At the border of acoustics and linguistics: bag-of-audio-words for the recognition of emotions in speech

Maximilian Schmitt, Fabien Ringeval, Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Schmitt, Maximilian, Fabien Ringeval, and Björn Schuller. 2016. “At the border of acoustics and linguistics: bag-of-audio-words for the recognition of emotions in speech.” In *Understanding speech processing in humans and machines: 17th Annual Conference of the International Speech Communication Association (INTERSPEECH 2016), San Francisco, California, USA, 8-12 September 2016; Volume 1*, 495–99. ISCA. https://doi.org/10.21437/interspeech.2016-1124.

INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

At the Border of Acoustics and Linguistics: Bag-of-Audio-Words for the Recognition of Emotions in Speech

Maximilian Schmitt[1], *Fabien Ringeval*[1], *Björn Schuller*[1,2]

[1]Chair of Complex & Intelligent Systems, University of Passau, Germany
[2]Department of Computing, Imperial College London, UK
{maximilian.schmitt,fabien.ringeval,bjoern.schuller}@uni-passau.de

Abstract

Recognition of natural emotion in speech is a challenging task. Different methods have been proposed to tackle this complex task, such as acoustic feature brute-forcing or even end-to-end learning. Recently, bag-of-audio-words (BoAW) representations of acoustic low-level descriptors (LLDs) have been employed successfully in the domain of acoustic event classification and other audio recognition tasks. In this approach, feature vectors of acoustic LLDs are quantised according to a learnt codebook of audio words. Then, a histogram of the occurring 'words' is built. Despite their massive potential, BoAW have not been thoroughly studied in emotion recognition. Here, we propose a method using BoAW created only of mel-frequency cepstral coefficients (MFCCs). Support vector regression is then used to predict emotion continuously in time and value, such as in the dimensions arousal and valence. We compare this approach with the computation of functionals based on the MFCCs and perform extensive evaluations on the RECOLA database, which features spontaneous and natural emotions. Results show that, BoAW representation of MFCCs does not only perform significantly better than functionals, but also outperforms by far most of recently published deep learning approaches, including convolutional and recurrent networks.

Index Terms: speech analysis, speech emotion recognition, bag-of-audio-words, computational paralinguistics

1. Introduction

Emotion recognition in speech (ERS) is a research field of growing interest, as it has found many real-life applications during the last decade, especially for human-computer interaction (HCI), and computer-mediated face-to-face interaction (FF-HCI). Indeed, having access to the affective state of the user makes HCI not only more efficient, but also more human-like [1]. Moreover, as emotion plays a prominent role in persuasion, FF-HCI can benefit from affective computing with, e. g., negotiation training systems [2].

Systems that perform automatic ERS generally follow a similar procedure. Time-continuous acoustic low-level descriptors (LLDs), such as spectral-, cepstral-, and source-related LLDs, are firstly extracted. They are computed over very short segments (usually 30 ms) of the audio signal, which are not meaningful w. r. t. the emotional state of the speaker, as emotion is a suprasegmental phenomenon. Therefore, functionals, such as moments and percentiles, are computed from the LLDs with a larger segment of the audio signal, e. g., a speaker turn, or a sliding window of several seconds. A discriminative classifier, such as Support Vector Machines (SVMs) [3], can then be used to perform emotion prediction.

Recently, *end-to-end* learning was proposed for ERS, by combining convolutional neural networks with memory enhanced Deep Neural Networks (DNNs), and thus omitting the feature extraction step [4]. However, this process is fully based on machine learning, and meaningful interpretations of the features learnt by the DNNs are hard to obtain. Another approach has recently emerged – bag-of-audio-words (BoAW) –, in order to estimate an understandable, yet useful, representation of the LLDs, taking benefits of the accomplishments obtained in the field of natural language processing, where documents are classified based on linguistic features. An interesting motivation behind BoAW is that the recognition system is more robust as the feature vectors are quantised as a first step. The codebooks are usually generated using clustering methods [5]. Random sampling of the training set has also been proposed [6].

