From Hard to Soft: Towards more Human-like Emotion Recognition by Modelling the Perception Uncertainty

Jing Han*
Chair of Embedded Intelligence for Health Care and Wellbeing,
University of Augsburg, Germany
jing.han@informatik.uni-augsburg.de

Zixing Zhang
Chair of Complex and Intelligent Systems,
University of Passau, Germany
zixing.zhang@uni-passau.de

Maximilian Schmitt†
Chair of Embedded Intelligence for Health Care and Wellbeing,
University of Augsburg, Germany
maximilian.schmitt@informatik.uni-augsburg.de

Maja Pantic
iBUG – intelligent Behaviour Understanding Group,
Imperial College London, UK
m.pantic@imperial.ac.uk

Björn Schuller‡
Chair of Embedded Intelligence for Health Care and Wellbeing,
University of Augsburg, Germany
schuller@ieee.org

ABSTRACT

Over the last decade, automatic emotion recognition has become well established. The gold standard target is thereby usually calculated based on multiple annotations from different raters. All related efforts assume that the emotional state of a human subject can be identified by a 'hard' category or a unique value. This assumption tries to ease the human observer's subjectivity when observing patterns such as the emotional state of others. However, as the number of annotators cannot be infinite, uncertainty remains in the emotion target even if calculated from several, yet few human annotators. The common procedure to use this same emotion target in the learning process thus inevitably introduces noise in terms of an uncertain learning target. In this light, we propose a 'soft' prediction framework to provide a more human-like and comprehensive prediction of emotion. In our novel framework, we provide an additional target to indicate the uncertainty of human perception based on the inter-rater disagreement level, in contrast to the traditional framework which is merely producing one single prediction (category or value). To exploit the dependency between the emotional state and the newly introduced perception uncertainty, we implement a multi-task learning strategy. To evaluate the feasibility and effectiveness of the proposed soft prediction framework, we perform extensive experiments on a time- and value-continuous spontaneous audiovisual emotion database including late fusion results. We show that the soft prediction framework with multi-task learning of the emotional state and its perception uncertainty significantly outperforms the individual tasks in both the arousal and valence dimensions.

CCS CONCEPTS

• **Information systems** → *Sentiment analysis*; • **Human-centered computing** → *Human computer interaction (HCI)*;

KEYWORDS

Emotion recognition; Perception uncertainty modelling; Multi-task learning; Long short-term memory

1 INTRODUCTION

Automatic Emotion Recognition (AER) is of extreme importance for achieving natural and friendly Human–Machine Interaction systems in the real world, since it enables machines to well understand humans' spontaneous affective state just as human beings do [2, 6]. Over the past decade, numerous research efforts have been made to build an effective and robust recognition model, leading to a great achievement in the field of AER [13, 17, 25].

However, one of the long-standing concerns in AER is how to appropriately indicate the emotional state, which is mainly caused by the human observer's subjectivity when perceiving the emotional state of others [15]; indeed, AER differs from many other pattern recognition tasks that hold a ground truth, such as face recognition and automatic speech recognition. To obtain the authentic emotional state for emotion modelling, the widely employed approach in the AER research community is to obtain a *gold standard*. It requires multiple annotators to perceive a same human expression by audio and/or video, and then merges these perceptions (or, annotations) made from these annotators into a unique ('hard') label, i. e., a category for a classification task or a value for a regression task.

*The author is further affiliated with Chair of Complex and Intelligent Systems, University of Passau, Germany.
†The author is further affiliated with Chair of Complex and Intelligent Systems, University of Passau, Germany.
‡The author is further affiliated with the Chair of Complex and Intelligent Systems, University of Passau, Germany, and with GLAM – Group on Language, Audio, and Music, Department of Computing, Imperial College London, UK.

MM'17, October 23–27, 2017, Mountain View, CA, USA.
 ISBN 978-1-4503-4906-2/17/10…$15.00
DOI: https://doi.org/10.1145/3123266.3123383

To fuse multiple annotations into a gold standard, several approaches have been introduced with the purpose of fostering a reasonable estimate of emotion. Among them, the most straightforward way is by performing majority voting or calculating the mean or median values among all available annotations to represent the gold standard [19, 31]. Despite the simplicity, this method is not stable when annotators show huge disagreement [31]. For this reason, more sophisticated approaches have been deployed. In [9], the authors introduced Evaluator Weighted Estimator (EWE), which considers inter-evaluator agreement to weight individual annotations and meanwhile filters out unreliable evaluators to improve the robustness of the results. Additionally, in [49], the authors presented Canonical Time Warping (CTW) for accurate spatio-temporal alignment of facial expressions, which accommodates for subject variability and allows temporal local transformations. Following this work, more advanced derivations have recently been proposed and investigated for emotion recognition [26, 50].

Even though the aforementioned approaches try to alleviate the human observer's annotation subjectivity via calculating the gold standard from several annotators, uncertainty remains in the emotion target. From the human perspective, we often apply various adverbs of degree including "occasionally", "probably", and "definitely" to extend and enrich our views when perceiving emotions. Furthermore, we also utilise diverse modal verbs such as "might", "could", and "ought" to express the degree of *uncertainty* accordingly. Unfortunately, a similar measure in an AER system has seemed to be missing in the literature so far. In other words, AER lacks an additional descriptor to indicate the degree of perception uncertainty, to picture the whole emotion analysis.

