Interspeech 2018
2-6 September 2018, Hyderabad

LSTM based Cross-corpus and Cross-task Acoustic Emotion Recognition

Heysem Kaya[1], *Dmitrii Fedotov*[2,3], *Ali Yeşilkanat*[4], *Oxana Verkholyak*[2], *Yang Zhang*[5], *Alexey Karpov*[2]

[1]Department of Computer Engineering, Namık Kemal University, Çorlu, Tekirdağ, Turkey
[2]St. Petersburg Institute for Informatics and Automation of Russian Academy of Sciences, Russia
[3] Institute of Communications Engineering, Ulm University, Ulm, Germany
[4] Department of Computer Engineering, Boğaziçi University, İstanbul, Turkey
[5] Noah's Ark Lab, Huawei Technologies, Shenzhen, China

hkaya@nku.edu.tr, dmitrii.fedotov@uni-ulm.de, ali.yesilkanat@boun.edu.tr, overkholyak@gmail.com, zhangyang86@huawei.com, karpov@iias.spb.su

Abstract

Acoustic emotion recognition is a popular and central research direction in paralinguistic analysis, due its relation to a wide range of affective states/traits and manifold applications. Developing highly generalizable models still remains as a challenge for researchers and engineers, because of multitude of nuisance factors. To assert generalization, deployed models need to handle spontaneous speech recorded under different acoustic conditions compared to the training set. This requires that the models are tested for cross-corpus robustness. In this work, we first investigate the suitability of Long-Short-Term-Memory (LSTM) models trained with time- and space-continuously annotated affective primitives for cross-corpus acoustic emotion recognition. We next employ an effective approach to use the frame level valence and arousal predictions of LSTM models for utterance level affect classification and apply this approach on the ComParE 2018 challenge corpora. The proposed method alone gives motivating results both on development and test set of the Self-Assessed Affect Sub-Challenge. On the development set, the cross-corpus prediction based method gives a boost to performance when fused with top components of the baseline system. Results indicate the suitability of the proposed method for both time-continuous and utterance level cross-corpus acoustic emotion recognition tasks.

Index Terms: speech emotion recognition, cross-corpus emotion recognition, context modeling, LSTM, computational paralinguistics

1. Introduction

Studies in emotion recognition play a central role in both computational paralinguistics and affective computing. Being a central theme is partly due to conceptual and psychophysical relation of emotion to a range of affective states and traits. Studies in emotion recognition and in the general research area of affective computing are gaining momentum towards maturity, thanks to sharing of resources and common protocol challenges such as the Computational Paralinguistics (ComParE) Challenge series.

Acoustic emotion recognition in realistic conditions is challenging due to a range of factors such as speaker, gender, language and recording environment variations. Thus, cross-corpus acoustic emotion recognition that aims to cope with these issues became a popular research direction. Proposed adaptation approaches range from normalization strategies [1, 2] to covariate shift compensation via employing transfer learning methods [3], from Canonical Correlation Analysis based multi-view approaches [4] to denoising auto-encoder based domain adaptation schemes [5, 6].

Aforementioned methods show methodological and experimental improvement over state-of-the-art, however all of them were tested across corpora using utterance level emotion recognition setting. Moreover, there is a growing research body on dimensional affect recognition with Long-Short-Term-Memory (LSTM) Recurrent Neural Networks (RNN) [7, 8, 9].

Inspired from the outstanding context modeling capability of LSTM-RNN, when annotations with sufficient frequency are provided, in this paper we investigate the suitability of cross-corpus and cross-task acoustic emotion recognition. That is, we seek to benefit from LSTM models trained on time- and space-continuously annotated corpora, on a different corpus with utterance level categorical emotion annotations. We device a simple but effective approach to the problem and combine the predictions of the proposed approach with the top components of the baseline system provided by the ComParE 2018 challenge organizers. Additionally, we carry out extensive experiments on cross-corpus dimensional affect prediction on three public corpora prior to application of the proposed approach on the challenge corpus.

In line with our recent experience on paralinguistic and multi-modal affective computing [10], we employ least squares based classifiers such as Kernel Extreme Learning Machines (KELM) and Partial Least Squares (PLS) regression based classifiers for modeling utterance-level feature representations. Furthermore, we employ their weighted versions to cope with the class imbalance problem, typically observed in challenge corpora [11].

