Dynamic Difficulty Awareness Training for Continuous Emotion Prediction

Zixing Zhang, *Member, IEEE*, Jing Han, *Student Member, IEEE*, Eduardo Coutinho, and Björn Schuller, *Fellow, IEEE*

***Abstract*—Time-continuous emotion prediction has become an increasingly compelling task in machine learning. Considerable efforts have been made to advance the performance of these systems. Nonetheless, the main focus has been the development of more sophisticated models and the incorporation of different expressive modalities (e.g., speech, face, and physiology). In this paper, motivated by the benefit of difficulty awareness in a human learning procedure, we propose a novel machine learning framework, namely, dynamic difficulty awareness training (DDAT), which sheds fresh light on the research—directly exploiting the difficulties in learning to boost the machine learning process. The DDAT framework consists of two stages: information retrieval and information exploitation. In the first stage, we make use of the reconstruction error of input features or the annotation uncertainty to estimate the difficulty of learning specific information. The obtained difficulty level is then used in tandem with original features to update the model input in a second learning stage with the expectation that the model can learn to focus on high difficulty regions of the learning process. We perform extensive experiments on a benchmark database REmote COLlaborative and affective to evaluate the effectiveness of the proposed framework. The experimental results show that our approach outperforms related baselines as well as other well-established time-continuous emotion prediction systems, which suggests that dynamically integrating the difficulty information for neural networks can help enhance the learning process.**

***Index Terms*—Emotion prediction, difficulty awareness learning, dynamic learning.**

I. Introduction

TIME-CONTINUOUS *emotion prediction* systems have received widespread interest in the machine learning (ML) community over the past decade [1]–[3]. One of the main reasons for this interest is the fact that time-continuous emotion predictions can analyse subtle and complex affective states of humans over time and play a central role in smart conversational agents that aim to achieve a natural and intuitive interaction between humans and machines [2], [4]–[7]. Great efforts have been made in this field, and most of them can generally be classified into two strands. One strand mainly focuses on designing or implementing increasingly sophisticated and robust prediction models, such as long short-term memory (LSTM)-based recurrent neural networks (RNNs) [1], [8], convolutional neural networks (CNNs) [9]–[13], and end-to-end learning frameworks [14]. Another strand mainly focuses on the integration of multiple modalities (e.g., audio and video) and modelling techniques [15], [16].

Apart from those studies, other research has recently found that emotional training data can be practically learnt in different degrees [17], [18]. That is, some data can be easily learnt given a specific model, whilst some data are relatively tough. In this light, some promising approaches have been proposed in machine learning to optimise the learning procedure. For example, the most conventional approach is associated with boosting [19], [20], which dynamically updates the weights of those samples that are hard to recognise or are even falsely recognised. Additionally, a more recent and promising approach refers to curriculum learning, which was firstly introduced in [21]. Curriculum learning presents the data from easy to hard during the training process so that the model can better avoid being caught in local minima in the presence of non-convex training criteria. Curriculum learning has become even more popular with the advance of deep learning. For emotion prediction, a handful of related studies have been reported very recently [18], [22], [23], which have shown the efficiency of curriculum learning.

However, one of the major disadvantages of these approaches is their non-friendliness to sequence-based pattern recognition tasks, such as the one we are facing. That is, in the learning process, the samples, whether or not they were presented within a sequence, are considered individually and independently. The ignored context information, nevertheless, indeed plays a vital role in sequence-based pattern recognition [24]. To this end, we propose a novel learning framework, *Dynamic Difficulty*

This work was supported in part by the TransAtlantic Platform "Digging into Data" collaboration Grant (ACLEW: Analysing Child Language Experiences Around The World), in part by the UK's Economic and Social Research Council through the research Grant HJ-253479 (ACLEW), and in part by the European Union's Horizon 2020 Programme through Research and Innovation Action 645094 (SEWA) and 645378 (ARIA-VALUSPA).

Z. Zhang is with the Group on Language, Audio and Music, Imperial College London, London, SW7 2AZ, U.K. (e-mail: zixing.zhang@imperial.ac.uk).

J. Han is with the ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg 86159, Germany (e-mail: jing.han@informatik.uni-augsburg.de).

E. Coutinho is with the Department of Music, University of Liverpool, Liverpool, L69 3BXl U.K., and also with the ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg 86159, Germany (e-mail: e.coutinho@liverpool.ac.uk).

B. Schuller is with the Group on Language, Audio and Music, Imperial College London, London, SW7 2AZ U.K., and the ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg 86159, Germany (e-mail: bjoern.schuller@imperial.ac.uk).

Awareness Training (DDAT), for time-continuous emotion prediction in this article. In contrast to the previous approaches, such as the aforementioned boosting and curriculum learning, the proposed DDAT can be well integrated into conventional context-sensitive models (e.g., LSTM-RNNs), enabling the models to ultimately exploit the context information. To the best of our knowledge, this is the first effort at exploring the difficulty information in sequence-based pattern recognition, such as the present case of time-continuous emotion prediction.

The underlying assumption of DDAT is that a model is able to deliver better performance if we explicitly let the model know the learning difficulty of the samples along with time. This assumption is in line with the finding that humans normally pay more attention to the tasks that are inherently difficult so as to perform better [25], [26].

To implement DDAT, we consider two strategies, i.e., utilising the *Reconstruction Error* (RE) of the input data or the *perception uncertainty* (PU) level of emotions as dynamic indicators of the difficulty to drive the learning process. Then, we integrate the difficulty indicator with original data for further learning, such that it endows the models with a difficulty learning awareness. This process is also partially inspired by the awareness techniques proposed for robust speech recognition [27], [28], where the noise types are considered to be auxiliary information for acoustic modelling.

In ML, RE normally serves as an objective function of an auto-encoder (AE) when extracting high-level representations. A well-designed AE is considered to reconstruct well the input from its learnt high-level representations [29]. Recently, the RE has also been exploited for tasks, such as anomaly detection [30], [31] and classification [32]. For anomaly detection, an AE is trained on normal samples beforehand to serve as a novel event detector. When feeding a new sample into the AE, the obtained RE compared with a predefined threshold decides whether it is abnormal [30], [31]. For classification, several class-specific AEs are pre-trained separately. When feeding an unknown sample into these AEs simultaneously, the values of the corresponding RE are then interpreted as indicators of class membership [32]. Notably, all these works hypothesise that data with the same label have similar data distributions. That is, the mismatched data potentially result in higher REs than those of the matched data. This motivates us to employ the RE as a learning difficulty index because it is well known in ML that mismatched data severely promote the complexity of modelling [33].

In addition, PU is a term employed in subjective pattern recognition tasks to refer to the inter-annotator disagreement level when calculating a gold standard in an annotation process [34]. For emotion prediction, it has been frequently determined that the PU has a high correlation with the learning difficulty of a recognition model. For example, the reported work in [35] and [36] found that the emotion prediction systems perform better in low-uncertainty regions than in high-uncertainty regions. Likewise, the findings in [17] showed that the elimination of the samples labelled with a high uncertainty from the training set leads to a better emotion prediction model. This finding provokes us to use the PU as another learning difficulty index. It is also worth noting that the principle of PU-based strategy constrains its application to subjective pattern recognition tasks. Despite the fact that the concept of 'uncertainty' was employed in previous emotion prediction work, it was calculated among multiple predictions from variable systems [37], which significantly differs from the definition of PU in this article, or merely utilised for multi-task learning [38] (cf. Section II).

Motivated by the above analysis and following our previous tentative work [39], where only the RE was investigated for emotion prediction in speech, we demonstrate in this paper that the proposed DDAT framework can aid the ML models in detection of 'moments' in the learning process that are of higher difficulty in the context of audiovisual time-continuous emotion prediction. More specifically, the contributions of the present article include the following: (i) proposing a new framework that exploits knowledge about the learning difficulty of the samples during the learning process for time-continuous emotion prediction; (ii) introducing and analysing two specific strategies (i.e., based on RE or PU) to implement this framework; (iii) presenting a dynamic tuning approach to further dynamically tune the predictions; and (iv) comprehensively evaluating the effectiveness of the proposed framework on a benchmarked audiovisual emotion prediction database.

The remainder of this article is organised as follows. In Section II, we briefly review past and related studies. In Section III, we present a detailed description of the structure and algorithm of the proposed DDAT framework. Then, in Section IV, we offer an extensive set of experiments conducted to exemplify the effectiveness and robustness of the DDAT framework along with a discussion. Finally, we present our conclusions and future research directions in Section V.

II. Related Work

For continuous emotion prediction, plenty of novel approaches have been proposed and investigated over the past decade. Some approaches expect to design or implement a more sophisticated and robust prediction model [1], [8]–[11], [14]. Given that context information is crucial for estimating sequential patterns (continuous emotion prediction in our case), recurrent neural networks (RNNs), especially the ones implemented with long short-term memory (LSTM) cells, were introduced [1], and they are still amongst current state-of-the-art models [40]. One of the main advantages of LSTM-RNNs is that they can model long-range dependencies between sequences [24], [41], and therefore, they are efficient in capturing the temporal information of emotional expression [1]. More recently, the so-called end-to-end network architecture has been emerging as a promising network structure, which can automatically derive representations directly from raw (unprocessed) data, rather than manually extracting hand-crafted features. For example, in [14], Tzirakis *et al.* jointly trained the CNNs at the front end and the LSTM-RNNs at the back end, where the CNNs mainly take charge of extracting representations from raw audio signals and the concatenated LSTM-RNNs are responsible for capturing the temporal information. A similar framework has also been shown in [42].

