Bags in Bag: Generating Context-Aware Bags for Tracking Emotions from Speech

Jing Han, Zixing Zhang, Maximilian Schmitt, Zhao Ren, Fabien Ringeval, Björn Schuller

Bags in Bag: Generating Context-Aware Bags for Tracking Emotions from Speech

Jing Han[1]*, Zixing Zhang*[2]*, Maximilian Schmitt*[1]*, Zhao Ren*[1]*, Fabien Ringeval*[3]*, Björn Schuller*[1,2]

[1]ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing,
University of Augsburg, Germany
[2]GLAM – Group on Language, Audio & Music, Imperial College London, UK
[3]Laboratoire d'Informatique de Grenoble, Université Grenoble Alpes, France

jing.han@informatik.uni-augsburg.de

Abstract

Whereas systems based on deep learning have been proposed to learn efficient representations of emotional speech data, methods such as Bag-of-Audio-Words (BoAW) have yielded similar or even better performance while providing understandable representations of the data. In those representations, however, context information is overlooked as the BoAW include only local information. In this paper, we propose to learn a novel representation 'Bag-of-Context-Aware-Words' that encapsulates the context with neighbouring frames of BoAW; segment-level BoAW are extracted in the first layer which are then utilised to create a final instance-level bag. Such a hierarchical structure of BoAW enables the system to learn representations with context information. To evaluate the effectiveness of the method, we perform extensive experiments on a time- and value-continuous spontaneous emotion database: RECOLA. The results show that, the best segment length for valence is twice of that for arousal, suggesting that the prediction of the emotional valence requires more context information than the prediction of arousal, and the performance obtained on RECOLA with the proposed Bag-of-Context-Aware-Words outperforms all previously reported results.

Index Terms: speech analysis, emotion recognition, bag-of-audio-words, context-aware representations

1. Introduction

Emotion Recognition from Speech (ERS) plays an essential role in establishing natural and friendly human-machine communication [1] in various applications such as healthcare [2, 3], education [4, 5], robotics [6, 7], and call-centres [8, 9]. Most traditional ERS systems have been focused on extracting statistical features of acoustic Low-Level Descriptors (LLDs) such as pitch, log energy, formants, and Mel Frequency Cepstral Coefficients (MFCCs). These statistical features are then fed into various classifiers based on generative models such as naive Bayes [10] or discriminative models such as Support Vector Machines [11].

Recently, representation learning methods based on deep learning have been proposed to learn an appropriate set of high-level information directly from the raw speech signal, instead of computing statistical measures of expert-based LLDs. In the so-called *end-to-end* learning [12, 13], a system jointly learns an emotion inference task, usually with a fully-connected recurrent network taking features as input, along with a feature learning task, usually with a convolutional neural network taking a portion of the raw signal as input. Even though *end-to-end* learning performs well for ERS [13, 14], the learnt representations are hard to interpret or understand.

In contrast, another novel approach, Bag-of-Audio-Words (BoAW), has been proposed for ERS, with the aim to estimate a meaningful and robust representation based, e.g., on MFCCs and log-energy as LLDs [15]. In [15], these LLDs are quantised, and histograms are computed with a random-selected codebook as final representations which give one of the best reported recognition performances on the popular spontaneous emotional dataset RECOLA [16]. Moreover, BoAW has been applied successfully in several other paralinguistic information retrieval tasks, such as sound event classification [17], music genre classification [18], and copy detection [19].

While BoAW has produced meaningful and robust representations for ERS, it does not take context information into consideration when creating the representations; since emotional content is involved in multiple coherent frames, context information is vital and needs to be dealt with care in an ERS system. Contrary to the conventional BoAW approach, we propose an approach to generate Bag-of-Context-Aware-Words (BoCAW) representations in a hierarchical architecture, to preserve the context information while learning the representations. More specifically, BoAW is applied twice but within different temporal scales; a small local window containing a number of context frames is first utilised, and then a global analysis window containing all frames of one instance is explored.