In this paper, we propose a novel framework, aiming at modelling a more human-like and comprehensive emotion analysis. Distinct from conventional frameworks which merely estimate the emotional state as a *'hard prediction'*, we extend it to a more humanoid *'soft prediction'*, by offering an additional descriptor (or, indicator/label) to describe the uncertainty of human perception. That is, not only the *emotional state* but also its corresponding *perception uncertainty* are jointly provided for each observed sample. Besides of this, another motivation of including the perception uncertainty for AER comes with the fact that, it can well reflect the difficulty and complexity of the samples for machine learning, which was demonstrated in [4] and [30]. In this sense, perception uncertainty can be interpreted as an indirect confidence measure for each prediction [30]. To the best of our knowledge, this is the first effort to model the emotional state as well as the perception uncertainty in the domain of AER.

Specifically, we employ an inter-rater disagreement level to simulate the human perception uncertainty, with an assumption that, for each sample, the personal perception uncertainty is highly correlated with the inter-rater disagreement level. This assumption is plausible given that individuals with higher confidence are more likely to show less disagreement with others [4]. Besides, this assumption also supposes all annotators are reliable enough.

The contributions of this paper mainly include: i) proposing a novel 'soft-prediction' framework, which aims to shape a humanoid emotion prediction; ii) training jointly a model with two targets (the emotional state and the newly proposed perception uncertainty) in a multi-task learning paradigm, so as to improve the performance of each task; iii) investigating the soft prediction framework for both audio and video modalities.

In the remainder of this paper, we first briefly introduce the related work in Section 2. Afterwards, in Section 3 we describe in detail the soft prediction framework with modelling the perception uncertainty. We further perform extensive experiments on spontaneous audiovisual emotion recognition in both the hard and soft manners in Section 4. Finally, we draw the conclusion and point out the potential research directions in Section 5.

2 RELATED WORK

In the literature, there are some, but not many, relevant works which attempt to leverage the perception uncertainty in terms of inter-rater disagreement level. In [16], Karpouzis et al. claimed to compare the disagreement level among all raters with the one between the automatic estimation and the gold standard. From this comparison, one can assess whether the established model outperforms human raters on average in terms of the perception uncertainty. To more fairly compare these two kinds of disagreement level, various evaluation metrics were introduced, such as the sign agreement metric [24] and the intra-class correlation coefficient [35].

Moreover, in [18, 36, 42], the authors built training sets with a low level of perception uncertainty by different methods, to improve the reliability of labelled data which further enhanced the performance of emotion recognition classifiers to some extent. Rather than for the naive data selection, perception uncertainty has also been employed as an informative degree indicator of an instance for active learning and cooperative learning [46]. For example, researchers in [47] built a regressor based on the perception uncertainty, which was then used to automatically predict the disagreement level of unlabelled data. Those data with median-level uncertainties were then picked up for manual annotation. Additionally, perception uncertainty has been considered as well during the human annotation process to decide how many annotators are necessary to label one sample, which is termed dynamic active learning [45]. That is, once the inter-rater agreement level reaches a predefined level, the manual annotation process ceases to reduce human labelling effort. Also, previous work in [39] has indicated that, the inter-rater agreement level can be predicted to a certain degree.

One most closely related work in the literature is found in [8], where perception uncertainty was taken into account as an *auxiliary* task, trained together with the normal emotion recognition task in a multi-task learning paradigm. That is, the perception uncertainty was merely utilised to improve the performance of emotion recognition. By contrast, in our work, the perception uncertainty is considered as an *individual* task, with a comprehensive analysis on both audio and video modalities, and can be learnt either independently or jointly.

The proposed framework is somewhat relevant to another term of Confidence Measure (CM), which was advocated for keeping track of the reliability of recognition results. For example, in [44], researchers applied the obtained accuracy on training data of each classifier to capture the reliability of each classifier for a given database condition. In [38], the authors utilised the probabilistic outputs

of classifiers as the CM to estimate weights when combining the decisions from multiple classifiers to achieve a final prediction for every instance. Moreover, in [7], scoring models were developed to describe the agreement levels for all intended emotional states separately, and then used to evaluate the reliability of the recognition results. However, all these approaches were designed to measure the confidence level of a model to provide a 'correct label', with no awareness of the uncertainty of the given label itself.

3 PERCEPTION UNCERTAINTY MODELLING

To estimate the soft emotion predictions, the inter-rater disagreement level is calculated to quantify the perception uncertainty apart from the conventional unique emotion judgement. In other words, for each prediction, two indicators will be given, i. e., the emotional state and the perception uncertainty, to profile the emotion in a more humanoid format. Note that, in this paper we specifically focus on the time- and value-continuous spontaneous emotion recognition from audiovisual signals in an arousal and valence dimensional space. Particularly, we employ the Recurrent Neural Networks (RNN) equipped with Long Short-Term Memory (LSTM) blocks as a baseline regressor because of its powerful learning capability of long-range context and its great success in continuous recognition of emotion [8, 11, 40]. In addition, its bidirectional version, shorted as BLSTM-RNN, is utilised to capture both the past and the future contextual information.

3.1 Soft Emotion Prediction

In this work, we provide a novel format to describe the emotion prediction, i. e., a pair of indicators including the *emotional state* E and the *perception uncertainty* σ. Specifically, given a feature vector extracted from one instance as the input, two outputs will be obtained for the corresponding dimension: $(E^{(A)}, \sigma^{(A)})$ for arousal, or $(E^{(V)}, \sigma^{(V)})$ for valence.