The remainder of the paper is organized as follows. We introduce the proposed framework and brief its components in the next section. In Section 3, we present experiments across three dimensional corpora in arousal and valence dimensions. Section 4 summarizes the experiments on the Ulm State-of-Mind Speech (USoMS) corpus. Section 5 concludes with discussion.

2. Proposed Framework

In our proposed framework, we combine discretized predictions of LSTM models trained on time- and space-continuously annotated corpora with the components of the baseline system [12]. The pipeline is illustrated in Figure 1, where the dashed lines correspond to processing of external dimensional corpora.

We first try to reproduce the baseline systems using SVM classifier and further improve learner performance with least

10.21437/Interspeech.2018-2298

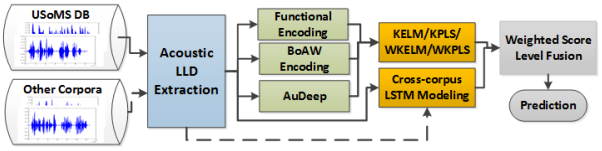

Figure 1: *Pipeline of the proposed framework.*

squares based weighted kernel classifiers. The feature set components of the baseline system comprise functional features extracted via openSMILE [13], Bag of Audio Words extracted with openXBOW [14] and the suprasegmental features from sequence to sequence autoencoder auDeep [15].

For LSTM based predictions, we train models on three corpora, namely RECOLA [16], SEMAINE [17] and CreativeIT [18] for both arousal and valence dimensions, although the target task in the Self-Assessed Affect Sub-Challenge is a three-state valence classification task. We hypothesize that the frame-level affect predictions can be summarized over the utterance using mean functional and then thresholded for discretization. The thresholds to separate target classes are optimized for UAR on the challenge training set. Note that this is a cross-corpus and cross-task affect recognition setting (from frame level dimensional affect to utterance level categorical emotions) and the resulting predictions are ordinal classes of valence. The final score-level fusion of the utterance level categorical predictions with other systems is done after converting the categorical labels $\mathbf{y}$ into a one vs. all code matrix $\mathbf{S}$:

$$\mathbf{S}_{t,l} = \begin{cases} 1 & \text{if } y^t = l, \\ 0 & \text{if } y^t \neq l. \end{cases} \quad (1)$$

2.1. Least Squares based Weighted Kernel Classifiers

In addition to the Support Vector Machines (SVM) used in the challenge baseline systems, we use Kernel ELM and PLS regression motivated from their fast and accurate learning capability and state-of-the-art results on recent paralinguistic/multimodal challenge corpora [19, 20]. We obtain linear kernels from the dataset and use them in PLS and ELM, optimizing the hyper-parameters on the development set. For handling the imbalanced data, we employ a variant of ELM dubbed *Weighted ELM* [21]. Inspired from it, in 2017 ComParE challenge, we applied this simple and efficient scheme to KPLS, introducing WKPLS [11], which gives higher weights to minority class instances in model learning.

2.2. Fusion

In line with experience on former challenge corpora [10, 11, 20], we investigate feature level fusion and two variants of score level fusion. The first is simple weighted fusion (SF) of scores, where the classifier confidence scores S^A and S^B are fused using weight $\gamma \in [0, 1]$:

$$S^{fusion} = \gamma * S^A + (1 - \gamma) * S^B. \quad (2)$$

Secondly, we apply weighted score fusion (WF) for each model and class. Let M and L denote the number of models and classes, respectively. The optimal fusion weights $W_{i,j}^{fusion} \in [0, 1], 1 \leq i \leq M, 1 \leq j \leq L, \sum_{i=1}^{M} W_{i,j}^{fusion} = 1$, are searched over a pool of randomly generated matrices.

3. Cross-Corpus Experiments for Dimensional Affect Prediction

This section presents the data, experimental setting and results for time- and space-continuous cross-corpus acoustic emotion prediction task for both the training and test corpora.

3.1. Data Description

To perform cross-corpora emotion recognition we used three well known corpora annotated in arousal and valence dimensions, namely RECOLA [16], SEMAINE [17] and CreativeIT [18].

RECOLA database consists of spontaneous interactions in French collected during solving of a collaborative task. Available version of the database has 23 recordings from different speakers with duration of 5 minutes each. Annotations from 6 raters are provided for each recording with frequency of 25 Hz. Participants are aged between 18 and 25 years old with mean of 21.3, almost equally distributed in gender (10 males, 13 females) and have three different mother tongues: French (17), Italian (3) and German (3).

