A Generic Human–Machine Annotation Framework Based on Dynamic Cooperative Learning

A Generic Human-Machine Annotation Framework Based on Dynamic Cooperative Learning

Yue Zhang, Andrea Michi, Johannes Wagner, Elisabeth André, Björn Schuller, Felix Weninger

***Abstract*—The task of obtaining meaningful annotations is a tedious work, incurring considerable costs and time consumption. Dynamic active learning and cooperative learning are recently proposed approaches to reducing human effort of annotating data with subjective phenomena. In this work, we introduce a novel generic annotation framework, with the aim to achieve the optimal trade-off between label reliability and cost reduction by making efficient use of human and machine work force. To this end, we use dropout to assess model uncertainty and thereby to decide which instances can be automatically labelled by the machine and which ones require human inspection. Additionally, we propose an early stopping criterion based on inter-rater agreement in order to focus human resources on those ambiguous instances that are difficult to label. In contrast to the existing algorithms, the new confidence measures are not only applicable to binary classification tasks, but also regression problems. The proposed method is evaluated on the benchmark datasets for non-native English prosody estimation, provided in the INTERSPEECH Computational Paralinguistics Challenge. In the result, the novel dynamic cooperative learning algorithm yields .424 Spearman's correlation coefficient compared to .413 with passive learning, while reducing the amount of human annotations by 74 %.**

***Index Terms*—Human-Machine Systems, Active Learning, Semi-supervised Learning, Confidence Measures, Inter-rater Agreement**

I. INTRODUCTION

WITHIN the research fields intersecting with machine learning, it is widely acknowledged that "there is no data like more data" (Bob Mercer, 1985). Deep learning has achieved new state-of-the-art performances on a vast array of recognition tasks. However, these gains are often difficult to translate into real-world settings as they require large hand-labelled training sets. With the massive growth of data created every instant, manual annotation has become the major bottleneck of data processing due to prohibitive costs and high time consumption. Thus, in recent years, much research effort has been undertaken to minimise human labelling effort while ensuring label quality.

To leverage the massive amounts of unlabelled data, various machine learning techniques have been devised, including semi-supervised learning (SSL) [1], Active Learning (AL) [2] and various combinations thereof [3], [4], [5]. The general principle behind these methods is to iteratively train a model by adding new (machine or human) labelled instances to the training set. Dispensing with the need for human annotation, SSL techniques (e. g., self-training [6]) adopt the machine-predicted labels with high model confidence. Complementary to SSL, AL query strategies (e. g., uncertainty sampling [7]) select the most informative instances, e. g., those with low model confidence, to be inspected by humans. Cooperative learning is a technique that distributes the work among human and machine labellers [5]. It is easy to see that in all the above mentioned approaches, an accurate *uncertainty measure* is key to model performance and cost efficiency. For (binary) classification tasks, model confidence can be assessed based on posterior probabilities of support vector machines (SVMs) or other classifiers [5]. However, this confidence measure is limited to binary classification and cannot be generalised to multi-class or regression problems in a straightforward way.

Label subjectivity, as with most affective and paralinguistic phenomena, poses a particular challenge for data annotation due to the necessity for laborious inter-rater agreement procedures. To approximate the "gold standard" (pseudo-truth) from a collection of possibly noisy annotations, majority voting (for nominal class labels) or measures of central tendency such as median or mean (for ordinal/ interval label scales) are usually used. However, the set of assumptions underlying these standard settings do not necessarily apply to real-world problems, namely that the instances are equally ambiguous and thus require the same number of queries. The drawback of majority voting among a fixed number of annotators has been addressed in the recent works [8], [9], introducing dynamic active learning (DAL) as a novel approach to fine-tuning the number of human annotations on a per instance level. The core idea is to allocate increased human resources to controversial instances while making less queries if a consensus can be easily reached. To this end, an early-stopping criterion has been introduced, setting a minimum number of votes for either category (referred to as *agreement level* in the work [10]). However, again, this is only applicable to discrete categories.

Thus, in this work, we introduce a novel human-machine annotation framework that can handle both classification and regression problems. The main contributions of this work lie in four aspects:

- We propose a novel dynamic cooperative learning (DCL) algorithm, which combines dynamic active learning [8] with cooperative learning [5] to minimise the amount of human annotations.

Y. Zhang is now with the Affective Computing Group, MIT Media Lab, Massachusetts Institute of Technology, Cambridge, MA, USA. The work was done while at Imperial College London, UK (e-mail: yuefw@mit.edu).

A. Michi is now with Google DeepMind, London, UK. The work was done while at Imperial College London, UK (e-mail: andrea.michi12@imperial.ac.uk).

J. Wagner and E. André are with the University of Augsburg, Germany (e-mail: {andre,johannes.wagner}@informatik.uni-augsburg.de).

B. Schuller is with the Department of Computing, Imperial College London, UK (e-mail: bjoern.schuller}@imperial.ac.uk).

F. Weninger is with Nuance Communications, Burlington, MA, USA (e-mail: felix@weninger.de).

- We generalise the DAL algorithm to regression problems, using dropout training to capture model uncertainty [11] and a variance-based within-group agreement index [12].
- We introduce a generic human-machine annotation framework, which incorporates the state-of-the-art learning algorithms (e. g., passive learning, active learning, semi-supervised learning), as instances of the DCL algorithm.
- We compare the different learning schemes in terms of model accuracy and cost reduction on a subjective task.

The remainder of the paper is organised as follows: Section II reviews the literature on methods to assess prediction uncertainty and label reliability. Section III elaborates on the human-machine annotation framework. Section IV evaluates the new method on the task of non-native English prosody recognition. Section V concludes the work with an outlook on future research.

II. Related Work

Prediction uncertainty and inter-rater agreement play a key role in the two-tier decision process as outlined in Section I. First, one needs to decide whether machine or human should label an instance. Second, one needs to know if sufficient annotations have been obtained to determine the gold standard.

A. Prediction Uncertainty

Capturing prediction uncertainty to estimate generalisation error is crucial in probabilistic machine learning [13], [14]. In classification models such as SVMs and deep neural networks (DNNs), pseudo-posteriors output by softmax or maximum entropy functions ([15], [16]) are often interpreted as model uncertainty, which is, however, erroneous in the general case [11].

Regression models output a single vector that regresses to the mean of the training data, but does not directly capture model uncertainty [11]. To represent uncertainty in deep learning without sacrificing either computational complexity or test accuracy, Gal and Ghahramani [17] propose to utilise dropout training in DNNs as a Bayesian approximation of a deep Gaussian process [18], [19]. As demonstrated in their work, the dropout technique achieves considerable improvement in terms of predictive log-likelihood and root mean square error (RMSE) compared to other uncertainty measurement methods such as probabilistic back-propagation [20] and variational inference [21]. An approach similar to dropout employs neural network ensembles in active learning settings for regression [22]. Based on the principle of query by committee [23], [24], the variation of the output of ensemble members is used to select new training data. A major drawback of the ensemble method is, however, the high computational cost.

