Context Modeling for Cross-Corpus Dimensional Acoustic Emotion Recognition: Challenges and Mixup

Dmitrii Fedotov[1(✉)], Heysem Kaya[2], and Alexey Karpov[3]

[1] Institute of Communications Engineering, Ulm University, Ulm, Germany
dmitrii.fedotov@uni-ulm.de
[2] Department of Computer Engineering,
Tekirdağ Namık Kemal University, Çorlu, Turkey
kaya.heysem@gmail.com
[3] St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, St. Petersburg, Russia
karpov@iias.spb.su

Abstract. Recently, focus of research in the field of affective computing was shifted to spontaneous interactions and time-continuous annotations. Such data enlarge the possibility for real-world emotion recognition in the wild, but also introduce new challenges. Affective computing is a research area, where data collection is not a trivial and cheap task; therefore it would be rational to use all the data available. However, due to the subjective nature of emotions, differences in cultural and linguistic features as well as environmental conditions, combining affective speech data is not a straightforward process. In this paper, we analyze difficulties of automatic emotion recognition in time-continuous, dimensional scenario using data from RECOLA, SEMAINE and CreativeIT databases. We propose to employ a simple but effective strategy called "mixup" to overcome the gap in feature-target and target-target covariance structures across corpora. We showcase the performance of our system in three different cross-corpus experimental setups: single-corpus training, two-corpora training and training on augmented (mixed up) data. Findings show that the prediction behavior of trained models heavily depends on the covariance structure of the training corpus, and mixup is very effective in improving cross-corpus acoustic emotion recognition performance of context dependent LSTM models.

Keywords: Cross-corpus emotion recognition
Time-continuous emotion recognition · Data augmentation

1 Introduction

Automatic affect recognition is a popular research topic, which brings researchers from psychological and technical areas together [19,24]. It can be beneficial in

A. Karpov et al. (Eds.): SPECOM 2018, LNAI 11096, pp. 155–165, 2018.
https://doi.org/10.1007/978-3-319-99579-3_17

a variety of applications in areas of human-computer interaction (HCI) and human-human interaction (HHI). Emotional component in an HCI system allows it to perceive the emotional state of speaker and adjust the response to increase the quality of interaction.

Although emotion recognition has been a hot topic for a long period and a high amount of research was conducted, the problem is far from being solved. Less than two decades ago, emotion recognition has left the laboratory conditions and faced the real-world data and problems; such as cultural, linguistic and environmental differences [10,22]. Combination of different corpora, which could solve the problem of data shortage, could not be applied in a straightforward manner in the context of acoustic emotion recognition. The main difficulty lies in the subjective nature of emotions, resulting in diverse and controversial annotations. Despite these issues, data combination and augmentation may lead to a dramatic increase in performance of affect recognition systems.

In this paper, we dealt with problems of cross-corpus time-continuous dimensional emotion recognition and proposed ways to overcome them. We observed that a pure cross-corpus emotion recognition may not work properly if data have different label distributions. We also showed that this problem can be partially solved by combining and augmenting data.

This paper is structured as follows: we introduce the related work in Sect. 2; provide information on corpora used, data preprocessing techniques and methodology in Sect. 3; present results of different cross-corpora emotion recognition settings in Sect. 4; and conclude the paper in Sect. 5.

2 Related Work

Most of the previous research on emotion recognition dealt with acted, categorically labeled corpora, providing information at utterance-level [1,7,11].

Continuously annotated databases of spontaneous interactions provide more naturalistic data, but also introduce several challenges, such as diversity in annotations [16,17], reaction lags between actual appearance of an emotion and its annotation [12] and amount of contextual information the system needs [5,6].

Problem of cross-corpus emotion recognition was investigated by several research groups. Schuller et al. studied this problem with acted, categorically annotated databases [22]. Performance of the proposed methodology was poor if some differences in environmental conditions were present. For some of the emotions, classification accuracy of used Support Vector Machine (SVM) based model was below the chance level. Authors also showed that normalization strategy has a crucial role in the cross-corpus scenario and concluded that speaker-level normalization leads to the best performance, compared to other approaches.

