Exploring perception uncertainty for emotion recognition in dyadic conversation and music listening

Jing Han, Zixing Zhang, Zhao Ren, Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Han, Jing, Zixing Zhang, Zhao Ren, and Björn Schuller. 2020. "Exploring perception uncertainty for emotion recognition in dyadic conversation and music listening." *Cognitive Computation* 13: 231–40. https://doi.org/10.1007/s12559-019-09694-4.

Exploring Perception Uncertainty for Emotion Recognition in Dyadic Conversation and Music Listening

Jing Han[1] · Zixing Zhang[2] · Zhao Ren[1] · Björn Schuller[1,2]

Abstract
Predicting emotions automatically is an active field of research in affective computing. Considering the property of the individual's subjectivity, the label of an emotional instance is usually created based on opinions from multiple annotators. That is, the labelled instance is often accompanied with the corresponding inter-rater disagreement information, which we call here the perception uncertainty. Such uncertainty information, as shown in previous studies, can provide supplementary information for better recognition performance in such a subjective task. In this paper, we propose a multi-task learning framework to leverage the knowledge of perception uncertainty to ameliorate the prediction performance. In particular, in our novel framework, the perception uncertainty is exploited in an explicit manner to manipulate an initial prediction dynamically, in contrast to merely estimating the emotional state and perception uncertainty simultaneously, as done in a conventional multi-task learning framework. To evaluate the feasibility and effectiveness of the proposed method, we perform extensive experiments for time- and value-continuous emotion predictions in audiovisual conversation and music listening scenarios. Compared with other state-of-the-art approaches, our approach yields remarkable performance improvements in both datasets. The obtained results indicate that integrating the perception uncertainty information can enhance the learning process.

Introduction

Automatic affect recognition is a multidisciplinary research field, spanning anthropology, cognitive science, linguistics, psychology, and computer science [4, 6, 34, 40, 47]. In particular, in order to incorporate cognitive capabilities into machines, detecting and understanding emotional states of humans in interactions is of broad interest in both academic and commercial communities [37, 50]. Additionally, estimating emotions in music automatically can lead to better user experiences, in a variety of music-related tasks, such as contextual music recommendation [25], emotion-based playlist generation [46], and music therapy [29].

Jing Han
jing.han@informatik.uni-augsburg.de

[1] ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, 86159 Augsburg, Germany

[2] GLAM – Group on Language, Audio & Music, Imperial College London, London SW7 2AZ, UK

In this respect, significant efforts have been made to investigate innovative technologies to facilitate various real-world applications to handle affective information. For this purpose, a variety of modalities have been studied, including but not limited to facial expressions [50], hand gestures [41], speech [56], text [1], and physiological signals such as electrocardiogram (ECG) [26] and electroencephalogram (EEG) [32].

Moreover, with the recent advent of deep learning techniques, we have witnessed fruitful theoretical and empirical works, which enable machines to recognise meaningful patterns of emotions [18, 31, 42, 55]. For instance, the first attempt to apply Long Short-Term Memory Recurrent Neural Networks (LSTM-RNNs) for long-range context modelling in Speech Emotion Recognition (SER) was by Wöllmer et al. [54]. More recently, the combination of CNNs and LSTM-RNNs, firstly constructed in [51], has been shown to be a promising approach for estimating dimensional emotions in an end-to-end manner. More recently, memory- or attention-enhanced RNNs were proposed and have shown to be efficient when modelling conventional emotions [21, 36].

Despite the great progress that has been made so far in the development of automatic emotion recognition approaches, a number of challenges remain to reach full applicability of current human–machine interaction systems. One among these is how to tailor existing algorithms to address a number of special requirements arising from the task at hand, i.e. emotion recognition. Up to now, most of these techniques have been originally proposed for other tasks, and then certainly shown beneficial throughout many applications. However, it is arguably inadequate to expect the same benefit arising in emotion recognition, by applying these frameworks directly without considering the property of the task. For this reason, these algorithms and structures have to be adjusted accordingly to fulfil the needs in several aspects such as effectiveness and efficiency, as well as reliability and robustness, in the context of emotion recognition.

To this end, aiming at incorporating the subjective property of the task at hand, several research have been carried out by leveraging the perception uncertainty retrieved on multiple annotations of the same sample. The foundation of this idea may stem from findings in [12] and [10] where it has been demonstrated that the emotion prediction systems perform better in regions with lower uncertainty. Therefore, a variety of approaches have been investigated in emotion recognition, such as estimating emotions together with the perception uncertainty [20], dynamically optimising the learning sequence accordingly [16], and exploiting the perception uncertainty information as a difficulty indicator to promote the modelling [59].

Motivated by the success of these research, in this paper, we propose a novel emotion recognition framework in which the ultimate emotion prediction is delivered, by regulating an initial emotion prediction with its corresponding uncertainty degree dynamically. Specifically, we employ a multi-task learning framework together with a dynamic tuning operation. On the one hand, the emotion prediction uncertainty will be estimated together as the conventional emotional state. On the other hand, this uncertainty will be exploited to adjust the emotional state predictions during inference.

In the remainder of this paper, we first briefly introduce the related work in the section "Related Work". Then, in the section "Perception Uncertainty Exploration", we detail the novel perception uncertainty exploration framework. Afterwards, extensive experiments on two databases for emotion recognition in dyadic conversation and music listening are carried out in the section "Experiments". Finally, our conclusions and future research directions are provided in the section "Conclusion".

Related Work

When considering the subjective property of the task of interest, i.e. emotion recognition, the inter-rater disagreement level among multiple annotators is found to be helpful [39, 57]. Take curriculum learning as one example, during the training process, instances with a lower disagreement degree can be learnt firstly. In such a context, better systems have been obtained in the affective computing literature [16, 33]. Likewise, results from [58] indicate that eliminating instances with a high disagreement level during training leads to improved performance for speech emotion recognition.