Such a hierarchical structure is conceptually similar to a Deep Belief Net (DBN), where features with various granularities can be extracted from each layer of the DBN [20]. However, we extend the concept to BoAW, which is simpler to train without massive hyper-parameter tuning, and also easier for theoretical analysis than DBN. In addition, BoCAW is further related to Dual-Layer Bag-of-Frames (DLBoF) proposed in [18]. The DLBoF framework attempts to model a piece of music with a two layer structure, where frame-level characteristics and segment-level semantics can be captured and integrated together for music information retrieval tasks.

In the present work, however, we utilise only the segment-level features from the second layer, and we demonstrate by an empirical analysis that these features can ameliorate the performance of ERS. Furthermore, we explore and find the proper length of the local window that best fits a time- and value-continuous ERS system for the emotional dimensions arousal and valence, respectively. To the best of our knowledge, this is the first work that learns features from a hierarchical BoAW structure for ERS. Also, this bridges the gap between frame-based features and long-term emotional speech by introducing segment-level words with context information, so as to enhance the regular BoAW approach.

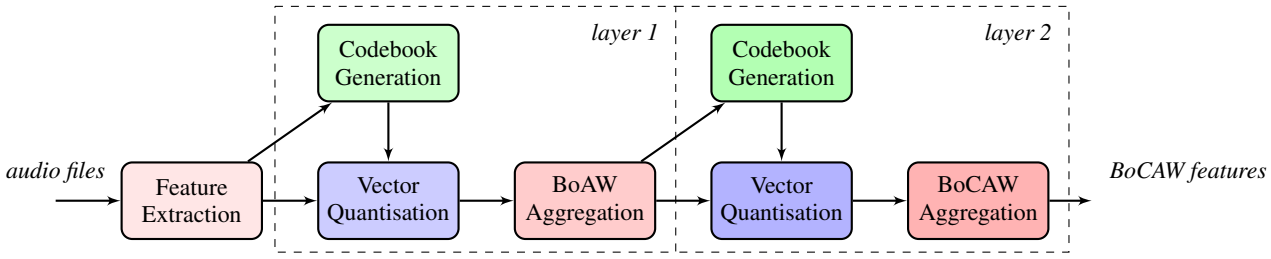

Figure 1: *Diagram of the bags-in-bag approach pipeline.*

2. Generation of Context-Aware Bags

The framework of our proposed BoCAW feature generation approach is depicted in Figure 1 and consists of two layers. In the first layer, sub-bag (or segment-level) features are generated by applying the conventional BoAW approach to input frame-level features within a sliding window. Note that, each sliding window contains several successive frames of features and can be much smaller than the total length of an instance to be analysed. After that, in the second layer, these sub-bag features are further utilised as context-aware words to generate a final context-aware bag for each given instance. Herein, BoAW is again applied as in the first layer but within a global window, the length of which equals to the total length of the instance to be analysed. We implement the whole framework with our open-source toolkit OPENXBOW [21]; the details of each system component are described below.

2.1. Bag of Audio Words Model

Let us denote a frame-level feature vector as $x_n \in \mathbb{R}^D$ such that $n = 1, ..., N$, where N is the total number of frames from the entire training set, and D is the dimension of the vector. Therefore, $\{x_n\}_{n \in N_i}$ denotes a set of features for a given audio file which is composed of N_i frames. Next, as shown in Figure 1, once the series of $\{x_n\}$ is extracted from all audio files, the traditional BoAW approach is conducted in layer 1, which contains the following three steps:

Codebook generation: a codebook C is a set of codewords c learnt from the feature space $\{x_n\}$, and the codebook generation problem can be formulated as:

$$C = \{c_k\}_{k=1}^{K}, c_k \in \mathbb{R}^D, \tag{1}$$

where c_k denotes the k-th codeword, and in total K codewords form the codebook $C \in \mathbb{R}^{D \times K}$.

Normally, C can be created by a clustering algorithm such as k-means. In addition, random sampling has also been proposed in [22] and utilised with success for ERS [15]. In this work, we build the frame-level codebook C_1 in layer 1 by random sampling, which is much faster than k-means but delivers comparative performances at the same time.