SEMAINE database was collected during an interaction in English between users and Sensitive Artificial Listener (SAL), which may be represented by operator (solid SAL), partially controlled by operator (semi-automated SAL) and completely functioning through dialogue system (automated SAL). In our research we chose only users' recording in Solid SAL scenario. The data consists of 24 recordings from 20 speakers with average duration of 18.6 minutes. Speakers have an average age of 30.3 years, while 60% of them are males and 40% - females. Annotations are available from up to 8 raters with frequency of 50 Hz.

CreativeIT database consists of two types of interactions: repetition of one phrase and semi-spontaneous dialogue with predefined aims of both speakers. To increase applicability of the data, only the second type of interaction was used in our research. Selected data consists of 31 recordings from 15 speakers with average duration of 4.3 minutes. The database is annotated by 3 raters with frequency of 60 Hz.

Distributions of instances in arousal and valence dimensions are shown in Figure 2, where we observe different skewness and kurtosis statistics for each corpus. While RECOLA corpus annotations are highly centered around zero and peaked, SEMAINE and CreativeIT corpora have annotation distributions that are skewed in opposite directions compared to a each other. Moreover, while the arousal valence annotations are positively correlated in RECOLA and SEMAINE, in CreativeIT they are negatively correlated. As will be shown in experimental results, this complicates the cross-corpus recognition for CreativeIT both as source and target corpus.

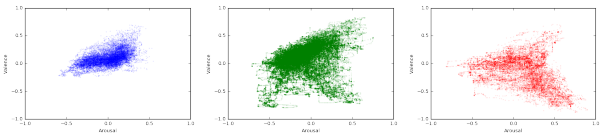

Figure 2: *Distribution of instances in arousal (x-axis) and valence (y-axis) dimensions for RECOLA, SEMAINE and CreativeIT, respectively.*

3.2. Data Processing

To create a model suitable for all the data available, we reduced annotation rate to 25 Hz and transformed labels of the training corpora using min/max scaler. Time-continuous labels were shifted backwards, correcting reaction lag of annotators [22, 23]. From raw audio recordings, 130 features (65 LLDs and their first derivatives) were extracted with openSMILE [13] using ComParE 2016 configuration [24]. Following the previous research, features were normalized using z-score transformation at a speaker level [2], then sparsed to cover 8 seconds of context with a time window size of 16 frames [25].

3.3. Experimental Setting and Results

Experiments are carried out in a purely cross-corpus setting, where we train on one, validate on another (used to optimize the number of epochs) and test on the third corpus. Further for avoiding over-fitting, in all experiments we fix the learning rate to 0.001, which is based on previous experimentation on RECOLA [25].

In the training of LSTM models, we used only five epochs using all instances of the source corpus. In order to assess the cross-corpus prediction performance across-corpora having different mean, variance and skewness statistics as observed in Figure 2, we use the Cross-Correlation (CC, also known as Pearson's Correlation) measure.

In Table 1, comparative cross-corpus performances of LSTM models are given. From the table, three patterns emerge. First, we observe that cross-corpus arousal scores are always positive, showing strong and significant correlations between predictions and the ground truth as opposed to predominantly poor results in the valence dimension. This is a common pattern as long as acoustic emotion recognition is concerned. Secondly, we see that the arousal predictions even with a single epoch training are statistically significant ($p < 10^{-50}$). The final pattern is that due to its different correlation structure between the affect primitives, results with CreativeIT corpus are either poorer (for arousal) or strongly negative (for valence) compared to the other two corpora. We proceed with cross-corpus cross-task predictions targeting the challenge corpus, keeping in mind the anomaly related to CreativeIT corpus.

4. Experiments on Ulm State-of-Mind Speech Corpus

Ulm State-of-Mind Speech Corpus (USoMS) corpus contains emotional speech utterances from 100 subjects (85f) having a mean age of 22.3 years. The participants self-reported their mood in arousal and valence dimensions before and after they told two positive and two negative narratives. The scores of valence dimension are discretized and used as a three class (low/medium/high) classification problem. It is important to note that the classes are strongly imbalanced and the baseline system employs an instance upsampling strategy to overcome this issue. For further details on the corpus and the challenge setting, the reader is referred to the paper on challenge [12].

4.1. Experiments with Baseline Approaches and Features

Since 2017 ComParE challenge [26], the baseline systems not only amount to a single type of feature but use multiple subsystems and their fusion. Moreover, the optimal test set scores are taken as baselines, instead of optimizing the parameters on development and using the corresponding parameters to retrain the combination of the training and development sets. This makes it almost impossible to outperform baseline performances without using the components of the baseline system. Thus, we first try to reproduce the baseline system and trial different classifiers on its feature sets.

