Leveraging Unlabeled Data for Emotion Recognition With Enhanced Collaborative Semi-Supervised Learning

Zixing Zhang, Jing Han, Jun Deng, Xinzhou Xu, Fabien Ringeval, Björn Schuller

Leveraging Unlabelled Data for Emotion Recognition with Enhanced Collaborative Semi-Supervised Learning

Zixing Zhang, *Member, IEEE*, Jing Han, *Student Member, IEEE*, Jun Deng, Xinzhou Xu, Fabien Ringeval, and Björn Schuller, *Fellow, IEEE*

***Abstract*—One of the major obstacles that has to be faced when applying automatic emotion recognition to realistic human–machine interaction systems is the scarcity of labelled data for training a robust model. Motivated by this concern, this article seeks to utmost exploit unlabelled data that are pervasively available in the real-world and easy to be collected, by means of novel Semi-Supervised Learning (SSL) approaches. Conventional SSL methods such as self-training, suffer from their inherent drawback of error accumulation, i. e., the samples that are misclassified by the system are continuously employed to train the model in the following learning iterations. To address this major issue, we first propose an enhanced learning strategy, by which we re-evaluate the previously automatically labelled samples in each learning iteration, in order to update the training set by correcting the mislabelled samples. We further exploit multiple modalities and models in the SSL system, by using collaborative SSL, where all modalities and models are considered simultaneously; samples are selected by means of minimising the joint entropy. This strategy is supposed to not only improve the performance of the model for data annotation and consequently enhance the trustability of the automatically labelled data, but also to elevate the diversity of selected data. To evaluate the effectiveness of the proposed approaches, we performed extensive experiments on the RECOLA database, which includes multimodal recordings of spontaneous affective interactions of dyads. The empirical results show that the proposed approaches significantly outperform recently well-established SSL methods.**

***Index Terms*—enhanced semi-supervised learning, collaborative learning, audiovisual emotion recognition**

I. Introduction

Automatic emotion recognition has attracted wide attention in artificial intelligence over the past decade, since it plays an essential role in achieving natural and friendly human–machine interactions [1]–[5]. However, one major obstacle that impedes its broad applications in real-life settings is the lack of sufficient labelled data in terms of *quantity* and *diversity*, which is regarded to be of high importance to build a robust and efficient recognition model [6]–[8].

Because of the public availability of massive unlabelled data that can be easily collected via pervasive electronic devices [8], [9], one natural solution comes to leveraging the value of these data in an effective way. *Semi-Supervised Learning* (SSL) has been emerged as a promising approach since it aims to efficiently make use of machines (i. e., recognition models) to *automatically* 'annotate' unlabelled data, with (almost) no need of manual intervention. Over the past few years, some efforts have been made and have shown the benefits of SSL for emotion recognition.

In [10], Wu et al. introduced a graphic-based SSL model for emotion recognition from music, by which the supervision knowledge (or the label information) is propagated from the labelled data to the unlabelled data by calculating the acoustic and tag similarity among songs. In [11], Schels et al. employed a density estimation of all available data to transfer the label information to unlabelled data. Similar work was further reported in [12], but for the text-based emotion classification.

In contrast to these *transductive* SSL approaches where both labelled and unlabelled data are considered to perform a prediction on the unlabelled data, more research efforts need to follow an inductive SSL paradigm, mainly due to the fact that the powerful capability of discriminative models (e. g., Neural Networks) for emotion recognition has been frequently shown over the past decade [13]. In the *inductive* paradigm, a predictive model is pre-built only on the labelled data and then used for predicting the unlabelled data. As an example, Zhang et al. [14] employed a typical inductive SSL approach called *self-training* to explore the unlabelled data from different databases for emotion recognition from speech. In addition, *co-training* was proposed to exploit two views (feature sets) for emotion recognition. For example, Zhang et al. [15], [16] split the acoustic features into two groups (e. g., energy- or spectral-related), each of which is regarded as one 'view' for emotion recognition from speech. Likewise, Li et al. [17] took the personal and impersonal (i. e., the sentence whose subject is not a person) opinions as two 'views' for emotion recognition from text. Recently, deep neural network-based SSL has emerged a great potential method owing to its capability to distil high-level representations. Most recently, Deng et al. [18] introduced a shared-hidden-layer framework with multi-task learning, which consists of two tasks – recon-

Z. Zhang is with GLAM – Group on Language, Audio & Music, Imperial College London, UK (e-mail: zixing.zhang@imperial.ac.uk).

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

J. Deng is with audEERING GmbH, Gilching, Germany (e-mail: jdeng@audeering.com).

X. Xu is with School of Internet of Things, Nanjing University of Posts and Telecommunications, China (e-mail: xinzhou.xu@njupt.edu.cn).

F. Ringeval is with Laboratoire d'Informatique de Grenoble, Université Grenoble Alpes, France (e-mail: fabien.ringeval@imag.fr).

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

structing inputs (autoencoder path) and predicting emotions (classification path). It is expected that the knowledge can be transferred from unlabelled data to labelled data through the autoencoder path.

However, most of these studies merely focused on a signal modality, i. e., either audio [19], video [20], [21], or text [17]. Nowadays, recognising emotion via multiple modalities emerges to be prominent [22]–[27], not only due to the broad usage of cameras and microphones as aforementioned, but also due to the fact that the combination of various modalities can often offer better performance than unimodality for emotion recognition [23]–[25], [28], [29]. Nevertheless, multimodal information is often ignored in most previous SSL research. Different from previous studies, in this article we intend to make efficient use of multiple modalities in SSL for emotion recognition.

Furthermore, traditional SSL approaches often suffer from a problem of *performance degradation*. That is, when adding more automatically annotated data to the training set often results in worse, rather than better, performance of recognition models [30]–[32]. Largely because the automated annotations (model predictions) are often not totally correct, the mislabelled samples (i. e., error or noise) are potentially taken into account when updating training models and sequentially accumulated in the follow-up learning iterations, leading to a gradual decrease of model performance [30]–[32]. The occurrence of this issue is supposed to highly relate to two factors: *model goodness*, and *correctness and diversity of selected data* when updating training data [33]. A poorly performed model reduces the reliability of the automated annotations, and increases the risk of adding mislabelled samples into the updated training set. In addition, as to the intrinsic prediction inclination of a model, the *diversity* of selected data in SSL might be limited [32], [34]. Adding more selected data from one model probably leads to a higher mismatched distribution between the updated training set and test set [32].

To address the performance degradation problem of SSL, many efforts have been made in the context of machine learning. In [20] and [32], Cohen et al. used unlabelled data to search for a better structure of Bayesian Network. This algorithm can effectively alleviate the problem, but it is only designed for probabilistic models. In [35], Nigam et al. suggested to assign different weights to unlabelled data according to their prediction probabilities (i. e., confidence). Their approach then trains a new model using the combination of original labelled and new weighted-unlabelled data, and iterates. This method effectively reduces the detrimental effect of poorly labelled data by machines [35]. Further, rather than such a *soft-weighted* strategy, its *binary* version was frequently used as well. That is, only a few most confidently predicted data are added to the labelled data set [30]. Besides, another enhanced version was introduced by Li et al. [36], by which the unlabelled data are actively identified with the help of some local information in a neighbourhood graph. By doing this, it keeps those mislabelled data from being added to the training set; hence, a less noisy training set is obtained [36].

In this article, we propose a novel SSL approach called enhanced collaborative SSL (ecSSL), with the purpose to address the performance degradation problem by leveraging multiple modalities and models with a re-evaluation process on selected data. Compared with previous work, the proposed approach can utmost upgrade the goodness of the recognition model as well as the 'correctness' and diversity of selected data. In general, our main contributions can be summarised as follows.

- We exploit the complementary of multiple modalities (i. e., audio and video) and classification models for SSL. This combination is crucial and assumed to offer at least two benefits: to build an enhanced and robust emotion recognition model, and to select more accurate and diverse data in the SSL process. Taking advantage of multiple models is originally motivated by the work presented in [37], [38], where different machine learning models can be learnt mutually.
- We propose to sequentially re-evaluate previously selected data to increase the correctness of selected data. It is supposed to correct possibly mislabeled data in previous iterative learning stages and this further enhances the overall confidence of the system predictions.
- We demonstrate the superiority of the proposed ecSSL approach on a multimodal database and provide insightful analysis.

The remainder of this article is organised as follows. In Section II, we describe the proposed enhanced collaborative SSL in detail. Then, we perform extensive empirical evaluations on the RECOLA database in Section III. Finally, we draw conclusions and point out some potential research directions in Section IV.

II. Enhanced Collaborative Semi-Supervised Learning

Let $\mathcal{L} = \{(\mathbf{x}_i, y_i), i = 1, \ldots, n_l\}$ denote the small set of labelled data and $\mathcal{U} = \{\mathbf{x}_i, i = 1, \ldots, n_u\}$ denote the large set of unlabelled data, where $\mathbf{x} \in \mathcal{X}$ indicates the feature vector in the input feature space; $y \in \mathcal{Y}$ indicates the label of the emotional label space; and n_l and n_u are the total number of labelled and unlabelled data, respectively. It is assumed that n_l is lower than n_u ($n_l \ll n_u$) due to the limited availability of labelled data as discussed in Section I.

In this article, we conduct SSL in an inductive paradigm. To select the data in each SSL iterations, we follow up the classic strategy based on prediction confidence (*aka* prediction uncertainty). Only the samples predicted most confidently are selected. To evaluate the confidence value, we employ the entropy $E(p)$ as a measure, which is calculated from the discrete probability distribution of predictions in our classification case, as

$$E(p) = \sum_{i=1}^{C} p_i log(p_i), \tag{1}$$

where p_i indicates the prediction probability for class i, and C is the number of classes. In this sense, a higher confidence value refers to a lower entropy. Henceforth, we use the entropy of the prediction probability $E(\cdot)$ as a criterion for data selection.