Being at the border of acoustics and linguistics, BoAW is a well-established technique in the field of acoustic event classification [5, 6, 7, 8], and has also been successfully applied for various others tasks, e. g., monitoring [9], song detection [10], multimedia copy detection [11], and multimedia event classification [12, 13, 14]. However, BoAW has rarely been applied in the field of ERS. In [15], the IS 2009 Emotion challenge feature set [16] was used to construct BoAW. As the feature vectors are quite large, they were split into several sub-vectors before being quantised according to corresponding sub-codebooks. Then, only the assigned indexes were quantised in another step according to a high-level codebook. Even though the reported results outperformed the state-of-the-art on a two-class emotion recognition task (negativity vs non-negativity), the use of functionals in a BoAW framework is neither common nor well-founded. Indeed, the histogram generation itself already describes the distribution of the features. Moreover, the use of a hierarchical creation of the bag from indexes is questionable, as their order and distance does not contain any information.

In this light, we investigate BoAW for a time-continuous prediction task of spontaneous emotions from speech. We will show that, this approach can provide highly competitive results, even better than recent DNNs based *end-to-end* learning [4], by using only MFCCs and energy as LLDs, and linear SVMs for regression. The major contributions of this study are the following: (i) to our best knowledge, time- and value-continuous ERS using BoAW is investigated for the first time, (ii) a relatively simple approach, based on standardised LLDs, is presented achieving still best results for the prediction of the emotional valence from speech on the RECOLA corpus, and (iii) BoAW representations are compared to and fused with a representation using functionals.

The remainder of this article is structured as follows: first, a detailed description of the BoAW system is introduced (Sec-

http://dx.doi.org/10.21437/Interspeech.2016-1124

tion 2); then, we define the database and experimental setups (Section 3). We next comment on the evaluation results (Section 4), before concluding (Section 5).

2. Methodology

2.1. Feature extraction

Even though they do not incorporate explicitly prosodic information, MFCCs have proven to be relevant for ERS [17]. As acoustic LLDs, we thus computed 12 MFCCs and the logarithmic signal energy on 25 ms long frames, with a frame rate of 10 ms, using our open-source *openSMILE* toolkit [18]; a preemphasis ($k = 0.97$) was done beforehand on the acoustic signal. All features are then standardised to zero mean and unit standard deviation with an online approach, i. e., mean and standard deviation values are computed on the training partition and used to standardise all features of training, validation, and test partition.

2.2. Bag-of-audio-words

The codebook generation is performed on the training partition. We investigate two different methods: random sampling [6], and k-means++ clustering [19]. Random sampling can be regarded as the initialisation step of k-means++ clustering, where the codebook vectors are picked iteratively from the entire training partition, with audio words having a larger distance to the already selected words have a higher probability of being chosen next. However, codebook generation based on kmeans++ clustering resulted in only little performance improvement (about 2 %), for some specific configurations. Therefore, we decided to only use random sampling to generate the codebook on all experiments, as it is much faster to compute compared to kmeans++.

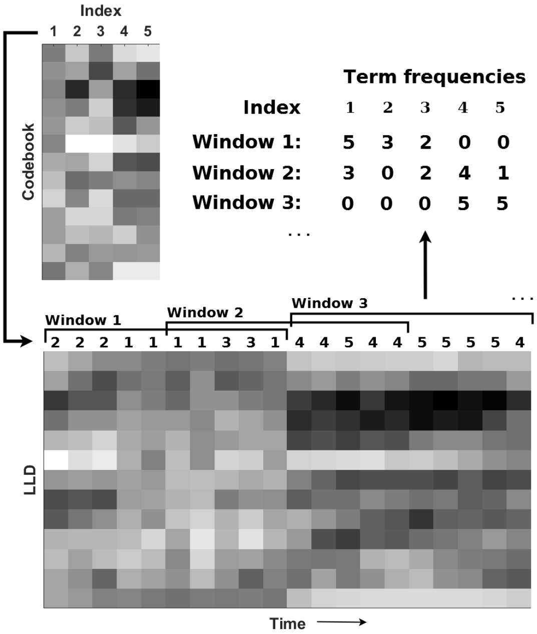
Regarding the codebook size, i. e., the number of audio words, there is no general best practice. It is, however, obvious that the optimum codebook size depends on the number and type of LLDs, but also on the task. In general, larger codebooks are thought to be more discriminative, whereas smaller codebooks should generalise better [12], especially when background noise is present in the data, as smaller codebooks are more robust against incorrect assignments.