Beyond sorting or discarding samples accordingly, approaches have been proposed to characterise the subjective property of emotion perception via the inter-rater disagreement level [15, 20, 28]. On one hand, it can be deemed as an auxiliary task, aiming at improving the performance of emotion recognition by providing complimentary information [15, 28]; on the other hand, the work in [20] distilled it as an auxiliary descriptor, namely perception uncertainty (PU), to deliver a 'soft' emotion prediction jointly with the conventional emotional state prediction. In addition, recently, the success achieved by the soft labelling approach encourages other appealing works in the literature [2, 8, 9, 27].

Motivated by these research, we propose a framework to further advance the emotion modelling by leveraging the perception uncertainty dynamically via multi-task learning. Although improved performances were obtained by optimising the learning process with the PU information, none of the aforementioned works exploit the PU knowledge during inference. In contrast, in this contribution, the PU is exploited not only to facilitate the emotion modelling process but also to adjust the predictions accordingly during inference.

Moreover, this work is also relevant to the dynamic difficulty awareness training approach [59], in which the estimated PUs are exploited as additional inputs to present the learning difficulty information explicitly for emotion recognition. In [59], it has been attempted to modify the original predictions by linearly taking the uncertainties into account to deliver the final predictions in a late fusion. In contrast, we consider the PU as a weighted index to regulate the emotion estimation throughout the predicting process.

Perception Uncertainty Exploration

In this section, we will detail a novel approach, the Perception Uncertainty Exploration (PUE), approach step by step.

First, we outline the proposed framework in the section "System Overview". Afterwards, we demonstrate how to quantify the perception uncertainty via inter-rater disagreement degree in the section "Perception Uncertainty Modelling". Then, a brief discussion of a vanilla multi-task learning structure will be given in the section "Multi-Task Learning", followed by a description of the proposed method via dynamic tuning in detail in the section "Dynamic Tuning".

System Overview

An overview of the PUE framework is illustrated in Fig. 1. In particular, the structure consists of three different layers, i.e. the *shared hidden layers* to learn embeddings from input features, the *task-specific layers* to estimate initial emotion predictions together with their perception uncertainty, and a *dynamic tuning operation* to provide the final emotion prediction which is regulated by the uncertainty information.

Mathematically, given an M-dimensional input feature $\mathbf{x} \in \mathbb{R}^M$, the shared layers are learned to generate a corresponding N-dimensional embedding $\mathbf{e} \in \mathbb{R}^N$, i.e. the output from the last shared hidden layer in Fig. 1. Therefore, the effect of these layers can be represented as a mapping function $\Phi : \mathbb{R}^M \rightarrow \mathbb{R}^N$, and $\mathbf{x} \mapsto \mathbf{e}$. After that, two disconnected task-specific layers are constructed to carry out two tasks separately. In our case, the two selected tasks are the *initial emotion recognition* task and the *perception uncertainty prediction* task. These two tasks can be trained jointly under a multi-task learning strategy. Furthermore, to efficiently aggregate the uncertainty information, a dynamic tuning operation is applied to produce the final emotion predictions.

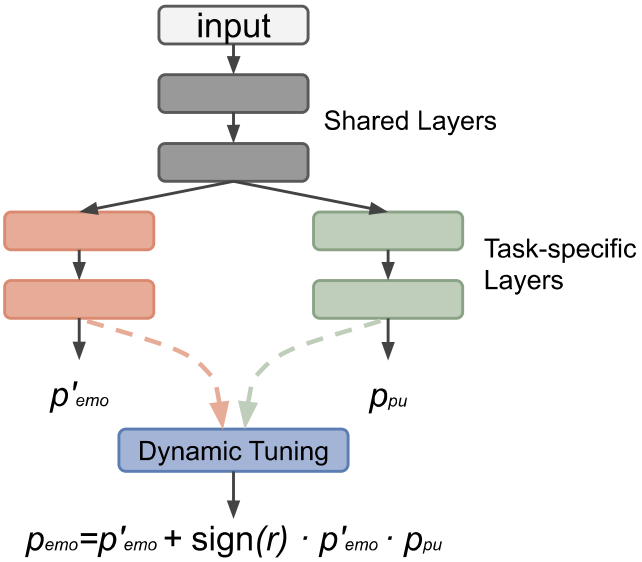

Fig. 1 Framework of the Perception Uncertainty Exploration framework for continuous emotion recognition

Perception Uncertainty Modelling

Before presenting the PUE approach, the definition of PU will be given firstly in the following paragraphs.

In this paper, PU is defined as an indicator of the uncertainty level of the perception of an emotional state for a given observed sample. As mentioned in the section "Introduction", emotion prediction is a subjective task that differs from many other objective pattern recognition tasks, such as object detection [22] and speaker identification [43], where there is a ground truth. In contrast, to obtain a gold standard for a subjective task like emotion recognition, it is common that a number of raters are required to annotate the same sample to minimise the individual bias in perception and rating as much as possible. In this context, PU can be inferred by calculating the inter-rater disagreement level, with an assumption that for each sample, PU is highly correlated with the inter-rater disagreement level [20, 38].

For this reason, given an emotional instance, its corresponding PU p_{pu} can be represented by the standard deviation of a total of n annotations (same with the definition in [59]) as:

$$p_{pu} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(p_{emo,i} - \bar{p}_{emo})^2}, \tag{1}$$

where $p_{emo,i}$ is the ith annotation of the instance with $i = 1, ..., n$, and $\bar{p}_{emo}$ denotes the mean value given all n annotations:

$$\bar{p}_{emo} = \frac{1}{n}\sum_{i=1}^{n} p_{emo,i}. \tag{2}$$

In previous works, the PU information has been successfully exploited for emotion prediction tasks. On one hand, it is indicated in [20] that a more human-like and comprehensive emotion prediction can be generated via reporting the confidence together with the emotional state. On the other hand, the PU information can be exploited as a learning difficulty indicator to dynamically supervise the learning process [59].