Further, as mentioned, for emotion recognition it is a common case that the prepared samples are in an imbalanced category distribution. Those imbalanced training data probably lead to a model prediction bias: The samples pertaining to the dominant categories (e. g., neutral speech) are easily classified with high confidence [39]. Such a prediction bias consequentially gives rise to a vicious circle in which the dominant categories are recognised increasingly better, while the opposite observation holds for the less represented categories [16]. According to the findings presented in [40], we employ the same number of samples per class to build the initial labelled-training-set. Moreover, we equally select the samples per class in each learning iteration. Compared with the 'traditional' SSL methods that are only based on the prediction confidence, the proposed *balanced selection* can effectively avoid the selection bias towards the dominant categories [16], [40].

A. Self-Training and Co-Training

As mentioned in Section I, self-training and co-training are the two widely used inductive SSL approaches for emotion recognition. For self-training [31], a classifier is firstly trained with an 'original' human-labelled data set $\mathcal{L}$. After that, the classifier is used to recognise the unlabelled data set $\mathcal{U}$. Typically, the unlabelled data $\mathcal{S}$ that are recognised with high confidence (or low entropy $E(y_{\mathbf{x}})$), together with their predicted labels, are added to the original training set ($\mathcal{L} \cup \mathcal{S}$), and removed from the unlabelled data set ($\mathcal{U} \setminus \mathcal{S}$). The classifier is then retrained with the updated training set. This process is repeated several times until a predefined stopping criterion is met.

To cease the learning process, several criteria can be implemented: e. g., (i) no performance improvement is shown on the evaluation set, (ii) a predefined iteration number is matched, or (iii) no target data remains in the unlabelled data set. Note that, in this article, the second stopping criterion is chosen throughout all of the experiments for an easy performance comparison.

Compared with self-training, where the classifier uses its own prediction to teach itself, co-training [41] tries to exploit the mutual information between two models trained on different feature domains ('views') – $\mathcal{X}_{v1}$ and $\mathcal{X}_{v2}$, each of which uses its predictions to teach not only itself but also the other one. Specifically, each 'view' is used to create one 'good' classifier h_{v1} or h_{v2}, and each classifier is tested on the unlabelled data set $\mathcal{U}$. The unlabelled data ($\mathcal{S} = \mathcal{S}_{v1} \cup \mathcal{S}_{v2}$) predicted with high confidence values (or low entropy $E(y_{\mathbf{x}})$) are then added (together with the new label) to the training set ($\mathcal{L} \cup \mathcal{S}$) and removed from the unlabelled data set ($\mathcal{U} \setminus \mathcal{S}$). Afterwards, the two classifiers are retrained from the updated training set based on the corresponding feature domain. The whole process repeats several times as self-training does.

Co-training relies on two assumptions [41]: (a) sufficiency – each 'view' is sufficient for classification on its own. That is, the two hypotheses $f_{v1} : \mathcal{X}_{v1} \mapsto \mathcal{Y}$ and $f_{v2} : \mathcal{X}_{v2} \mapsto \mathcal{Y}$ are good enough for recognition; (b) conditional independence – the 'views' are conditionally independent given the class label [41], that is, $p(y_i|\mathbf{x}) \leftarrow p(y_i|\mathbf{x}_{v1})p(y_i|\mathbf{x}_{v2})$, where $\mathbf{x} = [\mathbf{x}_{v1}, \mathbf{x}_{v2}]$.

Algorithm 1: Enhanced Semi-Supervised Learning.

Initialise:
n_l: number of initial labelled training samples;
n_u: number of unlabelled samples;
n: incremental number of selected samples per learning iteration;
h: classification model;
$\mathbf{x}$: feature set, i. e., $\mathbf{x}_a$, $\mathbf{x}_v$, or $\mathbf{x}_{av}$

for $i = 1, ..., I$ **do** *% iterate learning process*
  Train classifier $h^i := f(\mathcal{L}^i(\mathbf{x}, y))$;
  Predict $(y'_{\mathbf{x}}, E(y'_{\mathbf{x}})) \leftarrow h^i(\forall \mathbf{x} \in \mathcal{U})$;
  % re-evaluate the whole original unlabelled set
  Split $\mathcal{U} = \{\mathcal{U}^c, c = 1, \ldots, C\}$, where $\forall \mathbf{x} \in \mathcal{U}^c$, $y'_{\mathbf{x}} = c$;
  for $c = 1, ..., C$ **do** *% equally selected per class by the strategy of minimum entropy*
    Set $n^i = i \times \lfloor n/C \rfloor$;
    Copy $\mathcal{S}^c$ from $\mathcal{U}^c$, $size(\mathcal{S}^c) = n^i$, and satisfy $\max_{\forall \mathbf{x}^c \in \mathcal{S}^c} E(y'_{\mathbf{x}^c}) \leq \min_{\forall \mathbf{x}'^c \in (\mathcal{U}^c \setminus \mathcal{S}^c)} E(y'_{\mathbf{x}'^c})$;
    $\mathcal{S}^i = \bigcup \mathcal{S}^c$;
  end
  $\mathcal{L}^{i+1} = \mathcal{L}^0 \cup \mathcal{S}^i$;
end

B. Enhanced SSL

One main drawback of SSL is error accumulation, as mentioned in Section I. For traditional SSL, the data selected by the machine are fully trusted and pooled into the training set. However, some of these data are inevitably mislabelled in practise, and result in a noisy training set (cf. Section I).

To tackle this problem, we propose to not always trust the automatically labelled data, and call this approach *enhanced SSL*. The pseudocode describing the algorithm is shown in Algorithm 1. The core idea of this approach is to retain the previously selected data in the original unlabelled data set at *each* learning iteration. In doing this, the previously selected data will be re-evaluated by the following enhanced model. Therefore, it is possible to correct mislabelled data in future iterations with an improved model. Naturally, the previously selected samples may not be selected again in the following learning process, i. e., $\mathcal{S}^i \not\subset \mathcal{S}^j$, $i < j$.

Specifically, given the incremental number of selected samples per learning iteration n, the i-th learning iteration will select $i \times n$ samples in total, while the unlabelled data collection $\mathcal{U}$ remains the size of n_u, in our case.

C. Modality-based Collaborative SSL

The proposed *collaborative SSL* (cSSL) in this article can be considered an extension of co-training, where the views involve not only the *feature domains* (i. e., *modality-based cSSL*), but also the *recognition models* (i. e., *model-based*

cSSL, discussed in Section II-D). When integrated with the enhanced SSL, the new algorithm is named as *enhanced cSSL*.

The pseudocode describing the algorithm of enhanced cSSL based on multimodality is displayed in Algorithm 2. Compared with self-training, modality-based cSSL (e. g., audio, video, text, and physiology) employs multiple modalities as independent 'views' for training different models. Compared with co-training, it can implement multiple, rather than two, modalities in the learning system, which is similar to multi-view learning with less restriction in terms of conditional independence (For more details, the reader is referred to [42]).

Besides, in contrast to conventional co-training where different views individually select the samples that are classified with lowest entropies and then fuse them together (i. e., *minimum-individual-entropy* strategy) [15], [41], cSSL takes a *minimum-joint-entropy* strategy. That is, all predictions obtained by various views for each sample will be merged as one by means of majority voting. Particularly, in the even cases, the final decision is assigned to the category classified with the least entropy. This algorithm improvement can not only avoid the prediction-conflict caused by different views but also potentially increase the automated annotation correctness of the selected data [43]. Furthermore, the final entropy is calculated by averaging all entropies obtained by different views. These merged predictions and entropies will be then relied on for the following data selecting operation.

For the sake of simplicity, in this article we took audio and video as two representative modalities. In this case, the parameter P in Algorithm 2 equals to two, and both audio and video feature vectors can serve as different 'views', i. e., $\mathbf{x}_a \in \mathcal{X}_a = \mathcal{X}_1$, and $\mathbf{x}_v \in \mathcal{X}_v = \mathcal{X}_2$. The complete feature vector can be expressed as $\mathbf{x} = [\mathbf{x}_a, \mathbf{x}_v]$.

D. Model-based Collaborative SSL

In contrast to modality-based cSSL, the model-based cSSL seeks the benefits from multiple diverse classifiers, which are trained on the same feature sets. The pseudocode of its enhanced approach is shown in Algorithm 2 as well.

When combining multiple models (classifiers) into a strong one, it normally requires the individual ones to be sufficiently effective and diverse [44]. Again, for the sake of simplicity, we choose two models for evaluation in this article (i. e., $Q = 2$ in Algorithm 2). The two models are *Support Vector Machines* (SVM) and *Recurrent Neural Networks* (RNN), each of which are widely applied to emotion recognition [13], [23], [38]. In detail, SVM is a convex optimisation function, the characteristics of which offer it the capability to capture the global optimisation. Moreover, SVM is learnt by minimising an upper bound on the expected risk, as opposed to the neural networks that are trained by minimising the errors on all training data, which endows SVM a superior ability to generalise [45]. By contrast, the RNN model is easily trapped in a local minimum which can be hardly avoided and has a risk of overfitting, whilst it is good at capturing the context. Particularly, a memory-enhanced variation of RNN, namely *Long Short-Term Memory RNN* (LSTM-RNN), holds a much more powerful capability of learning long-range contextual information.

Algorithm 2: Enhanced Collaborative Semi-Supervised Learning based on *Multi-Modality* or *Multi-Model*.