Once the codebook has been generated, acoustic LLDs within a certain window of the speech signal are quantised, i. e., assigned to the closest (Euclidean distance) audio word in the codebook. Then, a histogram ('bag') is created from the frequencies, where each audio word w is closest to the features of an input frame within the window, the term frequency $TF(w)$. This is exemplified for a short audio excerpt in Figure 1. As it might be the case that one input frame has a low distance to several audio words and hence the assignment is ambiguous, we take multiple assignments into account, i. e., the TF of the N_a nearest audio words w is incremented by 1. Thus, no soft thresholding, or Gaussian encoding is applied, as proposed in [14], as preliminary experiments have not led to convincing results.

As in the standard *bag-of-words* approach from document classification, the common logarithm is taken to compress the range of the term frequencies: $TF'(w) = \lg(TF(w) + 1)$. The whole BoAW framework – *openXBOW* – has been implemented in Java and is publicly available as an open-source toolkit [20].

2.3. Support vector regression

In order to perform time-continuous prediction of emotional dimensions (arousal and valence), we used SVMs based regression. As shown in [12], and confirmed in our preliminary experiments, Gaussian and polynomial kernels do not perform better than linear kernels with BoAW. In addition, we tested the *histogram intersection kernel* [21, 12], but the performance was similar to that with a linear kernel, for a larger computational effort. To speed up the training and tuning of the hyperparameters, we thus used the *Liblinear* toolkit [22], with the default solver, i. e., L2-regularized / L2-loss regression with the dual formulation of the SVMs optimisation problem, and a unit bias. The complexity parameter is optimised in the range $C = [10^{-5} - 10^{0}]$ with a logarithmic scale. In order to compensate for scaling and bias issues in the predictions, but also noise in the data, we used the same post-processing chain as employed in [4].



Term frequencies					
Index	1	2	3	4	5
Window 1:	5	3	2	0	0
Window 2:	3	0	2	4	1
Window 3:	0	0	0	5	5
. . .					

Figure 1: *BoAW generation exemplified for the case of single audio word assignment.*

3. Experiments

3.1. Database

We evaluate the proposed method on the RECOLA (Remote Collaborative and Affective Interactions) corpus [23], as it contains fully spontaneous and natural affective behaviours. It includes 46 multimodal (audio, video, and physiological data) recordings of French speaking participants involved in a dyadic collaborative task. Even though the proposed *openXBOW* framework could be applied similarly to video and physiological data, we only used the audio recordings in this study.

Affective behaviour of the participants was evaluated by six different annotators (3 female, 3 male), for the first five minutes of each recordings, i. e., all annotators consistently annotated all recordings. Annotation was performed for arousal and valence separately, on a continuous scale ranging from -1 to $+1$. Obtained labels were then resampled to a constant 40 ms frame rate.

The gold standard, i. e., the emotion perceived by all raters, was estimated by considering inter-evaluator agreement [24, p. 25]. As the emotion shown by the participants and the one reported by the annotators are not contemporaneous, due to the reaction time of the raters, this delay must be taken into account when using a non-context aware machine learning algorithm [25]. Therefore, all annotations are shifted backward in time before training a model. In our experiments, we optimised the time shifts with values ranging from 0 to 8 s, and a 0.4 s step.