The study of normalization effect on cross-corpus emotion recognition performance was extended and cascaded normalization techniques, which are comprised of speaker, value and instance level normalization, were recently introduced and tested in [9]. The proposed approach achieved increased performance reducing cross-corpus differences with respect to suprasegmental acoustic features.

Resent study focused on cross-corpus recognition of self-assessed affect. Cross-corpus predictions of affective primitives were used as a data for extracting functionals and then combined with predictions of other sub-systems to improve performance [8].

These studies provided a starting point for the paper in-hand and a speaker normalization technique was used. Cross-corpus emotion recognition with time-continuous data is poorly studied, which served as motivation to conduct our research.

3 Data and Methodology

Three corpora of spontaneous, emotionally-rich interactions are used in this study: RECOLA [20], SEMAINE [13] and CreativeIT [14]. All corpora are annotated at frame level using two affective scales: arousal (activation) and valence (positivity). A brief overview of the used corpora is presented in Table 1.

Table 1. Overview of used corpora.

Corpus	Duration (min)	Recordings	Participants	Gender (m/f)	Age μ (σ)	Annotation rate (Hz)
RECOLA	115	23	23	10/13	21.4 (2.0)	25
SEMAINE	435	24	20	8/12	30.4 (10.4)	50
CreativeIT	132	31	15	7/8	N/A	60

3.1 RECOLA

RECOLA (Remote COLlaborative and Affective interactions) database was collected during spontaneous dyadic interactions between people while solving a cooperative problem. From 46 people participating in the database collection, 34 gave their consent to share the data publicly available and recordings from 23 users are presented in the current version of the database, shared with research community. Each recording has duration of five minutes, yielding 115 min of speech in total.

Participants are aged between 18 and 25 years and have different mother tongues although spoke French during the database collection process: 17 of them have French as a mother tongue, 3 – Italian and 3 – German. The corpus was recorded in four modalities: audio, video, electrocardiogram and electro-dermal activity. Recordings were continuously annotated by 6 equally gender-distributed persons via *ANNEMO* (ANNotating EMOtions) annotation tool [20] in two affective scales (arousal and valence) and five social behavior scales (agreement, dominance, engagement, performance, rapport).

3.2 SEMAINE

SEMAINE (Sustained Emotionally coloured Machine-human Interaction using Nonverbal Expression) database was collected within a project, where the aim was to build a system that could engage a person in a sustained conversation with a Sensitive Artificial Listener (SAL) agent. Three scenarios are used in the context of this project: Solid SAL, where the agent's role was played by a real human-operator; Semi-Automatic SAL, where system spoke phrases chosen by a human operator from a pre-defined list; and Automatic SAL, where the system chose phrases and non-verbal signals by itself. Only data collected from users (not operators) in Solid SAL scenario were used in this study.

The corpus consists of 24 recordings in English from 20 speakers, whose age range from 20 to 58 years. Recordings have durations from 11 to 30 min resulting in the total corpus length of 435 min. The corpus was recorded in two modalities: audio and video and annotated via *FeelTrace* annotation tool [3] in different dimensions and emotional labels: valence, arousal, power, anticipation, intensity, fear, anger, happiness, sadness, disgust, contempt and amusement.

3.3 CreativeIT

CreativeIT database was collected to serve as a multidisciplinary resource for theatrical performance improvement and emotion recognition. It was recorded by actors, coordinated by a director with an expert qualification in Active Analysis introduced by Stanislavsky. Two scenarios were used during the database collection: two-sentence exercise, where actors were permitted to use only one predefined phrase each; and paraphrase of script, where actors were following general script without any constraints on words and expressions. Only the paraphrase part of corpus was used in this study as it meets the conditions of spontaneous interaction closely.