It is also noteworthy that the model confidence does not necessarily reflect model uncertainty (if an out-of-distribution example is extrapolated far from the decision hyperplane, the logistic regression model would have an unjustified high confidence [25]). For ease of exposition though, we use the term confidence to refer to traditional confidence measures as well as model certainty.

B. Label Quality

The annotation procedures are often evaluated by means of interrater agreement indices, signifying the absolute rater consensus based on the within-group rating dispersion [12]. In comparison, interrater reliability refers to the relative rater consistency via correlation coefficients[1] [26]. In the study [27], the system performance is modelled as a function of the number of annotators, as well as various parameters (e. g., the difficulty of an annotation task). For the recognition of non-native English prosody, the authors have devised a rule of thumb: The gain in label quality is the highest from one to five labellers, and still noticeable from six to some ten [27]. Although this rule cannot be generalised to other subjective tasks, their study provides evidence that the number of annotators is a key variable for striking a balance between system performance and costs.

Besides the quantity of annotations, the quality of gold-standard assessment heavily depends on the rater reliability. This is amplified by the emergence of crowd-sourcing [28], which induces a shift from small-scale, expensive expert coding towards large-scale, low-cost labelling by clickworkers, who usually have no formal training at the task at hand. Sheng et al. [29] proposed to revise the annotation for some data points in the case of noisy labelling. Their study shows that re-labelling can be conducive to quality control especially with crowdsourcing. Raykar et al. [30], [31] proposed a maximum likelihood estimator that jointly learns the classifier, the experts' performance, and the actual true label. Donmez et al. [32] proposed an algorithm for estimating the reliability of multiple labellers and filtering out the best one(s) for active learning.

The dynamic active learning algorithm differs from the previous body of research in augmenting active learning with adaptive query strategies based on inter-rater agreement and reliability. Depending on the agreement level, it successively acquires new labels to compute the gold-standard, thereby systematically adjusting the amount of annotations for each individual instance. Moreover, it gives precedence to the most reliable annotators, whose reliability is appraised on a held-out dataset. In this work, we posit that there is an hitherto unexploited potential for labelling efficiency in combining cooperative learning and dynamic active learning to dynamic cooperative learning.

III. Proposed Framework

While dynamic active learning targets the human annotation part, cooperative learning aims to efficiently distribute the work among human and machine. This section describes the generic human-machine annotation framework, which can be configured to perform different learning algorithms within the following degrees of freedom: 1) Human-machine arbitration based on confidence measure; 2) Enabling early stopping based on inter-rater agreement; 3) Sorting the annotators by reliability.

[1]Note that the words "annotator", "rater" and "labeller" are used interchangeably in this paper.

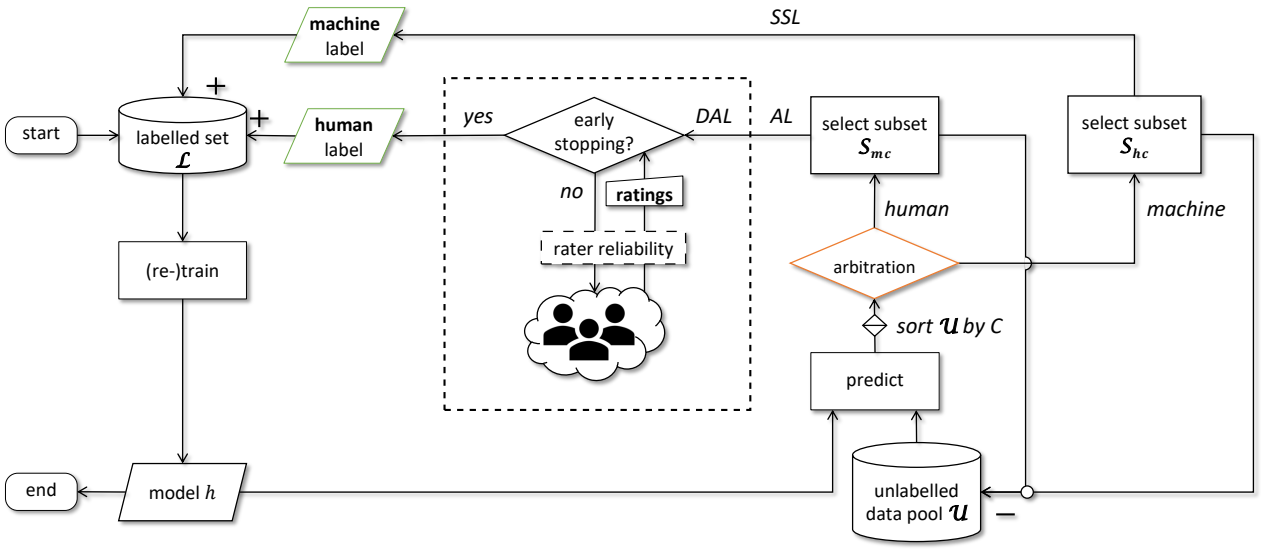


Fig. 1. Flowchart of the human-machine annotation framework with support for various learning paradigms including semi-supervised learning (SSL), active learning (AL), dynamic active learning (DAL), cooperative learning (CL), and dynamic cooperative learning (DCL); the instances predicted with high confidence (hc) are automatically labelled by machine, whereas the instances with medium confidence (mc) are subject to human inspection.

A. Dynamic Cooperative Learning

The functional diagram of the proposed framework is shown in Fig. 1. The learning cycle starts with training a model h on a small labelled training set $\mathcal{L}$. Next, h is applied to a large unlabelled dataset $\mathcal{U}$. According to the model confidence C, the instances are ranked into quantiles. In the human-machine arbitration step, the learner has the choice to adopt the predictions of h with high confidence (SSL, self-training), or to ask human annotators to inspect the uncertain regions (AL, uncertainty sampling), or to do both (cooperative learning). The rationale for medium certainty sampling is that it helps avoid selecting noisy data points [5], [10]. The main difference to the previous work [5] is the usage of a probabilistic confidence measure (dropout) instead of a confidence measure based on logistic regression (cf. Section I). In each iteration, the new labelled instances are added to the training set $\mathcal{L}$ and removed from $\mathcal{U}$. Then, the model h is re-trained and tested on a speaker-disjunctive test set $\mathcal{T}$. Finally, the learning process finishes by producing a model h_{end} when all data from $\mathcal{U}$ are annotated if not otherwise specified. Considering the fact that model performance usually starts to deteriorate at some point, the learning process can be stopped after having reached the best accuracy on a validation set, or if the number of remaining data points in $\mathcal{U}$ falls below a threshold $U_{\min}$. The DCL algorithm is described in Fig. 2.