Multi-Task Learning

After obtaining both the gold standard as our emotional states p_{emo} and the inter-rater disagreement level as the

perception uncertainty p_{pu}, the two tasks can be learned in a multi-task learning (MTL) structure, as shown in Fig. 1. While these two tasks are correlated with each other, during the training phase in a MTL context, the conventional emotion prediction task can benefit from the auxiliary task, i.e. PU modelling, by learning two tasks jointly in a supervised manner. In other words, the outputs from the two task-specific subnetworks are both concerned when updating the parameters in the shared front-end hidden layers.

To this end, given an input sample $\{\mathbf{x}_t\}$ with $t = 1, ..., T$, the network is optimised by minimising the loss function as:

$$\mathcal{J}(\boldsymbol{\theta}) = L_{emo}(\cdot) + \beta \cdot L_{pu}(\cdot) + \lambda \cdot R(\boldsymbol{\theta}), \tag{3}$$

where $L_{emo}(\cdot)$ and $L_{pu}(\cdot)$ denote the loss functions for emotion prediction and PU prediction, respectively, β is a predefined hyperparameter to regulate the contributions of $L_{pu}(\cdot)$, and λ is another hyperparameter which controls the importance of the L2 regularisation term $R(\boldsymbol{\theta})$.

Dynamic Tuning

As outlined in the sections "Related Work" and "Perception Uncertainty Modelling", with a vanilla MTL structure, knowledge of the PU is utilised to facilitate the training of the emotion estimation model. Albeit the notable advantages, one may note that, once the training is completed, the PU path is disregarded and does not contribute for the emotion estimation anymore. In other words, the PU path is not exploited at all during inference. In contrast to a vanilla MTL framework, here we propose to utilise the PU information to regulate the emotion estimation for both the training and evaluation phases. It is expected that the PU estimation can provide auxiliary information for the emotion estimation.

Specifically, as shown in Fig. 1, another dynamic tuning operation is built to aggregate the initial emotion prediction and the perception uncertainty in an explicit way. Formally, the function of the layer can be expressed via:

$$p_{emo} = p'_{emo} + p'_{emo} \cdot p_{pu}, \tag{4}$$

where p'_{emo} and p_{pu} stand for the output from the task-specific layers for emotion states and uncertainty degree, respectively, and p_{emo} denotes the output from the dynamic tuning operation. With this approach, the final emotion estimation p_{emo} is partially controlled by the PU information. In other words, the complementary information from PU is further employed in the evaluation phase.

Note that, a presupposition of this adjustment is that the emotion and the uncertainty are positively correlated. In the case where the two tasks are negatively correlated, Eq. 4 should be reformulated as:

$$p_{emo} = p'_{emo} - p'_{emo} \cdot p_{pu}. \tag{5}$$

For the sake of simplicity, Eqs. 4 and 5 can be unified to be:

$$p_{emo} = p'_{emo} + \text{sign}(r) \cdot p'_{emo} \cdot p_{pu}, \tag{6}$$

where r indicates the Pearson correlation coefficient (PCC) between the emotional state values and the perception uncertainty, and is determined by the prior knowledge with respect to the given task.

In this circumstance, the objective function in Eq. 3 can be reformulated as:

$$\mathcal{J}(\boldsymbol{\theta}) = L_{emo}(\cdot) + \alpha \cdot L'_{emo}(\cdot) + \beta \cdot L_{pu}(\cdot) + \lambda \cdot R(\boldsymbol{\theta}), \tag{7}$$

where $L_{emo}(\cdot)$ and $L'_{emo}(\cdot)$ denote the loss functions for the final emotion prediction (outputs of the dynamic tuning operation) and the initial emotion estimation (outputs of the emotion prediction task-specific layer), respectively, L_{pu} again stands for the loss for the PU modelling, α and β are predefined hyperparameters to control the contributions of $L'_{emo}(\cdot)$ and $L_{pu}(\cdot)$, respectively, and λ is deployed to modulate the regularisation.

Furthermore, a triplet loss function is investigated, with an aim to further advance the embedding learning process. In general, the triplet loss forces to project the original input features into a latent space where instances with similar semantics are pulled together while instances with dissimilar semantics are pushed away. Consequently, the similarity of instances with the same semantic information is preserved in the learned embedding space. More information on the implementation of the triplet can be found in [19].

When integrating the triplet constraint into the training approach, the objective function in Eq. 7 will then be rewritten as:

$$\begin{aligned}\mathcal{J}(\boldsymbol{\theta}) &= L_{emo}(\cdot) + \alpha \cdot L'_{emo}(\cdot) + \beta \cdot L_{pu}(\cdot) \\ &+ \gamma \cdot L_{tri}(\cdot) + \lambda \cdot R(\boldsymbol{\theta}),\end{aligned} \tag{8}$$

where L_{tri} is the triplet loss, and the hyperparameter γ is introduced to weight the contribution of the triplet loss.

Note that, the values of α, β, and γ are optimised on the development set, by achieving the best performance for the final emotion prediction. After training, the framework can be applied to estimate final emotional states given input features, by integrating the two estimations from the two separate task-specific subnetworks by Eqs. 4 or 5 accordingly.

Experiments

To evaluate the feasibility and effectiveness of our approach, we carried out extensive experiments on two

emotional datasets for continuous emotion prediction in dyadic conversation and music listening, respectively. In particular, the RECOLA dataset [45] was utilised for audiovisual emotion regression, and the emoMusic dataset [48] for music emotion prediction. In this section, we first provide a brief introduction to the two datasets and the selected feature sets (see section "Data and Features"). Then, the experimental setup and evaluation measurements are explained in detail for the sake of experiment replication together with a performance comparison (see section "Implementation and Evaluation"). After that, experimental results and discussions on the two datasets are reported in the sections "Experimental Results and Discussion for RECOLA" and "Experimental Results and Discussion for emoMusic", respectively.