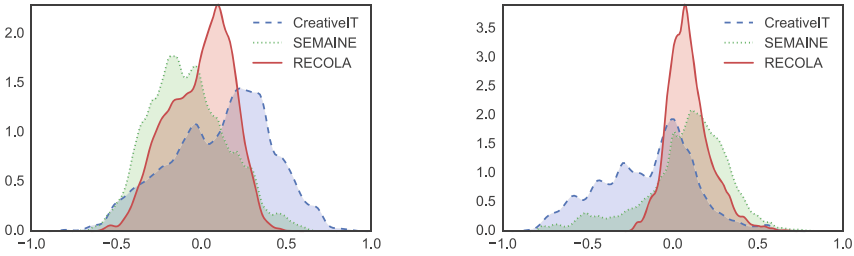
Selected part of the corpus consists of 31 recordings in English from 15 participants. Duration of recordings ranges from 2 to 7 min, with a total of 132 min. In addition to audio data from close-up microphones, motion capture data is available for each recording, representing body language of actors during interactions. Recordings were annotated via *FeelTrace* annotation tool [3] by three groups of evaluators: theater experts, actors and naive audience in different dimensional groups, such as emotional descriptors (arousal, valence) and theatrical performance ratings (naturalness, creativity).

3.4 Features and Labels

For cross-corpus emotion recognition, the audio modality was used in this study, as it is presented in each corpus described above. Audio features were extracted with *openSMILE* tool [4]. They consist of 65 low-level descriptors (LLDs) and their first order derivatives [21]. Feature step size was set to 0.01 sec. resulting in a feature extraction rate of 100 Hz. As corpora have different annotation rates (see Table 1) they were brought to the same data frequency to be able to share

the same prediction models. The lowest annotation frequency of 25 Hz, present in RECOLA, was used to subsample other two corpora.

Extracted features were speaker-level z-normalized, as it was previously shown to have a better performance in cross-corpus experiments [9]. Annotations of two main affective dimensions - arousal and valence - were used in this study as labels. Distributions of labels for corpora described above are presented in Fig. 1.



(a) Arousal **(b)** Valence

Fig. 1. Label distributions in three emotional corpora.

The label distribution of RECOLA is narrower in both affective dimensions, than remaining corpora. It can be a result of its pure spontaneous nature. Although all corpora used in this study are designed to be naturalistic, SEMAINE can simulate four personality prototypes, which affect operators' behavior and hence, the user. Even though actors participating in collection of CreativeIT database were not restricted lexically to choose the words for interaction, they had to follow the general scenario and the role. These conditions could have led to more idiosyncratic nature of emotions in both SEMAINE and CreativeIT.

3.5 Modeling

In this study, recurrent neural network with long short-term memory (LSTM-RNN) was used for context modeling. The model comprises of two layers with 80 and 60 neurons with ReLU activation function [15], respectively, each followed a dropout layer with $p = 0.3$ [23]. The models were optimized by root mean square propagation (RMSprop) using the concordance correlation coefficient as a metric function. We use the LSTM implementation provided by Keras toolkit [2].

Our recent study has revealed that performance of time-continuous emotion recognition has a strong relation with the amount of acoustic context used in recurrent neural network (RNN) models regardless of the number of time steps [5]. The required amount of context could be set by combination of two parameters: number of time steps fed into RNN model and a sparsing coefficient, which is responsible for decreasing the amount of data in each sample by skipping

frames. Regardless of sparsing coefficient, the step size between samples is one frame, hence there is no loss in total amount of information. The amount of context in seconds is then represented as:

$$C = \frac{SC \times TW}{FR}, \tag{1}$$

where SC is the sparsing coefficient that determines the amount of frames to skip, TW is the time window size and FR is the frame rate in Hz.

Based on our previous research [5], a context size of 7.68 s, which is obtained from the combination of *SC=12* and *TW=16*, was selected for this study. The same procedure of sparsing applies to respective labels. Sequence-to-sequence modeling is used in this study, thus features of TW previous frames were used to predict the corresponding labels for these frames. After prediction phase, the values of labels obtained for the same frame at different time steps were averaged to smooth final prediction.