3.2. Evaluation

In order to ensure speaker-independence in the experiments, the corpus was split into 3 disjoint partitions: training (16 subjects), validation (15 subjects), and testing (15 subjects), by stratifying on gender and mother tongue; we used the exact same partitions as in [4]. All hyper-parameters were optimised on the validation partition, and then applied on the test partition. Training of the models was performed with data computed only every 800 ms. In contrast, the evaluation on both validation and test set is done at the original rate, i. e., every 40 ms. As SVMs cannot learn long-term contextual information, we optimised the size of the sliding window used to compute acoustic LLDs within the range of 4 s to 12 s, with 2 s step.

To evaluate the agreement level between the predictions of emotion and the gold standard, the standard metric is the concordance correlation coefficient (CCC) [26], cf. Equation 1:

$$\rho_c = \frac{2\sigma_{xy}^2}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}, \tag{1}$$

with $\mu_x = \mathrm{E}(\mathbf{x})$, $\mu_y = \mathrm{E}(\mathbf{y})$, $\sigma_x^2 = \mathrm{var}(\mathbf{x})$, $\sigma_y^2 = \mathrm{var}(\mathbf{y})$ and $\sigma_{xy}^2 = \mathrm{cov}(\mathbf{xy})$. In contrast to the largely used Pearson's correlation coefficient (PCC), CCC takes also the bias, i. e., $(\mu_x - \mu_y)^2$, into account. Just as PCC, CCC takes values between -1 and 1, where 1 implies optimum prediction.

3.3. BoAW

We performed an iterative search on the parameter space consisting of delay (D), window size (W_s), codebook size (C_s), number of assignments (N_a) and the complexity parameter of SVR (C). Preliminary experiments have shown that a delay of 3.2 s, and a window size between 6 and 8 s, provide the best results. Thus, in the first round of optimisation, we kept the delay constant and varied the window size only between 6 and 8 s. Results obtained during this optimisation phase are given in Table 1.

In order to further optimise complexity, window size and delay, we chose only three combinations of N_a and C_s that prove to work well: the optimum parameters ($N_a = 200$, $C_s = 5000$), a well-suited codebook size in case of single-assignment ($N_a = 1$, $C_s = 200$) and the configuration ($N_a = 20$, $C_s = 1000$), as a trade-off between good performance for the prediction of valence and computational effort. It must be noted that, a larger codebook size leads to a higher computational complexity. Table 2 provides the best results with those three sets of parameters. Additionally, we show the evolution of performance over different delays and window sizes in Figure 2; we used here the same configuration ($N_a = 20$, $C_s = 1000$) to save computation time.

3.4. Comparison with functionals

To generate a point of comparison between BoAW and functionals, we computed the mean and the standard deviation for all 13 LLDs, instead of BoAW. The same optimisation procedure as used for BoAW was then performed on the obtained features. The results are given in Table 3.

Table 2: *Optimised parameters and results for the validation and test partitions for single-assignment and multi-assignment.*

N_a	C_S	Dimension	D [s]	W_s [s]	C	CCC Valid	Test
1	200	Arousal	4.0	8.0	10^{-3}	.768	.716
20	1000	Arousal	3.6	8.0	10^{-4}	.789	.738
200	5000	Arousal	3.2	6.0	10^{-5}	**.793**	**.753**
1	200	Valence	4.8	12.0	10^{-2}	.490	.417
20	1000	Valence	4.4	10.0	10^{-3}	.550	**.430**
200	5000	Valence	5.2	12.0	10^{-1}	**.558**	.378

Table 3: *Results with functionals of LLDs.*

Dimension	D [s]	W_s [s]	CCC Valid	Test
Arousal	4.0	8.0	.790	.720
Valence	4.0	10.0	.459	.402

3.5. Feature fusion

In order to estimate the complementarity between the two representations of the LLDs (functionals and BoAW), we performed early fusion of the features, i. e., we concatenated the 26 features obtained with the functionals with the best BoAW models obtained in Table 2. A delay of 4 s and a window size of 8 s and 10 s, for arousal and valence, respectively, were chosen as this configuration appeared to work best on average. To get a fair comparison between the two approaches, we provide results with and without standardisation of all features in Table 4.