Vector Quantisation: once the codebook C_1 has been generated, each x_n can be assigned to its closest (Euclidean distance) codeword c_k in C_1, and be encoded as the corresponding index k. This process is referred to as the vector quantisation step, and can be formulated as $\phi_n = f(x_n, C_1)$, where the function $f : \mathbb{R}^D \rightarrow \mathbb{R}^K$ encodes each feature x_n into the codebook space, resulting in a corresponding k-dimension feature $\phi_n \in \mathbb{R}^K$ while its k-th coefficient $\phi_{n,k}$ with $k = 1, ..., K$ is defined as follows:

$$\phi_{n,k} = \begin{cases} 1, & \text{if } k = \text{argmin}_k \left\| x_n - c_k \right\|_2^2 \\ 0, & \text{otherwise} \end{cases}. \tag{2}$$

However, during the quantisation step, there might be the case that one feature vector is nearly equidistant to several codewords, and therefore single assignment is ambiguous. Hence, instead of choosing the nearest codeword, it is also possible to assign a feature to a certain number N_a of closest codewords. This variant can be referred to as multiple assignments, and is used here as it has proven to perform better than a single assignment [15].

BoAW Aggregation: given an audio segment s spanning several frames, a 'bag' can then be created by simply computing a histogram of codewords. More specifically, a histogram representation h_s is formed to describe the distribution of the features, i. e., how and to what extent each codeword has contributed to represent s. This process is referred to as the BoAW aggregation step, and can be formulated as $h_s = g(\{\phi_n\}_{n \in N_s})$, where a pooling function $g : \mathbb{R}^{K \times N_s} \rightarrow \mathbb{R}^K$ aggregates occurrences of each codeword represented by ϕ_n in all N_s frames for a given segment s. Thus, its k-th component $h_{s,k}$ can be computed as:

$$h_{s,k} = \sum_{n=1}^{N_s} \{\phi_{n,k}\}. \tag{3}$$

At this point, with the learnt first-layer codebook C_1, original features $\{x_n\}$ are encoded into segment-level features $\{h_s\}$ which contain context information. Thus, we refer to these features as context-aware words.

2.2. Bag of Context-Aware Words Model

In regular BoAW, the histogram representation h_s that covers the entire instance (i. e., a very long segment) is the final high-level representation, which then can be exploited for audio classification or regression tasks. In contrast, we propose to treat series of the histogram feature $\{h_s\}$ generated from much shorter segments as mid-level representations, and apply a second BoAW layer to form final high-level representations. Our concern is that, typical emotion patterns may exist among a sequence of several coherent frames, and therefore features generated based on segments rather than frames may perform better for ESR.

As depicted in Figure 1, after layer 1, each given audio file a is now represented by a set of context-aware words $\{h_s\}$ with $h_s \in \mathbb{R}^K$. Next, in layer 2, similar as demonstrated in Sec. 2.1, we build a second-layer codebook C_2 by random sampling K codewords over all h_s from training files. After that, vector quantisation with MA is applied to convert h_s into indices of the N_a closest codewords from C_2. At the end, for each a, a final BoCAW representation h_a is computed by counting the occurrences of corresponding second-layer codewords for all segments in it. This process is referred to as the BoCAW aggregation step.

Note that audio files of variable lengths can be encoded

into BoCAW features with an equal and fixed length, and in the meanwhile short-term temporal information is preserved in these features.

3. Experiments and Results

To evaluate the effectiveness of the proposed bags of context-aware words approach for ERS, we conducted extensive experiments on a widely used database in the affective computing community.

3.1. Data and Features

We chose RECOLA [16], a standard database that was previously used in the Audio/Visual Emotion Challenge (AVEC) for dimensional emotion recognition in 2015 and 2016 [23, 24]. This database was created with the aim to study socio-affective behaviours from multimodal data in the context of remote collaborative tasks. More specifically, the corpus consists of spontaneous and natural interactions from 46 French-speaking participants involved in a dyadic collaborative task. Multimodal signals including audio, video, and peripheral physiology recordings such as electro-cardiogram and electro-dermal activity were recorded continuously and synchronously over time. In this study, however, only audio recordings were utilised for the emotion recognition task. To obtain the annotations, time- and value-continuous dimensional affect ratings in terms of arousal and valence were performed by six annotators. The obtained annotations were then resampled with a constant frame rate of 40 ms to align with the frame rate of the recordings. The 'gold standard' was then estimated by averaging all six annotations while considering the inter-evaluator agreement as a weighting factor [25].