B. Generic Learning Algorithm

Table I represents the existing learning algorithms as instantiations of the generic DCL algorithm (cf. Fig. 2). The configuration is done via human-machine arbitration based on model confidence, and the minimum required number of queries $Q_{\min}$ per instance. SSL employs automatic machine labelling with high model confidence, thus reducing the human effort to zero ($\mathcal{J}_H = \emptyset$). On the contrary, AL only relies on human annotation, enlisting all the available raters in $\mathcal{R}$, while the machine predictions are unused ($\mathcal{J}_M = \emptyset$). As described in Section II, the difference between AL and DAL is that the number of queries per instance can be less than the number of available annotators R, hence $Q_{\min} < R$. CL employs both machine and human labels based on model confidence, however, requires the same number of annotators for each instance $Q_{\min} = R$. On top, the proposed DCL algorithm enables early-stopping if a certain level of agreement has been reached. In addition, we compare to passive learning (PL) that randomly samples instances for manual labelling.

C. Deep Rectifier Neural Networks

In this study, we employ deep rectifier neural networks as learning models. Mathematically, an H-layer DNN with output $\mathbf{y} = N(\mathbf{x}, \mathbf{w})$ is defined as a composition of multiple non-linear transformations of an input feature vector $\mathbf{x}$

$$N(\mathbf{x}, \mathbf{w}) = \mathcal{H}(\mathbf{W}_H \mathbf{h}) = \mathcal{H}(\mathbf{W}_H \mathcal{G}(\mathbf{W}_{H-1}(\cdots \mathcal{G}(\mathbf{W}_1 \mathbf{x})))), \tag{1}$$

with per-layer weight matrices $\mathbf{W}_1, \ldots, \mathbf{W}_H$ stacked into a column vector $\mathbf{w}$, an output layer activation function $\mathcal{H}$ and a hidden layer activation function $\mathcal{G}$. In case of deep rectifier neural networks, $\mathcal{G}$ is defined as

$$\mathcal{G}(x) = \max(0, x). \tag{2}$$

The parameters $\mathbf{w}$ are optimised by means of error back-propagation and stochastic gradient descent (SGD) on a set of training vectors $\mathcal{X}$ with the corresponding ground-truth labels $\mathcal{Y}$. The rectified linear activation function is biologically inspired [33] and, on a practical level, it mitigates the vanishing gradient problem in DNN training, thus allowing for dispensing with pre-training [34].

Algorithm: *Generic Form of the DCL Algorithm*

Input: Original training set $\mathcal{L} = \{(\mathbf{x}_i, y_i) \mid y_i \neq \perp\}$
Unlabelled set $\mathcal{U} = \{(\mathbf{x}_i, y_i) \mid y_i = \perp\}$
Pool of raters $\mathcal{R}$
Min. number of queries $Q_{\min}$
Max. number of queries $Q_{\max} \leq |\mathcal{R}|$
Agreement level α for early stopping
Min. number of unlabelled instances $U_{\min}$
Output: Labelled data $\{(\mathbf{x}_j, y_j) \mid \mathbf{x}_j \in \mathcal{U}\}$
Final model h_{end}

Do:
$\mathcal{U} := \{\mathbf{x}_j \mid y_j = \perp\}$
$\mathcal{L} := \{\mathbf{x}_i \mid y_i \neq \perp\}$
$h := \text{Train}(\mathcal{L})$
$\tilde{\mathcal{Y}}, C := \text{Predict}(h, \mathcal{U})$ // Prediction + confidence measure
$\mathcal{J}_M, \mathcal{J}_H := \text{Select}(\mathcal{U}, C)$ // Arbitration + instance selection
For $j \in \mathcal{J}_M$: // Machine labelling
$y_j = \tilde{y}_j$
For $j \in \mathcal{J}_H$: // Human labelling
Do:
$r \in \text{Randomise}(\mathcal{R})$ // Optional: sort $\mathcal{R}$ by reliability
$y_{j,r} = \text{Query}(r, x_j)$ // Ask rater r about instance $\mathbf{x}_j$
$\mathcal{Y}_j := \mathcal{Y}_j \cup \{y_{j,r}\}$ // Set of human labels
If $(\text{Agreement}(\alpha, \mathcal{Y}_j) \wedge |\mathcal{Y}_j| \geq Q_{\min})$: **Break**
While $(|\mathcal{Y}_j| < Q_{\max})$
$y_j := \text{Gold-standard}(\mathcal{Y}_j)$ // e. g., mean, majority vote
While $|\mathcal{U}| > U_{\min}$
$h_{\text{end}} := h$

Fig. 2. Pseudo-code description of the generic Dynamic Cooperative Learning (DCL) algorithm, which combines Semi-Supervised Learning (SSL) and Dynamic Active Learning (DAL) by means of human-machine collaboration.

D. Generic Confidence Measure

To capture prediction uncertainty of DNN models, we use the dropout based technique proposed by Gal and Ghahramani [17]. Dropout is a widely used regularisation technique for reducing overfitting during training. It randomly drops units (along with their connections) from the neural network, thereby preventing units from co-adapting too much [35]. At test time, dropout can also be used to measure the prediction uncertainty as the output variance across T forward passes, each within a different "thinned" network.

Mathematically, we compute the Monte Carlo estimates of the mean and variance of the distribution

$$p(\mathbf{y}|\mathbf{x}, \mathcal{X}, \mathcal{Y}) = \int p(\mathbf{y}|\mathbf{x}, \mathbf{w}) p(\mathbf{w}|\mathcal{X}, \mathcal{Y}) d\mathbf{w}. \tag{3}$$

For the reason that $p(\mathbf{w}|\mathcal{X}, \mathcal{Y})$ can be generally interpreted, it is estimated by a distribution $q(\mathbf{w})$ over weight matrices with randomly applied dropout. It has been shown in the study [17] that this is a reasonable approximation: In fact, neural network training with dropout and L2 regularisation is equivalent to minimising the Kullback-Leibler (KL) divergence between q and a deep Gaussian process $p(\mathbf{w}|\mathcal{X}, \mathcal{Y})$ (marginalised over its finite rank covariance function parameters) [17]. Sampling T sets of weight vectors $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(T)}$ from $q(\mathbf{w})$, the Monte Carlo estimate of the mean $\overline{\mathbf{y}}$ is obtained by

$$\overline{\mathbf{y}} = \frac{1}{T} \sum_{t=1}^{T} N(\mathbf{x}, \mathbf{w}^{(t)}). \tag{4}$$