The comparative results of the four approaches that are presented in the challenge paper and the corresponding results we obtained are summarized in Table 2. We should note that the development set results are indicative of the test results only when openSMILE functional features (OS_{FUN}) are used. The reproduced results are similar or better than those reported in the baseline paper in two cases (OS_{FUN} and OS_{BOAW}, which is BoAW representation of the same set of 130 LLDs) and way lower in the other two, neural network based approaches.

Concerning the systems used to generate the test set baselines, no fusion system uses the END2YOU (CNN + LSTM) approach [12]. In line with this finding, we focused on the other three approaches and modeled them with alternative, least squares regression based classifiers. Table 3 summarizes best development set UAR performances of baseline features trained with classifiers mentioned in Section 2.1. We see that for each feature type better performances could be reached with proposed classifiers, and fused features can improve UAR up to 62.11%. However, when we analyze confusion matrices of each classifier outputs, we observe that KPLS consistently favors majority classes, keeping minority class recall close to zero and as such not contributing to classifier fusion. We thus combine the scores of the remaining three classifiers, trained on combined OS_{FUN} and OS_{BOAW} feature set and obtain a development set UAR of 64.0%. This is subsequently fused with the cross-corpus LSTM predictions for test set submissions.

4.2. LSTM based Cross-corpus and Cross-task Affect Prediction

After evaluating the components of the baseline system, we proceed with the proposed cross-corpus and cross-task affect prediction based method. Here, we obtain frame-level predictions, apply smoothing of the overlapping predictions from different analysis windows and subsequently use mean functional to summarize the affective primitives over the utterance. Finally, the utterance level valence scores are converted into low/medium/high categories with thresholds optimized on the target set. Table 4 summarizes UAR scores optimized on the training set of the USoMS corpus. In general, we obtain higher than chance level UAR performance in all but one case, and the highest UAR (60.1%) on the development set is better than the best individual classifier reported in the challenge paper (56.5%). Another interesting finding is that, models trained with arousal as target variable were also found to give good performance on the development set, reaching UAR of 52.1%.

4.3. Fusion and Test Set Results

The proposed method aims to combine LSTM based cross-corpus and cross-task prediction and within-corpus affect recognition using fixed-length suprasegmental feature representation methods. To this end, we optimized models separately and combined them at the score level.

For the test set submissions, we wished to see the individual performance of cross-corpus approach. In our first submission, we used the combination of predictions trained on SEMAINE corpus with arousal and valence targets, with thresholds optimized on the training set of USoMS. This resulted in development and test set UAR scores of 62.0% and 47.4%, respectively.

Table 1: *Cross-correlation (CC) results of LSTM models in cross-corpus setting. Models are optimized on the validation corpus.*

			Arousal				Valence			
			1 epoch		Optimal Epoch		1 epoch		Optimal Epoch	
Train	Validation	Test	Val. CC	Test CC	Val. CC	Test CC	Val. CC	Test CC	Val. CC	Test CC
RECOLA	SEMAINE	CreativeIT	0.411	0.390	0.411	0.390	0.256	-0.154	0.256	-0.154
	CreativeIT	SEMAINE	0.390	0.411	0.390	0.411	-0.154	0.256	-0.142	0.241
SEMAINE	RECOLA	CreativeIT	0.473	0.375	0.520	0.375	0.201	-0.110	0.219	-0.084
	CreativeIT	RECOLA	0.375	0.473	0.408	0.514	-0.110	0.201	-0.044	0.165
CreativeIT	RECOLA	SEMAINE	0.188	0.337	0.188	0.337	-0.029	-0.059	-0.029	-0.059
	SEMAINE	RECOLA	0.337	0.188	0.346	0.048	-0.059	-0.029	-0.031	-0.063

Table 2: *Development set UAR results (%) of baseline paper's approaches with hyper-parameters optimized on the test set (Benchmark) and their corresponding reproduced results by our team (Reproduced)*

Approach	Benchmark	Reproduced
OS_{FUN} + SVM	56.50	57.20
OS_{BOAW} + SVM	52.50	55.50
AuDeep + SVM	49.90	40.20
END2YOU (CNN+LSTM)	49.70	41.40

Table 3: *Development set UAR (%) scores of baseline feature sets modeled with proposed classifiers.*