Data and Features

For emotion recognition in dyadic multimodal interactions, the RECOLA dataset was utilised. This multimodal corpus is widely used for audiovisual dimensional emotion recognition, and also a benchmark database previously applied in a series of AVEC challenges since 2015 [44, 52]. It consists of recordings of spontaneous and natural interactions from 27 French-speaking individuals, aiming at studying socio-affective behaviours in the context of remote collaborative tasks. In particular, varied multimodal signals, i.e. audio, video, and physiological data, were collected continuously and synchronously [45]. In the current tentative study, only audio and video recordings will be investigated. Moreover, detailed time- and value-continuous dimensional emotion annotations in terms of arousal and valence are given with a constant frame rate of 40 ms for the first 5 min of each recording, by averaging six annotators, and meanwhile taking the inter-evaluator agreement into consideration [45]. Following the previous partitions in the AVEC challenge, the dataset is further equally divided into three disjoint partitions, by balancing the gender, age, and mother tongue of the participants. Thus, each partition contains nine unique audiovisual recordings, resulting in 67.5K segments in total for each partition, i.e. the training, development, or test set. For more details of the data distribution, please refer to Table 1.

For music emotion recognition (MER), the emoMusic dataset was selected, which was a publicly available benchmark for MER and first introduced during the MediaEval 2013 "Emotion in Music" task [49]. This music corpus spans 1000 45-second clips, by selecting 1000 songs from the Free Music Archive. After that, annotations were collected via more than 300 crowdworkers using the Amazon's Mechanical Turk platform. Especially, for each single clip, continuous arousal and valence annotations were generated with a constant sampling rate of 2 Hz, by averaging annotations of at least 10 annotators from the crowdsourcing platform. Moreover, the continuous annotations are between −1 and +1 and the first 15 s are excluded due to the instability of the annotations at the start of the clips [48]. Also note that, after removing redundant songs from the initial version, the reduced corpus now consists of 744 songs in total, and is further split into two disjoint parts, i.e. the development set with 619 clips and the test set with 125 clips. As a result, the remaining segments are 37,759 and 7,625 for the development and test sets, respectively.

Table 1 Three partitions of the RECOLA database

No.	Train	Development	Test
Female	6	5	5
Male	3	4	4
French	6	7	7
Italian	2	1	2
German	1	1	0
Age μ (σ)	21.2 (1.9)	21.8 (2.5)	21.2 (1.9)

In order to investigate the correlation between the emotion annotation (arousal or valence) and the inter-evaluator disagreement level, we computed the PCC r between the generated 'gold standard' and its corresponding perception uncertainties. The obtained results are presented in Table 2. Interestingly, one can notice that in RECOLA the two targets are partially positive linearly correlated (.215 for arousal and .103 for valence), while in emoMusic the correlations are negative (− .296 for arousal and − .203 for valence). That means, in human conversation, individuals tend to disagree more on instances with stronger emotions; however, for emotions in music, people demonstrate higher consistency on instances with richer emotional content. One reason might be that, some annotators tend to stick to 'what is common' unconsciously, resulting in a higher standard deviation of 'what is rare'. In most human interactions, neutral is common, while emotional representations are more ordinary in music as its main intention is to achieve consensus emotionally among listeners. Therefore, the PU information should be exploited differently for the

Table 2 Obtained Pearson correlation coefficients (PCCs) between the absolute value of the emotional states and the corresponding perception uncertainties for *RECOLA* and *emoMusic*, with respect to *arousal* and *valence*, respectively

PCC	Arousal	Valence
RECOLA	.215	.103
emoMusic	− .296	− .203

two tasks, when dynamic fine-tuning the initial emotion estimations (see section "Dynamic Tuning").

For acoustic features, we used the extended Geneva Minimalistic Acoustic Parameter Set (*eGeMAPS* [13]) for both datasets. The feature extraction can be done with our open-source openSMILE toolkit [14]. In particular, 88 supra-segmental features were extracted by applying various statistical functionals, such as mean and moments, over 23 frame-level low-level descriptors, such as MFCCs and energy. This handcrafted feature set has been successfully utilised in previous studies and achieved robust prediction performance in a wide range of audio tasks, especially in emotion recognition.

For video recordings in the RECOLA dataset, two types of facial descriptors are investigated, i.e. *appearance* and *geometry* based, which are standard features provided in the AVEC challenges [52], for a fair comparison with other methods in the literature. This resulted in 168 appearance and 632 geometric visual features. For more details on the feature extraction process, the reader is referred to [52].

Moreover, we applied online standardisation to the above-mentioned feature sets, respectively. Specifically, the means and variances of features were calculated on the training set, which were then applied over the development and test sets for standardisation.

Implementation and Evaluation

To implement the proposed PU modelling framework for continuous emotion recognition, we employed a deep RNN structure with gated recurrent units (GRUs). As an alternative to long short-term memory cells, GRU cells can capture long-term dependencies in sequence-based tasks and ameliorate the vanishing gradient problem as well.

For the RECOLA experiments, the number of hidden layers and the number of units per layer were defined, following our previous work on the same database after a grid search evaluation strategy [59]. In particular, in all systems, the number of hidden layers for the shared subnetwork and the task-specific subnetworks was set as 2, respectively. In addition, each hidden layer consisted of 120 GRU cells. During network training, we utilised the Adam optimisation algorithm with an initial learning rate of 0.001. Moreover, to facilitate the training process, the mini-batch size during training was 128. Finally, an early stopping strategy was deployed as no improvement of the prediction performance on the development partition has been observed during 20 epochs or a predefined maximum number of training epochs (100 runs in all our cases) has been reached.