When training the framework, the *emotional state* $E^{(i)}$, where $i \in \{A, V\}$, is acquired by performing a gold standard calculation algorithm of EWE [9] as mentioned in Section 1. The selection of EWE is mainly due to its superior performance to the method of using arithmetic mean (or median) [31] and its less complex than the algorithm of CTW [49]. Mathematically, the emotional state $E^{(i)}$ is computed by

$$E_n^{(i)} = \frac{1}{\sum_{k=1}^{K} r_k^{(i)}} \sum_{k=1}^{K} r_k^{(i)} e_{n,k}^{(i)}, \tag{1}$$

where the subscript k denotes the rater with $k = 1, ..., K$, $e_{n,k}^{(i)}$ is the annotation in any of the dimensions $i \in \{A, V\}$ of rater k for the instance n with $n = 1, \ldots, N$, and $r_k^{(i)}$ represents a rater-dependent weight and is calculated by

$$r_k^{(i)} = \frac{\sum_{n=1}^{N} (e_{n,k}^{(i)} - \mu_k^{(i)})(e_n^{\mathrm{MLE},(i)} - \mu^{\mathrm{MLE},(i)})}{\sqrt{\sum_{n=1}^{N} (e_{n,k}^{(i)} - \mu_k^{(i)})^2} \sqrt{\sum_{n=1}^{N} (e_n^{\mathrm{MLE},(i)} - \mu^{\mathrm{MLE},(i)})^2}}, \tag{2}$$

with

$$\mu_k^{(i)} = \frac{1}{N} \sum_{n=1}^{N} e_{n,k}^{(i)}, \tag{3}$$

and

$$\mu^{\mathrm{MLE},(i)} = \frac{1}{N} \sum_{n=1}^{N} e_n^{\mathrm{MLE},(i)}, \tag{4}$$

where $e_n^{\mathrm{MLE},(i)}$ denotes the Maximum Likelihood Estimator (MLE) of the instance n, which is equivalent to the mean value given K annotations:

$$e_n^{\mathrm{MLE},(i)} = \frac{1}{K} \sum_{k=1}^{K} e_{n,k}^{(i)}. \tag{5}$$

The other indicator, the *perception uncertainty* $\sigma^{(i)}$, is then represented by the standard deviation of the K annotations according to

$$\sigma_n^{(i)} = \sqrt{\frac{1}{K-1} \sum_{k=1}^{K} (e_{n,k}^{(i)} - e_n^{\mathrm{MLE},(i)})^2}. \tag{6}$$

3.2 Bidirectional Long Short-term Memory Recurrent Neural Network

In general, the BLSTM-RNN structure is composed of one input layer, one or multiple hidden layers, and one output layer [14]. The bidirectional hidden layers separately process the input sequences in a forward and a backward order and connect them to the same output layer. Compared with conventional RNN, it adopts LSTM blocks to replace the neurons in the hidden layers. Each block consists of a self-connected memory cell and three gate units, namely input, output, and forget gate. These three gates allow the network to learn when to read, write, or reset the value in the memory cell, respectively. Such a structure grants BLSTM-RNN to learn past and future context in both short and long range. For a more in-depth explanation of BLSTM-RNN, the reader is referred to [14].

3.3 Multi-task Learning

To investigate the dependency between the emotional state ($E^{(A)}$ or $E^{(V)}$) and the perception uncertainty ($\sigma^{(A)}$ or $\sigma^{(V)}$), in this paper we apply multi-task learning. In contrast to single-task learning which has one output node, multi-task learning has more nodes to match multiple targets (two in our case). With such a learning strategy, the network might be able to improve the performance of the task of interest. On the one hand, the network might better predict the emotional state, as it could learn to pay more attention to the samples with higher uncertainty. On the other hand, it might better model the perception uncertainty, as it may benefit from a general understanding of the emotional state.

Even though the empirical results presented in [8] indicate that the performance of perception uncertainty was not improved by multi-task learning, the perception uncertainty was merely regarded as a secondary task in model optimisation. Therefore, we propose to train two individual networks, with each one concentrating on one primary target. The only difference between them is the objective function during training. When calculating the objective function, different weights are assigned to the Mean Squre Error (MSE) regarding the primary target and to the MSE regarding the auxiliary target.

Specifically, when training the networks in a multi-task learning manner, a weighted average loss function $\mathcal{J}(\theta)$ is calculated by:

$$\mathcal{J}(\theta) = w_E \cdot MSE_E + w_\sigma \cdot MSE_\sigma, \tag{7}$$

with the following restriction

$$w_E + w_\sigma = 2, \tag{8}$$

where θ stands for the network parameters, MSE_E and MSR_σ represent the MSEs of the tasks of emotional state and perception uncertainty respectively, which are calculated by their estimations and their corresponding gold standards, and w_E and w_σ denote the weights of each task to regulate their contributions to $\mathcal{J}(\theta)$. The values of w_E and w_σ are optimised on the development set, by achieving a best performance of the selected primary task.

3.4 Audiovisual Late Fusion

Since the audio and video modalities are able to provide complementary information mutually, fusion strategies are therefore frequently exploited to further improve the prediction performance [11, 41]. In this light, we conduct a late fusion strategy to combine the output predictions (either the emotional states or the emotion uncertainties) from audio and video modalities. In this study, the late fusion is performed with a Simple Linear Regression (SLR) model:

$$y = \epsilon + \sum \gamma_i \cdot y_i, \tag{9}$$

where y_i denotes the original prediction with the modality i (audio or video in our case), ϵ and γ_i are the parameters estimated on the development set, and y is the final prediction.

4 EXPERIMENTS

This section is devoted to empirically investigating the proposed soft prediction approach for emotion recognition.

4.1 Data and Features

For our experiments, we utilised the German Video-chat Database within the Automatic Sentiment Analysis in the Wild (SEWA) project. This database was collected by undertaking spontaneous video-chats with 64 subjects (32 pairs), leading to a total duration of approximately 178 minutes. Specifically, each pair of subjects had a remote discussion after watching four given advertisement, and each discussion session lasted about three minutes. The discussions were recorded at a video sampling rate between 20 and 30 fps and at an audio sampling rate of either 44.1 or 48.0 KHz, depending on the recording devices used by the subjects. The dataset is available to researchers for non-commercial use at *https://db.sewaproject.eu*. Along with the audiovisual episodes and the annotations, acoustic and visual features are provided as well.