In order to ensure speaker-independence, the dataset was further divided into three disjoint partitions, i. e., training (16), development (15), and test (15), by balancing the gender, age, and mother tongue of the participants. Note that, we employed the same partitions as in [15, 26–30]. As a result, the total numbers of overlapping segments in the train, development, and test partitions are 120.0 k, 112.5 k, and 112.5 k, respectively.

Furthermore, to extract acoustic features, we used our open-source OPENSMILE toolkit [31] to extract 13 LLDs, i. e., MFCC 0-12 and the logarithmic energy, with a frame window size of 25 ms and a step size of 10 ms, as the inputs of the proposed framework, which are exactly the same as the ones of a previous framework proposed in [15].

3.2. Implementation and Evaluation

Before learning the BoCAW representations, an online standardisation was conducted on all LLDs. Specifically, the mean and variance of each LLDs was calculated on the training set, which were then applied over the development and test sets for standardisation.

To demonstrate the effectiveness of the proposed method for ERS, we utilised Support Vector Regression (SVR) implemented in the LIBLINEAR toolkit [32] with a linear kernel and trained with an L2-regularised L2-loss dual solver. The complexity C was optimised on the development set in the range of $[10^{-5}, 10^{0}]$.

We also performed a grid-search over the parameters of the BoCAW which include the local window size (W_1) and the time step size (Ts_1) of layer 1. More specifically, a best setting was determined on the best performance achieved on the development set by a grid search over $[0.01, 0.02, 0.05, 0.1, 0.2, 0.4, 0.8, 1.6]$ for W_1 and $[0.05, 0.1, 0.2, 0.4]$ for Ts_1 when training (Ts_1 for development and test is fixed to be 40 ms to match the granularity of the annotations). Furthermore, for a fair comparison with our previous work in [15], we fixed other hyper-parameters of the model, maintaining the settings for the codebook size (C_s=1000), number of assignments (N_a=20), and the global window size (W_2=8.0 s) and the time step size (Ts_2=800 ms for training to achieve a fast process or 40 ms for development and test to match the granularity of the annotations) of layer 2.

Additionally, we performed the same post-processing chain on all predictions as in [15, 24, 28]: smoothing, centring, scaling, and time-shifting. All the modification parameters were optimised on the development set and then applied on the test set.

To evaluate the performance of our methods, we use *Concordance Correlation Coefficient* (CCC), which is a standard evaluation metric for time- and value-continuous prediction of emotion; it measures the agreement between the gold standard and the predictions. Given two time series x and y, their CCC is calculated as follows:

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}, \tag{4}$$

where ρ is the *Pearson's Correlation Coefficient* (PCC) between two time series (e.g., prediction and gold-standard); μ_x and μ_y are the means of each time series; and σ_x^2 and σ_y^2 are the corresponding variances. In contrast to the PCC, CCC takes not only the linear correlation, but also the bias between the two temporal series, i.e., $(\mu_x - \mu_y)^2$, into account. Hence, the value of CCC is within the range of $[-1, 1]$, where ± 1 represents perfect concordance and discordance while 0 means no correlation. In other words, a higher CCC indicates a better performance.

To further assess the significance level of performance improvement, a statistical evaluation was carried out over all predictions obtained with the BoCAW approach and with other benchmark methods such as BoAW, by means of the Fisher's r-to-z transformation [33]. Unless stated otherwise, a p value lower than .05 indicates statistical significance.

3.3. Results and Discussion

In our experiments, we carried out two regression tasks, i.e., arousal and valence prediction from speech. Table 1 presents result performances in terms of CCC for the proposed BoCAW features. Results are reported on both the development and test sets over different window sizes W_1 for arousal and valence, respectively. It can be seen from the table that, the best results on the development set for the arousal and valence dimensions is .800 and .603, respectively, and the best results obtained on the test set are .757 for arousal and .497 for valence.