Furthermore, we applied a grid-search strategy for the three hyperparameters α, β, and γ in Eq. 8 which control the contribution of the initial emotion prediction loss, the PU prediction loss, and the triplet loss, respectively. More specifically, the best setting was determined on the best performance achieved on the development set by a grid search over $[.1, .2, .5, 1.0]$ for α and over $[0, .01, .02, ..., .09, .1, .2, .5, 1.0]$ for β and γ, respectively. In addition, we executed the same annotation delay compensation strategy and the post-processing chain on all predictions, following the suggestions by the AVEC challenge in [52]. Specifically, the post-processing parameters were optimised on the development set and then applied to the test set to refine the obtained predictions. Thus, these settings varied from task to task.

Finally, to evaluate the performance of the models, we computed the official metric of the AVEC challenges for dimensional emotion recognition tasks, namely Concordance Correlation Coefficient (CCC) [52] which can be computed via:

$$r_c = \frac{2r\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}, \tag{9}$$

where r denotes PCC between two objects (e.g., prediction and gold standard), μ_x and μ_y are the means of each object, and σ_x^2 and σ_y^2 stand for the corresponding variances. The CCC metric falls into the range of $[-1, 1]$, where $+1$ represents perfect concordance, -1 total discordance, and 0 no concordance at all. In general, a higher CCC indicates a better prediction performance.

Experimental Results and Discussion for RECOLA

For the task of emotion regression in dyadic conversation, the performances in terms of CCC are presented in Table 3. For a fair performance comparison, the results of the corresponding baseline systems, our proposed PU modelling system, as well as that of other state-of-the-art benchmarks on RECOLA, are listed for each selected feature set in detail. In particular, our single-task learning baseline system (denoted as *STL* in Table 3) is a plain four-layer GRU-RNN deep structure with 120 cells per layer, aiming to provide time- and value-continuous predictions in the arousal and valence dimensions, respectively. As can be seen in Table 3, our STL baseline performances are already quite competitive, when comparing their performance with that of other recent works in the literature on the same database.

Moreover, in order to explore the auxiliary information from the perception uncertainty, a multi-task learning framework (denoted as MTL in Table 3) was further evaluated. From the obtained results, one may notice that the MTL systems outperform the corresponding STL ones in all audio feature-based recognition tasks and six out of eight cases when using video features (appearance or geometric).

Table 3 Performance comparison in terms of Concordance Correlation Coefficient (*CCC*) for emotion regression tasks on the *dev(elopment)* and *test* sets of RECOLA in *arousal* and *valence*, respectively

CCC	Audio-eGeMAPS				Video-appearance				Video-geometric			
	Arousal		Valence		Arousal		Valence		Arousal		Valence	
	Dev	Test	Dev	Test	Dev	Test	Dev	Test	Dev	Test	Dev	Test
Proposed												
STL	.766	.605	.504	.381	.512	.411	.545	*.525*	.499	.399	.619	.529
MTL	.779	.620	.520	.405	.514	*.456*	.549	.493	*.542*	.327	.647	.531
MTL with PU modelling	.779	.623	*.534**	*.420**	.522*	.440	.566*	*.525**	.528	*.407**	*.651*	*.545**
Other state of the art												
PU-based DDAT [59]	.811	.664	.498	.407	.518	.438	.514	.431	.513	.397	.632	.501
DNNs [33]	.573	.517	.129	.044	.387	.220	.306	.206	.312	.296	.362	.216
Curriculum learning (DNN) [33]	.687	.591	.159	.174	.417	.343	.446	.419	.394	.267	.300	.269
Curriculum learning (GRU-RNN) [59]	.754	.611	.501	.357	.491	.391	.557	.492	.444	.336	.609	.500
Strength modelling [17]	.755	*.666*	.476	.364	.350	.196	*.592*	.464	–	–	–	–
Feature selection + offset [23]	.800	–	.398	–	*.587*	–	.441	–	.173	–	.441	–
SVR + offset [52]	.796	.648	.455	.375	.483	.343	.474	.486	.379	.272	.612	.507
CNN + LSTM-RNN [5]	*.846*	–	.450	–	.346	–	.511	–	–	–	–	–

Results are reported for the proposed PU modelling approach, together with the single-task learning (*STL*) framework, and the multi-task learning (MTL) framework, as well as other state-of-the-art works (*DDAT*, dynamic difficulty awareness training). The best achieved CCCs are italicised. Three feature sets (audio-eGeMAPS, video-appearance, and video-geometric) are evaluated for all methods. Cases where the proposed PU modelling has a statistical significance ($p < .05$) of performance improvement over MTL by means of Fisher's r-to-z transformation are marked by $*$

This meets our expectation and further consolidates findings from our previous works [20, 59].

Furthermore, by modulating the emotion estimations with PU directly, our PU modelling delivered significant performance improvements continuously over a large margin, when compared with STL. Besides, it is interesting to notice that the MTL with the PU modelling approach is superior to the MTL-only method, which indicates the effectiveness of the PU-based amendment. In particular, on the test set, the best CCC of .623 is reached with audio-eGaMAPS features for arousal, while the best CCC of .545 for valence is achieved on the video-geometric feature set. In addition, we observe that the best results of PU modelling frameworks are consistently achieved when $\alpha = 1$. This may imply that the loss from the initial emotion estimation is essential to regulate the training process of a PU modelling framework. Similarly, we observe that during the optimisation procedure, best performance was in most cases achieved when β and γ were non-zeros, indicating that taking these terms away from the proposed objective function will lead to a performance decrease.