To annotate the dataset, value- and time-continuous dimensional affect ratings with respect to arousal and valence were performed by six German-speaking raters for all recording sequences. The obtained annotations were then resampled at a constant frame rate of 100 ms, and the 'gold standard' to present the emotional state was created with EWE [9] based on the resampled annotations. Besides, the 'gold standard' to denote the prediction uncertainty was created by computing the standard deviation of all six raters' annotations as described in Section 3.1. Further, we performed z-normalisation to rescale the prediction uncertainty to the range of [0, 1]. The distributions of the perception uncertainties of the corresponding emotional state in terms of arousal and valence are illustrated in Fig. 1, respectively.

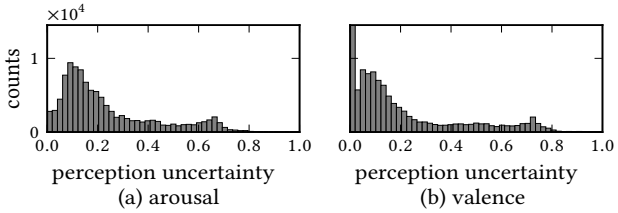


Figure 1: Distribution of perception uncertainties with respect to the arousal (a) and valence (b) dimensions.

Moreover, in order to ensure speaker-independence in the experiments, the 64 recordings were divided into three partitions, i. e., 34 recordings for the training set, 14 ones for development (or validation), and the remaining 16 ones for test. Therefore, the total number of the annotated frames in the training, development, and test set is 55 072, 22 307, and 27 597, respectively.

To graphically demonstrate the meaning of emotional state as well its perception uncertainty, we took two frame-pairs for example from a randomly selected subject in the test set for arousal (see Fig. 2) and valence (see Fig. 3), respectively. Each pair is with a similar emotional state value but different perception uncertainties. When comparing the frame-pairs in rows in Fig. 2 and Fig. 3, one may note that the right frames with higher perception uncertainties sound to show more ambiguity than the left frames with lower uncertainties from the human perception perspective.

To obtain the acoustic features, we used the established 65 Low-level Descriptors (LLDs) set from the Interspeech 2013 Computational Paralinguistic ChallengeE (ComParE) [32], which were extracted with a frame window size of 20 ms or 60 ms (for different LLDs) at a step size of 10 ms. The ComParE LLD set consists of spectral (relative spectra auditory bands 1-26, spectral energy, spectral slope, spectral sharpness, spectral centroid, etc.), cepstral (Mel frequency cepstral coefficient 1-14), prosodic (loudness, root mean square energy, zero-crossing rate, F_0 via subharmonic summation, etc.), and voice quality (probability of voicing, jitter, shimmer and harmonics-to-noise ratio). Then, the arithmetic mean and the coefficient of variance were computed over the sequential LLDs with a window size of 6 s at a step size of 100 ms to align with the annotations, resulting in an acoustic feature vector of 130 dimensions for each functional window.

Finally, to obtain the visual features, the extraction of 49-point per-frame facial landmark locations were conducted, in line with the work described in [34]. The detected face regions consist of the left and right eyebrows (five points each), the left and right eyes (six points each), the nose (nine points), the inner mouth (six points), and the outer mouth (twelve points). To reduce the variance of these landmark points, we normalised these points and then downsampled the normalised features to an interval of 100 ms again to align with the annotations.

(a) $E^{(A)} = 0.01$, $\sigma^{(A)} = 0.01$

(b) $E^{(A)} = 0.01$, $\sigma^{(A)} = 0.12$

(c) $E^{(A)} = 0.20$, $\sigma^{(A)} = 0.17$

(d) $E^{(A)} = 0.20$, $\sigma^{(A)} = 0.37$

Figure 2: Illustration of two pairs of frames ((a) vs. (b), (c) vs. (d)) with comparable emotional states but distinct perception uncertainties in *arousal*.

(a) $E^{(V)} = 0.08$, $\sigma^{(V)} = 0.17$

(b) $E^{(V)} = 0.08$, $\sigma^{(V)} = 0.79$

(c) $E^{(V)} = 0.79$, $\sigma^{(V)} = 0.47$

(d) $E^{(V)} = 0.69$, $\sigma^{(V)} = 0.70$

Figure 3: Illustration of two pairs of frames ((a) vs. (b), (c) vs. (d)) with comparable emotional states but distinct perception uncertainties in *valence*.

4.2 Implementation and Evaluation

To construct the BLSTM-RNNs, we employed two hidden layers, with 240 LSTM cells in each layer. In the training process, the learning rate and momentum was set to be 10^{-5} and 0.90, respectively. Moreover, zero mean Gaussian noise with a standard deviation of 0.2 was added to the input activations to improve generalisation. All weights of the neural networks were randomly initialised in the range from -0.1 to 0.1. All these parameters were optimised on the development set.Also, the early stopping strategy was used as no decease of MSE on the development set was observed in 20 successive epochs or the predefined maximum number of training epochs (150 in this case) had been executed. To implement the BLSTM-RNN models, we utilised the publicly available toolkit of CURRENNT [43] for the sake of reproducibility. It should be noted that, an online standardisation was carried out on the features for both development and test partitions, i. e., the means and variances of the features were calculated on the training partition and then used on the two other partitions for standardisation.