Meanwhile, some other approaches attempt to overcome the drawbacks of individual models by means of integrating multiple different modalities or models in an ensemble strategy [15], [16]. One common approach when considering multiple modalities is *early* (aka *feature*-level) fusion of unimodal information. This is typically achieved by concatenating all the features from multiple modalities into one combined feature vector, which is then used as the input information for the models [16], [43]–[45]. A benefit of early fusion is that it can provide better discriminative ability to the model by exploiting the complementary information that exists among different modalities. For example, acoustic features empirically outperform visual features for arousal estimation, whereas the opposite occurs for valence estimation [43]. Another frequently employed approach is *late* (aka *decision*-level) fusion, which involves the combination of predictions obtained from diverse learners (models) to determine the final prediction. To build the diverse learners, Wei *et al.* [46] created an ensemble of LSTM-RNN learners that were trained on different modalities (e.g., audio and video), whereas Qiu *et al.* [47] developed a variety of topology structures of deep belief networks (DBN). To combine the predictions from multiple learners, a straightforward approach applies (un-)weighted averaging, such as simple linear regression (SLR) [45], [48]. Another common approach is stacking, whereby the predictions from different learners are stacked and used as inputs of a subsequent non-linear model that is trained to make a final decision [46], [47], [49]. In order to leverage the individual advantages of different models, Han *et al.* [16] further proposed a *strength modelling* framework that concatenates two different models in a hierarchical architecture. In this approach, the prediction yielded by the first model is concatenated with the original input features, and this expanded feature vector is then set as the input to the next model.

All of the outlined approaches above merely focus on either extending the capability or overcoming the drawbacks of the learning model. Difficulty information in the learning process, however, has seldom been exploited to date, to the best of our knowledge.

Moreover, DDAT relates to *multi-task learning* (MTL) as well [6], [38], [50], [51]. In [51], Deng *et al.* reconstructed the inputs with an AE as an auxiliary task for emotion prediction in a semi-supervised manner, and they demonstrated that the AE can distill representative high-level features from large-scale unlabelled data. In [38], Han *et al.* proposed utilisation of the PU as an auxiliary task for continuous and dimensional emotion prediction, and they found that this information helps improve performance. In [50], Nicolaou *et al.* introduced an output-associative framework to learn the correlations and patterns among different emotional dimensions (i.e., arousal *and* valence). In this framework, the arousal and valence predictions from independent models are fused together and fed into a consequential model for a final prediction (i.e., arousal *or* valence). The effectiveness of this approach has been replicated in [52] and [53].

Analogous to MTL, the present DDAT framework considers the tasks of reconstructing inputs or predicting perception uncertainty to be auxiliary tasks. Nevertheless, the RE and the PU are further assumed to be the learning difficulty indicators, and the model inputs are dynamically updated in order to endow the model with a difficulty-aware learning capability.

Algorithm 1: Dynamic Difficulty Awareness Training

```
Initialise:
$h$: neural networks;
$\mathbf{x}$: feature vector, $\mathbf{x} = [x_1, x_2, \ldots, x_r]$ $I, J$: predefined training epochs

if ontology-driven then
    auxiliary task $\mathcal{T} \leftarrow$ reconstructing input;
else if content-driven then
    auxiliary task $\mathcal{T} \leftarrow$ predicting perception uncertainty;
end

% retrieving difficulty information stage
for $i = 1, ..., I$ do
    optimise $h$ by minimising a joint loss function
      $\mathcal{J}(\boldsymbol{\theta}_0) = w_1 * L_{emt}(\cdot) + w_2 * L_{aux}(\cdot) + \lambda R(\boldsymbol{\theta}_0)$,
      where $w_1$ and $w_2$ regulate the contributions of the
      emotion prediction $L_{emt}(\cdot)$ and the auxiliary task
      prediction $L_{aux}(\cdot)$;
    evaluate $h$ on the development set for emotion
      prediction: $CCC_{val,i}$;
    if $CCC_{val,i} > CCC_{val,i-1}$ then
        save $h$;
    end
end
obtain the difficulty attention $\mathbf{d}$ based on the chosen
  auxiliary task $\mathcal{T}$;

% exploiting difficulty information stage
for $j = 1, ..., J$ do
    update the input feature vector: $\mathbf{x}' = [\mathbf{x}, \mathbf{d}]$;
    optimise $h$ by minimising the loss function $\mathcal{J}(\theta_{emt})$
      for emotion prediction;
    evaluate $h$ on the development set for emotion
      prediction: $CCC_{val,j}$;
    if $CCC_{val,j} > CCC_{val,j-1}$ then
        save $h$;
    end
end
```

III. Dynamic Difficulty Awareness Training

In this section, we describe the DDAT framework. Let $\mathbf{x} \in \mathcal{X}$ denote the feature vector in the input feature space, and $y \in \mathcal{Y}$, the label in the emotion label space. For a sequential pattern recognition task in our case, $\mathbf{x}_t$ thus indicates a feature vector at the t-th frame extracted from an utterance.

A. System Overview

The pseudo-code describing the proposed algorithm is presented in Algorithm 1. It consists of two main stages: (i) *retrieving* difficulty information and (ii) *exploiting* difficulty information. In the first stage, in order to extract and indicate the information related to the difficulty of the learning process, we propose two different strategies: *ontology*- and *content-driven* strategies.

The *ontology-driven* strategy focuses on the model itself. Specifically, we determine the difficulty of the task through the reconstruction of the input information, assuming that the RE is a proxy for its learning capability in a given moment.

On the contrary, the *content-driven* strategy focuses on the data and assumes that different data can be learnt to different degrees. That is, some data can be easily learnt with a specific model, whereas other data can be difficult. This approach partially stems from curriculum learning [21], which has demonstrated that each datum cannot be equally learnt so as to be well-organised for model training. In the field of emotion prediction, a few studies have shown that the difficulty-level of the data to be learnt is closely related to its PU [35], [36], as discussed in Section I. Inspired by these studies, we employ the PU to represent the difficulty and complexity of the samples.

In the second stage, we concatenate the original features $\mathbf{x}_t$ with the difficulty vector $\mathbf{d}_t$ retrieved by one of the aforementioned two strategies, update the inputs via $[\mathbf{x}_t, \mathbf{d}_t]$ and re-train the regression model for continuous emotion prediction. Due to the fact that $\mathbf{d}_t$ varies along with time, the extended difficulty vector provides dynamic awareness when modelling $\mathbf{x}$ in a continuum.

B. Multi-Task Learning

MTL is a process of learning multiple tasks *concurrently*. Typically, there is one main task and one or more auxiliary tasks. By attempting to model the auxiliary tasks together with the main task, the model learns shared information among tasks, which may be beneficial to learning the main task. Mathematically, the objective function in MTL can be formatted as:

$$\mathcal{J}(\boldsymbol{\theta}_0) = \sum_{m=1}^{M} w_m L_m(\mathbf{x}, y_m; \boldsymbol{\theta}_m) + \lambda R(\boldsymbol{\theta}_0), \qquad (1)$$

where M denotes the number of tasks and $L_m(\cdot)$ represents the loss function of the task m, which is weighted by w_m. $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_m$ represent, respectively, the general model parameters and the specific ones with respect to task m, and λ is a hyperparameter that controls the importance of the regularisation term $R(\boldsymbol{\theta}_0)$.

In this article, in order to infer the difficulty of the information being modelled in the first stage of the DDAT framework, we use an MTL structure to jointly learn continuous emotion prediction together with the reconstruction of the input features or the PU prediction. The rationale is twofold: On the one hand, the model makes better use of MTL for continuous emotion prediction. The benefit of MTL has been shown by several studies for emotion prediction, as described in Section II. On the other hand, the model takes one network, rather than two [39], to explore the difficulty of the learning process.

C. Ontology-Based Difficulty Information Retrieval

Figure 1 illustrates the framework of RE-based DDAT, where the difficulty indicator $\mathbf{d}$ is generated from the reconstruction process of the inputs. As described in Section III-B, the employed network is trained in an MTL context, and so the output includes two paths–the emotion prediction path and the AE path. The former is trained in a supervised fashion, whereas the latter is trained in an unsupervised manner. Thus, there are two tasks to be conducted when training the network, i.e., predicting emotions and reconstructing inputs. Specifically, given a time sequence as input $\mathbf{x}$, the network is optimised by minimising the loss function as

$$\mathcal{J}(\boldsymbol{\theta_0}) = w_1 * L_{emt}(\cdot) + w_2 * L_{re}(\cdot) + \lambda R(\boldsymbol{\theta_0}), \qquad (2)$$

where $L_{emt}(\cdot)$ and $L_{re}(\cdot)$ denote the loss functions for emotion prediction and input reconstruction, respectively. To calculate them, we take the mean square error (MSE) for both learning paths, i.e., for emotion prediction,

$$L_{emt}(\cdot) = \sum_{t=1}^{T} ||\hat{y}_t - y_t||^2; \qquad (3)$$

and for the input reconstruction,

$$L_{re}(\cdot) = \sum_{t=1}^{T} ||\hat{\mathbf{x}}_t - \mathbf{x}_t||^2, \qquad (4)$$

where $\mathbf{x}_t$ and y_t are a sample and its annotation at time t from an input sequence with a period of time T, respectively. $\hat{\mathbf{x}}_t$ and $\hat{y}_t$ denote the network predictions to reconstruct its inputs $\mathbf{x}_t$ and estimate the emotions y_t, respectively.