Finally, we reported the state-of-the-art results obtained on RECOLA with the same feature sets. It is noticeable that our systems get the best results on the test set except for arousal with audio features (ours is .623 while the best CCC .666 is obtained by strength modelling). The rationale for this is that, in strength modelling [17], the advantages of RNN and SVR are both exploited, while in our model only an RNN was used. However, in this specific case, an SVR model brings more benefits for the audio arousal regression than an NN-based model (.648 for SVR and .611 for curriculum learning with RNN). In this regard, one future investigation might be integrating the PU modelling with the strength modelling, to further boost the capability of a system for emotion modelling.

Experimental Results and Discussion for emoMusic

For music emotion recognition, experiments on the emoMusic database were conducted to further justify the robustness and efficiency of the proposed paradigm. Details of the obtained results are presented in Table 4. Note that, while the correlation between the emotional states and the corresponding uncertainties are negative (cf. Table 2 for both arousal and valence), Eq. 5 was applied to modulate the final predictions with the help of their PUs.

As shown in Table 4, similar observations could be seen as in the previous tasks. Comparing the plain STL and MTL systems, one may notice that, by training a network under the MTL strategy, better performances have been

Table 4 Performance comparison in terms of Concordance Correlation Coefficient (*CCC*) on the test set of the emoMusic dataset for arousal and valence predictions

Methods	Arousal	Valence
Proposed		
STL	.753	.558
MTL	.761	.562
MTL with PU modelling	.779	.597
Other state-of-the-art		
BLSTM-RNN [3]	.300	.060
Sum of CCC objective [53]	.719	.582
TCCC objective [53]	*.790*	*.648*

Results obtained after the post-processing chain are given, for the proposed PU modelling system, together with the corresponding single-task learning (*STL*) and multi-task learning (*MTL*) baselines, as well as other state-of-the-art works (*BLSTM-RNN*, bidirectional long short-term memory recurrent neural networks). The best results are italicised

achieved for both arousal and valence. This observation is in accordance with our previous findings and our expectation that estimating the emotions and PU levels simultaneously could bring benefit to the emotion recognition task.

Furthermore, when modifying the learning strategy via utilising the PU estimations explicitly, our method delivers further performance improvements over a large margin consistently on all cases. In particular, after the post-processing procedures, the performance of our proposed system reaches CCC values of .779 and .597 for arousal and valence, respectively. These results demonstrate that the PU modelling approach significantly outperforms both the plain STL and MTL baselines ($p < .05$ by means of Fisher's r-to-z transformation). Moreover, results achieved by other state-of-the-art works are also presented in Table 4. In particular, it can be observed that by replacing the conventional root mean square error loss with CCC-based objective function, performance in both arousal and valence is increased. Hence, further investigations need to be performed to integrate CCC optimisation to the present PU modelling system.

In addition, to further assess the effect of the PU modelling for the music emotion regression task, we conducted experiments with the positive regulation with Eq. 4 too. In this context, the best CCCs were obtained at all times when β in Eq. 8 was set to 0. This further meets our expectation that when two targets are negatively correlated, the PU-based emotion modification should be operated with Eq. 5 in place of Eq. 4. In other words, given a new task, the correlation between the task itself and its PU degrees (positive or negative) should be understood first before deploying the framework. Thus, analysing the correlation between a subjective task and its perception uncertainty based on human cognition might be a promising avenue for future work, and would promote further applications of our method in other subjective recognition tasks.

Conclusion

In this paper, we propose a multi-task learning framework for emotion recognition in human conversation and music listening. By integrating the perception uncertainty information into the emotion estimation during both the training and evaluation phases, we have altered the initial emotion prediction in order to produce its corresponding final prediction. When conducting intensive experiments on two emotional datasets for human interaction and music emotion recognition, respectively, impressive performance improvements have been observed in both tasks. To the best of our knowledge, though there are many existing studies in exploiting the inter-rater disagreement level, this is for the first time that the disagreement level is applied to revise the prediction during inference in a dynamic way.

In the future, we will consider the sign function in Eq. 6 as a hyperparameter rather than a pre-defined parameter in the network training phase. This improvement will make the framework more flexible when dealing with the unknown correlationship between the emotional state and perception uncertainty for each utterance. In addition, to further justify the effectiveness and robustness of the approach, we plan to evaluate it on additional large-scale emotional datasets, where annotations from multiple raters are provided, such as SEWA [30]. Moreover, given that our framework can be deployed to other subjective recognition tasks, we would like to examine its generalisation properties on more tasks, such as sentiment analysis [11], personality estimation [35], and engagement detection [7]. Lastly, Bayesian learning-based approaches will be explored to model the uncertainty and learn interpretable representations of emotional instances in future [24].

Funding Information This study was partially supported by the TransAtlantic Platform "Digging into Data" collaboration grant (ACLEW: Analysing Child Language Experiences Around The World), with the support of the UK's Economic & Social Research Council through the research Grant No. HJ-253479, and by the European Union's Horizon H2020 Research and Innovation programme under Marie Skłodowska-Curie grant agreement No. 766287 (TAPAS).