Additionally, annotation delay compensation was performed to compensate for the temporal delay between the observable cues, as seen in the recordings, and the corresponding emotion reported by the annotators [21]. In this study, we identified this delay to be four seconds which was duly compensated, by shifting the 'gold standard' back in time with respect to the audio-visual features for both arousal and valence.

Further, following the post-processing procedure of predictions in [12, 28], we applied the same chain of post-processing operations on the output predictions: smoothing, centring, scaling, and time-shifting. Likewise, all the post-processing parameters were optimised on the development set.

For multi-task learning, we investigated different values of w_E and w_σ in Eq. (7), ranging from 0 to 2 with an interval of 0.1. For late fusion, we employed the SLR algorithm (see Eq. (7)) that is implemented in the WEKA toolkit with default parameters [10]. Note that, all the parameters for multi-task learning and late fusion were again optimised on the development set.

To estimate the performance of the proposed framework, we employ Concordance Correlation Coefficient (CCC) [20] as a main metric since it has been first proposed in [27] and then widely used for continuous emotion recognition [28, 41, 48]. Formally, the CCC is calculated by

$$CCC = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}, \tag{10}$$

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

Additionally, in this work we also use Root Mean Square Error (RMSE) as another metric, which is defined as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(\hat{\theta}_n - \theta_n)^2}, \tag{11}$$

where $\hat{\theta}_n$ and θ_n are the prediction and the gold standard, respectively, for the instance n with $n = 1, ..., N$. RMSE has also been frequently reported in the literatures to assess the capability of a system for analysing continuous emotion [11, 24, 37]. In contrast to CCC, a smaller value of RMSE means a better system performance.

To further assess the significance level of performance improvement, a statistical evaluation was carried out over the whole predictions when comparing two different approaches by means of Fisher's r-to-z transformation [5]. Unless stated otherwise, a p value less than .05 indicates significance.

Table 1: *Concordance Correlation Coefficient* (CCC) of the soft predictions (i. e., the emotional states [E] and the perception uncertainties [σ]) via individual *audio* and *video* modalities, and their *late fusion* (audio+video), on the development (*dev.*) and test sets in the dimensions of arousal (A) and valence (V). Models were trained in *single*- or *multi*-task learning paradigm. The best achieved CCCs are highlighted.

CCC		*dev.*		*test*		*dev.*		*test*	
modality	*task*	$E^{(A)}$	$\sigma^{(A)}$	$E^{(A)}$	$\sigma^{(A)}$	$E^{(V)}$	$\sigma^{(V)}$	$E^{(V)}$	$\sigma^{(V)}$
audio	*single*	0.281	0.103	0.234	0.185	0.298	0.075	0.267	0.015
	multi	0.356	0.181	0.275	**0.246**	0.396	0.180	0.292	0.089
video	*single*	0.386	0.204	0.295	0.171	0.456	0.266	0.402	0.120
	multi	0.477	**0.276**	0.373	0.167	**0.588**	**0.317**	0.505	**0.153**
audio+video	*single*	0.505	0.195	0.386	0.193	0.502	0.261	0.478	0.111
	multi	**0.559**	0.273	**0.450**	0.200	0.575	0.235	**0.515**	0.110

Table 2: *Root Mean Square Error* (RMSE) of the soft predictions (i. e., the emotional states [E] and the perception uncertainties [σ]) via individual *audio* and *video* modalities, and their *late fusion* (audio+video), on the development (*dev.*) and test sets in the dimensions of arousal (A) and valence (V). Models were trained in *single*- or *multi*-task learning paradigm. The best achieved RMSEs are highlighted.

RMSE		*dev.*		*test*		*dev.*		*test*	
modality	*task*	$E^{(A)}$	$\sigma^{(A)}$	$E^{(A)}$	$\sigma^{(A)}$	$E^{(V)}$	$\sigma^{(V)}$	$E^{(V)}$	$\sigma^{(V)}$
audio	*single*	0.139	0.201	0.116	0.158	0.140	0.193	0.128	0.254
	multi	0.140	0.207	0.117	0.162	0.160	0.192	0.155	0.255
video	*single*	0.126	0.180	0.111	0.166	0.124	0.176	0.109	0.245
	multi	0.122	0.175	0.112	0.171	0.138	0.189	0.126	0.254
audio+video	*single*	0.119	0.178	0.111	**0.158**	0.119	**0.175**	0.102	0.247
	multi	**0.115**	**0.173**	**0.105**	0.159	**0.113**	0.176	**0.100**	**0.241**