In practice, the sampling from $q(\mathbf{w})$ is implemented as T forward passes using random dropout masks for the weights. The optimal number of T passes mainly depends on the size of the model, the data provided, and the percentage of dropout. A small value for T speeds up the computation, however, it can affect the accuracy of the results and the merit of the uncertainty measure. In analogy, the covariance $\Sigma_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$ is estimated as

$$\Sigma_{NN} = \zeta + \frac{1}{T} \sum_{t=1}^{T} N(\mathbf{x}, \mathbf{w}^{(t)}) N(\mathbf{x}, \mathbf{w}^{(t)})^{\intercal} - \overline{\mathbf{y}}\overline{\mathbf{y}}^{\intercal}, \tag{5}$$

where ζ is a constant that depends on the model precision (uncertainty of the output). Within the context of DAL, the automatically predicted label $\tilde{y}_j$ is equivalent to $\overline{\mathbf{y}}$. In the case of regression, high variances are directly interpretable as high uncertainty (low confidence of the DNN in its prediction). This is convenient since only the rank of the instances w.r.t. the confidence measure is important, whereas the value of ζ, as well as the actual range of the variances – which depend on the data – is irrelevant in our context. An advantage of the dropout method is that it can be applied for any network topologies, and in particular, for both regression and classification output layers. For multi-class problems, the methods presented in the work [17] such as predictive entropy [36] can be applied to extract uncertainty information from the softmax output layer.

TABLE I
INSTANTIATIONS OF THE GENERIC LEARNING FRAMEWORK. $\mathcal{J}_M$ AND $\mathcal{J}_H$ ARE THE SETS OF DATA POINTS LABELLED BY MACHINE OR HUMAN; $Q_{\min}$ DENOTES THE MINIMUM NUMBER OF QUERIES PER INSTANCE. $R = |\mathcal{R}|$ IS THE NUMBER OF AVAILABLE ANNOTATORS.

Algorithm	$\mathcal{J}_M$	$\mathcal{J}_H$	$Q_{\min}$
PL	$\emptyset$	Random Sampling	$= R$
SSL	High conf.	$\emptyset$	N/A
AL	$\emptyset$	Medium/ Low Confidence	$= R$
DAL	$\emptyset$	Medium/ Low Confidence	$< R$
CL (AL + SSL)	High Conf.	Medium/ Low Confidence	$= R$
DCL (DAL + SSL)	High Conf.	Medium/ Low Confidence	$< R$

E. Generic Early Stopping Criterion

The crux of the dynamic query strategy is *early stopping*, which depends on the labelling scheme, i. e., nominal level or ordinal/ interval scale. For binary tasks, the agreement level α is defined as the minimum number of votes to be obtained for either category. Given the number R of available annotators, meaningful values for α range from $2, \ldots, \lfloor \frac{R+1}{2} \rfloor$, where the upper limit of the interval is rounded down if R is an even number. Since regression labels are computed as the mean of all ratings, there does not exist such a deterministic condition since the resulting value can still change. Nevertheless, the concept of agreement levels can be generalised by using the inter-rater agreement indices as detailed in the following.

As a generic measure of label reliability, we use the within-group agreement index r_{wg} [37] defined as

$$r_{wg} = 1 - \frac{\sigma_Y^2}{\sigma_E^2}, \tag{6}$$

where σ_Y^2 is the sample variance of the ratings and σ_E^2 is the expected variance of the random ratings drawn from a null distribution, which is generally assumed to be a uniform distribution [12]. The early stopping is triggered if the value of r_{wg} rises above a threshold value of α (cf. Fig. 2). This bears some similarity to the early stopping technique sometimes used in neural network training, where the training procedure is interrupted once the generalisation capability of the network is deemed to stagnate. Just as its deep learning analogy, our early stopping of the rating procedure remains a heuristic – there is no guaranteed convergence of r_{wg}. Hence, the convergence of r_{wg} requires that the label subjectivity is limited to a certain extent, say, the agreement-based procedure is not suitable for overly ambiguous phenomena such as perceived voice likeability [38]. For example, the degree of nativeness, speech emotion, and interest recognition represent mildly subjective tasks [25]. It remains to mention that in the sequential annotation process, whether an annotator is asked or not depends on the previous outcomes. In practice, this be realised by managing queries in a queue-like system.

IV. Empirical Evaluation

In the experiments, we apply the different instantiations of the generic learning framework to a typical regression problem in computational paralinguistics. In the "Degree of Nativeness" task, the non-native prosody of English L2 speakers is to be recognised on a continuous scale. Located on the word level and above, prosodic speech phenomena encompass word accent position, syntactic-prosodic boundaries, sentence melody and rhythm [39].

To evaluate the proposed framework in realistic settings, we perform cross-corpus experiments, using data resources drawn from different acoustic settings, recording equipments, sound environment, and speech material. To simulate the small labelled set, the large unlabelled set, and the speaker-independent test set, we use the datasets provided in the INTERSPEECH Computational Paralinguistics Challenge (ComParE) [40]. It is noted that the sets of annotators are disjoint on these datasets, which further fosters realism.

A. Datasets

The Nativeness Corpus (NC) [41] serves as the small labelled set $\mathcal{L}$ for the first training iteration. It contains voice recordings from 54 non-native English speakers with varying degree of spoken language proficiency (gender: 28 female, 26 male; age: 31.3 $\pm$ 9.0; native languages: 22 German, 13 Chinese, 4 Arabic, and 15 other). Each speaker read aloud a set of 11 sentences taken from two standard stories in phonetics written in English ("The North Wind and the Sun" and "The Rainbow"). The dataset has 594 speech files, totalling 1.4 hours of speech. For the purpose of annotation, perceptive experiments were conducted using the web tool PEAKS [42]. A group of 27 native English speakers were instructed to rate the prosody on a 5-point Likert scale (1 – normal; 2 – acceptable; 3 – slightly unusual; 4 – unusual; and 5 – very unusual). The obtained labels range from 1.1 to 5.0, with an average of 2.9 and a standard deviation of 0.7.

To simulate the large unlabelled data pool $\mathcal{U}$, we use the automatic web-based learner-feedback (AUWL) corpus [43]. In AUWL, L2 learners of English practised pre-scripted dialogues by using a standard web-based dialogue training tool, but their own recording hardware, which resulted in different sound qualities. In total, 5.5 hours of speech were obtained from 31 speakers (gender: 13 female, 18 male; age: 36.5 $\pm$ 15.3 years; native languages: 16 German, 4 Italian, 3 Chinese, 3 Japanese, and 5 other). Similar to the annotation scheme as described above, five phoneticians rated the sentence prosody for each of the 3 732 speech files. The reference prosody scores are derived by computing the arithmetic mean of the ratings, with an average of 1.7 and a standard deviation of 0.5 (range 1.0–3.8) [44].