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

Ethical Approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Agarwal B, Poria S, Mittal N, Gelbukh A, Hussain A. Concept-level sentiment analysis with dependency-based semantic parsing: a novel approach. Cogn Comput. 2015;7(4):487–99.
2. Albanie S, Nagrani A, Vedaldi A, Zisserman A. Emotion recognition in speech using cross-modal transfer in the wild. In: Proc. ACM international conference on multimedia (MM). Seoul; 2018. p. 292–301.
3. Aljanaki A, Yang YH, Soleymani M. Developing a benchmark for emotional analysis of music. PloS One. 2017;12(3):e0173,392.
4. Beatty A. Anthropology and emotion. J R Anthropol Instit. 2014;20(3):545–63.
5. Brady K, Gwon Y, Khorrami P, Godoy E, Campbell WM, Dagli CK, Huang TS. Multi-modal audio, video and physiological sensor learning for continuous emotion prediction. In: Proc. 6th international workshop on audio/visual emotion challenge (AVEC). Amsterdam; 2016. p. 97–104.
6. Cambria E. Affective computing and sentiment analysis. IEEE Intell Syst. 2016;31(2):102–7.
7. Chorianopoulou A, Tzinis E, Iosif E, Papoulidi A, Papailiou C, Potamianos A. Engagement detection for children with autism spectrum disorder. In: Proc. international conference on acoustics, speech and signal processing (ICASSP). Calgary; 2017. p. 5055–9.
8. Chou H, Lee C. Every rating matters: joint learning of subjective labels and individual annotators for speech emotion classification. In: Proc. IEEE international conference on acoustics, speech and signal processing (ICASSP). Brighton; 2019. p. 5886–90.
9. Dang T, Sethu V, Ambikairajah E. Dynamic multi-rater gaussian mixture regression incorporating temporal dependencies of emotion uncertainty using kalman filters. In: Proc. IEEE International conference on acoustics, speech and signal processing (ICASSP). Calgary; 2018. p. 4929–33.
10. Dang T, Sethu V, Epps J, Ambikairajah E. An investigation of emotion prediction uncertainty using gaussian mixture regression. In: Proc. Annual conference of the international speech communication association (INTERSPEECH). Stockholm; 2017. p. 1248–52.
11. Dashtipour K, Poria S, Hussain A, Cambria E, Hawalah AY, Gelbukh A, Zhou Q. Multilingual sentiment analysis: state of the art and independent comparison of techniques. Cogn Comput. 2016;8(4):757–71.
12. Deng J, Han W, Schuller B. Confidence measures for speech emotion recognition: a start. In: Proc.the 10th ITG conference on speech communication. Braunschweig; 2012. p. 1–4.
13. Eyben F, Scherer K, Schuller B, Sundberg J, André E., Busso C, Devillers L, Epps J, Laukka P, Narayanan S, Truong K. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Trans Affect Comput. 2016;7(2):190–202.
14. Eyben F, Wöllmer M, Schuller B. openSMILE – the Munich versatile and fast open-source audio feature extractor. In: Proc. ACM international conference on multimedia (ACM MM). Florence; 2010. p. 1459–62.
15. Eyben F, Wöllmer M, Schuller B. A multitask approach to continuous five-dimensional affect sensing in natural speech. ACM Trans Interact Intell Syst. 2012;2(1):1–29.
16. Gui L, Baltrušaitis T, Morency L. Curriculum learning for facial expression recognition. In: Proc. 12th IEEE international conference on automatic face gesture recognition (FG). Washington; 2017. p. 505–11.
17. Han J, Zhang Z, Cummins N, Ringeval F, Schuller B. Strength modelling for real-world automatic continuous affect recognition from audiovisual signals. Image Vis Comput. 2017;65:76–86.
18. Han J, Zhang Z, Cummins N, Schuller B. Adversarial training in affective computing and sentiment analysis: recent advances and perspectives. IEEE Comput Intell Mag. 2019;14(2):68–81.
19. Han J, Zhang Z, Keren G, Schuller B. Emotion recognition in speech with latent discriminative representations learning. Acta Acust United Acust. 2018;104(5):737–40.
20. Han J, Zhang Z, Schmitt M, Schuller B. From hard to soft: towards more human-like emotion recognition by modelling the perception uncertainty. In: Proc. ACM International conference on multimedia (MM). Mountain View; 2017. p. 890–97.
21. Hazarika D, Poria S, Zadeh A, Cambria E, Morency L, Zimmermann R. Conversational memory network for emotion recognition in dyadic dialogue videos. In: Proc. the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies (NAACL-HLT). New Orleans; 2018. p. 2122–132.
22. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proc. IEEE conference on computer vision and pattern recognition (ICCV). Las Vegas; 2016. p. 770–78.
23. He L, Jiang D, Yang L, Pei E, Wu P, Sahli H. Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks. In: Proc. 5th international workshop on audio/visual emotion challenge (AVEC). Brisbane; 2015. p. 73–80.
24. He L, Liu B, Li G, Sheng Y, Wang Y, Xu Z. Knowledge base completion by variational Bayesian neural tensor decomposition. Cogn Comput. 2018;10(6):1075–84.
25. Kaminskas M, Ricci F. Contextual music information retrieval and recommendation: state of the art and challenges. Comput Sci Rev. 2012;6(2–3):89–119.
26. Katsigiannis S, Ramzan N. DREAMER: a database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices. IEEE J Biomed Health Inf. 2018;22(1):98–107.
27. Kim Y, Kim J. Human-like emotion recognition: multi-label learning from noisy labeled audio-visual expressive speech. In: Proc. IEEE International conference on acoustics, speech and signal processing (ICASSP). Calgary; 2018. p. 5104–08.
28. Kim Y, Provost EM. Leveraging inter-rater agreement for audio-visual emotion recognition. In: Proc. International conference on affective computing and intelligent interaction (ACII). Xi'an; 2015. p. 553–59.
29. Koelsch S. Music-evoked emotions: principles, brain correlates, and implications for therapy. Ann N Y Acad Sci. 2015;1337(1):193–201.
30. Kossaifi J, Walecki R, Panagakis Y, Shen J, Schmitt M, Ringeval F, Han J, Pandit V, Schuller B, Star K, Hajiyev E, Pantic M. SEWA DB: a rich database for audio-visual emotion and sentiment research in the wild. In: IEEE Transactions on pattern analysis and machine intelligence. No pagination. 2019.
31. Li X, Bing L, Lam W, Shi B. Transformation networks for target-oriented sentiment classification. In: Proc. Annual meeting of the association for computational linguistics (ACL). Melbourne; 2018. p. 946–56.
32. Liu N, Fang Y, Li L, Hou L, Yang F, Guo Y. Multiple feature fusion for automatic emotion recognition using EEG signals. In: Proc. IEEE International conference on acoustics, speech and signal processing (ICASSP). Calgary; 2018. p. 896–900.
33. Lotfian R, Busso C. Curriculum learning for speech emotion recognition from crowdsourced labels. IEEE/ACM Trans Audio Speech Lang Process. 2019;27(4):815–26.
34. Majid A. Current emotion research in the language sciences. Emot Rev. 2012;4(4):432–43.