Initialise:
n_l: number of initial labelled training samples;
n_u: number of unlabelled samples;
n: incremental number of selected samples per learning iteration;
h: classification model

for $i = 1, ..., I$ **do** *% iterate learning process*
  • *either based on* **multi-modality**
  for $p = 1, ..., P$ **do** *% use* P modalities
    Train classifier based on the p-th modality, $h^{ip} := f(\mathcal{L}^i(\mathbf{x}_p, y))$;
    Classification $(y'_{\mathbf{x}_p}, E(y'_{\mathbf{x}_p})) \leftarrow h^{ip}(\forall \mathbf{x}_p \in \mathcal{U})$;
  end
  Merge predictions $y'_{\mathbf{x}} \leftarrow M(y'_{\mathbf{x}_1}, \cdots, y'_{\mathbf{x}_p})$;
  Average entropies $\bar{E}(y'_{\mathbf{x}}) \leftarrow \frac{1}{P}\sum_{p=1}^{P} E(y'_{\mathbf{x}_p})$;

  • *or based on* **multi-model**
  for $q = 1, ..., Q$ **do** *% use* Q models
    Train the q-th classifier $h^{iq} := f_q(\mathcal{L}^i(\mathbf{x}, y))$;
    Classification $(y'^q_{\mathbf{x}}, E(y'^q_{\mathbf{x}})) \leftarrow h^{iq}(\forall \mathbf{x} \in \mathcal{U})$;
  end
  Merge predictions $y'_{\mathbf{x}} \leftarrow M(y'^1_{\mathbf{x}}, \cdots, y'^Q_{\mathbf{x}})$;
  Average entropies $\bar{E}(y'_{\mathbf{x}}) \leftarrow \frac{1}{Q}\sum_{q=1}^{Q} E(y'^q_{\mathbf{x}})$;

  Split $\mathcal{U} = \{\mathcal{U}^c, c = 1, \ldots, C\}$, where $\forall \mathbf{x} \in \mathcal{U}^c$, $y'_{\mathbf{x}} = c$;
  for $c = 1, ..., C$ **do**
    Set $n^i = i \times \lfloor n/C \rfloor$;
    Copy $\mathcal{S}^c$ from $\mathcal{U}^c$, $size(\mathcal{S}^c) = n^i$, and satisfy $\bar{E}(y'_{\mathbf{x}^c}) \leq \bar{E}(y'_{\mathbf{x}'^c})$;
    $\mathcal{S}^i = \bigcup \mathcal{S}^c$;
  end
  $\mathcal{L}^{i+1} = \mathcal{L}^0 \cup \mathcal{S}^i$;
end

Thus, it is supposed that combining the two models could provide an opportunity for them to learn the strength from each other and avoid the weaknesses. Encouraged by the success of such a combination for continuous emotion recognition [26], [38], we believe that this algorithm could further enhance the correctness and the diversity of the selected data in each learning iteration. Analogous to the modality-based cSSL, a minimum-joint-entropy strategy is taken as well for data selection in this case.

E. Enhanced Collaborative SSL based on Multi-Modality and -Model

An enhanced cSSL based on multi-modality and -model is illustrated in Fig. 1, which integrates the enhanced modality-based cSSL (cf. Section II-C) and the enhanced model-based cSSL (cf. Section II-D). By this approach, the data from the audio and video domains are respectively utilised to build

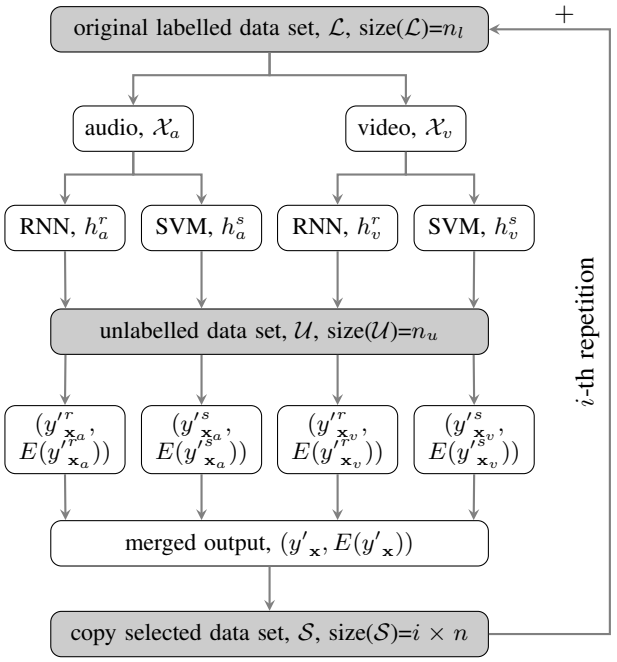


Fig. 1. Flowchart of enhanced collaborative Semi-Supervised Learning based on multi-modality (i. e., audio [a] and video [v]) and multi-model (i. e., RNN [r] and SVM [s]).

RNN and SVM models. For each sample, predictions via various modalities and models are merged to one by majority voting

$$y'_{\mathbf{x}} = M(y'^1_{\mathbf{x}_1}, \cdots, y'^q_{\mathbf{x}_p}, \cdots, y'^Q_{\mathbf{x}_P}), \tag{2}$$

where $y'^q_{\mathbf{x}_p}$ denotes the prediction from the q-th model by using the p-th modality. In case of a draw, the decision is then made by the category that holds the least entropy. Meanwhile, the joint prediction entropy is calculated by

$$\bar{E}(y'_{\mathbf{x}}) = \frac{1}{Q \cdot P} \sum_{q=1}^{Q} \sum_{p=1}^{P} E(y'^q_{\mathbf{x}_p}), \tag{3}$$

where $E(\cdot)$ indicates the prediction entropy. After that, the data selection process is conducted by the minimum-joint entropy strategy for each category as described in Section II-C, such that the sample $\mathbf{x}$ with pseudo-label c in the selected subset $\mathcal{S}$ satisfies

$$\underset{\forall \mathbf{x}^c \in \mathcal{S}^c}{\bar{E}(y'_{\mathbf{x}^c})} \leq \underset{\forall \mathbf{x}'^c \in (\mathcal{U}^c \setminus \mathcal{S}^c)}{\bar{E}(y'_{\mathbf{x}'^c})} . \tag{4}$$

It is worth noting that the size of the selected subset is incrementally increased to $i \times n$, whereas the unlabelled data set always remains the same size n_u, and the updated training set becomes $n_l + i \times n$, at the i-th learning iteration.

III. Experiments and Results

In this section, we perform an empirical evaluation of the proposed SSL approaches on the audiovisual RECOLA database for emotion recognition.

TABLE I
Distribution of speakers and instances per partition of the RECOLA database. spk: speakers, POS: positive, NEG: negative.

	# spk	*arousal*		
		# POS	# NEG	$\sum$
pool	23	623	344	967
test	11	366	149	515

A. RECOLA Database

The multimodal corpus REmote COLlaborative and Affective interactions (RECOLA) [46] (the standard database of the AVEC challenges for audiovisual emotion recognition in 2015 and 2016 [29]) 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. Spontaneous and natural interactions were proceeded from 46 French-speaking participants (27 females and 19 males with a mean age of 22 years and a standard deviation of 3 years) whilst solving a collaborative task conducted in dyads via video conferencing. In total, the database includes 9.5 hours multimodal recordings, i. e., audio, video, electrocardiogram, and electro-dermal activity, which were obtained synchronously and continuously over time. Due to the consent of the participants to share their data, the data set is reduced to a subset of 34 participants with an overall duration of 7.0 hours.

After the data collection process, six gender-balanced French-speaking assistants were asked for annotating the time-continuous ratings of emotional arousal for the first five minutes of all recordings via the ANNEMO web-based annotation toolkit. For the purpose of this study, these continuous ratings for arousal dimension are further discretised into a binary category – POSitive and NEGative. To do this, the continuous audiovisual signals were firstly split into sequential short segments (instances) via voice activity detection. Then, we assigned POS or NEG to each of them if the average rating value of the segment is above or under zero. These data were finally divided into pool set (unlabelled data set) and test set assuring a speaker independence. The details of the speaker and instance distribution of RECOLA used in this article are shown in Table I. More information on the RECOLA database can be found in [46].

B. Acoustic and Visual Features

Regarding the acoustic features, we kept in line with the standard statistical feature set for the past four INTERSPEECH Computational Paralinguistic ChallengEs (ComParE 2013-2017) [47]. This feature set is obtained by applying various functionals (segment level) on the Low-Level Descriptors (LLDs, frame level). Specifically, it contains 4 energy related LLDs (loudness, RASTA spectrum, RMS energy, and zero-crossing rate), 55 spectral related LLDs (spectrum bands, MFCC 1-14, spectral energy, spectral flux/centroid/entropy/slope, psychoacoustic sharpness, harmonicity, and spectral variance/skewness/kurtosis), and 6 voicing related LLDs (F_0, probability of voicing, logHNR, jitter, and shimmer). These 65 LLDs of speech with their first order

derivate leads to 130 LLDs in total (for more details, please refer to [48]). After that, 5 functionals (min, max, range, mean, and variance) are applied over each LLD contour. Thus, the complete acoustic feature set includes 650 attributes per segment.

Regarding the visual features, we extracted 20 LLDs and their first order derivate (40 LLDs in total) for each frame in the video recordings. The 20 LLDs contain 15 facial actions units (AU1-2, 4-7, 9, 11-12, 15, 17, 20, and 23-25), head-pose in three dimensions, and the mean and standard deviation of the optical flow in the region around the head (for more details, please refer to [49]). Similar to acoustic features, the same 5 functionals are applied over the extracted frame-based LLD contours per video segment, which leads to 200 visual attributes per segment in total.

C. Experimental Setup and Evaluation Metrics

Following on previous work [33], we kept taking the binary arousal recognition as a representative emotion recognition task. For SSL, we considered *audio*, *video*, *audio+video* (i. e., combined audio and video) as three independent modalities, respectively leading to an acoustic (650), a visual (200), and an audiovisual (850) feature set. As to the modality-based cSSL, the acoustic and visual feature sets were separately split into two pseudo 'views' (feature subsets) based on the property of the LLDs – the efficiency of this rule was frequently demonstrated in our previous work [15], [16]. That is, the acoustic feature set was divided by the rule of MFCC-related or not, and the visual feature set was partitioned by original or first derived delta features. For the audiovisual feature set, nevertheless, it was split as usual into individual acoustic and visual feature sets as two 'views'.

As to the model-based cSSL, we chose two of the most popular and robust models, i. e., RNNs and SVMs, as exemplary ones, since both of them i) are widely used for emotion recognition (see [29], [47], [50]); ii) are considered to be highly distinct in principle, and frequently employed in an ensemble learning paradigm [38], [51]. Specifically, the RNN model was constructed in the Tensorflow platform [52] with 40 hidden neurons of one hidden layer. To accelerate the RNN learning process, we employed a mini batch of eight instances as network inputs. Additionally, we trained the RNN models with Adam Stochastic Gradient Descent with a learning rate of 10^{-4}. Meanwhile, the SVM model used for our experiments was implemented with the LibSVM toolbox [53], and was optimised with a polynomial kernel and a fixed penalty factor of 0.05.