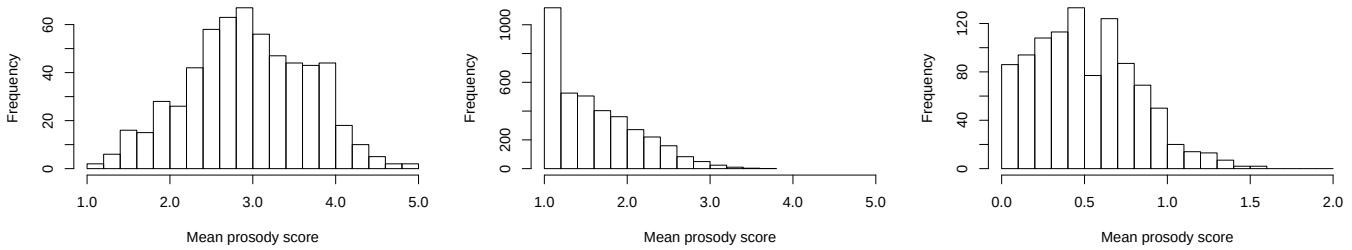
As in the Challenge, the computer-assisted pronunciation and dialogue training (C-AuDiT) corpus [43] is used as the test set. It comprises 2.7 hours of read speech recorded with standard headphones in a quiet office environment. The 999 speech files contain sentences in English targeting different phonetic patterns such as intonation of phrase accent and tongue twisters. These were read aloud by 58 speakers (gender: 31 female, 27 male, native languages: 26 German, 10 French, 10 Spanish, 10 Italian, and 2 Hindi). The recordings were annotated by 21 native English speakers on three-point scales from 0 for 'good' to 2 for 'bad' (0.5 $\pm$ 0.3, range 0.0–1.6). Although this scale differs from the one used for the training set, it is still valid for evaluation (on a common set of 290 sentences, the association between the two scales corresponds to a Spearman's rank correlation coefficient (CC) of $\rho = 0.73$).

Fig. 3 depicts the distribution of the prosody scores in the data partitions (training set, unlabelled set, and test set). It can be observed that the ratings in the NC database are approximately normally distributed, while the distributions of ratings in both the AUWL and the C-AuDiT databases are positively skewed, i. e., imbalanced towards 'good' scores. This shows a typical training/test data mismatch encountered in cross-corpus settings.

In our evaluation, the number of queries per instance only varies on the AUWL database, which represents the unlabelled set $\mathcal{U}$. In particular, for the evaluation on the test set, the gold standard label, serving as a reference for the predictions, is fixed, i. e., it does not depend on the number of raters requested during iterative training, thus allowing for a fair comparison of the methods employing static or dynamic numbers of queries.

B. Acoustic Features for Prosody Prediction

The ComParE set of supra-segmental acoustic features [45], [46] serves as the standard feature set for paralinguistic analysis. It contains 6 373 static features obtained from the computation of various functionals over low-level descriptor (LLD) contours. For its extraction, we used openSMILE in its 2.1 release [45]. Important subgroups of the ComParE feature



(a) NC (b) AUWL (c) C-AuDiT

Fig. 3. Distribution of prosody scores in the NC (training set $\mathcal{L}$), AUWL (unlabelled set $\mathcal{U}$), and C-AuDiT (test set $\mathcal{T}$) datasets.

set comprise prosodic, Mel Frequency Cepstral Coefficients (MFCCs), spectral, and voice quality features. Of particular importance for assessing the sentence melody and rhythm are the prosodic features based on loudness, energy, pitch, and pauses. Besides, MFCC features are among the most common speech features for automatic speech recognition and paralinguistic recognition tasks. A detailed analysis of the relevance of these feature groups for prosody prediction can be found in our previous work [47].

For dimensionality reduction, we use the correlation-based feature selection (CFS) [48] to eliminate redundant features in the ComParE set. The CFS algorithm searches for features which are highly correlated with the target label, yet uncorrelated with each other. Typically, CFS discards well over half of the features without sacrificing performance [48]. According to the BestFirst algorithm [49], feature subsets S are constructed in a forward search starting from an empty set by greedy hillclimbing augmented with backtracking. The merit M of a feature subset S with k features is given by

$$M(S) = \frac{k\ \mathrm{CC}_{\mathrm{cf}}}{\sqrt{k + k(k-1)\mathrm{CC}_{\mathrm{ff}}}}. \tag{7}$$

The backtracking level (number of iterations without improvement of M) is set to five. Feature selection is based only on the labelled data $\mathcal{L}$, resulting in 1 084 acoustic features. While using the reduced feature set did not improve performance in our experiments compared to the full feature set, it does decrease the number of parameters in the rectifier DNN and thereby accelerates the training and evaluation process.

C. Experimental Setup

The deep rectifier network consists of four hidden layers with 1 000, 750, 500, and 150 units as well as an output layer with a single neuron. This topology was determined by cross-validation on the training set. The loss function is given by the mean squared error. Optimisation is done via SGD with a constant learning rate of 10^{-4} and a batch size of 64 samples. The initial network is trained for 200 epochs on the labelled set $\mathcal{L}$. The features of the training set and the unlabelled set are jointly standardised to zero mean and unit variance; the test set is standardised using the same scales and offsets. Re-training after each labelling step is done for 30 epochs. To enable the comparison of the different instantiations of the framework, we save and reuse the initial model (i. e., its architecture and weight parameters). In this way, we ensure that the learning curves all start at the same point and that the experiments are reproducible. The dropout rate is set to 20 % both in the training and prediction stage. $T = 500$ forward passes with random dropout are used to obtain prosody estimates and confidences on the unlabelled set $\mathcal{U}$. The network topology as well as the training hyper-parameters were determined by a cross-validation experiment on the labelled set $\mathcal{L}$.

In each labelling iteration, 200 instances are labelled by using the proposed framework; in the case of cooperative learning, 100 instances with high confidence are machine-labelled and 100 instances with medium confidence are human-labelled. For early stopping, we set the agreement threshold α to 0.75, representing high agreement as the r_{wg} coefficient is bounded by 1, and $Q_{\min}$ to 2 (the smallest number for which r_{wg} is meaningful). These numbers were validated in a preliminary experiment on the AUWL database. The variance of the null distribution σ_E^2 in Eq. (6), which is a discrete uniform distribution on $\{1, 2, 3, 4, 5\}$, is $((5-1+1)^2-1)/12 = 2$. Furthermore, $U_{\min}$ is set so as to label a maximum of 2 000 instances from the AUWL database, i. e., $U_{\min} = 3\,732 - 2\,000 = 1\,732$.