Moreover, our results also show that predictions of arousal and valence are differently influenced by the length of W_1. Therefore, to better illustrate the effect of W_1 for the prediction of emotions, we compute the performance (in CCC) averaged over all four selected time step sizes Ts_1 for each predefined window size, as shown in Figure 2. When W_1=0.01 s, i.e., only one frame is included in each segment on layer 1, then, the steps conducted on layer 1 equal to quantising original features, delivering only a slight improvement. When the window size increases, i.e., an increasing number of frames are contained in a segment, the performance of emotion prediction improves until a point where information of different emotional nature are

Table 1: *Performances in terms of Concordance Correlation Coefficient (CCC) of the proposed BoCAW features with various window sizes of the first layer (W_1), for both arousal and valence regressions, evaluated on the devlopment and test partitions. Note that, for each W_1, only the best performance among four examined time step sizes (Ts_1) is reported, by calculating the averaged predictions of arousal and valence on the development set. The best results achieved are highlighted. The symbol of $*$ indicates the significance of the performance improvement over the bag-of-audio-words (BoAW) method.*

settings		*arousal*		*valence*	
$W_1(s)$	$Ts_1(s)$	*dev*	*test*	*dev*	*test*
BoAW [15]		.789	.738	.550	.430
0.01	0.1	.791	.746*	.557*	.432
0.02	0.1	.791	.753*	.581*	**.497***
0.05	0.1	**.800***	.750*	.572*	.463*
0.1	0.2	.797*	**.757***	**.603***	.465*
0.2	0.4	.787	.752*	.546	.455*
0.4	0.2	.780	.747*	.543	.492*
0.8	0.4	.775	.738	.540	.459*
1.6	0.4	.765	.733	.532	.423

contained in the window, and thus performances starts to decrease. To this end, we need to identify a proper analysis window size W_1 for the task at hand. We can observe from the figure that, the best window size is $0.05\,s$ for arousal, whereas the best performance for valence is obtained with a longer window ($0.1\,s$). This result is coherent with other findings in the literature [15, 25], and confirms that more context information is essential for valence than arousal when generating context-aware bags. Interestingly, as for human annotators, people are also slower to give valence ratings, compared to arousal [34].

Additionally, to further highlight the advantages of the BoCAW approach, we compared the best performance it achieved on the RECOLA dataset with seven others systems from the state-of-the-art. In [26], CCC was exploited as the cost function instead of standard mean squared error when training a deep model, whereas an end-to-end framework that learns representations directly from raw signals was implemented in [27]. Besides, by compensating the weakness of a model itself or incorporating the strength of different models, the Reconstruction-Error-based (RE-based) learning framework and prediction-based learning framework were proposed in [29] and [30], respectively. More recently, an adversarial training approach was investigated for emotion regression problems in [28] for the first time. The last framework to compare with is obviously BoAW [15], which is the fundamental of this work as well. A comparison of the best performances of all above mentioned approaches and our BoCAW on the RECOLA dataset is presented in Table 2. It can be seen that, when using BoCAW representations, the CCC performance is significantly improved for both arousal and valence predictions (statistical evaluation via a Fisher's r-to-z transformation as outlined in Section 3.2) compared to original BoW framework. Also, to the best of our knowledge, the BoCAW features we proposed yield the best results to date on the RECOLA database from speech.

4. Conclusions

This paper proposes a hierarchical framework that ameliorates Bag-of-Audio-Words (BoAW) with context information maintained on segment-level features, named as Bag-of-Context-Aware-Words (BoCAW). In this framework, BoAW is first applied on a sequence of segments, and then, these segment-level features are fed into a second BoAW layer to extract an higher-level representation of the information captured in the first layer. Evaluations have been conducted on the RECOLA database to assess the system performance. Results show that, the proposed BoCAW obtains state-of-the-art performance for ERS while providing understandable representations.

Further, the proposed BoCAW representations based on segments are also applicable to other pattern recognition tasks where a specific pattern lasts a period of time, such as laughter detection [35], engagement recognition [36], acoustic scene classification [37], and language identification [38].