It is expected that $L_{re}(\cdot) \rightarrow 0$ if the model is sufficiently powerful and robust. However, empirical experiments have shown that the results are far from this expectation. Previous findings frequently indicate that a higher distribution mismatch between the given data and the entire training dataset is inclined to produce a higher RE [30], [31], [54], [55]. Therefore, the RE somewhat implies the difficulty degree of the model to learn such data or, in other words, reflects the difficulty of the data to be learnt by the model.

Once the model is trained, the difficulty of the learning process $(\mathbf{d})$ can be obtained by computing the distance between the input $\mathbf{x}$ and its corresponding reconstruction $\hat{\mathbf{x}}$. The distance can be either a vector $\mathbf{e}$ calculated by,

$$\mathbf{d} = \mathbf{e} = \mathbf{x} - \hat{\mathbf{x}}, \qquad (5)$$

or a scalar E summed over all attributes, i.e.,

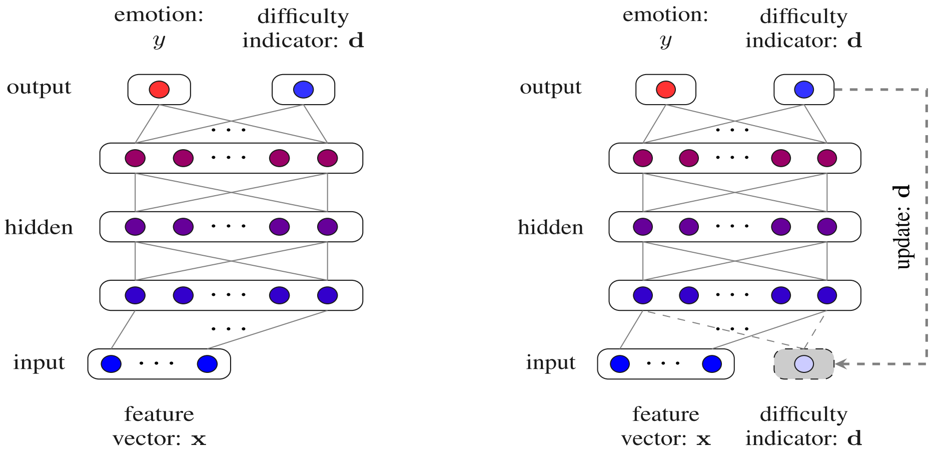
$$\mathbf{d} = [E] = \left[\sum_{i=1}^{r}(x_i - \hat{x}_i)\right], \qquad (6)$$

where $\mathbf{x} = [x_1, x_2, \ldots, x_r]$ and r is the dimension of the feature vector.

In the difficulty exploitation stage, we update the model input with the new vector, i.e., $\mathbf{x}' = [\mathbf{x}, \mathbf{e}]$ or $\mathbf{x}' = [\mathbf{x}, E]$. In doing this, the input feature vectors are of $2r$ or $r + 1$ dimensions when feeding back an error vector or scalar.

D. Content-Based Difficulty Information Retrieval

As mentioned earlier, *PU* is an indicator of the uncertainty level of the perception of an emotional state for a given observed sample. In the context of affective computing, we deem


emotion: y
difficulty indicator: $\mathbf{d}$
output
hidden
input
feature vector: $\mathbf{x}$
emotion: y
difficulty indicator: $\mathbf{d}$
output
hidden
input
update: $\mathbf{d}$
feature vector: $\mathbf{x}$
difficulty indicator: $\mathbf{d}$

(a) Difficulty information retrieving stage

(b) Difficulty information exploiting stage

Fig. 1. Dynamic Difficulty Awareness Training (DDAT) includes difficulty information (a) retrieving stage and (b) exploiting stage. Difficulty information is indicated by either the input reconstruction error (i.e., an error vector or the sum of all errors), or the emotion perception uncertainties.

that emotion prediction is a subjective task that differs from many other objective pattern recognition tasks, such as face recognition, that hold a ground truth [56]. In order to obtain a gold standard for a subjective task, it is required that a sufficient number of raters observe the same sample and that their ratings are collapsed in order to eliminate as much as possible individual variations in perception and rating. In this case, a possible way to infer uncertainty is by calculating the *inter-rater disagreement level*, which assumes that for each sample, the personal PU is highly correlated with the inter-rater disagreement level [38], [57].

In this study, the PU $u^{(i)}$, $i \in \{arousal, valence\}$, is represented by the standard deviation of the K annotations as

$$u_n^{(i)} = \sqrt{\frac{1}{K-1}\sum\nolimits_{k=1}^{K}(y_{n,k}^{(i)} - \bar{y}_n^{(i)})^2}, \tag{7}$$

where $\bar{y}_n^{(i)}$ denotes the mean value given K annotations:

$$\bar{y}_n^{(i)} = \frac{1}{K}\sum_{k=1}^{K} y_{n,k}^{(i)}. \tag{8}$$

The framework of PU-based DDAT is also illustrated in Fig. 1, where the difficulty indicator $\mathbf{d}$ is determined by the perception uncertainty. The designed network includes an emotion prediction path and a PU prediction path, both of which are jointly trained in a supervised manner. Therefore, the objective function of Eq. (1) can be re-formulated as

$$\mathcal{J}(\boldsymbol{\theta_0}) = w_1 \cdot L_{emt}(*) + w_2 \cdot L_{pu}(*) + \lambda \cdot R(\boldsymbol{\theta_0}). \tag{9}$$

$L_{pu}(*)$ stands for the loss functions for PU prediction, and it is expressed by

$$L_{pu}(*) = \sum_{t=1}^{T} ||\hat{u}_t - u_t||, \tag{10}$$

where u_t is a PU value for the sample at time t from input sequences with time T.

Once the network is optimised in the first learning stage, its input will then evolve to $\mathbf{x}' = [\mathbf{x}, u]$ with $r+1$ dimensions in the second learning stage.

E. Late Fusion and Dynamic Tuning

As discussed in Section II, late fusion approaches have been frequently shown to be effective for continuous emotion prediction [15], [16], [48] due to the fact that complementary information can be provided by the various modalities or models [15], [16], [48]. In this light, we conduct a late fusion to combine the emotion predictions from different modalities, learning models, or a combination thereof. The late fusion is performed with an SLR approach:

$$y = \epsilon + \sum \gamma_i \cdot y_i, \tag{11}$$

where y_i denotes the original prediction with the modality (i.e., audio or video) or model i (i.e., RE- or PU-based DDAT), ϵ and γ_i are the parameters estimated on the development set, and y is the fused prediction.

Despite the effectiveness of SLR, this conventional fusion approach simply assumes that the predictions $y_{i,t}$ in a continuum are considered to be equally important for each prediction stream y_i. This means that the parameter of γ_i remains a constant in time, given a set of y_i, and therefore, this approach ignores the changes of the reliability of the predictions along time. To address this problem, we further propose a *dynamic tuning* strategy according to the reliability of predictions in time.

Mathematically, we applied an additional SLR on the original prediction $y_{i,t}$ and the corresponding difficulty indicator $d_{i,t}$ at time t:

$$y'_{i,t} = \epsilon + \gamma_i \cdot y_{i,t} + \gamma_d \cdot d_{i,t}, \tag{12}$$

where $d_{i,t}$ is represented by E_t for the RE-based DDAT systems or u_t for the PU-based DDAT systems. Intuitively, the prediction is dynamically tuned by the difficulty information.

IV. Experiments and Results

To evaluate the effectiveness of the proposed methods, we conducted extensive experiments with the benchmark database of the AudioVisual Emotion Challenges (AVEC) from 2015 [58] and 2016 [48].

A. Databases and Features

The multimodal corpus REmote COLlaborative and Affective interactions (RECOLA) [59] (a standard database of the AVEC challenges for audiovisual time-continuous emotion prediction [48], [58]) was selected for our experiments due to its widespread use in this area. This database was created to study socio-affective behaviours from multimodal data in the context of remote collaborative tasks. It includes audiovisual (and physiological) recordings of spontaneous and natural interactions from 27 French-speaking participants whilst solving a collaborative task conducted in dyads via video conferencing. The corpus is comprised of audio, video, and peripheral physiology recordings that were obtained synchronously and continuously over time.

In order to ensure speaker-independence for ML experiments, the corpus was divided into three partitions–training, development (validation), and testing–with each partition containing nine collaborative sessions. This division is balanced in terms of gender, age, and mother tongue of the participants. The corpus contains value- and time-continuous annotations of two affective dimensions–arousal and valence–that were obtained from six French-speaking raters (three female) for the first five minutes of each audiovisual recording. The obtained labels were then resampled at a constant frame rate of 40 ms and averaged over all raters to create a 'gold standard' for each instance. Interrater disagreements were also computed for all instances [59]. For our experiments, we only made use of audio and video signals.