D. Implementation

Our implementation of the DAL method is written in the Python language. As input data format, our software uses the common Pandas DataFrame representation, which has a two-dimensional tabular structure with the rows corresponding to the name of audio files and the columns indicating the audio and label attributes. For DNN training and computing of uncertainty measures, the choice fell on Keras [50], an open source deep learning library built on top of Theano [51] and TensorFlow [52]. The simulations reported on in this paper use the TensorFlow CPU backend of Keras.

E. Results

For the purpose of this study, we retain the same evaluation metric as in the 2015 ComParE Degree of Nativeness task, i. e., the Spearman's rank correlation coefficient (CC). In this section, we analyse the results obtained by using the proposed confidence measures and the different learning algorithms (cf. Table I). As is evident from Table I, the DCL algorithm has the following features: usage of a generic confidence measure, selection of instances for machine labelling and/ or

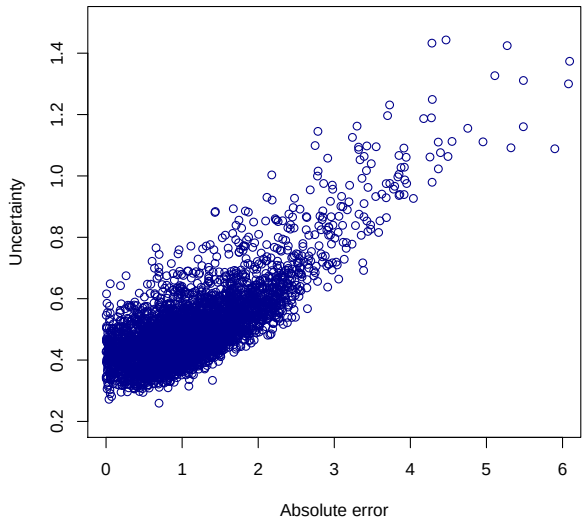


Fig. 4. Scatter diagram representing the correlation between the absolute prediction error and the estimated model uncertainty on the unlabelled set $\mathcal{U}$ (Spearman's CC = .658).

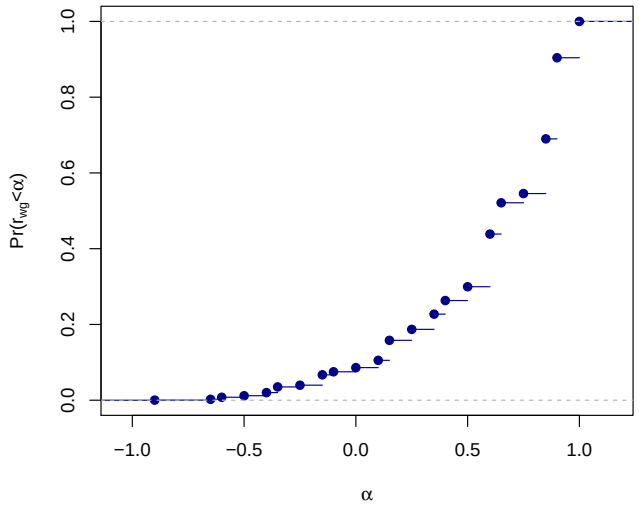


Fig. 5. Cumulative distribution function of r_{wg} measured on the ratings of the instances in the AUWL database.

human annotation, and dynamically stopping the annotation procedure. We perform an ablation study where we disable one or several of these features, and thereby show that the utmost annotation efficiency can only be obtained if the full-fledged version of the DCL algorithm is used. For instance, by comparing DCL against CL or DAL against AL, we can prove the importance of the agreement-based early stopping, and by comparing PL with AL, we can provide evidence for the effectiveness of the confidence measure.

As described in Section III-A, the key module in the generic framework in order to achieve the best accuracy and efficiency is the confidence-based arbitration. For instance, in SSL, if the model is confident in the case of erroneous predictions, it will train itself with incorrect label predictions, thus gradually deteriorate in performance due to error accumulation. Therefore, it is of crucial importance that the value of uncertainty reflects the magnitude of prediction errors. To substantiate the uncertainty measure, we train a DNN of the above-mentioned topology for 200 epochs on the NC and validate its predictions and uncertainties against the gold standard (from all raters) on the AUWL database. The corresponding scatter diagram in Fig. 4 depicts the positive correlation between the absolute prediction error and the estimated model uncertainty. The CC of the two variables is .658, which is sufficient for the purpose of uncertainty measure. The rationale is that in each iteration a subset of instances is selected based on their uncertainty quantiles (high, medium or low) and hence the exact order is less relevant for the purpose of sampling.

In Fig. 6, we compare the learning curves of the standard techniques (i. e., without the dynamic query adaptation) by plotting the Spearman's CC as a function of the total number of labelled instances. As a common feature, the sequential addition of labelled instances into the training set $\mathcal{L}$ (200 per iteration) leads to continuous improvement in the recognition accuracy. Here, the superior performance is achieved by AL with the MC strategy, markedly outperforming PL, SSL, and AL with the LC strategy in terms of accuracy and smoothness of curves. The relatively deficient performance (slow rise and unstable curve progression) of SSL can be explained by the fact that prosody recognition is considered to represent an appraisal-based subjective task; thus, involving the human in the loop naturally enhances the system performance.

Next, we investigate the early stopping based on the agreement level α, which essentially affects the label reliability and the annotation cost for each sample. First, we examine the functional relationship between a given agreement level α and the percentage of rating sequences that would trigger early stopping at this threshold, which can be approximated as

$$Pr(r_{wg} > \alpha) = 1 - Pr(r_{wg} \leq \alpha) = 1 - F_{r_{wg}}(\alpha), \qquad (8)$$

where $F_{r_{wg}}$ is the cumulative distribution function (CDF) of the r_{wg} measure (6). The monotonically increasing shape of the discrete CDF estimated on the instances of the AUWL database (each associated with five ratings on the 5-point Likert scale) is illustrated in Fig. 5. According to Eq. (8), we can estimate the probability of early stopping at 45 % for the threshold $\alpha = 0.75$. It is noted that this is an approximation, as not all rating combinations with a certain r_{wg} value could be encountered in an actual run of the DAL/ DCL algorithm. For instance, the sequence $(3, 3, 4, 5, 3)$ would not occur at $Q_{\min} = 2$ since early stopping would be initiated already after the first two queries.

