011, pp. 77–80.

[30] F. Ringeval and M. Chetouani, "Hilbert-Huang Transform for non-linear characterization of speech rhythm," in *ISCA Tutorial and Research Workshop on on Non-Linear Speech Processing (NOLISP 2009)*, Vic, Spain, June 2009.

[31] F. Ringeval, M. Chetouani, and B. Schuller, "Novel metrics of speech rhythm for the assessment of emotion," in *Proc. of INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association*, Portland (OR), USA, September 2012, pp. 346–349.