3.6 Mixup for Data Augmentation and Corpus Adaptation

To combine data from different corpora, a recently introduced methodology called *mixup* was used in this study [25]. *mixup* is a data augmentation technique, that constructs virtual training examples based on existing ones, using weights drawn from Beta distribution to regulate their contribution to the synthetic instance:

$$x_{new} = \lambda x_i + (1 - \lambda) x_j, \tag{2}$$

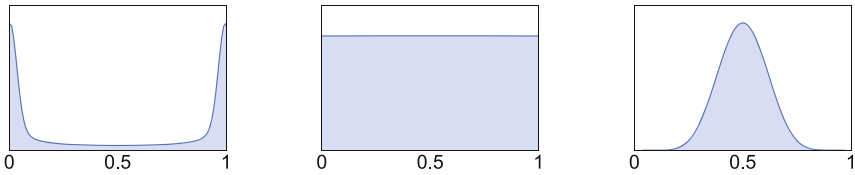
$$y_{new} = \lambda y_i + (1 - \lambda) y_j, \tag{3}$$

where $\lambda \sim \text{Beta}(\alpha, \alpha)$, α is a hyper-parameter for the Beta distribution, x_i, x_j are feature vectors, and y_i, y_j are label values/vectors. This kind of data augmentation encourages the model to behave more linearly in-between training examples, which can be useful for cross-corpus learning.

In this study, feature vectors x_i, x_j and corresponding labels y_i, y_j were taken from two different corpora. To create different sets of augmented data, hyper-parameter α of mixup routine was varied (see Fig. 2). Three values were tested: $\alpha = 0.1$, which provides slight changes to original data and a minor contribution of the second corpus; $\alpha = 1$, which provides an uniformly distributed level of contribution of both corpora to augmented data; and $\alpha = 10$, which creates most examples in the middle of feature-label space between two samples. To preserve sequential nature of data, streams were mixed up at the recording level with consecutive frames.

4 Experimental Results

In this paper, the problem of cross-corpus multi-dimensional emotion recognition is considered. To study the issues and particularities of time-continuous



(a) $\alpha = 0.1$ **(b)** $\alpha = 1$ **(c)** $\alpha = 10$

Fig. 2. Beta-distribution with three different values of parameter α.

and multidimensional emotion recognition, three experimental setups were used: single-corpus training, two-corpora training and training on augmented data. The performance of cross-corpus prediction was estimated using Pearson's correlation coefficient (ρ).

4.1 Single-Corpus Training

The first problem definition was to predict values on an unseen corpus using model trained on single corpus. Two models (for arousal and valence) were trained on whole data available for one corpus up to 5 epochs, then they were used to generate predictions for different corpora, including the training corpus itself (to show the ground-truth label distributions). Scatter plots of prediction in the single-corpus training settings are presented in Fig. 3.

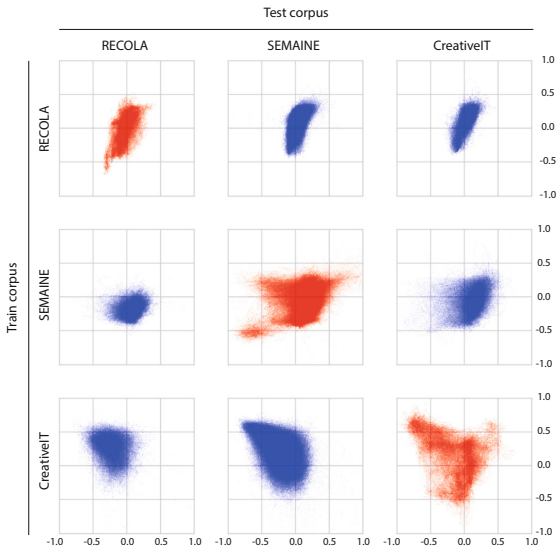


Fig. 3. Single-corpus training (x-axis – valence, y-axis – arousal).