Taking the AL (MC) method as a baseline, we now evaluate our advanced techniques to reduce human labelling work. Note that this is a strong baseline, as it already employs state-of-the-art AL based on Bayesian approximation similar to the method proposed in [11]. As can be seen in Fig. 7, CL, with joint human and machine effort, converges faster to a performance competitive with AL (MC). This is in accordance with the findings in [5]. Augmenting AL and CL with the dynamic query strategy based on early stopping (i. e., using DAL and DCL) further reduces the amount of

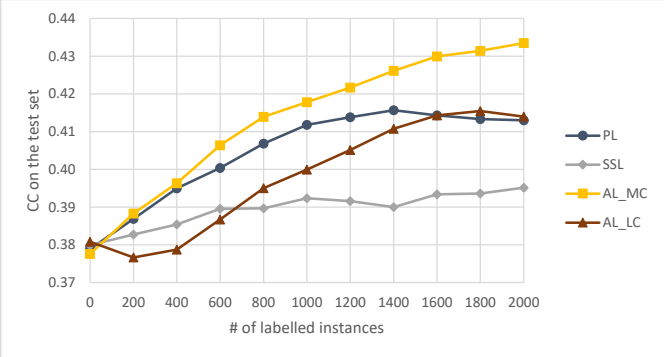


Fig. 6. Active Learning (AL) with the medium confidence (MC) or low confidence (LC) query strategy vs Semi-Supervised Learning (SSL) vs Passive Learning (PL): Performance in terms of the Spearman's CC values on the C-AuDiT test set vs the number of labelled instances from the AUWL database.

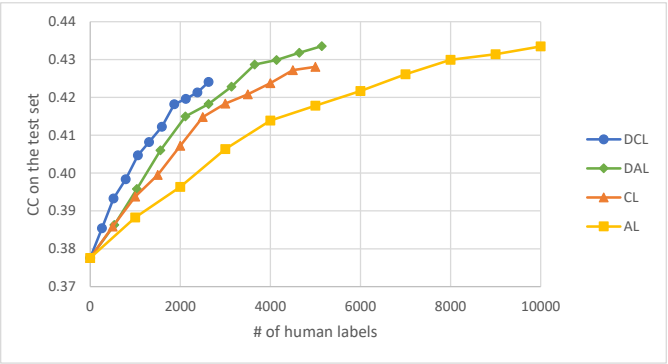


Fig. 7. Comparison between Dynamic Cooperative Learning (DCL)/ Dynamic Active Learning (DAL), which adapts the number of queries to an agreement level α, and standard Cooperative Learning (CL)/ Active Learning (AL), which sets a fixed number of annotations per instance: Performance in terms of the Spearman's CC values on the C-AuDiT test set vs the number of human labels. The medium confidence query strategy is used throughout.

human annotation effort required to reach similar performance. All the AL based methods shown in this graph increase the performance monotonically with the amount of human labels.

Table II allows interpreting the trade-off between the regression performance (Spearman's CC) and the number of queries to human annotators, i. e., annotation cost, in more detail. Comparing the numbers achieved by (D)AL and (D)CL, it can be seen that the proposed technique to dynamically adjust the number of raters based on agreement level provides an annotation cost reduction (CR) by roughly 1/2 at similar performance. In particular, the method is found to be complementary to the approach of CL, which reduces the human effort by combining AL with self-training. In the result, the proposed DCL algorithm is capable of a CR of almost 3/4, suggesting a multiplicative efficiency gain by combining CL with the dynamic query strategy. Finally, it is noteworthy that the empirically found cost reduction by 1/2 is reasonably close to the 45 % probability of early stopping, which was approximated a priori based on the consideration of the r_{wg} cumulative distribution (Fig. 5).

TABLE II
EFFICIENCY OF THE ITERATIVE LEARNING METHODS WITH STATIC AND DYNAMIC NUMBER OF ANNOTATORS, IN TERMS OF COST REDUCTION (CR) VS SPEARMAN'S CC ON THE C-AUDIT TEST SET AFTER 10 ITERATIONS, I. E., 2 000 LABELLED TRAINING INSTANCES FROM THE AUWL DATABASE.

	CC	# Queries	CR [%]
Static # of annotators per instance			
PL	.413	10 000	0
AL	.433	10 000	0
CL [5]	.428	5 000	50
Dynamic # of annotators per instance			
DAL	**.434**	5 143	49
DCL	.424	**2 630**	**74**

V. CONCLUSION

We presented a novel algorithmic framework for iterative human-machine annotation of large databases in subjective dimensions. As a highlight, the generic framework using the proposed dynamic cooperative learning technique is able to fully automatically distribute the annotation workload between humans and machines, combining uncertainty sampling (AL) and self-training (SSL). For arbitration, the model confidence is assessed by applying dropout to DNNs, which is suitable for nominal level and ordinal/ interval scale annotation tasks.

In realistic cross-corpus experiments, the empirical results on the INTERSPEECH ComParE task of scoring the language proficiency of non-native English speakers substantiate our proposition that DCL allows for considerably reducing the human annotation effort for 'mildly' subjective or ambiguous tasks, for which a consensus among raters can be presumed in a majority of cases. Comparing the performance of our DNN regressors on the C-AuDiT test set with the Challenge baseline (CC = .415), we find them to be competitive, despite them being trained with considerably less human intervention.

In the future, we plan on evaluating the learning paradigms on a wide range of computational paralinguistic tasks as well as pattern recognition tasks beyond the speech modality, extending the algorithm to integrate multi-modal learning methods such as [53]. In the same vein, our generic framework allows for integrating semi-supervised learning methods (cf., e. g., [54]) to provide more reliable machine labels, which is complementary to improving human-machine collaboration. Moreover, we will explore multi-task shared-hidden-layer DNNs for assigning multiple labels of interest at once. Finally, the DCL method will be implemented in state-of-the-art toolkits for the annotation of subjective tasks such as NOVA (NOnVerbal behavior Analyser) [55], which is designed for automated analysis of non-verbal signals in social interactions.

ACKNOWLEDGMENT

This work has received funding from the European Union's Horizon 2020 Framework Programme under the grant no. 645378 (ARIA-VALUSPA) and the European Union's Horizon 2020 Marie Skłodowska-Curie actions under the grant no. 797323 (HOL-DEEP-SENSE). The authors would like to thank Florian Hönig for providing the Nativeness datasets.

REFERENCES

[1] X. Zhu, "Semi-supervised learning literature survey," University of Wisconsin, Madison, WI, Tech. Rep. TR 1530, 2006.

[2] B. Settles, "Active learning literature survey," vol. 52, no. 55–66, p. 11 pages, 2010.

[3] X. Zhu, J. Lafferty, and Z. Ghahramani, "Combining active learning and semi-supervised learning using gaussian fields and harmonic functions," in *Proc. of ICML, Workshop on the Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining*. Washington, DC: AAAI Press, 2003, pp. 58–65.