The acoustic and visual features employed in our experiments are the same sets used to compute the AVEC 2015 and 2016 baselines for fair comparison with other methods. The acoustic features consist of the extended Geneva Minimalistic Acoustic Parameter Set (*eGeMAPS* [60]). Since the RECOLA database contains long time-continuous signals and annotations, two functionals (arithmetic mean and standard derivation) were applied over the sequential low-level descriptors (LLDs, e.g., pitch, loudness, energy, Mel Frequency Cepstral Coefficients, jitter, and shimmer) over a fixed window of 8 s with a 40 ms step. This resulted in a set of 88 acoustic features per segment.

In relation to the visual features, we utilised both the *appearance* and *geometric* standard features of the AVEC challenges. The appearance features were computed by using local Gabor binary patterns from three orthogonal planes through splitting the video into spatio-temporal video volumes. A feature reduction was then performed by applying a principal component analysis from a low-rank (up to rank 500) approximation, leading to 84 features representing 98% of the variance. To extract the geometric features, 49 facial landmarks were firstly extracted from each frame and then aligned with a mean shape from stable points (located on the eye corners and on the nose region). This resulted in 316 features per frame: i.e., 196 features were obtained by computing the difference between the coordinates of the aligned landmarks and those from the mean shape and between the aligned landmark locations in the previous and the current frame, and 71 were obtained by calculating the Euclidean distances (L2-norm) and the angles (in radians) between the points in three different groups. An additional 49 features correspond to the Euclidean distance between the median of the stable landmarks and each aligned landmark in a video frame.

Similar to the acoustic features, the arithmetic mean and the standard derivation were computed over the sequential visual features of each frame using a sliding window of 8 s with a step size of 40 ms. This process led to 168 appearance and 632 geometric visual features.

For full details on the database and feature sets, please refer to [48], [58]. Note that we obtained 67.5 k extracted segments in total for each partition (training, development, or test).

B. Experimental Setup and Evaluation Metrics

The implemented DDAT framework in our experiments consists of a deep RNN (DRNN) equipped with gated recurrent units (GRUs) [61]. GRUs are an alternative to long short-term memory units, which can also capture the long-term dependencies in sequence-based tasks and mitigate the effects of the vanishing gradient problem [61]. Compared to LSTM units, GRUs have fewer parameters due to the fact that they do not have separate memory cells and output gates, which results in a faster training process and a less-training-data demand for achieving a good generalisation. Most importantly, many empirical evaluations [62] have indicated that GRUs perform as competitively as LSTM units.

The DRNN structure was optimised in terms of the number of hidden layers and the number of GRUs per layer in the development phase. We applied a search grid that is comprised of $\{1, 3, 5, 7, 9\}$ hidden layers and $\{40, 80, 120\}$ hidden units per layer. For each learning strategy, we always choose the best performing network structure in order to alleviate the impact of the variation of network structures on the system performance. The training of the DRNNs was conducted using the Adam optimisation algorithm [63] with an initial learning rate of 0.001. To facilitate the training process, we set the size of mini-batch to four. Additionally, an online standardisation was applied to the input data by using the means and variations of the training set.

Additionally, as suggested in [48], annotation delay compensation was employed to compensate for the temporal delay between the observable cues and the corresponding annotations reported by the annotators [64]. We identified this delay to be 2.4 s, according to a series of experimental evaluations in [65], and shifted the gold standard back in time with respect to the features for all modalities and tasks in our experiments.

In order to evaluate the performance of the models, we took the official metric of the AVEC 2015 and 2016 challenges–the *Concordance Correlation Coefficient* (CCC) [58]:

$$r_c = \frac{2r\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}, \tag{13}$$

where r represents *Pearson's correlation coefficient* between two time series (e.g., prediction and gold-standard), μ_x and μ_y denote the mean of each time series, and σ_x^2 and σ_y^2 stand for the corresponding variances. Compared with PCC, the CCC considers not only the shape similarity between two series but also the value precision. This is especially relevant for estimating the performance of time-continuous emotion prediction models, as both the trends as well as absolute prediction values are relevant for describing the performance of a model. The CCC metric falls in the range of $[-1, 1]$, where $+1$ represents perfect concordance, -1 total discordance, and 0 no concordance at all.

To refine the obtained prediction, we further performed a chain of post-processing, including median filtering, centering, scaling, and time-shifting, as suggested in [48], [58]. The filtering window size W (ranging from 0.12 s to 0.44 s at a rate of 0.08 s) and the time-shifting delay D (ranging from 0.04 s to 0.60 s at a step of 0.04 s) were optimised using a grid search method. All the post-processing parameters were optimised on the development set and then applied to the test set. Therefore, those post-processing parameters had various settings for different tasks.

To compare the proposed DDAT approach with other related and state-of-the-art approaches, we further conducted *curriculum learning*, as introduced in Section I. We particularly selected the criterion of 'disagreement between annotators' (i.e., PU in this article) as an example because it is appropriate for the task at hand and also superior to other criteria [23]. To retain the optimised setups, we continued using the deep neural networks (DNNs) equipped with two hidden layers (1 024 nodes per layer) and split the whole training set into five parts based on the PU levels. Moreover, we implemented it with GRU-RNNs as well for a fair performance comparison between the curriculum learning and the proposed DDAT.

Finally, to statistically compare the various experiments conducted with the AVEC challenges baselines, we carried out the *Fisher r-to-z transformation* [66]. In detail, given two distributions X and Y [the pairs $(X_i, Y_i) \sim$ i.i.d.] that have a bivariate normal distribution with correlation, the Fisher transformation z is approximately normally distributed with mean

$$m = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) = \operatorname{arctanh}(r), \tag{14}$$

and standard error

$$\sigma = \frac{1}{\sqrt{N-3}}, \tag{15}$$

where N is the sample size and r is the true correlation coefficient. Theoretically, the Fisher transformation is exceptionally efficient for small sample sizes because the sampling distribution of the Pearson correlation is normally highly skewed.

After the Fisher transformation, a one-tailed test was performed to compare two distributions. A p-value lower than .05 indicates a significant difference. It is noted that r is replaced with r_c (CCC) due to the efficiency of r_c in this article.

C. Emotion Prediction With Dynamic Difficulty Awareness Training

The performance of the evaluated systems *before* and *after* post-processing the predictions for both arousal and valence targets is presented in Tables I and II, respectively. To investigate the proposed DDAT framework, we not only conducted the traditional single-task learning but also the MTL for comparison, with three different feature sets–one acoustic feature set (eGeMAPS) and two visual feature sets (appearance and geometric features), as described in Section IV-A. It is worthy to note that the network structure employed for each modality and learning approach was respectively optimised in the constrained parameter space, as mentioned in Section IV-B. Then, the best performing network structures were employed for performance comparison. Doing this largely alleviates the inconsistent impact on the system performance due to the variation of network structures. From the comparison of both Tables I and II, it can be seen that the post-processing of the model predictions generally leads to better performance. For instance, the best baselines for arousal and valence are respectively boosted from 0.617 to 0.652 CCC with acoustic features (eGeMAPS) and from 0.403 to 0.417 CCC with visual features (geometric). Similar observations can also be obtained in the MTL systems and the proposed DDAT systems; e.g., for the MTL systems, the CCCs are increased from 0.613 to 0.654 with the eGeMAPS feature set for arousal and from 0.487 to 0.488 with the geometric feature set for valence. Given these results, we henceforth focus on analysing the experiments with the post-processing step (cf. Table II).

For the baseline system, the obtained results are competitive to, or even better than, the benchmark of the emotion prediction subchallenge in the AVEC 2016 [48] over three information streams and two prediction tasks. These results support previous findings showing that GRUs can deliver competitive performance when compared to LSTM units [61], [62].

When training the networks jointly with input reconstruction (RE-based MTL) or perception uncertainty prediction (PU-based MTL), one can observe that the systems slightly outperform the baseline systems in nine out of twelve cases on the test set. This indicates that there is a substantial relationship between the two jointly learnt tasks. To be more specific, the representations from the last neural network hidden layer, which are learnt synchronously from the emotion prediction and other auxiliary tasks (i.e., reconstructing the input or predicting the perception uncertainty), potentially further benefit the emotion prediction.