Feature	KELM	WKELM	KPLS	WKPLS
OS_{FUN}	57.55	55.95	61.22	58.59
OS_{BOAW}	53.62	47.14	56.34	52.12
AuDeep	44.57	41.69	37.50	49.10
OS_{FUN}+OS_{BOAW}	58.46	54.09	62.11	60.98

We next combined predictions of models trained on SEMAINE for valence and CreativeIT for arousal and optimized the thresholds on the development corpus. This resulted in lower test set UAR of 45.8%. In the third submission, we again used SEMAINE models' predictions but optimized the thresholds using the combined training and development corpora, obtaining a test set UAR of 47.8%. This was combined with the predictions of three within-corpus classifiers (KELM, WKELM and WKPLS), which are trained on feature level fusion of OS_{FUN} and OS_{BOAW}. This fusion scheme boosted the development set UAR to 75.9%, also improving the test set UAR to 59.3%. However, this score remained below the test baseline UAR of 66.0%.

Table 4: *Comparative performances (%) of LSTM based cross-corpus predictions with thresholds optimized on the training set.*

		USoMS Training Set		USoMS Val. Set	
Source	**Dimension**	**WAR**	**UAR**	**WAR**	**UAR**
CreativeIT	Arousal	56.1	51.0	40.0	44.4
CreativeIT	Valence	44.7	37.4	40.2	32.4
RECOLA	Arousal	39.5	48.2	41.2	41.4
RECOLA	Valence	44.9	40.9	37.2	31.7
SEMAINE	Arousal	55.6	45.8	56.2	52.1
SEMAINE	Valence	52.5	46.9	48.9	60.1

5. Discussion and Conclusions

In this work, we introduced an LSTM based method for cross-corpus, dimensional-to-categorical acoustic affect recognition and embedded it in a fusion framework with within-corpus utterance level paralinguistic processing. The results of the proposed method were found to outperform singleton within-corpus suprasegmental feature based systems on the challenge development set, however rendered a lower performance on the test set. Moreover, when the proposed systems' predictions are combined with the components of the challenge baseline system, the development set performance is observed to boost. The final system is found to give lower performance compared to the challenge test set baseline, which may be attributed to both the mismatch of class distributions/ways of labeling the continuous corpora and the fact that the challenge baseline is somewhat over-optimized.

An important issue with three out of four baseline approaches (i.e. all but openSMILE functionals based) is that due to their inherent stochasticity it is impossible to reproduce the baseline features/results exactly. This is the case with not only neural networks, such as END2YOU and auDeep approaches, but also with BoAW representation as the LLDs are sampled randomly and cluster centroids of K-Means are also initialized randomly. A possible solution to this would be to share the BoAW and auDeep features along with the baseline release or to prefix a seed for random number generation.

An interesting result observed in cross-corpus analyses is that although the target task was three-class valence recognition, models trained on arousal dimension rendered cross-corpus UAR performances that are significantly higher than chance level. Moreover, models trained on RECOLA and CreativeIT with arousal had higher UAR on USoMS compared to those trained with valence. This may be attributed to generally positive correlations of these affect primitives as well as the fact that arousal can be more accurately modeled from speech acoustics. In both cases, it opens new research avenues to use combination of predicted arousal and valence as mid-level features for cross-corpus tasks targeting valence related classes.

6. Acknowledgements

The participation in the ComParE 2018 challenge with experiments on USoMS corpus (Section 4) was supported exclusively by the Russian Science Foundation (Project No. 18-11-00145). The rest research was supported by the Huawei Innovation Research Program (Agreement No. HO2017050001BM).

7. References

[1] B. Schuller, B. Vlasenko, F. Eyben, M. Wollmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-corpus acoustic emotion recognition: variances and strategies," *Affective Computing, IEEE Transactions on*, vol. 1, no. 2, pp. 119–131, 2010.

[2] H. Kaya and A. A. Karpov, "Efficient and effective strategies for cross-corpus acoustic emotion recognition," *Neurocomputing*, vol. 275, pp. 1028–1034, 2018.

[3] A. Hassan, R. Damper, and M. Niranjan, "On acoustic emotion recognition: compensating for covariate shift," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1458–1468, 2013.

[4] H. Sagha, J. Deng, M. Gavryukova, J. Han, and B. Schuller, "Cross lingual speech emotion recognition using canonical correlation analysis on principal component subspace," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5800–5804.