To carry out the SSL experiments, we first randomly and equally selected 20 instances per class from the pool set, i. e., $n_l = 40$ in total, with the annotations obtained from human raters as an initial training set, which resembles approximately 4% of the whole pool set. The remaining instances in the pool set were regarded as the unlabelled data set. At each SSL iteration, we *incrementally* selected $n = 40$ instances (20 instances per class based on the pseudo (automated) annotations by a pre-trained model). (Note that because the unlabelled data set always remains the same in each learning iteration, selecting a fixed number instances is equal to selecting a fixed ratio of the whole pool set.) More specifically, at the i-th learning iteration, 40 instances were selected in total for our baseline SSL approaches without the enhancement strategy, whilst $40 \times i$ instances were picked in total for the SSL with the enhancement strategy as the previously selected instances remain in the pool set for re-evaluation by an updated model (cf. Section II-B). Further, the learning iteration time was set to be $I = 20$ for better performance comparison. To ease the influence of random selection for the initial training set, we repeated the initial selection 20 times with different random initialisations ('seeds'), leading to 20 independent learning runs throughout all the following experiments.

For performance evaluation, we utilised the widely used metric in the context of emotion recognition – *Unweighted Average Recall* (UAR). It is calculated by the sum of recalls per class divided by the class number as

$$\text{UAR} = \frac{\sum_{i=1}^{K} Recall_i}{\text{K}}, \tag{5}$$

where K is the number of classes. Thus, UAR well reflects the overall accuracy in the presence of class imbalances. Further, to assess the statistical difference of the performance obtained between two approaches, we employed a *paired* t*-test* in what follows. Moreover, to estimate the diversity of selected data, we took euclidean distance measurement that is calculated by

$$D(X) = \sqrt{\left(\sum_{i=1,j=1}^{n} (\mathbf{x}_i - \mathbf{x}_j)^2 / n \right)}, \tag{6}$$

where n is the instance number in data set X.

D. Enhanced vs Non-Enhanced SSL

Fig. 2 and 3 illustrate the performance of *enhanced SSL* and *non-enhanced SSL* evaluated by the models of the RNN and SVM, respectively. Note that for the multi-modalities based cSSL, the audio and video feature sets are partitioned into two pseudo 'views' as mentioned in Section III-C. For the multi-model based cSSL, both the RNN and SVM models are jointly considered for data selection, but the learning process is assessed by either the RNN (cf. Fig. 2) or the SVM (cf. Fig. 3).

From the figures, it can be seen that the enhanced SSL (*black solid lines*) performs better than the non-enhanced SSL (*black dash lines*) in a majority of experimental settings either by the models of RNN (cf. Fig. 2) or SVM (cf. Fig. 3). Specifically, all the scenarios where the enhanced SSL significantly outperforms the non-enhanced SSL are indicated by $p < .05$ at the bottom of each subfigure.

To find out the reason behind the performance improvement, we further calculate the UAR of the predictions on the selected data set, which is presented by the *blue lines* in Fig. 2 and 3. From these subfigures, it is interesting to notice that the enhanced SSL (*solid lines*) is able to select more accurately predicted samples than the non-enhanced SSL (*dash lines*) in most settings. In addition, one can further observe that the performance gain obtained on the selected set highly relates to the gain on the test set. Intuitively, the figures show that

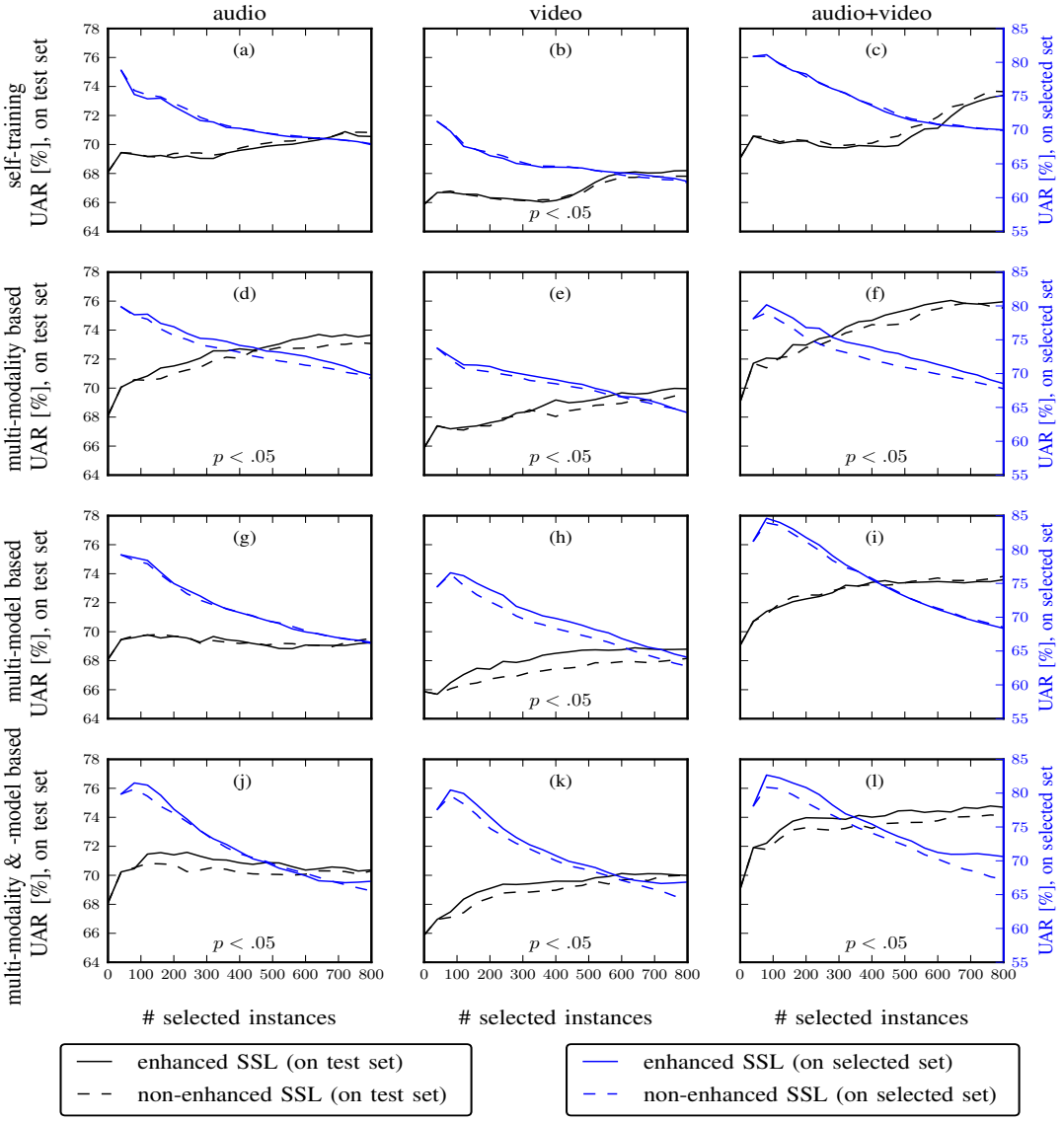


Fig. 2. Performance (averaged UAR over 20 independent runs) comparison between *enhanced* and *non-enhanced* (collaborative) Semi-Supervised Learning (SSL), evaluated by **Recurrent Neural Networks** (RNNs). The *left* (black) and *right* (blue) y-axes indicate the obtained performance on the *test* set and the *selected* set, respectively. Four subfigure-rows from top to bottom refer to the performance of traditional self-training, multi-modality based, multi-model based, and multi-modality and -model based collaborative SSL, respectively. Three subfigure-columns from left to right denote the performance on *acoustic* (audio), *visual* (video), and *audiovisual* (audio+video) features, respectively. (Note: the missing x-axes, left y-axes, and right y-axes are aligned with the bottom, left, and right ones, respectively.)

the cases where the selected set predicted more accurately (in black lines) are largely overlapped to the cases where the test set is recognised more precisely (in blue lines). Such a accuracy increase on the selected set potentially attributes to the fact that the updated models are likely to have corrected part of the previously selected samples that are misclassified by previous weak-models or have dismissed them in the subsequent data selection steps. These re-evaluation and re-selection operations on the pre-selected data set, therefore, partially mitigate the error accumulation problem of SSL and consequentially deliver a more efficient model. The conclusion is consistent with the assumption proposed in Section I and II-B.

Furthermore, the enhanced SSL strategy sounds to perform more effectively when integrating with cSSL approaches (cf. the subfigures in the second, third, and fourth rows of Fig. 2 and 3) than integrating with self-training (cf. the subfigures in the first row of Fig. 2 and 3). This implies that for the better models we obtain in the SSL process, a higher performance gain can be yielded by the enhanced SSL strategy.

E. Collaborative vs Non-Collaborative SSL

According to the findings presented in Section III-D, we henceforth concentrate on the enhanced SSL for analysing the collaborative learning strategy. In Fig. 4 and 5, we compare the proposed *collaborative* SSL approaches with the *non-collaborative* SSL (self-training), evaluated on the modalities of audio, video, or audio+video, and by the models of RNN