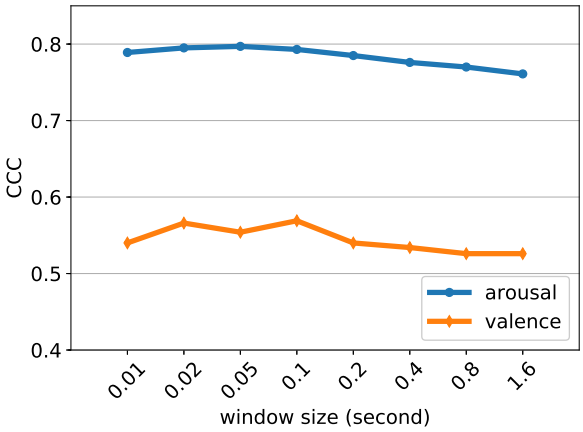

Figure 2: *The effect of the sub-bag's window size on the performance (CCC) when predicting arousal and valence separately. Performances are averaged over all examined time step sizes on the development partition.*

Table 2: *Performances in terms of CCC of the proposed method comparing with other state-of-the-art approaches on the RECOLA dataset. The best results achieved are highlighted. The symbol of $*$ indicates the significance of the performance improvement over the bag-of-audio-words (BoAW) method.*

approaches	arousal		valence	
	dev	test	dev	test
CCC-objective [26]	.412	.350	.242	.199
end-to-end [27]	.752	.699	.406	.311
RE-based [29]	.785	.729	.378	.360
prediction-based [30]	.774	.744	.440	.393
adversarial training [28]	.797	.737	.501	.455
BoAW [15]	.789	.738	.550	.430
BoCAW (*proposed*)	.800*	**.750***	**.603***	.465*

5. Acknowledgements

This work has been supported by the EU's Horizon 2020 Programme through the Innovation Action No. 645094 (SEWA), the EU's Horizon 2020 / EFPIA Innovative Medicines Initiative through GA No. 115902 (RADAR-CNS), and the UK's Economic & Social Research Council through the research Grant No. HJ-253479 (ACLEW).

6. References

[1] S. Ramakrishnan and I. M. El Emary, "Speech emotion recognition approaches in human computer interaction," *Telecommunication Systems*, vol. 52, no. 3, pp. 1467–1478, Mar. 2013.

[2] D. Tacconi, O. Mayora, P. Lukowicz, B. Arnrich, C. Setz, G. Troster, and C. Haring, "Activity and emotion recognition to support early diagnosis of psychiatric diseases," in *Proc. PervasiveHealth*, Tampere, Finland, 2008, pp. 100–102.

[3] M. S. Hossain and G. Muhammad, "Cloud-assisted speech and face recognition framework for health monitoring," *Mobile Networks and Applications*, vol. 20, no. 3, pp. 391–399, June 2015.

[4] A. Zhu and Q. Luo, "Study on speech emotion recognition system in e-learning," in *Proc. HCI*, Beijing, China, 2007, pp. 544–552.

[5] K. Bahreini, R. Nadolski, and W. Westera, "Towards multimodal emotion recognition in e-learning environments," *Interactive Learning Environments*, vol. 24, no. 3, pp. 590–605, Apr. 2016.

[6] J.-S. Park, J.-H. Kim, and Y.-H. Oh, "Feature vector classification based speech emotion recognition for service robots," *IEEE Transactions on Consumer Electronics*, vol. 55, no. 3, pp. 1590 – 1596, Aug. 2009.

[7] M. J. Salvador, S. Silver, and M. H. Mahoor, "An emotion recognition comparative study of autistic and typically-developing children using the zeno robot," in *Proc. ICRA*, Seattle, WA, 2015, pp. 6128–6133.

[8] L. Devillers and L. Vidrascu, "Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs," in *Proc. INTERSPEECH*, Pittsburgh, PA, 2006, pp. 801–804.

[9] P. Gupta and N. Rajput, "Two-stream emotion recognition for call center monitoring," in *Proc. INTERSPEECH*, Antwerp, Belgium, 2007, pp. 2241–2244.

[10] S. K. Bhakre and A. Bang, "Emotion recognition on the basis of audio signal using naive bayes classifier," in *Proc. ICACCI*, Jaipur, India, 2016, pp. 2363–2367.

[11] Y. Pan, P. Shen, and L. Shen, "Speech emotion recognition using support vector machine," *International Journal of Smart Home*, vol. 6, no. 2, pp. 101–108, Apr. 2012.

[12] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Proc. INTERSPEECH*, Singapore, Singapore, 2014, pp. 223–227.