4. Discussion

It is obvious that the optimal number of assignments depends on the codebook size. Results show that, multi-assignment (and thus larger codebooks) are more useful for the prediction of valence compared to arousal, for which BoAW representations are only beneficial, i. e., statistically significant ($p < 0.05$) w. r. t. Fisher's z-transform, compared to simple functionals, for the largest codebook size, cf. Table 2. However, results obtained on valence with BoAW are almost always significantly better than with functionals, for both validation and test partition, except for the test partition with the largest codebook size. Parameters thus must be tuned more carefully for valence than for arousal, which confirms that, prediction of the emotional valence is more challenging than for arousal. However, considering the lower performance obtained on the test partition for valence ($N_a = 200$, $C_s = 5000$), a codebook size larger than 10^3 seems not reasonable.

Early fusion of functionals and BoAW clearly improves the performance for valence on the test partition, which thus show their complementarity. Surprisingly, larger codebooks, which generally worked better in case of BoAW only, decreased the performance when fused with functionals. One possible reason might be the larger difference of the dimensions. Also standardisation does not seem to be beneficial.

A comparison of performance obtained on the audio recordings of the RECOLA database with the two best performing methods based on DNNs is given in Table 5. All of those approaches are significantly ($p < 0.05$) outperformed by the proposed BoAW for valence, while gaining almost the same performance for arousal. It must be stated, however, that the winner of the AV+EC 2015 Challenge [27], had less data available

Table 1: *Best CCC (arousal|valence) for the given codebook size (C_s) and number of assignments (N_a) on the validation partition.*

	N_a									
C_s	1	2	5	10	20	50	100	200	500	1000
10	.750|.358	.716|.332	.715|.222							
20	.751|.355	.750|.353	.744|.314	.739|.319						
50	.776|.447	.773|.463	.775|.410	.782|.393	.765|.425					
100	.771|.469	.777|.477	.786|.477	.784|.440	.784|.422	.768|.382				
200	**.766|.474**	.774|.502	.779|.491	.785|.458	.786|.431	.782|.388	.769|.399			
500	.761|.480	.760|.477	.779|.519	.787|.518	.790|.512	.788|.466	.789|.442	.784|.383		
1000	.763|.444	.760|.471	.777|.501	.783|.522	**.789|.539**	.789|.509	.787|.490	.788|.462	.785|.402	
2000	.746|.459	.752|.459	.770|.494	.779|.505	.783|.528	.787|.541	.790|.530	.790|.515	.788|.449	.789|.406
5000	.742|.423	.746|.423	.760|.482	.768|.493	.772|.504	.785|.525	.791|.540	**.793|.543**	.792|.514	.791|.491
10000	.747|.373	.750|.373	.761|.484	.761|.484	.764|.494	.780|.515	.787|.522	.790|.532	.791|.520	.791|.509