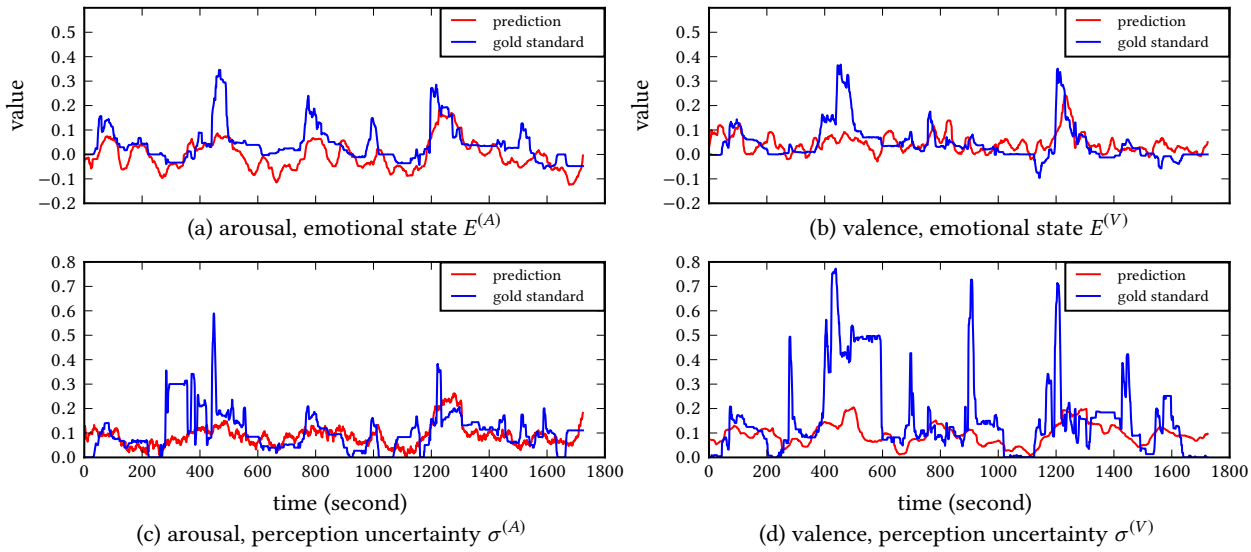
4.3 Experimental Results and Discussion

Table 1 presents the soft prediction results in terms of *CCC* via BLSTM-RNN models for the prediction of arousal and valence dimensions, whilst Table 2 demonstrates the results in terms of *RMSE*. We initiate our analysis on the performance of learning two indicators, i. e., the emotional state and the perception uncertainty independently. Generally speaking, the system performance with video modality outperforms the performance with audio in most cases on both the development and the test sets. This implicitly indicates the video cues are more informative than the audio ones on this database. Besides, this observation might be also partially due to the fact that, part of speech recordings are quiet due to the silence of subjects or only the speech of the partners; during these periods, annotations are given merely based on video signals. Therefore, Voice Activity Detection (VAD) and lip activity detection should be firstly considered to extract active audio segments in the future.

Further, when comparing single-task learning and multi-task learning, one may notice that, the performance of the latter is significantly superior to the former in most cases in terms of CCC. This suggests that the multi-task learning framework is able to exploit the dependency between the two indicators. Whereas, when comparing the corresponding RMSE results on audio or video only, we note that multi-task is not as good as single-task. A rationale behind this is that the weights w_E and w_σ in Eq. (7) were optimised by achieving a higher CCC rather than a lower RMSE.

Moreover, we find that the late fusion of the audio and video modalities (audio+video) significantly improves the performance of the emotional state recognition on the test set in terms of CCC (0.295 to 0.386 for arousal, 0.402 to 0.478 for valence). Similar findings are also confirmed in terms of RMSE. Whereas, for perception uncertainty, we notice that the improvement is not as high as the one for emotional state in terms of RMSE. This case is even worse in terms of CCC, mainly due to the employed late fusion strategy (see Section 3.4 for more detail) that aims to reduce the MSE utmost rather than increase the CCC. Meanwhile, owing to the data mismatch problem between the development and test sets, a performance improvement on the development set does not always guarantee a similar improvement on the test set. This conclusion is confirmed by the results presented in Table 1 and 2. This overfitting problem can be solved in the future by increasing the size and diversity of training data, and employing more advanced generalisation algorithms.

Overall, the best performance of the emotional state prediction on the test set is obtained at 0.450 of CCC for arousal, and 0.515 of CCC for valence when the target was generated by fusing the audio and video predictions learnt with a multi-task learning strategy. In addition, one can also notice that, the best performance



(a) arousal, emotional state $E^{(A)}$ (b) valence, emotional state $E^{(V)}$

(c) arousal, perception uncertainty $\sigma^{(A)}$ (d) valence, perception uncertainty $\sigma^{(V)}$

Figure 4: Illustration of *arousal* emotional state (a) and perception uncertainty (c) predictions, and *valence* emotional state (b) and perception uncertainty (d) predictions obtained via multi-task learning and late fusion strategies for one single subject from the test partition. The red lines denote the results of the automatic predictions, and the blue lines denote the gold standard.

of the perception uncertainty prediction on the test set is poorer, which implies that learning the patterns from the perception uncertainty is more difficult. To further demonstrate the performance of the soft prediction approach, Fig 4 illustrates the automatic predictions of arousal and valence obtained in the best settings for a single test subject. In general, the predictions generated by the proposed method can capture the trend of the gold standard. Besides, it is shown that the prediction uncertainties change more rapidly than the emotional state, which consequently gives rise to a rather tougher task. This might be raised by the various delay of annotations by each rater during the annotation process.

5 CONCLUSION

In this paper, we proposed a novel soft prediction method, towards providing a human-like emotion analysis for automatic emotion recognition systems. BLSTM-RNN regressors were utilised to predict the emotional state together with the perception uncertainty, via independently or jointly training paradigm. The experimental results evaluated on a time- and value-continuous spontaneous emotional database demonstrated that our method can achieve a promising performance. Moreover, fusing the predictions from audio and video modalities can further enhance the performance, indicating its effectiveness.

In the future, we will focus on evaluating the proposed method on more large-scale emotional datasets (e. g., IEMOCAP [3], SEMAINE [23], and RECOLA [29]) to further justify its effectiveness and robustness. More recently, deep learning algorithms have attracted tremendous attention and have achieved great success in the context of machine learning. This will continue enriching our research topics in the future, by considering diverse deep learning architectures for the soft prediction framework. Moreover, it is also possible to exploit soft prediction to tackle other subjective regression problems, or even classification. For instance, the soft prediction can be applied to tasks such as music emotion recognition [1], video recommendation systems [33], and predicting product ratings from review text [22].

ACKNOWLEDGEMENTS

This work was partially supported by the European Union's Horizon 2020 Programme through the Innovative Action No. 645094 (SEWA) and the European Union's 7th Framework Programme through the ERC Starting Grant No. 338164 (iHEARu).