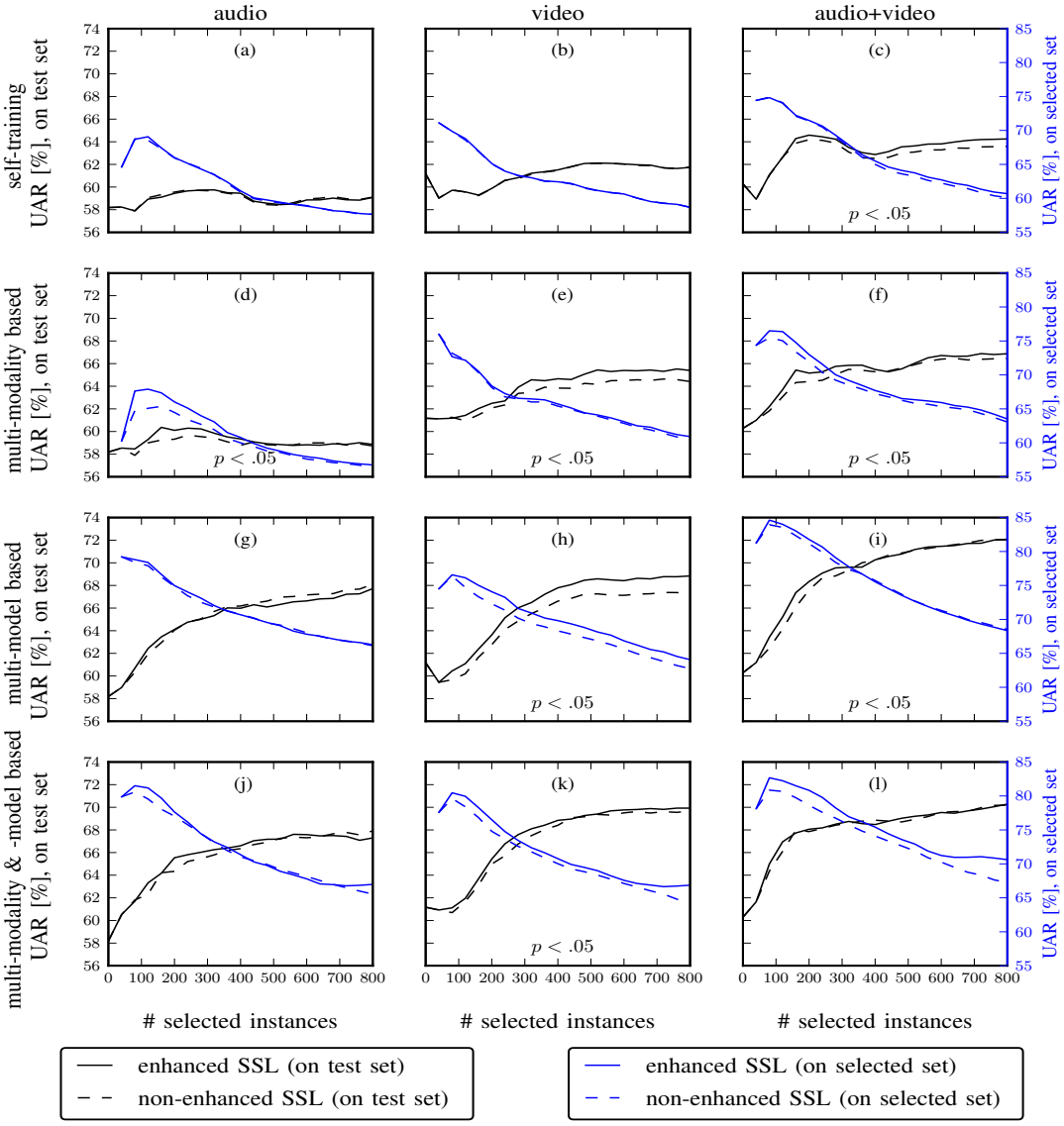


Fig. 3. Performance (averaged UAR over 20 independent runs) comparison between *enhanced* and *non-enhanced* (collaborative) Semi-Supervised Learning (SSL), evaluated by **Support Vector Machines** (SVMs). The *left* (black) and *right* (blue) y-axes indicate the obtained performance on the *test* set and the *selected* set, respectively. Four subfigure-rows from top to bottom refer to the performance of traditional self-training, multi-modality based, multi-model based, and multi-modality and -model based collaborative SSL, respectively. Three subfigure-columns from left to right denote the performance on *acoustic* (audio), *visual* (video), and *audiovisual* (audio+video) features, respectively. (Note: the missing x-axes, left y-axes, and right y-axes are aligned with the bottom, left, and right ones, respectively.)

(Fig. 4) and SVM (Fig. 5). Specifically, the subfigures in the first rows of Fig. 4 and 5 plot the averaged UARs (test set) over 20 independent runs in each learning iteration, achieved by three cSSL approaches and self-training. Generally speaking, all exemplary SSL approaches remarkably further the original UAR gained by the initial training model.

The *modality-based cSSL* (green lines) significantly outperforms self-training (red lines) in almost all chosen modalities and models by performing a paired t-test, which keeps in line with the findings reported in our previous work merely with audio as modality [15], [16]. Similar observations are further made for the *model-based cSSL* (blue lines), which implicitly indicate that employing multiple models in a mutual learning paradigm is quite helpful to boost the performance of SSL. We further discover that the modality-based approach performs better than the model-based one when using the classification model of the RNN, and vice versa for the SVM. This outcome is partially due to the initial UAR gap gained by the RNN model and the SVM model (more details will be found in Section III-F).

When combining the *modality- and model-based cSSL* (black lines), it can be seen that better performance can be delivered in three out of six cases (see the first rows of Fig. 4 and 5). To quantitatively analyse the performance improvement of cSSL, we calculated the averaged initial, last, maximum, mean (over the 20 learning iterations) UARs as well as their corresponding standard deviation across 20 independent runs for each SSL approach (see Table II).

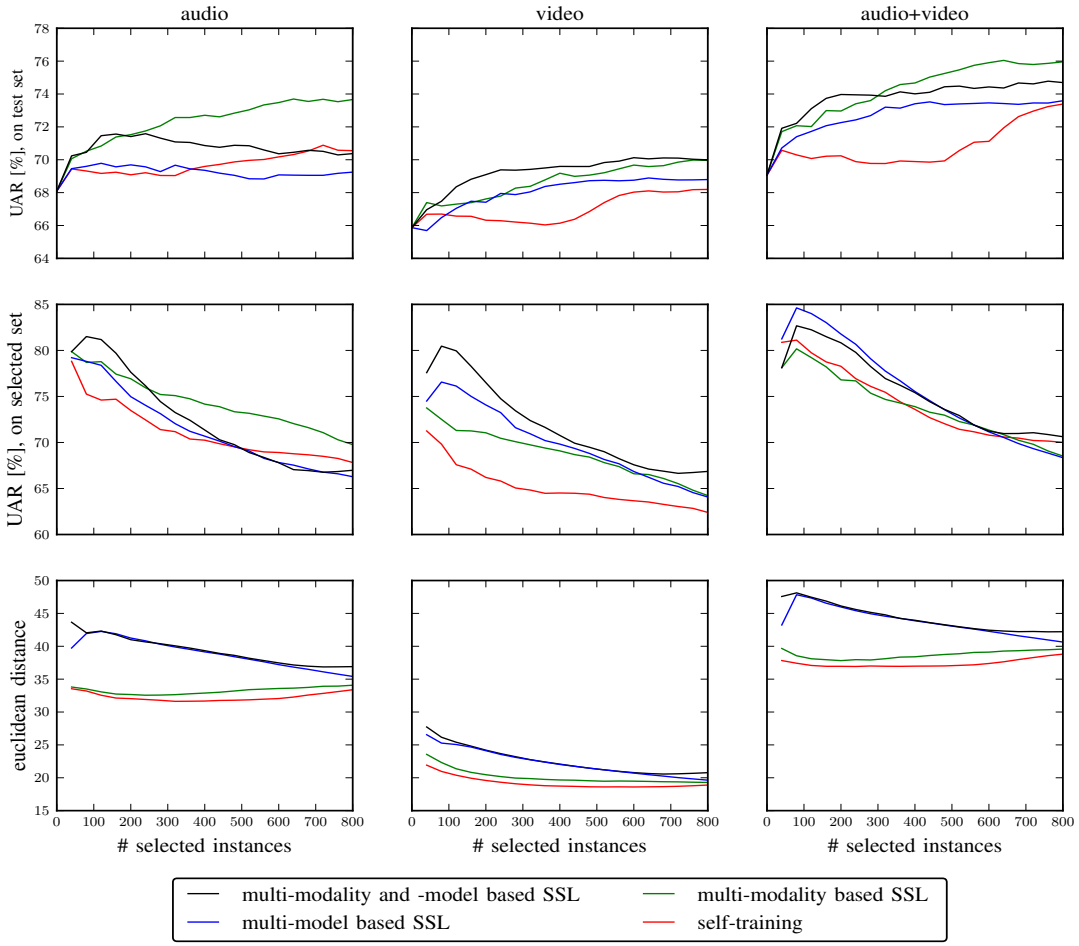


Fig. 4. Comparison between the proposed collaborative Semi-Supervised Learning approaches and self-training, evaluated by **Recurrent Neural Networks** (RNNs). Three rows from top to bottom denote the obtained UARs on the *test* set, the obtained UARs on the *selected* set, and the *euclidean distance* among the data of the selected set, respectively. Three columns from left to right indicate the obtained UARs or euclidean distance on *acoustic* (audio), *visual* (video), and *audiovisual* (audio+video) features, respectively. (Note: the missing x-axes and y-axes are aligned with the bottom and left ones, respectively.)

TABLE II

STATISTICAL PERFORMANCE (AVERAGED UARS AND CORRESPONDING STANDARD DEVIATION [STD]) COMPARISON BETWEEN THE ENHANCED *collaborative semi-supervise learning* (BASED ON MULTI-MODALITY AND/OR MULTI-MODAL) AND THE ENHANCED SELF-TRAINING (BASED ON UNIMODALITY AND UNIMODEL), EVALUATED BY A RECURRENT NEURAL NETWORK (RNN) AND A SUPORT VECTOR MACHINE (SVM). THE *initial*, *last*, *maximum*, AND *mean* OF THE UARS OVER THE 20 LEARNING ITERATIONS ARE SHOWN. ALL VALUES ARE AVERAGED ACROSS 20 INDEPENDENT RUNS.