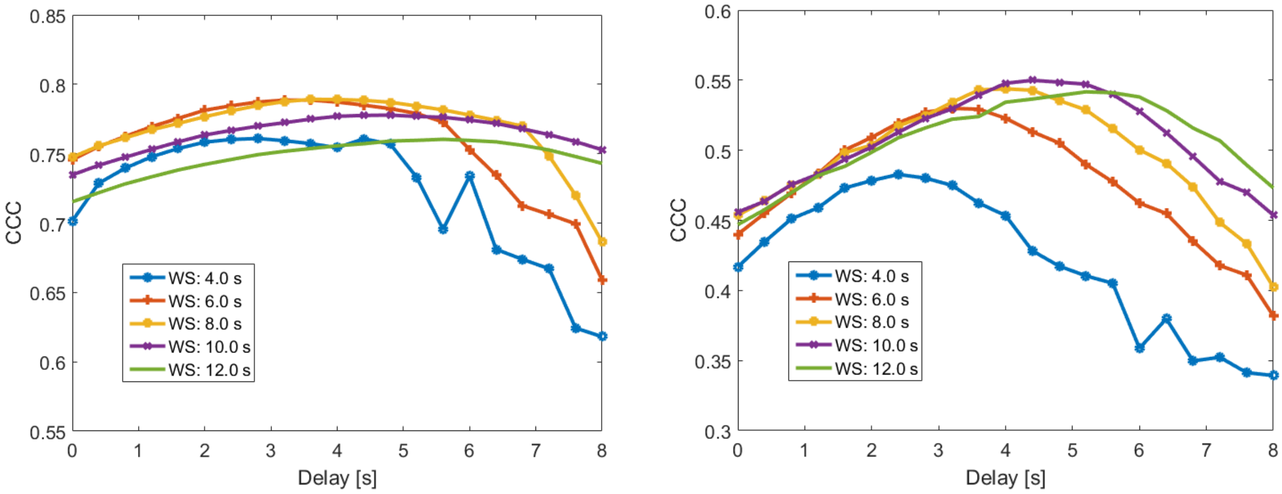


Figure 2: *Performance for arousal (left) and valence (right) with different window sizes and delays. ($N_a = 20$, $C_s = 1000$)*

Table 4: *Results with early fusion of functionals and BoAW. Delay D=4.0 s.*

N_a	C_s	Dimension	W_s	Std.	CCC	
			[s]	BoAW	Valid	Test
1	200	Arousal	8.0	no	**.799**	**.738**
20	1000	Arousal	8.0	no	.677	.511
1	200	Arousal	8.0	yes	.796	.728
20	1000	Arousal	8.0	yes	.535	.384
1	200	Valence	10.0	no	.518	.457
20	1000	Valence	10.0	no	.309	.234
1	200	Valence	10.0	yes	**.521**	**.465**
20	1000	Valence	10.0	yes	.245	.196

Table 5: *Performance comparison of recently published methods for speech-based emotion recognition on RECOLA.*

Model	Ref.	CCC			
		Arousal		Valence	
		Valid	Test	Valid	Test
BLSTM-RNN	[27]	**.800**		.398	
CNN (end-to-end)	[4]	.741	.686	.325	.261
Proposed (BoAW)	Table 2	.793	**.753**	**.550**	.430
Proposed (early fusion)	Table 4	.799	.738	.521	**.465**

to train their model (9 sessions), and some improvement could probably be obtained by training on more sessions.

5. Conclusions and outlook

In this paper, we have shown that, BoAW can significantly outperform best performing deep learning based approaches for ERS on the RECOLA database, by using only MFCC and energy LLDs. Moreover, we have shown that this representation is complementary with traditional functionals, as early fusion improved further the performance for valence.

Future work will comprise the investigation of methods taking structural information into account, such as the pyramid scheme [5] or n-grams [13], which are both well known in language processing. To exploit the linguistic information of the speech, the proposed features from the acoustic domain will be augmented by textual BoW by means of automatic speech recognition. The BoAW will be further evaluated in real-life conditions, i. e., on noisy data sets recorded 'in the wild'. Moreover, long short-term memory recurrent neural networks (LSTM-RNN) will be exploited for the regression task, instead of SVMs, as they are capable of modelling long-term dependencies between features and emotional behaviour.

6. Acknowledgements

The research leading to these results has received funding from the EC's 7th Framework Programme through the ERC Starting Grant No. 338164 (iHEARu) and the European Union's Horizon 2020 Programme through the Innovative Action No. 645094 (SEWA).

7. References

[1] M. Pantic and L. J. Rothkrantz, "Toward an affect-sensitive multimodal human-computer interaction," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1370–1390, 2003.

[2] B. Barry and R. L. Oliver, "Affect in dyadic negotiation: A model and propositions," *Organizational Behavior and Human Decision Processes*, vol. 67, no. 2, pp. 127–143, 1996.

[3] O.-W. Kwon, K. Chan, J. Hao, and T.-W. Lee, "Emotion recognition by speech signals." in *Proc. of INTERSPEECH*, Geneva, Switzerland, September 2003, pp. 125–128.

[4] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. of ICASSP*, Shanghai, China, 2016.

[5] R. Grzeszick, A. Plinge, and G. A. Fink, "Temporal acoustic words for online acoustic event detection," in *Proc. of the 37th German Conf. Pattern Recognition*, Aachen, Germany, 2015.