We further implemented the curriculum learning approach as well as its baseline by means of DNNs [23] and GRU-RNNs. From Table I, it can be seen that the DNNs perform unsurprisingly worse than GRU-RNNs, mainly due to their limited capability of capturing the context information [24]. When feeding the data to the training model from a low-difficulty level to a high-difficulty level, the performance of the models is remarkably boosted in all scenarios. Nevertheless, it is still not competitive with the DDAT models in most cases. Moreover, it is observed that the GRU-RNN-based curriculum learning

TABLE I
System Performance (Concordance Correlation Coefficient; CCC) **before** Post-Processing the Model Predictions for the Conventional Single-Task Learning (Baseline) Framework, the Multi-Task Learning (MTL) Framework, and the Proposed Dynamic Difficulty Awareness Training (DDAT) Framework Using Reconstruction Error (RE, a Vector or a Scalar of sum) and Perception Uncertainty (PU) Variants. These Results Pertain to the Experiments Conducted on the *dev*elopment and *test* Partitions for Both *aro*usal and *val*ence Targets. Three Feature Sets (Audio-EGeMAPS, Video-Appearance, and Video-Geometric) Were Employed to Evaluate all Approaches. The Cases WQhere DDAT has a Statistical Significance of Performance Improvement Over MTL are Marked by the "⋆" Symbol

CCC	audio-eGeMAPS				video-appearance				video-geometric			
	aro		val		aro		val		aro		val	
	dev	test	dev	test	dev	test	dev	test	dev	test	dev	test
baseline	.743	.617	.460	.380	.501	.416	.481	.391	.407	.256	.598	.403
Multi-task learning (MTL)												
RE-based	.743	.590	.513	.298	.472	.434	.512	.351	.429	.283	.632	.487
PU-based	.727	.613	.485	.417	.459	.426	.444	.342	.442	.284	.630	.444
Proposed Dynamic Difficulty Awareness Training (DDAT)												
RE-based (vector)	.745	.605⋆	.485	.374⋆	.473	.429	.482	.326	.509⋆	.299⋆	.627	.464
RE-based (sum)	.783⋆	.671⋆	.495	.410⋆	.487⋆	.464⋆	.507	.460⋆	.478⋆	.359⋆	.633	.467
PU-based	.769⋆	.623⋆	.493	.397	.478⋆	.457⋆	.476⋆	.412⋆	.450	.336⋆	.629	.500⋆
Other state of the art												
DNNs [23]	.216	.362	.003	.004	.265	.173	.294	.174	.017	.011	.193	.128
Curriculum learning (DNN) [23]	.445	.533	.018	.006	.402	.292	.437	.444	.199	.116	.271	.245
Curriculum learning (GRU-RNN)	.711	.600	.494	.335	.406	.366	.547	.479	.392	.291	.582	.497

TABLE II
System Performance (Concordance Correlation Coefficient; CCC) **after** Post-Processing the Model Predictions for the Conventional Single-Task Learning (Baseline) Framework, the Multi-Task Learning (MTL) Framework, and the Proposed Dynamic Difficulty Awareness Training (DDAT) Framework Using Reconstruction Error (RE, a Vector or a Scalar of sum) and Perception Uncertainty (PU) Variants. These Results Pertain to the Experiments Conducted on the *dev*elopment and *test* Partitions for Both *aro*usal and *val*ence Targets. Three Feature Sets (Audio-eGeMAPS, Video-Appearance, and Video-Geometric) Were Employed to Evaluate all Approaches. The Best Results Achieved on the Test set are in Bold. The Cases Where DDAT has a Statistical Significance of Performance Improvement Over MTL are Marked by the "⋆" Symbol

CCC	audio-eGeMAPS				video-appearance				video-geometric			
	aro		val		aro		val		aro		val	
	dev	test	dev	test	dev	test	dev	test	dev	test	dev	test
baseline	.783	.652	.473	.400	.528	.403	.493	.404	.523	.314	.620	.417
Multi-task learning (MTL)												
RE-based	.788	.629	.519	.331	.512	.425	.529	.366	.502	.324	.632	.488
PU-based	.803	.654	.506	.416	.502	.406	.468	.418	.508	.327	.643	.452
Proposed Dynamic Difficulty Awareness Training (DDAT)												
RE-based (vector)	.806⋆	.676⋆	.517	.378⋆	.533⋆	.434	.520	.329	.559⋆	.355⋆	.634	.473
RE-based (sum)	.807⋆	**.694**⋆	.508	**.422**⋆	.539⋆	.437⋆	.528	**.457**⋆	.544⋆	**.400**⋆	.639	.471
PU-based	.811	.664⋆	.498	.407	.518⋆	**.438**⋆	.514⋆	.431⋆	.513	.397⋆	.632	**.501**⋆
Other state of the art												
DNNs [23]	.573	.517	.129	.044	.387	.220	.306	.206	.312	.296	.362	.216
Curriculum learning (DNN) [23]	.687	.591	.159	.174	.417	.343	.446	.419	.394	.267	.300	.269
Curriculum learning (GRU-RNN)	.754	.611	.501	.357	.491	.391	.557	.492	.444	.336	.609	.500
Strength modelling [16]	.755	.666	.476	.364	.350	.196	.592	.464	–	–	–	–
End-to-end [14] [a]	.786	.715	.428	.369	.371	.435	.637	.620	–	–	–	–
Feature selection + offset [49] [b]	.800	–	.398	–	.587	–	.441	–	.173	–	.441	–
SVR + offset [48] [c]	.796	.648	.455	.375	.483	.343	.474	.486	.379	.272	.612	.507
SC + CNN + LSTM [40] [d]	.846	–	.450	–	.346	–	.511	–	–	–	–	–

Note: "–" indicates that the corresponding CCC is not provided.
a acoustic and visual features automatically extracted by deep neural network models
b AVEC '15 challenge winner method
c AVEC '16 baseline method
d AVEC '16 challenge winner method

outperforms the DNN-based system mainly due to the learning capability of GRUs.

The performance of the MTL systems is further enhanced by the proposed DDAT framework, as shown in Table II. In particular, the performance of the DDAT system for arousal and valence regressions respectively reaches CCC values of 0.694 and 0.422 with the audio-eGeMAPS feature set, 0.438 and 0.457 with the video-appearance feature set, and 0.400 and 0.501 with the video-geometric feature set. These results demonstrate that the DDAT systems significantly outperform ($p < .05$ via Fisher r-to-z transformation) the baseline method as well as the MTL approach (except in the case of valence regression with the audio-eGeMAPS feature set).

Moreover, the systems using the proposed DDAT framework consistently outperform the curriculum learning approach, and they are competitive with, and in some cases even superior

TABLE III
Obtained PCCs Between Each Other Among The Performance Improvement (Δ_c), the Reconstruction Error (ϵ), and the Perception Uncertainty (μ)

	aro		val	
	dev	test	dev	test
PCC(ϵ, Δ_c)				
audio-eGeMAPS	.128	.180	.078	.114
video-appearance	.139	.371	.304	.263
video-geometric	.140	.155	.044	.090
PCC(μ, Δ_c)				
audio-eGeMAPS	.150	.181	.150	.181
video-appearance	.205	.173	.383	.440
video-geometric	.101	-.104	.310	.103
PCC(ϵ, μ)				
audio-eGeMAPS	.150	.072	.040	-.024
video-appearance	-.077	.060	-.077	.059
video-geometric	.127	.071	.050	.021

to, most other state-of-the-art methods, such as the strength modelling [16] and the 'sparse coding (SC) + CNN + LSTM' systems (AVEC 2016 winner) [40]. Despite the fact that the proposed systems are slightly worse than the end-to-end system, which automatically extracts the representations from raw audio and video signals that retain complete pattern information, the DDAT framework can be incorporated with the end-to-end system in the future.

When comparing the two approaches used in the RE-based DDAT experiments, we find that adding the overall sum of the error [cf. Fig. 1(b)] leads to a better performance than adding the error vector [cf. Fig. 1(a)]. This is possibly attributable to the redundant dimensionality of the error vector, which meanwhile yields much noise in the network training. When comparing the RE-based DDAT and the PU-based DDAT, it is noticeable that the two approaches perform similarly. This suggests that both approaches achieve the same goal but in different ways. That is, both approaches successfully explore the difficulty information in the pattern learning process, whereas the RE-based and PU-based DDAT approaches measure the difficulty information by the data reconstruction-capability and by the data perception-uncertainty, respectively. Moreover, it is worth mentioning that the RE-based DDAT approach, in contrast to the PU-based DDAT, not only fits the subjective pattern recognition tasks (e.g., emotion prediction in this work) but also holds the potential to be applied to objective tasks (e.g., phoneme prediction).

To investigate the contribution of the extracted difficulty information to the system performance improvement, we further calculated the correlation (in terms of PCC) between the values of the difficulty indicator (i.e., the obtained RE or the PU) between the performance improvement. Specifically, the performance improvement Δ_c was computed as $\Delta_c = |\hat{y}_{bs} - y| - |\hat{y}_{DDAT} - y|$, given the target (gold standard) y and the prediction of the DDAT system $\hat{y}_{DDAT}$ (or the baseline system $\hat{y}_{bs}$).

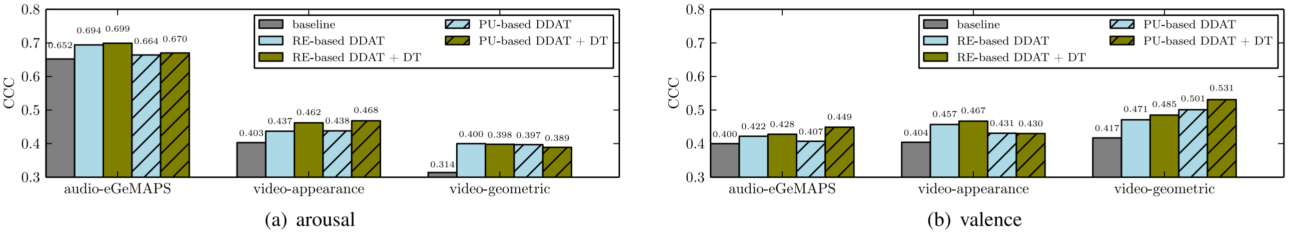
The first three rows of Table III show the obtained PCCs between the RE and the performance improvement [i.e., PCC (ϵ, Δ_c)]. These positive PCCs suggest that the difficulty information can help improve the model performance in the learning process. This conclusion confirms our previous findings in [39]. Note that, in [39], the selected database has subjects that are different from the one in this article. Similar observations can be found when calculating the PCCs between the PU and the performance improvement [i.e., PCC(μ, Δ_c), as shown in the second three rows]. The PCCs are boosted to .384 and .440 in the development and test sets in the case of valence when using appearance-based visual features. In more detail, it can be seen that when using the RE-based DDAT approach, the achieved PCCs for arousal prediction are relatively higher than the ones for valence prediction in most cases. Nevertheless, an opposite observation is made when using the PU-based DDAT approach. This is probably due to the fact that arousal is more sensitive than valence to the expression strength or scale that potentially results in higher RE, whilst the valence is more associated with the subtle variations that easily mislead the judgement of annotators [15], [56].