	UAR [%]	RNN				SVM			
	$\text{Average}_{\text{std}}$	initial	last	maximum	mean	initial	last	maximum	mean
self-training	audio	$68.1_{\pm 5.8}$	$70.6_{\pm 5.9}$	$70.9_{\pm 6.1}$	$69.7_{\pm 5.6}$	$58.2_{\pm 4.4}$	$59.1_{\pm 14.7}$	$59.7_{\pm 11.7}$	$58.9_{\pm 11.8}$
	video	$65.9_{\pm 2.6}$	$68.2_{\pm 1.9}$	$68.2_{\pm 1.9}$	$67.0_{\pm 1.8}$	$61.2_{\pm 4.5}$	$61.7_{\pm 16.3}$	$62.1_{\pm 15.9}$	$61.1_{\pm 13.3}$
	audio+video	$69.1_{\pm 5.1}$	$73.4_{\pm 4.1}$	$73.4_{\pm 4.1}$	$70.8_{\pm 4.6}$	$60.3_{\pm 5.4}$	$64.3_{\pm 13.8}$	$64.6_{\pm 11.5}$	$63.3_{\pm 11.8}$
multi-modality based SSL	audio	$68.1_{\pm 5.8}$	$73.7_{\pm 4.6}$	$73.7_{\pm 4.3}$	$72.3_{\pm 4.6}$	$58.2_{\pm 4.4}$	$58.9_{\pm 12.2}$	$60.4_{\pm 9.6}$	$59.2_{\pm 10.6}$
	video	$65.9_{\pm 2.6}$	$70.0_{\pm 1.6}$	$70.0_{\pm 1.7}$	$68.6_{\pm 2.1}$	$61.2_{\pm 4.5}$	$65.4_{\pm 11.3}$	$65.5_{\pm 11.4}$	$63.9_{\pm 9.6}$
	audio+video	$69.1_{\pm 5.1}$	$76.0_{\pm 2.0}$	$76.0_{\pm 2.1}$	$74.2_{\pm 3.1}$	$60.3_{\pm 5.4}$	$66.9_{\pm 12.4}$	$66.9_{\pm 12.3}$	$65.2_{\pm 10.9}$
multi-model based SSL	audio	$68.1_{\pm 5.8}$	$69.2_{\pm 5.3}$	$69.8_{\pm 5.7}$	$69.2_{\pm 5.5}$	$58.2_{\pm 4.4}$	$67.7_{\pm 4.6}$	$67.7_{\pm 4.6}$	$64.9_{\pm 4.9}$
	video	$65.9_{\pm 2.6}$	$68.8_{\pm 3.6}$	$68.9_{\pm 3.6}$	$68.0_{\pm 3.7}$	$61.2_{\pm 4.5}$	$68.8_{\pm 4.1}$	$68.8_{\pm 4.1}$	$66.0_{\pm 4.2}$
	audio+video	$69.1_{\pm 5.1}$	$73.6_{\pm 2.6}$	$73.6_{\pm 2.6}$	$72.7_{\pm 3.4}$	$60.3_{\pm 5.4}$	$72.1_{\pm 2.3}$	$72.1_{\pm 2.3}$	$69.0_{\pm 3.0}$
multi-modality & -model based SSL	audio	$68.1_{\pm 5.8}$	$70.4_{\pm 5.9}$	$71.6_{\pm 4.6}$	$70.7_{\pm 5.4}$	$58.2_{\pm 4.4}$	$67.3_{\pm 5.2}$	$67.6_{\pm 5.2}$	$65.6_{\pm 4.6}$
	video	$65.9_{\pm 2.6}$	$70.0_{\pm 1.4}$	$70.1_{\pm 1.4}$	$69.2_{\pm 1.7}$	$61.2_{\pm 4.5}$	$69.9_{\pm 0.8}$	$69.9_{\pm 0.8}$	$67.2_{\pm 1.9}$
	audio+video	$69.1_{\pm 5.1}$	$74.7_{\pm 1.7}$	$74.8_{\pm 1.7}$	$73.8_{\pm 2.4}$	$60.3_{\pm 5.4}$	$70.3_{\pm 1.7}$	$70.3_{\pm 1.7}$	$68.0_{\pm 2.9}$

Generally speaking, for RNN or SVM, the modality-based or the model-based cSSL can yield better performance than self-training according to the averaged maximum and mean UARs. For example, the highest maximum UARs were achieved at 76.0 % on average by applying the multi-modality based SSL to audio+video modality when using RNN, and achieved at 72.1 % on average by applying multi-model based SSL to audio+video modality when using SVM. When further fusing

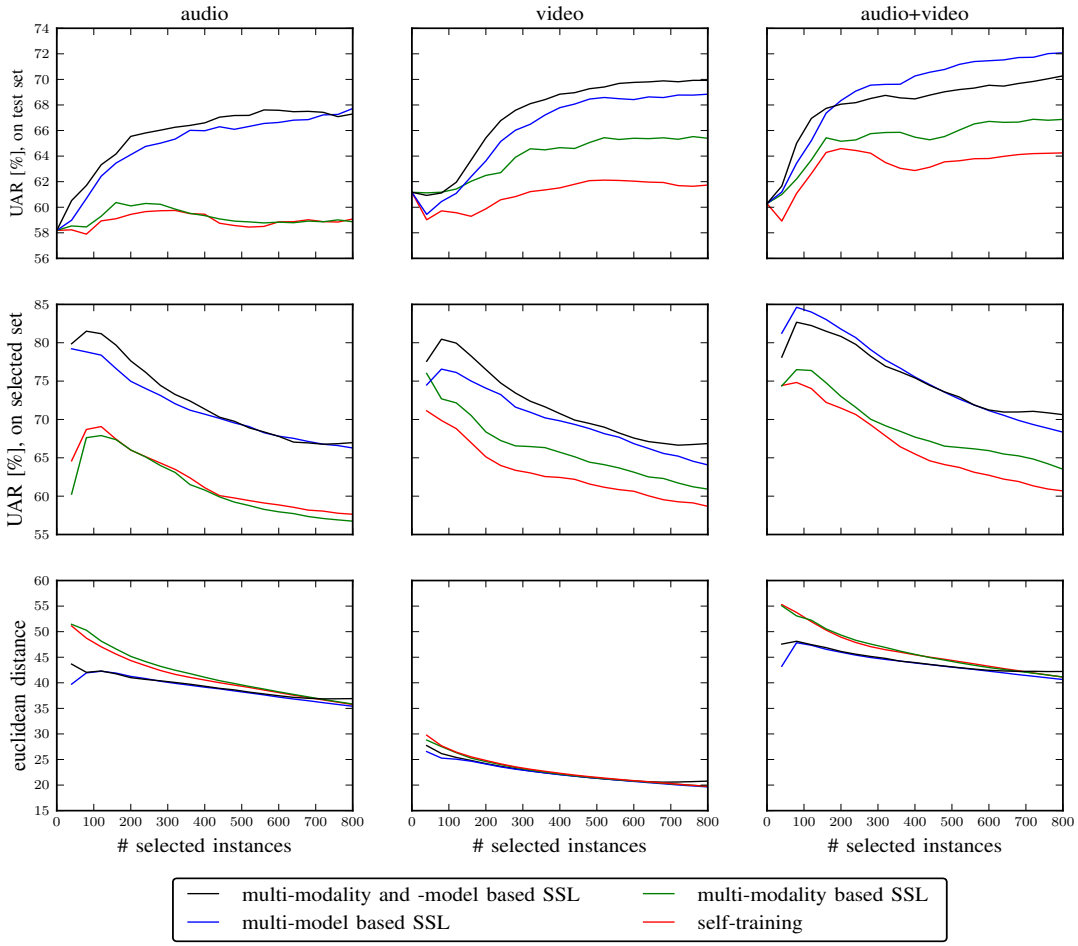


Fig. 5. Comparison between the proposed collaborative Semi-Supervised Learning approaches and self-training, evaluated by **Support Vector Machines** (SVMs). Three rows from top to bottom denote the obtained UARs on the *test* set, the obtained UARs on the *selected* set, and the *euclidean distance* among the data of the selected set, respectively. Three columns from left to right indicate the obtained UARs or euclidean distance on *acoustic* (audio), *visual* (video), and *audiovisual* (audio+video) features, respectively. (Note: the missing x-axes and y-axes are aligned with the bottom and left ones, respectively.)

the modality- and model-based cSSL, we can observe that the models become more robust as the obtained UARs in 20 independent runs are with lower standard deviation. This is important in realistic applications since the SSL process is often undertaken only limited times, normally once. However, we obverse that the models cannot always achieve the highest UARs throughout all experimental scenarios, for example, 74.8 % and 70.3 % of UARs were obtained by using RNN or SVM as classifiers, respectively, for the audio+video modality, which are lower than the best results delivered by multi-modality or multi-modal based SSL. These exceptions possibly attribute to the limited sample number of the database we employed for experiments. Despite of this observation, it can be seen that the fused approach outperforms the ones based on either multi-modality or multi-model for audio, video, or audio+video modalities in four out of six cases when using RNN or SVM. Therefore, the fused multi-modality & -model based SSL is particularly attractive when without knowing which modality or model fits the data best.

We further compared the enhanced cSSL (ecSSL) with two traditional SSL approaches (i. e., *Label Spreading* (LS) [54] and *Label Propagation* (LP) [55]), as well as two recently proposed deep-learning based SSL approaches (i. e., based on either *Generative Adversarial Network* (GAN) [56] or *AutoEncoder* (AE) [18]). The former two approaches belong to transductive SSL, which take the distribution of the unlabelled and labelled data into account as introduced in Section I. For more details, the readers can be referred to [54] and [55]. The later two approaches have recently attracted increasing interest due to the rise of deep learning. GAN was first proposed

TABLE III
PERFORMANCE COMPARISON IN TERMS OF UAR BETWEEN ENHANCED COLLABORATIVE SEMI-SUPERVISED LEARNING (ECSSL) AND TRADITIONAL APPROACHES IN 20 INDEPENDENT RUNS.