35. Majumder N, Poria S, Gelbukh A, Cambria E. Deep learning-based document modeling for personality detection from text. IEEE Intell Syst. 2017;32(2):74–9.
36. Majumder N, Poria S, Hazarika D, Mihalcea R, Gelbukh A, Cambria E. DialogueRNN: an attentive RNN for emotion detection in conversations. In: Proc. Thirty-Third AAAI conference on artificial intelligence (AAAI). Honolulu; 2019. p. 6818–25.
37. Malandri L, Xing FZ, Orsenigo C, Vercellis C, Cambria E. Public mood–driven asset allocation: the importance of financial sentiment in portfolio management. Cogn Comput. 2018;10(6):1167–76.
38. Mauss IB, Robinson MD. Measures of emotion: a review. Cogn Emotion. 2009;23(2):209–37.
39. Mower E, Metallinou A, Lee C, Kazemzadeh A, Busso C, Lee S, Narayanan S. Interpreting ambiguous emotional expressions. In: Proc. International conference on affective computing and intelligent interaction (ACII). Amsterdam; 2009. p. 1–8.
40. Niedenthal PM, Ric F. Psychology of emotion, 2nd ed. New York: Psychology Press; 2017.
41. Noroozi F, Kaminska D, Corneanu C, Sapinski T, Escalera S, Anbarjafari G. Survey on emotional body gesture recognition. IEEE Transactions on Affective Computing. No pagination. 2018.
42. Poria S, Cambria E, Gelbukh A. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In: Proc. International conference on empirical methods in natural language processing (EMNLP). Lisbon; 2015. p. 2539–44.
43. Principi E, Rotili R, Wöllmer M, Eyben F, Squartini S, Schuller B. Real-time activity detection in a multi-talker reverberated environment. Cogn Comput. 2012;4(4):386–97.
44. Ringeval F, Schuller B, Valstar M, Jaiswal S, Marchi E, Lalanne D, Cowie R, Pantic M. AV+EC 2015: the first affect recognition challenge bridging across audio, video, and physiological data. In: Proc. the 5th international workshop on audio/visual emotion challenge (AVEC). Brisbane; 2015. p. 3–8.
45. Ringeval F, Sonderegger A, Sauer JS, Lalanne D. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In: Proc. 10th IEEE International conference and workshops on automatic face and gesture recognition (FG). Shanghai; 2013. p. 1–8.
46. Sarda P, Halasawade S, Padmawar A, Aghav J. Emousic: emotion and activity-based music player using machine learning. In: Proc. International conference on computer communication and computational sciences (IC4S). Bangkok; 2018. p. 179–88.
47. Schuller B, Batliner A. Computational paralinguistics: emotion, affect and personality in speech and language processing. Hoboken: Wiley; 2013.
48. Soleymani M, Caro MN, Schmidt EM, Sha CY, Yang YH. 1000 songs for emotional analysis of music. In: Proc. 2nd ACM international workshop on crowdsourcing for multimedia (CrowdMM); 2013. p. 1–6.
49. Soleymani M, Caro MN, Schmidt EM, Yang YH. The mediaeval 2013 brave new task: emotion in music. In: Proc. MediaEval workshop; 2013. p. 1–2.
50. Sun X, Lv M. Facial expression recognition based on a hybrid model combining deep and shallow features. Cogn Comput. 2019;11(4):587–97.
51. Trigeorgis G, Ringeval F, Bruckner R, Marchi E, Nicolaou M, Schuller B, Zafeiriou S. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In: Proc. International conference on acoustics, speech and signal processing (ICASSP). Shanghai; 2016. p. 5200–4.
52. 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: Proc. the 6th international workshop on audio/visual emotion challenge (AVEC). Amsterdam; 2016. p. 3–10.
53. Weninger F, Ringeval F, Marchi E, Schuller B. Discriminatively trained recurrent neural networks for continuous dimensional emotion recognition from audio. In: Proc. International joint conference on artificial intelligence (IJCAI). New York; 2016. p. 2196–02.
54. Wöllmer M, Eyben F, Reiter S, Schuller B, Cox C, Douglas-Cowie E, Cowie R. Abandoning emotion classes – towards continuous emotion recognition with modelling of long-range dependencies. In: Proc. Annual conference of the international speech communication association (INTERSPEECH). Brisbane; 2008. p. 597–600.
55. Zhang L, Wang S, Liu B. Deep learning for sentiment analysis: a survey. Wiley Interdiscip Rev: Data Mining Knowl Discov. 2018;8(4):1–25.
56. Zhang Z, Coutinho E, Deng J, Schuller B. Cooperative learning and its application to emotion recognition from speech. IEEE/ACM Trans Audio Speech Lang Process. 2015;23(1):115–26.
57. Zhang Z, Cummins N, Schuller B. Advanced data exploitation for speech analysis – an overview. IEEE Signal Process Mag. 2017;34(4):107–29.
58. Zhang Z, Eyben F, Deng J, Schuller B. An agreement and sparseness-based learning instance selection and its application to subjective speech phenomena. In: Proc. 5th international workshop on emotion social signals, sentiment & linked open data, satellite of LREC. Reykjavik; 2014. p. 21–6.
59. Zhang Z, Han J, Schuller B. Dynamic difficulty awareness training for continuous emotion prediction. IEEE Trans Multimed. 2019;21(5):1289–301.