Furthermore, we calculated the PCCs between the obtained RE and PU, as shown in the last three rows in Table III. Generally speaking, most of these PCCs are around zero, indicating the obtained RE is largely independent of the PU. This further implies that the proposed RE-based and PU-based DDAT strategies capture the different underlying phenomena. Thus, the combination of the two approaches is expected to deliver better performance. The related experiments and corresponding results are given in Section IV-D.

D. Dynamic Tuning and Late Fusion

Figure 2 illustrates the performances of the DDAT models with and without dynamic-tuning of the predictions. Compared with the predictions without dynamic-tuning, the dynamically-tuned predictions yield gains in most cases. For instance, the best achieved CCC for arousal prediction increased from 0.684 to 0.699, using the RE-based DDAT system with the audio-eGeMAPS feature set, whereas for valence prediction it increased from 0.511 to 0.531, using the PU-based DDAT system with the video-geometric feature set. The exceptions include the arousal predictions for both RE- and PU-based DDAT systems using the video-geometric feature set and the valence predictions for the PU-based DDAT system using the video-appearance feature set. In both cases, the differences remain minimal and insignificant via the aforementioned statistical test of Fisher z-to-r transformation.

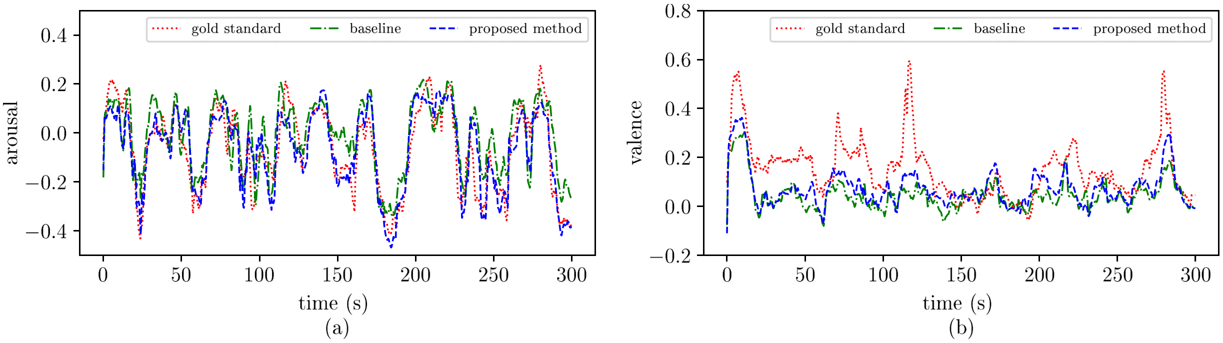
We then conducted a set of late fusions on the individual predictions produced by using different modalities and models. Table IV lists all scenarios (combinations) considered in our experiments as well the respective performance. As can be seen in the table, the best performance on the test set for both arousal and valence was obtained when fusing the predictions from all *modalities* and *models*. In this context, the best results on the test set have been achieved at 0.766 CCC for arousal and 0.660 CCC for valence. These results beat most of the latest reported results from the same data, and they are close to the best result presented in AVEC 2016 [40] (i.e., 0.770 and 0.687 of CCCs for



(a) arousal

(b) valence

Fig. 2. Performance comparison (CCC) between the single-task learning, the proposed dynamic difficulty awareness training approach based on reconstruction error (RE) or perception uncertainty (PU), and their dynamically-tuned (DT) versions. Results pertain to the test partition for both arousal (a) and valence (b) targets using three feature sets (audio-eGeMAPS, video-appearance, and video-geometric).



(a)

(b)

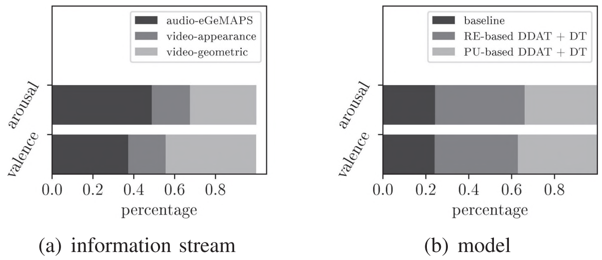
Fig. 3. Automatic prediction of arousal (a) and valence (b) via audiovisual signals obtained with the best late fusion model for a random subject (# 9) from the test partition.

TABLE IV

Late Fusion Performance (CCC) in Different Fusion Strategies (i.e., Modality-Based, Modality- and Model-Based, and Dynamically-Tuned Modality- and Model-Based) for the *dev*elopment and *test* Partitions of Both *aro*usal and *val*ence Regressions. The Predictions Are Generated From the Reconstruction-Error-Based DDAT Framework (P_{re}) or the Perception-Uncertainty-Based DDAT Framework (P_{pu}); Their Dynamically-Tuned Versions ($P_{re,dt}$ or $P_{pu,dt}$); or the Baseline Model (P_{bs}). The Best Results Achieved on the Test Set Are in Bold. Note That P_{re}, $P_{re,dt}$, P_{pu}, $P_{pu,dt}$, and P_{bs} Are the Fused Predictions From Diverse Information Streams (i.e., Audio-eGeMAPS, Video-Appearance, and Video-Geometric). The 1st–3rd, 4th–5th, and 6th–8th Result Rows Are Respectively Obtained Modality-Based, Modality- and Model-Based, and Dynamically-Tuned Modality- and Model-Based Late Fusion Strategies

various late fusion approaches					aro		val	
P_{re}	$P_{re,dt}$	P_{pu}	$P_{pu,dt}$	P_{bs}	dev	test	dev	test
modality-based								
				✓	.822	.690	.705	.584
✓					.853	.763	.738	.615
		✓			.838	.715	.738	.615
modality- and model-based								
✓		✓			.860	.761	.755	.639
✓		✓		✓	.864	.752	.766	.653
modality- and model-based (dynamically-tuned)								
	✓				.853	.761	.739	.621
			✓		.819	.721	.733	.631
	✓		✓		.856	**.766**	.756	.651
	✓		✓	✓	.863	.754	.766	**.660**
state of the art								
strength modelling [16]					.808	.685	.671	.554
end-to-end [14]					.731	.714	.502	.612
state of the art (+ physiology)								
feature selection + offset [49] [a]					.824	.747	.688	.609
SVR + offset [48] [b]					.820	.702	.682	.638
SC + CNN + LSTM [40] [c]					.862	.770	.750	.687

[a] AVEC '15 winner; [b] AVEC '16 baseline; [c] AVEC '16 winner



(a) information stream

(b) model

Fig. 4. Percentage of the contribution of each information stream (a) or model (b) for achieving the best arousal or valence predictions.

arousal and valence prediction), despite this system also utilising an additional modality (physiological features). An illustration of the performance of the best DDAT system compared to the baseline system and the gold standard is depicted in Fig. 3 (data from a random subject from the test partition). Generally, it can be seen that our predictions are closer to the gold standard, especially in the region that has relative peak values.

In order to analyse the importance of each modality and model, we calculated their contributions to the arousal and valence predictions of the respective best performing models. Fig. 4 depicts their contributions. For arousal prediction, the acoustic features play a more important role than the visual features, whereas the opposite happens for valence prediction. It is also expected that the RE-based and PU-based DDAT systems contribute more than the baseline systems to the final predictions. Furthermore, the PU-based DDAT system is slightly more important for valence prediction than it is for arousal prediction. This might be due to the fact that prediction of emotional valence is much more difficult than arousal for audio modality [2], [67], [68].

V. Conclusion and Future Work

In contrast to previous studies that aimed to explore the 'strength' or overcome the 'weakness' of modelling, we for the first time investigated exploiting the difficulty (weakness) information straightforwardly in the learning process for continuous emotion prediction. To extract the difficulty information, we proposed two strategies based on either the ontology of modelling or the content to be modelled. The two types of information separately measure the learning difficulty of a model by reconstructing its input, or the 'hardness' of the data to be learnt by predicting their perception uncertainty. This information indicated by an index was then concatenated into the original features to update the inputs. The proposed methods were systematically evaluated on a benchmark database RECOLA [48]. Experimental results have demonstrated that the proposed methods clearly improve the prediction performance of a model by evolving the difficulty information into its learning process.

Going beyond the traditional curriculum learning and boosting approaches that are specifically designed for discrete pattern recognition tasks, the proposed Dynamic Difficulty Awareness Training (DDAT) approaches can particularly learn well the sequential pattern, such as the continuous emotion prediction in this article. When involving either the input reconstruction error information or the emotion perception uncertainty information, we find that the neural networks can better perform. Nevertheless, it is worth noting that the perception uncertainty is merely defined for a subjective pattern recognition task. For an objective task, it might be reasonable to alternatively employ the prediction uncertainty.