UAR_{std}	audio	video	audio+video
state of the art			
label spreading [54]	$53.7_{\pm 0.0}$	$58.1_{\pm 0.0}$	$53.8_{\pm 0.0}$
label propagation [55]	$67.3_{\pm 0.0}$	$59.4_{\pm 0.0}$	$64.8_{\pm 0.0}$
GAN-based [56]	$68.1_{\pm 3.7}$	$67.5_{\pm 3.1}$	$71.5_{\pm 3.7}$
AE-based [18]	$70.4_{\pm 4.1}$	$65.3_{\pm 2.3}$	$70.3_{\pm 4.7}$
proposed			
ecSSL (SVM)	$67.6_{\pm 5.2}$	$69.9_{\pm 0.8}$	$70.3_{\pm 1.7}$
ecSSL (RNN)	$71.6_{\pm 4.6}$	$70.1_{\pm 1.4}$	$74.8_{\pm 1.7}$

in [57], where a deep generative model is learnt to model the data distribution of target, when training jointly with another discriminative model as two players in a minimax game. The GAN-based SSL is particularly designed to address the data sparsity problem – the generator aims to simulate sufficient data as real as possible to augment the training set, whereas the discriminator not only detects the sources where its input samples come from, but also performs a classification [56]. Besides, the AE-based SSL was reported in [18], where a multi-task learning framework was implemented. On the one hand, it classifies the emotions in a supervised manner; on the other hand, it simultaneously reconstructs the input in an unsupervised manner. The motivation of taking this framework is to explore the underlying representations shared among the unlabelled and labelled data, so that the knowledge can be transferred from the massive unlabelled data to the limited labelled data. For a fair performance comparison, we implemented the same network structure with the one used in our approach for both two recently proposed approaches, and the same learning rate and batch size when training the networks. The performance comparison is shown in Table III. When comparing with the two transductive SSL approaches (i. e., LS and LP), we find that ecSSL significantly improves the performance with the video or the fused audio+video modalities when performing a statistical one-tailed z-test ($p < 0.05$). When comparing with the deep-learning based SSL approaches (i. e., GAN-based and AE-based), we observe that the proposed approach also yields performance gain in a large margin by using RNN as a classifier.

F. Discussion

To demonstrate the observations shown in Section III-E, we further investigate the quality of the selected data set in terms of *accuracy* (second rows) and *diversity* (third rows) in both Fig. 4 and 5. As to the accuracy, it can be seen that all three proposed cSSL approaches can achieve higher averaged UARs than self-training on the selected data set in most, if not all, scenarios. Interestingly, the UAR curves obtained on the selected set for each SSL approach have an almost identical order with the UAR curves obtained on the test set, which again explicitly indicates the importance of prediction accuracy of the selected data as aforementioned (cf. Section III-D). Further, as we expected, the averaged UARs are to decrease when incrementally adding more automatically labelled instances by the machine in the SSL process. This clearly reveals the intrinsic problem of SSL where errors will be accumulated along with the learning iterations. As a consequence, the model performance will decrease when the detrimental effect that the selected data cause surpasses the benefit that they offer.

As to the diversity, Fig. 4 (third row) shows the averaged euclidean distance among all data-pairs in the selected set, by using the RNN classification model. Obviously, cSSL is capable of choosing diverse data, which potentially provide a plethora of feature variations and sufficiently cover the whole picture of a data distribution. More concretely, the model-based cSSL as well as its integrated approach with modality-based cSSL can provide much more diverse data than the modality-only-based cSSL. However, these observations are not seen in Fig. 5 (third row) where self-training provides relatively more diverse data. This might largely attribute to the principle of SVMs for classification: The data far away from the decision hyperplanes are often predicted with high confidence, which gives rise to a high diversity of the selected data.

To compare the performance of the modality-based cSSL and the model-based cSSL when using RNN or SVM recognition models, we discover that for the RNN recognition model (cf. Fig. 4), the preferably selected data by SVM are more *diverse* than the ones just provided by RNN; in the third subfigure-row of Fig. 4, the blue lines are obviously higher than the green lines. Nevertheless, for the SVM recognition model (cf. Fig. 5), the preferably selected data by the RNN are more precise than the ones just provided by the SVM; in the second subfigure-row of Fig. 5, the blue lines are obviously higher than the green lines. Therefore, combining the two models in a mutually learning paradigm can efficiently exploit the strengths of each model, whilst avoiding their weaknesses.

Moreover, to compare the SSL performance between uni-modality (i. e., audio or video, the first and second columns of Fig. 4 and 5) and multi-modality (i. e., audio+video, the third column of Fig. 4 and 5), one can notice that combining the multiple modalities is able to boost the performance in almost all cases. A more quantitative performance comparison can be found in Table II as well. These findings are in consistence with the ones reported by previous studies [38], [49].

IV. Conclusion

To leverage the ubiquitous unlabelled data for automatic emotion recognition, this article proposed enhanced collaborative Semi-Supervised Learning (SSL). Dissimilar to traditional SSL, it performs a data re-evaluation process on previously selected data (enhanced strategy) on one hand. On the other hand it takes a mutual learning process among multiple modalities and models (collaborative strategy). The proposed approaches have been systematically evaluated on the widely used audiovisual affective database RECOLA in various settings. The experimental results demonstrate that the proposed approaches significantly improve the system performance by enhancing the correctness and diversity of selected data.

More recently, deep learning algorithms have attracted tremendous attention and achieved a great success in the context of machine learning. This will form one of the main research directions in the future, by considering diverse deep learning architectures in the SSL systems.

Acknowledgment

This work was supported by the UK's Economic & Social Research Council through the research Grant No. HJ-253479 (ACLEW), and the European Union's Horizon 2020 Programme through the Research Innovation Action No. 645094 (SEWA).

REFERENCES

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, Jan. 2001.

[2] M. Chen, P. Zhou, and G. Fortino, "Emotion communication system," *IEEE Access*, vol. 5, pp. 326–337, Dec. 2016.

[3] M. S. Hossain and G. Muhammad, "An emotion recognition system for mobile applications," *IEEE Access*, vol. 5, pp. 2281–2287, Feb. 2017.

[4] B. G. Lee, T. W. Chong, B. L. Lee, H. J. Park, Y. N. Kim, and B. Kim, "Wearable mobile-based emotional response-monitoring system for drivers," *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 5, pp. 636–649, Oct. 2017.

[5] Y. Kim, T. Soyata, and R. F. Behnagh, "Towards emotionally-aware AI smart classroom: Current issues and directions for engineering and education," *IEEE Access*, vol. PP, no. 99, pp. 1–1, Jan. 2018.

[6] A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," *IEEE Intelligent Systems*, vol. 24, no. 2, pp. 8–12, Mar. 2009.

[7] J. Deng, S. Frhholz, Z. Zhang, and B. Schuller, "Recognizing emotions from whispered speech based on acoustic feature transfer learning," *IEEE Access*, vol. 5, pp. 5235–5246, Mar. 2017.

[8] Z. Zhang, N. Cummins, and B. Schuller, "Advanced data exploitation for speech analysis: An overview," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 107–129, July 2017.

[9] S. Hantke, T. Appel, F. Eyben, and B. Schuller, "iHEARu-PLAY: Introducing a game for crowdsourced data collection for affective computing," in *Proc. International Conference on Affective Computing and Intelligent Interaction (ACII)*, Xi'an, China, 2015, pp. 891–897.

[10] B. Wu, E. Zhong, D. H. Hu, A. Horner, and Q. Yang, "SMART: Semi-supervised music emotion recognition with social tagging," in *Proc. SIAM International Conference on Data Mining (SDM)*, Austin, TX, 2013, pp. 279–287.

[11] M. Schels, M. Kächele, M. Glodek, D. Hrabal, S. Walter, and F. Schwenker, "Using unlabeled data to improve classification of emotional states in human computer interaction," *Journal on Multimodal User Interfaces*, vol. 8, no. 1, pp. 5–16, Mar. 2014.

[12] M. Giulianelli, "Semi-supervised emotion lexicon expansion with label propagation and specialized word embeddings," *arXiv preprint arXiv:1708.03910*, Aug. 2017.

[13] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 9, pp. 1062–1087, Dec. 2011.

[14] Z. Zhang, F. Weninger, M. Wöllmer, and B. Schuller, "Unsupervised learning in cross-corpus acoustic emotion recognition," in *Proc. IEEE workshop on Automatic Speech Recognition and Understanding (ASRU)*, Big Island, HI, 2011, pp. 523–528.

[15] Z. Zhang, J. Deng, and B. Schuller, "Co-training succeeds in computational paralinguistics," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, pp. 8505–8509.

[16] Z. Zhang, E. Coutinho, J. Deng, and B. Schuller, "Cooperative learning and its application to emotion recognition from speech," *IEEE Transactions on Audio, Speech, Language Processing*, vol. 23, no. 1, pp. 115–126, Jan. 2015.

[17] S. Li, C.-R. Huang, G. Zhou, and S. Y. M. Lee, "Employing personal/impersonal views in supervised and semi-supervised sentiment classification," in *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, Uppsala, Sweden, 2010, pp. 414–423.

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

[19] A. Mahdhaoui and M. Chetouani, "Emotional speech classification based on multi-view characterization," in *Proc. IEEE International Conference on Pattern Recognition (ICPR)*, Istanbul, Turkey, 2010, pp. 4488–4491.

[20] I. Cohen, N. Sebe, F. G. Cozman, M. C. Cirelo, and T. S. Huang, "Learning bayesian network classifiers for facial expression recognition both labeled and unlabeled data," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Madison, WI, 2003, p. 595.

[21] I. Cohen, N. Sebe, F. G. Cozman, and T. S. Huang, "Semi-supervised learning for facial expression recognition," in *Proc. ACM SIGMM international workshop on Multimedia Information Retrieval (MIR)*, New York, NY, 2003, pp. 17–22.

[22] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, Jan. 2009.

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

[24] S. Petridis and M. Pantic, "Prediction-based audiovisual fusion for classification of non-linguistic vocalisations," *IEEE Transactions on Affective Computing*, vol. 7, no. 1, pp. 45–58, Jan. 2016.

[25] O. Celiktutan and H. Gunes, "Automatic prediction of impressions in time and across varying context: Personality, attractiveness and likeability," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 29–42, Jan. 2017.

[26] J. Han, Z. Zhang, F. Ringeval, and B. Schuller, "Prediction-based learning for continuous emotion recognition in speech," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, 2017, pp. 5005–5009.

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

[28] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 211–223, Apr. 2012.

[29] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "AVEC 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proc. International Workshop on Audio/Visual Emotion Challenge (AVEC)*, Amsterdam, Netherlands, 2016, pp. 3–10.

[30] X. Zhu, "Semi-supervised learning literature survey," Department of Computer Sciences, University of Wisconsin at Madison, Madison, WI, Tech. Rep. TR 1530, 2006.

[31] D. Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods," in *Proc. Annual meeting of the Association for Computational Linguistics (ACL)*, Stroudsburg, PA, 1995, pp. 189–196.

[32] I. Cohen, F. G. Cozman, N. Sebe, M. C. Cirelo, and T. S. Huang, "Semi-supervised learning of classifiers: Theory, algorithms, and their application to human-computer interaction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 12, pp. 1553–1566, Dec. 2004.