[13] G. Trigeorgis, F. Ringeval, R. Bruckner, E. Marchi, M. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. ICASSP*, Shanghai, China, 2016, pp. 5200–5204.

[14] S. Khorram, Z. Aldeneh, D. Dimitriadis, M. McInnis, and E. M. Provost, "Capturing long-term temporal dependencies with convolutional networks for continuous emotion recognition," in *Proc. INTERSPPECH*, Stockholm, Sweden, 2017, pp. 1253–1257.

[15] M. Schmitt, F. Ringeval, and B. Schuller, "At the border of acoustics and linguistics: Bag-of-audio-words for the recognition of emotions in speech," in *Proc. INTERSPEECH*, San Francisco, CA, 2016, pp. 495–499.

[16] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *Proc. EmoSPACE (FG)*, Shanghai, China, 2013, pp. 1–8.

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

[18] C.-C. M. Yeh, L. Su, and Y.-H. Yang, "Dual-layer bag-of-frames model for music genre classification," in *Proc. ICASSP*, Vancouver, Canada, 2013, pp. 246–250.

[19] 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. ACM CIVR*, Xi'an, China, 2010, pp. 89–96.

[20] H. Lee, P. Pham, Y. Largman, and A. Y. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *Proc. NIPS*, Vancouver, Canada, 2009, pp. 1096–1104.

[21] M. Schmitt and B. Schuller, "openXBOW-Introducing the Passau open-source crossmodal bag-of-words toolkit," *Journal of Machine Learning Research*, vol. 18, no. 96, pp. 1–5, Oct. 2017.

[22] 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. INTERSPEECH*, Lyon, France, 2013, pp. 2929–2933.

[23] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic, "AVEC 2015: The 5th international audio/visual emotion challenge and workshop," in *Proc. ACM MM*, Brisbane, Australia, 2015, pp. 1335–1336.

[24] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "AVEC 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proc. AVEC*, Amsterdam, The Netherlands, 2016, pp. 3–10.

[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, Nov. 2015.

[26] F. Weninger, F. Ringeval, E. Marchi, and B. Schuller, "Discriminatively trained recurrent neural networks for continuous dimensional emotion recognition from audio," in *Proc. IJCAI*, New York, NY, 2016, pp. 2196–2202.

[27] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, Dec. 2017.

[28] J. Han, Z. Zhang, Z. Ren, F. Ringeval, and B. Schuller, "Towards conditional adversarial training for predicting emotions from speech," in *Proc. ICASSP*, Calgary, Canada, 2018, pp. 6822–6826.

[29] J. Han, Z. Zhang, F. Ringeval, and B. Schuller, "Reconstruction-error-based learning for continuous emotion recognition in speech," in *Proc. ICASSP*, New Orleans, LA, 2017, pp. 2367–2371.

[30] ——, "Prediction-based learning for continuous emotion recognition in speech," in *Proc. ICASSP*, New Orleans, LA, 2017, pp. 5005–5009.

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

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

[33] J. Cohen, P. Cohen, S. G. West, and L. S. Aiken, *Applied multiple regression/correlation analysis for the behavioral sciences*. Abingdon, UK: Routledge, 2013.

[34] A. Nicolle and V. Goel, "Differential impact of beliefs on valence and arousal," *Cognition & Emotion*, vol. 27, no. 2, pp. 263–272, Feb. 2013.

[35] G. Gosztolya, "Optimized time series filters for detecting laughter and filler events," in *Proc. INTERSPEECH*, Stockholm, Sweden, 2017, pp. 2376–2380.

[36] Y. Huang, E. Gilmartin, and N. Campbell, "Speaker dependency analysis, audiovisual fusion cues and a multimodal BLSTM for conversational engagement recognition," in *Proc. INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3359–3363.

[37] J. Guo, N. Xu, L.-J. Li, and A. Alwan, "Attention based CLDNNs for short-duration acoustic scene classification," in *Proc. INTERSPEECH*, Stockholm, Sweden, 2017, pp. 469–473.

[38] S. Irtza, V. Sethu, E. Ambikairajah, and H. Li, "Investigating scalability in hierarchical language identification system," in *Proc. INTERSPEECH*, Stockholm, Sweden, 2017, pp. 2581–2585.