[6] S. Rawat, P. F. Schulam, S. Burger, D. Ding, Y. Wang, and F. Metze, "Robust audio-codebooks for large-scale event detection in consumer videos," in *Proc. of INTERSPEECH*, Lyon, France, August 2013, pp. 2929–2933.

[7] A. Plinge, R. Grzeszick, and G. A. Fink, "A bag-of-features approach to acoustic event detection," in *Proc. of the 39th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Florence, Italy: IEEE, 2014, pp. 3732–3736.

[8] H. Lim, M. J. Kim, and H. Kim, "Robust sound event classification using lbp-hog based bag-of-audio-words feature representation," in *Proc. of INTERSPEECH*, Dresden, Germany, September 2015, pp. 3325–3329.

[9] J. Joshi, R. Goecke, S. Alghowinem, A. Dhall, M. Wagner, J. Epps, G. Parker, and M. Breakspear, "Multimodal assistive technologies for depression diagnosis and monitoring," *Journal on MultiModal User Interfaces*, vol. 7, no. 3, pp. 217–228, 2013.

[10] M. Riley, E. Heinen, and J. Ghosh, "A text retrieval approach to content-based audio hashing," in *Proc. of the 9th International Conference on Music Information Retrieval (ISMIR)*, Philadelphia, PA, USA, September 2008, pp. 295–300.

[11] Y. Liu, W.-L. Zhao, C.-W. Ngo, C.-S. Xu, and H.-Q. Lu, "Coherent bag-of audio words model for efficient large-scale video copy detection," in *Proc. of the ACM International Conference on Image and Video Retrieval*. ACM, 2010, pp. 89–96.

[12] S. Pancoast and M. Akbacak, "Bag-of-audio-words approach for multimedia event classification," in *Proc. of INTERSPEECH*, Portland, USA, September 2012, pp. 2105–2108.

[13] ——, "N-gram extension for bag-of-audio-words," in *Proc. of the 38th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vancouver, Canada: IEEE, 2013, pp. 778–782.

[14] ——, "Softening quantization in bag-of-audio-words," in *Proc. of the 39th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Florence, Italy: IEEE, 2014, pp. 1370–1374.

[15] F. Pokorny, F. Graf, F. Pernkopf, and B. Schuller, "Detection of negative emotions in speech signals using bags-of-audio-words," in *Proc. of the 1st International Workshop on Automatic Sentiment Analysis in the Wild (WASA 2015) held in conjunction with ACII 2015*, AAAC. Xi'an, China: IEEE, September 2015, pp. 879–884.

[16] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Proc. of INTERSPEECH*, Brighton, UK, September 2009, pp. 312–315.

[17] A. Batliner and R. Huber, *Speaker Classification I*. Berlin, Heidelberg: Springer, 2007, ch. Speaker Characteristics and Emotion Classification, pp. 138–151.

[18] 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: ACM, October 2013, pp. 835–838.

[19] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proc. of the 18th annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.

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

[21] A. Barla, R. Odone, and A. Verr, "Histogram intersection kernel for image classification," in *Proc. of the International Conference on Image Processing (ICIP)*, vol. 3. Barcelona, Spain: IEEE, 2003, pp. 513–516.

[22] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *The Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[23] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *Face and Gestures 2013, Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE)*, 2013.

[24] B. Schuller, *Intelligent Audio Analysis*, ser. Signals and Communication Technology. Springer, 2013.

[25] 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 physiological data," *Pattern Recognition Letters*, vol. 66, pp. 22–30, November 2015.

[26] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic, "AV+EC 2015 - the first affect recognition challenge bridging across audio, video, and physiological data," in *Proc. of the 5th International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '15. Brisbane, Australia: ACM, October 2015, pp. 3–8.

[27] L. He, D. Jiang, L. Yang, E. Pei, P. Wu, and H. Sahli, "Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks," in *Proc. of the 5th International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '15. Brisbane, Australia: ACM, October 2015, pp. 73–80.