[33] Z. Zhang, F. Ringeval, B. Dong, E. Coutinho, E. Marchi, and B. Schuller, "Enhanced semi-supervised learning for multimodal emotion recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 2016, pp. 5185–5189.

[34] Z.-H. Zhou, "When semi-supervised learning meets ensemble learning," in *Proc. International Workshop on Multiple Classifier Systems (MCS)*, Reykjavik, Iceland, 2009, pp. 529–538.

[35] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using EM," *Machine learning*, vol. 39, no. 2-3, pp. 103–134, May 2000.

[36] M. Li and Z.-H. Zhou, "SETRED: Self-training with editing," in *Proc. Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, Hanoi, Vietnam, 2005, pp. 611–621.

[37] O. Kursun, H. Seker, F. G?rgen, N. Aydin, O. V. Favorov, and C. O. Sakar, "Parallel interacting multiview learning: An application to prediction of protein sub-nuclear location," in *Proc. International Conference on Information Technology and Applications in Biomedicine (ITAB)*, Larnaca, Cyprus, 2009, pp. 1–4.

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

[39] Z. Zhang and B. Schuller, "Active learning by sparse instance tracking and classifier confidence in acoustic emotion recognition," in *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Portland, OR, 2012, pp. 362–365.

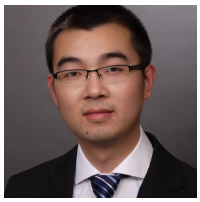
[40] Z. Zhang, F. Eyben, J. Deng, and B. Schuller, "An agreement and sparseness-based learning instance selection and its application to subjective speech phenomena," in *Proc. International Workshop on Emotion Social Signals, Sentiment & Linked Open Data, satellite of LREC 2014*, Reykjavik, Iceland,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

[41] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. Annual Conference on Computational Learning Theory (COLT)*, Madison, WI, 1998, pp. 92–100.

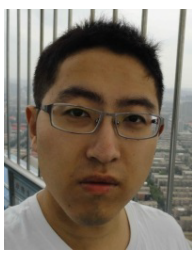
[42] S. Sun, "A survey of multi-view machine learning," *Neural Computing and Applications*, vol. 23, no. 7-8, pp. 2031–2038, Dec. 2013.
[43] G. James, "Majority vote classifiers: theory and applications," Ph.D. dissertation, Stanford University, Stanford, CA, 1998.
[44] Z.-H. Zhou, *Ensemble methods: foundations and algorithms*. Boca Raton, FL: CRC press, 2012.
[45] V. Vapnik, *The nature of statistical learning theory*, 2nd ed. Berlin/Heidelberg, Germany: springer, 1999.
[46] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *Proc. IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Shanghai, China, 2013, pp. 1–8.
[47] B. Schuller, S. Steidl, A. Batliner, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom, A. Warlaumont, G. Hidalgo, S. Schnieder, C. Heiser, W. Hohenhorst, M. Herzog, M. Schmitt, K. Qian, Y. Zhang, G. Trigeorgis, P. Tzirakis, and S. Zafeiriou, "The INTERSPEECH 2017 Computational Paralinguistics Challenge: Addressee, Cold & Snoring," in *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Stockholm, Sweden, 2017, pp. 3442–3446.
[48] F. Eyben, *Real-time speech and music classification by large audio feature space extraction*. Berlin/Heidelburg, Germany: Springer, 2016.
[49] F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J.-P. Thiran, T. Ebrahimi, D. Lalanne, and B. Schuller, "Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data," *Pattern Recognition Letters*, vol. 66, pp. 22–30, Nov. 2015.
[50] A. Dhall, R. Goecke, J. Joshi, J. Hoey, and T. Gedeon, "EmotiW 2016: Video and group-level emotion recognition challenges," in *Proc. ACM International Conference on Multimodal Interaction (ICMI)*, Tokyo, Japan, 2016, pp. 427–432.
[51] J. Wei, E. Pei, D. Jiang, H. Sahli, L. Xie, and Z. Fu, "Multimodal continuous affect recognition based on LSTM and multiple kernel learning," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, Chiang Mai, Thailand, 2014, pp. 1–4.
[52] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, Mar. 2016.
[53] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, p. 27, Apr. 2011.
[54] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, Vancouver, Canada, 2003, pp. 321–328.
[55] O. Delalleau, Y. Bengio, and N. L. Roux, "Efficient non-parametric function induction in semi-supervised learning," in *Proc. International Workshop on Artificial Intelligence and Statistics (AISTATS)*, Bridgetown, Barbados, 2005.
[56] J. T. Springenberg, "Unsupervised and semi-supervised learning with categorical generative adversarial networks," in *Proc. International Conference on Learning Representations (ICLR)*, San Juan, PR, 2016, 20 pages.
[57] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, Montreal, Canada, 2014, pp. 2672–2680.

Zixing Zhang (M'15) received his master degree in physical electronics from Beijing University of Posts and Telecommunications (BUPT), China, in 2010, and his PhD degree in computer engineering from Technical University of Munich (TUM), Germany, in 2015. Currently, he is a research associate with the Department of Computing at the Imperial College London (ICL), UK, from 2017. Before that, he was a postdoctoral researcher at the University of Passau, Germany, from 2015 to 2017. He has authored more than sixty publications in peer-reviewed books, journals, and conference proceedings to date, and has organised special sessions, such as at the IEEE 7th Affective Computing and Intelligent Interaction (ACII) and at the 42nd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) in 2017. Dr Zhang serves as a reviewer for leading-in-their fields journals such as IEEE T-NNLS, IEEE T-CYB, IEEE T-AC, IEEE T-MM, IEEE T-ASLP, Speech Communication, and Computer Speech & Language. His research interests lie in various machine learning techniques (e. g., semi-supervised learning and deep learning), which aim to leverage the value of large-scale unlabelled data for automatic speech analysis (e. g., emotion recognition).

Jing Han (S'16) received her bachelor degree (2011) in electronic and information engineering from Harbin Engineering University (HEU), China, and her master degree (2014) in Nanyang Technological University, Singapore. She is now working as a doctoral student with the Chair of Embedded Intelligence for Health Care and Wellbeing at the University of Augsburg, Germany, involved in the EU's Horizon 2020 programme SEWA. She reviews regularly for IEEE Transaction on Cybernetics and IEEE Signal Processing Letter. Her research interests are related to deep learning for multimodal affective computing.



Jun Deng (S'13-M'16) received his bachelor degree (2009) in electronic and information engineering from Harbin Engineering University (HEU) and his master degree (2011) in information and communication engineering from Harbin Institute of Technology (HIT), Harbin/China, his doctor degree (2016) for his study on Feature Transfer Learning for Speech Emotion Recognition, in electrical engineering and information technology from Technische Universität München (TUM), Germany. Currently, he is a lead researcher at audEERING GmbH, Gilching, Germany. Before that, he was a post-doctoral researcher at the Chair of Complex and Intelligent Systems at the University of Passau, Germany, where he was involved in the EU-FP7 starting grant project iHEARu and the Horizon 2020 project DE-ENIGMA. His research interests are machine learning methods such as transfer learning and deep learning with an application preference to affective computing. He reviews regularly for IEEE Transactions on Affective Computing, IEEE Signal Processing Letters, Speech Communication, and Computer Speech & Language.



Xinzhou Xu received the bachelor's degree from Nanjing University of Posts and Telecommunications, Nanjing/China, in 2009, the master's and the PhD degree from Southeast University, Nanjing/China, in 2012 and 2017, respectively. He is currently a Lecturer with the School of Internet of Things, Nanjing University of Posts and Telecommunications. Previously, he was with the Machine Intelligence & Signal Processing group, MMK, Technische Universität München (TUM), Munich/Germany (from 2014 to 2016), and the Chair of Complex and Intelligent Systems, University of Passau, Passau/Germany (from 2015 to 2016). His research interests include spoken signal processing, pattern recognition, machine learning, and affective computing.

Fabien Ringeval received his master degree in 2006, and his doctoral degree for his research on the automatic recognition of acted and spontaneous emotions from speech in 2011, from the Université Pierre et Marie Curie (UPMC), Paris, France. He is Associate Professor in Informatics at the Laboratoire d'Informatique de Grenoble (LIG), CNRS, Université Grenoble Alpes, France, since 2016. His research interests concern digital signal processing and machine learning, with applications on the automatic recognition of paralinguistic information (e.g., emotions, social and atypical behaviours) from multimodal data (e.g., audio, video and physiological signals). Dr. Ringeval (co-)authored more than 60 publications in peer-reviewed books, journals and conference proceedings in the field leading to more than 1 300 citations (h-index = 15). He co-organised several workshops and international challenges (ComParE 2013, AVEC 2015-2017), serves as Grand Challenge Chair for the 20th ACM International Conference on Multimodal Interaction (ICMI 2018), Publication Chair for the 7th AAAC International Conference on Affective Computing and Intelligent Interaction (ACII 2017), and as a reviewer for funding and projects (ANR, ANRT, NSERC), several IEEE and other leading international journals, conferences and workshops in the field.

Björn Schuller (M'05-SM'15-F'18) received his diploma in 1999, his doctoral degree for his study on Automatic Speech and Emotion Recognition in 2006, and his 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 TUM in Munich/Germany. He is Reader in Machine Learning in the Department of Computing at the Imperial College London/UK, Full Professor and head of the ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing at the University of Augsburg/Germany, and an Associate of the Swiss Center for Affective Sciences at the University of Geneva/Switzerland. He was previously full professor and head of the Chair of Complex and Intelligent Systems at the University of Passau/Germany. Professor Schuller is Fellow of the IEEE, President-emeritus of the Association for the Advancement of Affective Computing (AAAC), elected member of the IEEE Speech and Language Processing Technical Committee, and Senior Member of the ACM, and (co-)authored 5 books and more than 700 publications in peer-reviewed books, journals, and conference proceedings leading to more than overall 18 000 citations (h-index = 65). Schuller is co-Program Chair of Interspeech 2019, repeated Area Chair of ICASSP, and Editor in Chief of the IEEE Transactions on Affective Computing next to a multitude of further Associate and Guest Editor roles and functions in Technical and Organisational Committees.