REFERENCES

[1] Anna Aljanaki, Yi-Hsuan Yang, and Mohammad Soleymani. 2017. Developing a benchmark for emotional analysis of music. *PLOS ONE* 12, 3 (Mar. 2017), 1–22.
[2] Timothy W Bickmore and Rosalind W Picard. 2005. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction* 12, 2 (June 2005), 293–327.
[3] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation* 42, 4 (Nov. 2008), 335–359.
[4] Emily M Campbell, Dean F Sittig, Wendy W Chapman, Brian L Hazlehurst, and Aaron M Cohen. 2010. Understanding inter-rater disagreement: A mixed methods approach. In *Proc. American Medical Informatics Association Annual Symposium*. Washington, DC, 81–85.
[5] Jacob Cohen, Patricia Cohen, Stephen G West, and Leona S Aiken. 2013. *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge, Abingdon, UK.

[6] Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor. 2001. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine* 18, 1 (Jan. 2001), 32–80.

[7] Jun Deng, Wenjing Han, and Björn Schuller. 2012. Confidence measures for speech emotion recognition: A start. In *Proc. 10th ITG Symposium on Speech Communication*. Braunschweig, Germany, 1–4.

[8] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2012. A multitask approach to continuous five-dimensional affect sensing in natural speech. *ACM Transactions on Interactive Intelligent Systems* 2, 1, Article 6 (Mar. 2012), 29 pages.

[9] Michael Grimm and Kristian Kroschel. 2005. Evaluation of natural emotions using self assessment manikins. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*. Cancún, Mexico, 381–385.

[10] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The WEKA data mining software: An update. *ACM SIGKDD explorations newsletter* 11, 1 (June 2009), 10–18.

[11] Jing Han, Zixing Zhang, Nicholas Cummins, Fabien Ringeval, and Björn Schuller. 2016. Strength modelling for real-world automatic continuous affect recognition from audiovisual signals. *Image and Vision Computing* (Dec. 2016), no pagination.

[12] Jing Han, Zixing Zhang, Fabien Ringeval, and Björn Schuller. 2017. Reconstruction-error-based learning for continuous emotion recognition in speech. In *Proc. International Conference on Acoustics, Speech and Signal Processing*. New Orleans, LA, 2367–2371.

[13] Lang He, Dongmei Jiang, Le Yang, Ercheng Pei, Peng Wu, and Hichem Sahli. 2015. Multimodal affective dimension prediction using Deep Bidirectional Long Short-Term Memory Recurrent Neural Networks. In *Proc. International Workshop on Audio/Visual Emotion Challenge*. Brisbane, Australia, 73–80.

[14] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (Nov. 1997), 1735–1780.

[15] Brendan Jou, Subhabrata Bhattacharya, and Shih-Fu Chang. 2014. Predicting viewer perceived emotions in animated GIFs. In *Proc. ACM International Conference on Multimedia*. Orlando, FL, 213–216.

[16] Kostas Karpouzis, George Caridakis, Loic Kessous, Noam Amir, Amaryllis Raouzaiou, Lori Malatesta, and Stefanos Kollias. 2007. Modeling naturalistic affective states via facial, vocal, and bodily expressions recognition. 4451 (Mar. 2007), 91–112.

[17] Jonghwa Kim and Elisabeth André. 2009. Fusion of multichannel biosignals towards automatic emotion recognition. *Multisensor Fusion and Integration for Intelligent Systems* 35 (Mar. 2009), 55–68.

[18] Yelin Kim and Emily Mower Provost. 2015. Leveraging inter-rater agreement for audio-visual emotion recognition. In *Proc. International Conference on Affective Computing and Intelligent Interaction (ACII)*. Xian, China, 553–559.

[19] Ya Li, Jianhua Tao, Björn Schuller, Shiguang Shan, Dongmei Jiang, and Jia Jia. 2016. MEC 2016: The multimodal emotion recognition challenge of CCPR 2016. In *Proc. Chinese Conference on Pattern Recognition*. Chengdu, China, 667–678.

[20] Lawrence I-Kuei Lin. 1989. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 45, 1 (Mar. 1989), 255–268.

[21] Soroosh Mariooryad and Carlos Busso. 2015. Correcting time-continuous emotional labels by modeling the reaction lag of evaluators. *IEEE Transactions on Affective Computing* 6, 2 (Apr. 2015), 97–108.

[22] Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: Understanding rating dimensions with review text. In *Proc. ACM Conference on Recommender Systems*. Hong Kong, China, 165–172.

[23] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. 2012. The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing* 3, 1 (Jan. 2012), 5–17.

[24] Mihalis A Nicolaou, Hatice Gunes, and Maja Pantic. 2011. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing* 2, 2 (Apr. 2011), 92–105.

[25] Mihalis A Nicolaou, Hatice Gunes, and Maja Pantic. 2011. A multi-layer hybrid framework for dimensional emotion classification. In *Proc. ACM International Conference on Multimedia*. Scottsdale, Arizona, 933–936.

[26] Mihalis A Nicolaou, Vladimir Pavlovic, and Maja Pantic. 2012. Dynamic probabilistic CCA for analysis of affective behaviour. In *Proc. European Conference on Computer Vision*. Florence, Italy, 98–111.

[27] Fabien Ringeval, Florian Eyben, Eleni Kroupi, Anil Yuce, Jean-Philippe Thiran, Touradj Ebrahimi, Denis Lalanne, and Björn Schuller. 2015. Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. *Pattern Recognition Letters* 66 (Nov. 2015), 22–30.