In future, we will continue investigating the efficiency of the proposed DDAT in discrete pattern predictions. Additionally, we will investigate the approaches for which the difficulty information could be possibly used as the prediction weights. Moreover, end-to-end structures that are designed to automatically extract representations have attracted increasing attention, and they are starting to show promising performance. Therefore, an advanced end-to-end framework will be considered in our system as well. In more detail, with respect to the perception-uncertainty-based DDAT end-to-end system, we can simply replace the GRN-RNNs with an end-to-end network while all other inputs and outputs remain. With respect to the reconstruction-error-based end-to-end system, we may consider reconstructing the high-level representations rather than the raw signals when extracting the reconstruction error information (i.e., the difficulty indicator).

References

[1] M. Wöllmer *et al.*, "Abandoning emotion classes—Towards continuous emotion recognition with modelling of long-range dependencies," in *Proc. INTERSPEECH*, Brisbane, Australia, 2008, pp. 597–600.

[2] H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition," *Int. J. Synthetic Emotions*, vol. 1, no. 1, pp. 68–99, Jan. 2010.

[3] H. Gunes, B. Schuller, M. Pantic, and R. Cowie, "Emotion representation, analysis and synthesis in continuous space: A survey," in *Proc. Int. Conf. Autom. Face Gesture Recognit.*, Santa Barbara, CA, USA, 2011, pp. 827–834.

[4] R. W. Picard, *Affective Computing*. Cambridge, MA, USA: MIT press, 1997.

[5] Y.-H. Yang and J.-Y. Liu, "Quantitative study of music listening behavior in a social and affective context," *IEEE Trans. Multimedia*, vol. 15, no. 6, pp. 1304–1315, Oct. 2013.

[6] S. Zhao, H. Yao, Y. Gao, R. Ji, and G. Ding, "Continuous probability distribution prediction of image emotions via multitask shared sparse regression," *IEEE Trans. Multimedia*, vol. 19, no. 3, pp. 632–645, Mar. 2017.

[7] Z. Zhang *et al.*, "Leveraging unlabelled data for emotion recognition with enhanced collaborative semi-supervised learning," *IEEE Access*, vol. 6, pp. 22 196–22 209, 2018.

[8] B. Schuller *et al.*, "AVEC 2011—The first international audio/visual emotion challenge," in *Proc. 4th Int. Conf. Affect. Comput. Intell. Interact.*, Memphis, TN, USA, 2011, pp. 415–424.

[9] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Trans. Multimedia*, vol. 16, no. 8, pp. 2203–2213, Dec. 2014.

[10] H. Meng, N. Bianchi-Berthouze, Y. Deng, J. Cheng, and J. Cosmas, "Time-delay neural network for continuous emotional dimension prediction from facial expression sequences," *IEEE Trans. Cybern.*, vol. 46, no. 4, pp. 916–929, Apr. 2016.

[11] R. Xia and Y. Liu, "A multi-task learning framework for emotion recognition using 2D continuous space," *IEEE Trans. Affect. Comput.*, vol. 8, no. 1, pp. 3–14, Jan. 2017.

[12] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching," *IEEE Trans. Multimedia*, vol. 20, no. 6, pp. 1576–1590, Jun. 2018.

[13] H. Li, J. Sun, Z. Xu, and L. Chen, "Multimodal 2D+3D facial expression recognition with deep fusion convolutional neural network," *IEEE Trans. Multimedia*, vol. 19, no. 12, pp. 2816–2831, Dec. 2017.

[14] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE J. Sel. Topics Signal Process., Special Issue End-to-End Speech Lang. Process.*, vol. 11, no. 8, pp. 1301–1309, Dec. 2017.

[15] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Mar. 2009.

[16] J. Han, Z. Zhang, N. Cummins, F. Ringeval, and B. Schuller, "Strength modelling for real-world automatic continuous affect recognition from audiovisual signals," *Image Vis. Comput.*, vol. 65, pp. 76–86, Sep. 2017.

[17] Z. Zhang, F. Eyben, J. Deng, and B. Schuller, "An agreement and sparseness-based learning instance selection and its application to subjective speech phenomena," in *Proc. 5th Int Workshop Emotion Social Signals, Sentiment Linked Open Data, Satellite LREC*, Reykjavik, Iceland, 2014, pp. 21–26.

[18] L. Gui, T. Baltruaitis, and L. P. Morency, "Curriculum learning for facial expression recognition," in *Proc. 12th IEEE Int Conf. Autom. Face Gesture Recognit.*, Washington, DC, USA, 2017, pp. 505–511.

[19] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, 2014, pp. 1805–1812.

[20] L. L. Presti and M. L. Cascia, "Boosting Hankel matrices for face emotion recognition and pain detection," *Comput. Vis. Image Understanding*, vol. 156, pp. 19–33, Mar. 2017.

[21] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, Montreal, Canada, 2009, pp. 41–48.

[22] S. Braun, D. Neil, and S. Liu, "A curriculum learning method for improved noise robustness in automatic speech recognition," in *Proc. 25th Eur. Signal Process. Conf.*, Kos, Greece, 2017, pp. 548–552.

[23] R. Lotfian and C. Busso, "Curriculum learning for speech emotion recognition from crowdsourced labels," May 2018, arXiv: 1805.10339.

[24] A. Graves, *Supervised Sequence Labelling With Recurrent Neural Networks*. Berlin, Germany: Springer, 2012.

[25] M. I. Posner and S. E. Petersen, "The attention system of the human brain," *Annu. Rev. Neurosci.*, vol. 13, no. 1, pp. 25–42, 1990.

[26] D. A. Washburn and R. Putney, "Attention and task difficulty: When is performance facilitated?" *Learn. Motivation*, vol. 32, no. 1, pp. 36–47, Feb. 2001.

[27] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Vancouver, Canada, 2013, pp. 7398–7402.

[28] P. Karanasou, Y. Wang, M. J. Gales, and P. C. Woodland, "Adaptation of deep neural network acoustic models using factorised i-vectors," in *Proc. INTERSPEECH*, Singapore, 2014, pp. 2180–2184.

[29] P. Vincent, H. Larochelle, Y. Bengio, and P. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. Int. Conf. Mach. Learn.*, Helsinki, Finland, 2008, pp. 1096–1103.

[30] E. Marchi *et al.*, "A novel approach for automatic acoustic novelty detection using a denoising autoencoder with bidirectional LSTM neural networks," in *Proc. IEEE Int. Conf. Audio, Speech, Signal Process.*, Brisbane, Australia, 2015, pp. 1996–2000.

[31] H. Yang *et al.*, "Unsupervised extraction of video highlights via robust recurrent auto-encoders," in *Proc. IEEE Int. Conf. Comput. Vis.*, Santiago, Chile, 2015, pp. 4633–4641.

[32] S. Petridis and M. Pantic, "Prediction-based audiovisual fusion for classification of non-linguistic vocalisations," *IEEE Trans. Affect. Comput.*, vol. 7, no. 1, pp. 45–58, Jan. 2016.

[33] Z. Zhang, N. Cummins, and B. Schuller, "Advanced data exploitation for speech analysis—An overview," *IEEE Signal Process. Mag.*, vol. 34, no. 4, pp. 107–129, Jul. 2017.

[34] L. Devillers, L. Vidrascu, and L. Lamel, "Challenges in real-life emotion annotation and machine learning based detection," *Neural Netw.*, vol. 18, no. 4, pp. 407–422, May 2005.

[35] J. Deng, W. Han, and B. Schuller, "Confidence measures for speech emotion recognition: A start," in *Proc. 10th ITG Conf. Speech Commun.*, Braunschweig, Germany, 2012, pp. 1–4.

[36] T. Dang, V. Sethu, J. Epps, and E. Ambikairajah, "An investigation of emotion prediction uncertainty using gaussian mixture regression," in *Proc. INTERSPEECH*, Stockholm, Sweden, 2017, pp. 1248–1252.

[37] T. Dang *et al.*, "Investigating word affect features and fusion of probabilistic predictions incorporating uncertainty in AVEC 2017," in *Proc. 7th Annu. Workshop Audio/Vis. Emotion Challenge*, Mountain View, CA, USA, 2017, pp. 27–35.

[38] J. Han, Z. Zhang, M. Schmitt, and B. Schuller, "From hard to soft: Towards more human-like emotion recognition by modelling the perception uncertainty," in *Proc. ACM Int. Conf. Multimedia*, Mountain View, CA, USA, 2017, pp. 890–897.

[39] J. Han, Z. Zhang, F. Ringeval, and B. Schuller, "Reconstruction-error-based learning for continuous emotion recognition in speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, New Orleans, LA, USA, 2017, pp. 2367–2371.

[40] K. Brady *et al.*, "Multi-modal audio, video and physiological sensor learning for continuous emotion prediction," in *Proc. 6th Int. Workshop Audio/Vis. Emotion Challenge*, Amsterdam, The Netherlands, 2016, pp. 97–104.