Table 2. Pearson correlation scores (arousal/valence) for single-corpus training.

Train on	Test on		
	RECOLA	SEMAINE	CreativeIT
RECOLA	0.923/0.890	0.375/0.223	0.337/−0.024
SEMAINE	0.533/0.170	0.821/0.750	0.322/−0.065
CreativeIT	−0.027/0.009	0.306/−0.013	0.953/0.952

Distributions of each corpus labels can be seen as self-prediction (main diagonal). Figure 3 shows that models predict only in the limits of their own annotation distributions and exhibit the same tendencies regardless of the test data. This results in a low cross-corpus prediction performance, even in some cases leading to a negative correlation (see Table 2).

Negative correlations may also be attributed to the use of different annotation tools. ANNEMO tool has two separate bars for arousal and valence that are manipulated by the user independently. However, FeelTrace toolkit provides the two-dimensional emotion representation with basic emotions played on the graph, which are in some cases drastically converse (e.g. for "afraid") to other research [18].

4.2 Multi-corpus Training

The second research problem was to predict affect primitives on an unseen corpus using the model trained on two remaining corpora. Other experimental parameters were left the same as in the single-corpus training setting. We refer this multi-corpus training scheme as "combining".

The third research problem was to predict arousal and valence on one of the corpora using the model trained on fully synthetic data, generated from the remaining corpora with mixup routine. Comparative multi-corpus training results with combining and mixup strategies are presented in Table 3, where the improved performance of multi-corpus training over the best single corpus training performance on a target corpus is shown in **bold**.

Table 3. Pearson correlation scores (arousal/valence) for leave-one-corpus-out training results.

Train on	Test on	Combined	Mixed up (best α)
RECOLA + CreativeIT	SEMAINE	0.359/0.050	0.368 (1)/−0.012 (10)
RECOLA + SEMAINE	CreativeIT	**0.435**/−0.016	**0.431** (1)/−0.041 (0.1)
CreativeIT + SEMAINE	RECOLA	0.222/0.149	**0.695** (1)/**0.294** (10)

Compared to the single-corpus training, combination of data results in approximate averaging of performances of two corpora used for training. Only a

combination of SEMAINE and RECOLA provides better results for CreativeIT as the test corpus with arousal dimension. Mixup based data augmentation allows model to benefit more from differences in databases, creating synthetic samples that train a model having higher generalization ability. Thus, mixup dramatically improves over single-corpus training on two corpora, and renders a relatively slight performance decrease (from 0.375 to 0.368) on SEMAINE arousal dimension. The advantage of using mixup over simple combination is seen clearly on RECOLA corpus: while combining approach markedly underperforms the single-corpus performance, mixup improves it in both arousal and valence dimensions.

5 Conclusions and Future Work

In this paper, we studied problems of time-continuous multidimensional cross-corpus emotion recognition. In addition to the feature distribution problem that is present in other cross-corpus settings and could be partially solved by a speaker-level normalization, the dimensional approach introduces the challenge of different label distributions. It can be caused by initial database collection scenario, different annotation software or people's perception of emotions. Nevertheless, it may serve as a limiting factor for the system, may not let it predict outside originally trained distribution and may even result in converse behavior.

In future work, a cross-task approach will be introduced to the current research to increase coverage of arousal-valence space by using corpora with categorical annotation. The question of mapping emotion labels between corpora is still poorly studied, but an effective approach may increase amount of data available for different experimental settings, which will have a positive effect on the performance of the emotion recognition system.

Acknowledgments. This research is supported by the Russian Science Foundation (project No. 18-11-00145).