[28] Fabien Ringeval, Björn Schuller, Michel Valstar, Shashank Jaiswal, Erik Marchi, Denis Lalanne, Roddy Cowie, and Maja Pantic. 2015. AVEC 2015: The 5th international audio/visual emotion challenge and workshop. In *Proc. ACM International Conference on Multimedia*. Brisbane, Australia, 1335–1336.

[29] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. 2013. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *Proc. International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. Shanghai, China, 1–8.

[30] Björn Schuller. 2015. Speech analysis in the big data era. In *Proc. International Conference on Text, Speech, and Dialogue*. Pilsen,Czech Republic, 3–11.

[31] Björn Schuller, Simone Hantke, Felix Weninger, Wenjing Han, Zixing Zhang, and Shrikanth Narayanan. 2012. Automatic recognition of emotion evoked by general sound events. In *Proc. International Conference on Acoustics, Speech and Signal Processing*. Kyoto, Japan, 341–344.

[32] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, and others. 2013. The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. In *Proc. INTERSPEECH*. Lyon, France, 148–152.

[33] Rajiv Ratn Shah, Yi Yu, and Roger Zimmermann. 2014. ADVISOR: Personalized video soundtrack recommendation by late fusion with heuristic rankings. In *Proc. ACM International Conference on Multimedia*. Orlando, FL, 607–616.

[34] Jie Shen, Stefanos Zafeiriou, Grigoris G Chrysos, Jean Kossaifi, Georgios Tzimiropoulos, and Maja Pantic. 2015. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *Proc. International Conference on Computer Vision Workshop*. Santiago, Chile, 1003–1011.

[35] Patrick E Shrout and Joseph L Fleiss. 1979. Intraclass correlations: Uses in assessing rater reliability. *Psychological bulletin* 86, 2 (Mar. 1979), 420–428.

[36] Ingo Siegert, Ronald Böck, and Andreas Wendemuth. 2014. Inter-rater reliability for emotion annotation in human–computer interaction: comparison and methodological improvements. *Journal on Multimodal User Interfaces* 8, 1 (Mar. 2014), 17–28.

[37] Mohammad Soleymani, Anna Aljanaki, Yi-Hsuan Yang, Michael N Caro, Florian Eyben, Konstantin Markov, Björn Schuller, Remco Veltkamp, Felix Weninger, and Frans Wiering. 2014. Emotional analysis of music: A comparison of methods. In *Proc. ACM International Conference on Multimedia*. Orlando, FL, 1161–1164.

[38] Mohammad Soleymani, Maja Pantic, and Thierry Pun. 2012. Multimodal emotion recognition in response to videos. *IEEE Transactions on Affective Computing* 3, 2 (Apr. 2012), 211–223.

[39] Stefan Steidl, Anton Batliner, Björn Schuller, and Dino Seppi. 2009. The hinterland of emotions: facing the open-microphone challenge. In *Proc. International Conference on Affective Computing and Intelligent Interaction (ACII)*. Amsterdam, Netherlands, 1–8.

[40] George Trigeorgis, Fabien Ringeval, Raymond Bruckner, Erik Marchi, Mihalis A Nicolaou, Björn Schuller, and Stefanos Zafeiriou. 2016. Adieu Features? End-to-end speech emotion recognition using a Deep Convolutional Recurrent Network. In *Proc. International Conference on Acoustics, Speech and Signal Processing*. Shanghai, China, 5200–5204.

[41] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Dennis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. 2016. AVEC 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proc. International Workshop on Audio/Visual Emotion Challenge*. Amsterdam, Netherlands, 3–10.

[42] Bogdan Vlasenko and Andreas Wendemuth. 2015. Annotators' agreement and spontaneous emotion classification performance. In *Proc. INTERSPEECH*. Dresden, Germany, 1546–1550.

[43] Felix Weninger, Johannes Bergmann, and Björn Schuller. 2015. Introducing CURRENNT: The Munich open-source CUDA recurrent neural network toolkit. *Journal of Machine Learning Research* 16, 1 (Jan. 2015), 547–551.

[44] Zhihong Zeng, Jilin Tu, Ming Liu, and Thomas S Huang. 2005. Multi-stream confidence analysis for audio-visual affect recognition. In *Proc. International Conference on Affective Computing and Intelligent Interaction*. Beijing, China, 964–971.

[45] Yue Zhang, Eduardo Coutinho, Zixing Zhang, Caijiao Quan, and Björn Schuller. 2015. Dynamic active learning based on agreement and applied to emotion recognition in spoken interactions. In *Proc. International Conference on Multimodal Interaction*. Seattle, WA, 275–278.

[46] Zixing Zhang, Nicholas Cummins, and Björn Schuller. 2017. Advanced data exploitation in speech analysis: An overview. *IEEE Signal Processing Magazine* 34, 4 (July 2017), 107–129.

[47] Zixing Zhang, Jun Deng, Erik Marchi, and Björn Schuller. 2013. Active learning by label uncertainty for acoustic emotion recognition. In *Proc. INTERSPEECH*. Lyon, France, 2856–2860.

[48] Zixing Zhang, Fabien Ringeval, Jing Han, Jun Deng, Erik Marchi, and Björn Schuller. 2016. Facing realism in spontaneous emotion recognition from speech: Feature enhancement by autoencoder with LSTM neural networks. In *Proc. INTERSPEECH*. San Francisco, CA, 3593–3597.

[49] Feng Zhou and Fernando De la Torre. 2009. Canonical time warping for alignment of human behavior. In *Proc. Advances in Neural Information Processing Systems*. Vancouver, Canada, 2286–2294.

[50] Feng Zhou and Fernando De la Torre. 2016. Generalized canonical time warping. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38, 2 (Feb. 2016), 279–294.