[41] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[42] X. Ma, H. Yang, Q. Chen, D. Huang, and Y. Wang, "DepAudioNet: An efficient deep model for audio based depression classification," in *Proc. 6th Int. Workshop Audio/Vis. Emotion Challenge*, Amsterdam, The Netherlands, 2016, pp. 35–42.

[43] F. Ringeval *et al.*, "Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data," *Pattern Recognit. Lett.*, vol. 66, pp. 22–30, Nov. 2015.

[44] L. Chao, J. Tao, M. Yang, Y. Li, and Z. Wen, "Long short term memory recurrent neural network based multimodal dimensional emotion recognition," in *Proc. 5th Int. Workshop Audio/Vis. Emotion Challenge*, Brisbane, Australia, 2015, pp. 65–72.

[45] Z. Huang *et al.*, "An investigation of annotation delay compensation and output-associative fusion for multimodal continuous emotion prediction," in *Proc. 5th Int. Workshop Audio/Vis. Emotion Challenge*, Brisbane, Australia, 2015, pp. 41–48.

[46] J. Wei *et al.*, "Multimodal continuous affect recognition based on LSTM and multiple kernel learning," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, Siem Reap, Cambodia, 2014, pp. 1–4.

[47] X. Qiu, L. Zhang, Y. Ren, P. N. Suganthan, and G. Amaratunga, "Ensemble deep learning for regression and time series forecasting," in *Proc. Comput. Intell. Ensemble Learn.*, Orlando, FL, USA, 2014, pp. 1–6.

[48] M. Valstar *et al.*, "AVEC 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proc. 6th Int. Workshop Audio/Vis. Emotion Challenge*, Amsterdam, The Netherlands, 2016, pp. 3–10.

[49] L. He *et al.*, "Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks," in *Proc. 5th Int. Workshop Audio/Vis. Emotion Challenge*, Brisbane, Australia, 2015, pp. 73–80.

[50] M. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Trans. Affect. Comput.*, vol. 2, no. 2, pp. 92–105, Apr. 2011.

[51] J. Deng, X. Xu, Z. Zhang, S. Frühholz, and B. Schuller, "Semi-supervised autoencoders for speech emotion recognition," *IEEE/ACM Trans. Audio, Speech, Lang Process.*, vol. 26, no. 1, pp. 31–43, Jan. 2018.

[52] M. A. Nicolaou, H. Gunes, and M. Pantic, "Output-associative RVM regression for dimensional and continuous emotion prediction," *Image Vis. Comput.*, vol. 30, no. 3, pp. 186–196, Mar. 2012.

[53] S. Parthasarathy and C. Busso, "Jointly predicting arousal, valence and dominance with multi-task learning," in *Proc. INTERSPEECH*, Stockholm, Sweden, 2017, pp. 1103–1107.

[54] Y. Xia, X. Cao, F. Wen, G. Hua, and J. Sun, "Learning discriminative reconstructions for unsupervised outlier removal," in *Proc. IEEE Int. Conf. Comput. Vis.*, Santiago, Chile, 2015, pp. 1511–1519.

[55] Z. Zhang *et al.*, "Facing realism in spontaneous emotion recognition from speech: Feature enhancement by autoencoder with LSTM neural networks," in *Proc. INTERSPEECH*, San Francisco, CA, USA, 2016, pp. 3593–3597.

[56] B. Schuller and A. Batliner, *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*. Hoboken, NJ, USA: Wiley, 2013.

[57] I. B. Mauss and M. D. Robinson, "Measures of emotion: A review," *Cognition Emotion*, vol. 23, no. 2, pp. 209–237, Feb. 2009.

[58] F. Ringeval *et al.*, "AV+EC 2015: The first affect recognition challenge bridging across audio, video, and physiological data," in *Proc. 5th Int. Workshop Audio/Vis. Emotion Challenge*, Brisbane, Australia, 2015, pp. 3–8.

[59] F. Ringeval, A. Sonderegger, J. S. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit.*, Shanghai, China, 2013, pp. 1–8.

[60] F. Eyben *et al.*, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, Apr. 2016.

[61] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," in *Proc. Workshop Syntax, Semantics Struct. Statist. Transl.*, Doha, Qatar, 2014, pp. 103–111.

[62] R. Jozefowicz, W. Zaremba, and I. Sutskever, "An empirical exploration of recurrent network architectures," in *Proc. 32nd Int. Conf. Mach. Learn.*, Lille, France, 2015, pp. 2342–2350.

[63] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, San Diego, CA, USA, 2015.

[64] S. Mariooryad and C. Busso, "Correcting time-continuous emotional labels by modeling the reaction lag of evaluators," *IEEE Trans. Affect. Comput.*, vol. 6, no. 2, pp. 97–108, Apr. 2015.

[65] F. Ringeval *et al.*, "AVEC 2017–Real-life depression, and affect recognition workshop and challenge," in *Proc. 7th Int. Workshop Audio/Vis. Emotion Challenge*, Mountain View, CA, USA, 2017, pp. 3–10.

[66] J. Cohen, P. Cohen, S. G. West, and L. S. Aiken, *Applied Multiple Regression/correlation Analysis for the Behavioral Sciences*. Abingdon, U.K.: Routledge, 2013.

[67] Z. Zhang, E. Coutinho, J. Deng, and B. Schuller, "Cooperative learning and its application to emotion recognition from speech," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 23, no. 1, pp. 115–126, Jan. 2015.

[68] E. Coutinho and B. Schuller, "Shared acoustic codes underlie emotional communication in music and speech—Evidence from deep transfer learning," *PloS One*, vol. 13, no. 1, 2018, Art. no. e0191754.

Zixing Zhang (M'15) received the master's degree in physical electronics from the Beijing University of Posts and Telecommunications, Beijing, China, in 2010, and the Ph.D. degree in computer engineering from the Technical University of Munich, Munich, Germany, in 2015. He is currently a Research Associate with the Department of Computing Imperial College London, London, U.K., since 2017. Before that, he was a Postdoctoral Researcher with the University of Passau, Germany, from 2015 to 2017. He has authored about seventy publications in peer-reviewed books, journals, and conference proceedings to date, and has organised special sessions, such as with the IEEE 7th Affective Computing and Intelligent Interaction in 2017 and the 43rd IEEE International Conference on Acoustics, Speech, and Signal Processing in 2018. He serves as a reviewer for leading-in-their fields journals such as the IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Cybernetics, IEEE Transactions on Automatic Control, IEEE Transactions on Multimedia, IEEE/ACM Transactions on Audio, Speech, and Language Processing, Speech Communication, and Computer Speech and Language. His research interests include lie in deep learning, weekly supervised learning, and transfer learning for intelligent and robust speech analysis, such as emotion recognition.

Jing Han (S'16) received the bachelor's degree in electronic and information engineering from Harbin Engineering University, Harbin, China, in 2011, and the master's degree from Nanyang Technological University, Singapore, in 2014. She is currently working toward the Doctoral degree with the ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg, Germany, involved in the EU's Horizon 2020 programme SEWA. She reviews regularly for IEEE Transactions on Cybernetics and the IEEE Signal Processing Letters. Her research interests include related to deep learning for multimodal affective computing and health care.

Eduardo Coutinho received the Graduate dagree from the University of Porto, Porto, Portugal, in 2003, and the Ph.D degree in computer and affective sciences, in 2009. Since then, he was with the interdisciplinary fields of music psychology and affective sciences, University of Sheffield, Sheffield, U.K., the Swiss Center for Affective Sciences, Geneva, Switzerland, the Technical University of Munich, Munich, Germany, and Imperial College London, London, U.K. He is currently a Lecturer in Music Psychology with the University of Liverpool, Liverpool, U.K. and a Research Associate in the ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing with the University of Augsburg, Augsburg, Germany. In his research, he focuses on the study of emotional experiences with music and the links between the communication of emotion by music and the tone of voice. He is currently also engaged with the development of methods and tools that permit the use music for improving different aspects of well-being in everyday life. In 2013, he was the recipient of the Knowledge Transfer Award from the Swiss National Center of Competence in Research in Affective Sciences, and in 2014, he was the recipient of the Young Investigator Award from the International Neural Network Society.

Björn Schuller (M'05–SM'15–F'18) received the Diploma degree, in 1999, the Doctoral degree for his study on Automatic Speech and Emotion Recognition, in 2006, and the habilitation and Adjunct Teaching Professorship in the subject area of Signal Processing and Machine Intelligence, in 2012, all in electrical engineering and information technology from the Technical University of Munich, Munich, Germany. He is a Professor in Machine Learning with the Department of Computing, the Imperial College London, London, UK, where he heads the Group on Language, Audio and Music, Full Professor and Head of the ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing with the University of Augsburg, Augsburg, Germany, and CEO of audEERING. He was previously Full Professor and Head of the Chair of Complex and Intelligent Systems with the University of Passau, Passau, Germany. Professor Schuller is President-emeritus of the Association for the Advancement of Affective Computing, elected member of the IEEE Speech and Language Processing Technical Committee, and Senior Member of the ACM. He has coauthored 5 books and more than 700 publications in peer-reviewed books, journals, and conference proceedings leading to more than overall 19 000 citations (h-index = 66). Schuller is co-Program Chair of Interspeech 2019, repeated Area Chair of ICASSP, and Editor in Chief for the IEEE Transactions on Affective Computing next to a multitude of further Associate and Guest Editor roles and functions in Technical and Organisational Committees.