[5] J. Deng, Z. Zhang, F. Eyben, and B. Schuller, "Autoencoder-based unsupervised domain adaptation for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1068–1072, 2014.

[6] J. Deng, X. Xu, Z. Zhang, S. Frühholz, and B. Schuller, "Semisupervised autoencoders for speech emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 31–43, 2018.

[7] A. Tawari and M. M. Trivedi, "Speech emotion analysis: Exploring the role of context," *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 502–509, 2010.

[8] M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling," in *Proc. INTERSPEECH 2010, Makuhari, Japan*, 2010, pp. 2362–2365.

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

[10] H. Kaya and A. A. Karpov, "Fusing acoustic feature representations for computational paralinguistics tasks," in *INTERSPEECH*, San Francisco, USA, Proceedings, 2016, pp. 2046–2050.

[11] ——, "Introducing weighted kernel classifiers for handling imbalanced paralinguistic corpora: Snoring, addressee and cold," in *INTERSPEECH*, Stockholm, Sweden, Proceedings, 2017, pp. 3527–3531.

[12] B. Schuller, S. Steidl, A. Batliner, P. Marschik, H. Baumeister, F. Dong, F. B. Pokorny, E.-M. Rathner, K. D. Bartl-Pokorny, C. Einspieler, D. Zhang, A. Baird, S. Amiriparian, K. Qian, Z. Ren, M. Schmitt, P. Tzirakis, and S. Zafeiriou, "The INTERSPEECH 2018 computational paralinguistics challenge: Atypical & self-assessed affect, crying & heart beats," in *INTERSPEECH*. Hyderabad, India, Proceedings: ISCA, 2018.

[13] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent Developments in openSMILE, the Munich open-source Multimedia Feature Extractor," in *Proceedings of the 21st ACM International Conference on Multimedia*. ACM, 2013, pp. 835–838.

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

[15] M. Freitag, S. Amiriparian, S. Pugachevskiy, N. Cummins, and B. Schuller, "audeep: Unsupervised learning of representations from audio with deep recurrent neural networks," *arXiv preprint arXiv:1712.04382*, 2017.

[16] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 2013, pp. 1–8.

[17] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 5–17, 2012.

[18] A. Metallinou, C.-C. Lee, C. Busso, S. Carnicke, and S. Narayanan, "The USC CreativeIT database: A multimodal database of theatrical improvisation," *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*, p. 55, 2010.

[19] H. Kaya, A. A. Karpov, and A. A. Salah, "Fisher vectors with cascaded normalization for paralinguistic analysis," in *INTERSPEECH*, Dresden, Germany, Proceedings, 2015, pp. 909–913.

[20] H. Kaya, F. Gürpınar, and A. A. Salah, "Video-based emotion recognition in the wild using deep transfer learning and score fusion," *Image and Vision Computing*, vol. 65, pp. 66–75, 2017.

[21] W. Zong, G.-B. Huang, and Y. Chen, "Weighted extreme learning machine for imbalance learning," *Neurocomputing*, vol. 101, pp. 229 – 242, 2013.

[22] S. Mariooryad and C. Busso, "Correcting time-continuous emotional labels by modeling the reaction lag of evaluators," *IEEE Transactions on Affective Computing*, vol. 6, no. 2, pp. 97–108, 2015.

[23] A. Mencattini, E. Martinelli, F. Ringeval, B. Schuller, and C. Di Natale, "Continuous estimation of emotions in speech by dynamic cooperative speaker models," *IEEE Transactions on Affective Computing*, vol. 8, no. 3, pp. 314–327, 2017.

[24] B. W. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. C. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The INTERSPEECH 2016 computational paralinguistics challenge: Deception, sincerity & native language," in *INTERSPEECH, Proceedings*, 2016, pp. 2001–2005.

[25] D. Fedotov, D. Ivanko, M. Sidorov, and W. Minker, "Contextual dependencies in time-continuous multidimensional affect recognition," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018, pp. 1220–1224.

[26] B. Schuller, S. Steidl, A. Batliner, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom, A. Warlaumont, G. Hidalgo, S. Schneider, C. Heiser, W. Hohenhorst, M. Herzog, M. Schmitt, K. Qian, Y. Zhang, G. Trigeorgis, P. Tzirakis, and S. Zafeiriou, "The INTERSPEECH 2017 Computational Paralinguistics Challenge: Addressee, Cold & Snoring," in *INTERSPEECH*, Stockholm, Sweden, Proceedings, 2017, pp. 3442–3446.