References

1. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F., Weiss, B.: A database of German emotional speech. In: Ninth European Conference on Speech Communication and Technology (2005)
2. Chollet, F., et al.: Keras (2015). https://keras.io
3. Cowie, R., Douglas-Cowie, E., Savvidou*, S., McMahon, E., Sawey, M., Schröder, M.: 'FEELTRACE': An instrument for recording perceived emotion in real time. In: ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion (2000)
4. Eyben, F., Wöllmer, M., Schuller, B.: Opensmile: the Munich versatile and fast open-source audio feature extractor. In: Proceedings of the 18th ACM International Conference on Multimedia, pp. 1459–1462. ACM (2010)
5. Fedotov, D., Ivanko, D., Sidorov, M., Minker, W.: Contextual dependencies in time-continuous multidimensional affect recognition. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC) (2018)

6. Gunes, H., Pantic, M.: Automatic, dimensional and continuous emotion recognition. Int. J. Synth. Emotions **1**(1), 68–99 (2010)
7. Haq, S., Jackson, P.J.: Multimodal emotion recognition. Machine audition: principles, algorithms and systems, pp. 398–423 (2010)
8. Kaya, H., Fedotov, D., Yeşilkanat, A., Verkholyak, O., Zhang, Y., Karpov, A.: LSTM based cross-corpus and cross-task acoustic emotion recognition. In: INTERSPEECH 2018. ISCA (2018, in press)
9. Kaya, H., Karpov, A.A.: Efficient and effective strategies for cross-corpus acoustic emotion recognition. Neurocomputing **275**, 1028–1034 (2018)
10. Lim, N.: Cultural differences in emotion: differences in emotional arousal level between the east and the west. Integr. Med. Res. **5**(2), 105–109 (2016)
11. Makarova, V., Petrushin, V.A.: RUSLANA: A database of Russian emotional utterances. In: Seventh International Conference on Spoken Language Processing (2002)
12. Mariooryad, S., Busso, C.: Analysis and compensation of the reaction lag of evaluators in continuous emotional annotations. In: Affective Computing and Intelligent Interaction (ACII), pp. 85–90. IEEE (2013)
13. McKeown, G., Valstar, M., Cowie, R., Pantic, M., Schroder, M.: The SEMAINE database: annotated multimodal records of emotionally colored conversations between a person and a limited agent. IEEE Trans. Affect. Comput. **3**(1), 5–17 (2012)
14. Metallinou, A., Lee, C.C., Busso, C., Carnicke, S., Narayanan, S.: The USC CreativeIT database: a multimodal database of theatrical improvisation. In: Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality, p. 55 (2010)
15. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning (ICML 2010), pp. 807–814 (2010)
16. Nicolaou, M.A., Gunes, H., Pantic, M.: Automatic segmentation of spontaneous data using dimensional labels from multiple coders. In: Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality. German Research Center for AI (DFKI) (2010)
17. Nicolle, J., Rapp, V., Bailly, K., Prevost, L., Chetouani, M.: Robust continuous prediction of human emotions using multiscale dynamic cues. In: Proceedings of the ACM International Conference on Multimodal Interaction, pp. 501–508 (2012)
18. Paltoglou, G., Thelwall, M.: Seeing stars of valence and arousal in blog posts. IEEE Trans. Affect. Comput. **4**(1), 116–123 (2013)
19. Petta, P., Pelachaud, C., Cowie, R.: Emotion-Oriented Systems: The HUMAINE Handbook. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-15184-2
20. Ringeval, F., Sonderegger, A., Sauer, J., Lalanne, D.: Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), pp. 1–8. IEEE (2013)
21. Schuller, B., Steidl, S., Batliner, A., Epps, J., Eyben, F., Ringeval, F., Marchi, E., Zhang, Y.: The INTERSPEECH 2014 computational paralinguistics challenge: cognitive & physical load. In: Fifteenth Annual Conference of the International Speech Communication Association (2014)
22. Schuller, B., Vlasenko, B., Eyben, F., Wollmer, M., Stuhlsatz, A., Wendemuth, A., Rigoll, G.: Cross-corpus acoustic emotion recognition: variances and strategies. IEEE Trans. Affect. Comput. **1**(2), 119–131 (2010)

23. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)
24. Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres Torres, M., Scherer, S., Stratou, G., Cowie, R., Pantic, M.: Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, pp. 3–10. ACM (2016)
25. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)