[4] G. Tur, D. Hakkani-Tür, and R. E. Schapire, "Combining active and semi-supervised learning for spoken language understanding," *Speech Communication*, vol. 45, no. 2, pp. 171–186, 2005.

[5] Z. Zhang, E. Coutinho, J. Deng, and B. Schuller, "Cooperative learning and its application to emotion recognition from speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 115–126, 2014.

[6] C. Rosenberg, M. Hebert, and H. Schneiderman, "Semi-supervised self-training of object detection models," in *Proc. of IEEE Workshop on Motion and Video Computing*. Breckenridge, CO: IEEE, 2005, pp. 29–36.

[7] D. Lewis and W. Gale, "A sequential algorithm for training text classifiers," in *Proc. 17th International ACM SIGIR conference on Research and development in information retrieval*. Dublin, Ireland: Springer, 1994, pp. 3–12.

[8] Y. Zhang, E. Coutinho, Z. Zhang, C. Quan, and B. Schuller, "Dynamic active learning based on agreement and applied to emotion recognition in spoken interactions," in *Proc. of ICMI*. Seattle, WA: ACM, 2015, pp. 275–278.

[9] Y. Zhang, E. Coutinho, Z. Zhang, M. Adam, and B. Schuller, "On rater reliability and agreement based dynamic active learning," in *Proc. of ACII*. Xi'an, P. R. China: IEEE, 2015, pp. 70–76.

[10] Y. Zhang, E. Coutinho, Z. Zhang, C. Quan, and B. Schuller, "Agreement-based dynamic active learning with least and medium certainty query strategies," in *Proc. of ICML, Workshop on Advances in Active Learning: Bridging Theory and Practice*. Lille, France: IMLS, 2015, p. 5 pages.

[11] Y. Gal, "Uncertainty in deep learning," Ph.D. dissertation, University of Cambridge, 2016.

[12] J. M. LeBreton and J. L. Senter, "Answers to 20 questions about interrater reliability and interrater agreement," *Organizational Research Methods*, vol. 11, no. 4, pp. 815–852, 2008.

[13] M. Krzywinski and N. Altman, "Points of significance: visualizing samples with box plots," *Nature Methods*, vol. 11, no. 2, pp. 119–120, 2014.

[14] Z. Ghahramani, "Probabilistic machine learning and artificial intelligence," *Nature*, vol. 521, no. 7553, pp. 452–459, 2015.

[15] T.-F. Wu, C.-J. Lin, and R. C. Weng, "Probability estimates for multi-class classification by pairwise coupling," *Journal of Machine Learning Research*, vol. 5, pp. 975–1005, 2004.

[16] C. White, J. Droppo, A. Acero, and J. Odell, "Maximum entropy confidence estimation for speech recognition," in *Proc. of ICASSP*, vol. 4. Hawaii, HI: IEEE, 2007, pp. 809–812.

[17] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. of ICML*. New York, NY: IMLS, 2016, pp. 1050–1059.

[18] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*, ser. Adaptive Computation and Machine Learning. Cambridge, MA: MIT Press, 2006.

[19] A. Damianou and N. Lawrence, "Deep Gaussian Processes," in *Proc. of Workshop on Artificial Intelligence and Statistics*. Arizona, USA: JMLR, 2013, pp. 207–215.

[20] J. M. Hernández-Lobato and R. P. Adams, "Probabilistic backpropagation for scalable learning of Bayesian neural networks," in *Proc. of ICML*, vol. 37. Lille, France: JMLR.org, 2015, pp. 1861–1869.

[21] A. Graves, "Practical variational inference for neural networks," in *Advances in Neural Information Processing Systems 24*. Curran Associates, Inc. 2011, pp. 2348–2356.

[22] A. Krogh, J. Vedelsby *et al.*, "Neural network ensembles, cross validation, and active learning," *Advances in neural information processing systems*, vol. 7, pp. 231–238, 1995.

[23] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby, "Information, prediction, and query by committee," *Advances in neural information processing systems*, pp. 483–483, 1993.

[24] R. Burbidge, J. J. Rowland, and R. D. King, "Active learning for regression based on query by committee," in *Proc. of International Conference on Intelligent Data Engineering and Automated Learning (IDEAL)*. Birmingham, UK: Springer-Verlag, 2007, pp. 209–218.

[25] Y. Zhang, "Machine learning techniques for holistic computational paralinguistics," Ph.D. dissertation, Imperial College London, London, U.K., 2018.

[26] P. D. Bliese, D. Chan, and R. E. Ployhart, "Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis," *Multilevel Theory, Research, and Methods in Organizations*, pp. 349–381, 2000.

[27] F. Hönig, A. Batliner, K. Weilhammer, and E. Nöth, "How many labellers? Modelling inter-labeller agreement and system performance for the automatic assessment of non-native prosody," in *Proc. of Workshop on Speech and Language Technology in Education (SLaTE)*. Tokyo, Japan: ISCA, 2010, pp. 137–140.

[28] J. Howe, "The rise of crowdsourcing," *Wired Magazine*, vol. 14, no. 6, p. 5 pages, 2006.

[29] V. Sheng, F. Provost, and P. Ipeirotis, "Get another label? Improving data quality and data mining using multiple, noisy labelers," in *Proc. 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Las Vegas, NV, 2008, pp. 614–622.

[30] V. Raykar, S. Yu, L. Zhao, A. Jerebko, C. Florin, G. H. Valadez, L. Bogoni, and L. Moy, "Supervised learning from multiple experts: whom to trust when everyone lies a bit," in *Proc. 26th International Conference on Machine Learning*. Montreal, Canada: ACM, 2009, pp. 889–896.

[31] V. Raykar, S. Yu, L. Zhao, G. Valadez, C. Florin, L. Bogoni, and L. Moy, "Learning from crowds," *Journal of Machine Learning Research*, vol. 11, pp. 1297–1322, 2010.

[32] P. Donmez, J. G. Carbonell, and J. Schneider, "Efficiently learning the accuracy of labeling sources for selective sampling," in *Proc. 15th ACM SIGKDD International conference on Knowledge Discovery and Data Mining*. Paris, France: ACM, 2009, pp. 259–268.

[33] R. H. Hahnloser, R. Sarpeshkar, M. A. Mahowald, R. J. Douglas, and H. S. Seung, "Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit," *Nature*, vol. 405, no. 6789, pp. 947–951, 2000.

[34] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: An overview," in *Proc. of ICASSP*. Vancouver, Canada: IEEE, 2013, pp. 8599–8603.

[35] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15 pages, no. 1, pp. 1929–1958, 2014.

[36] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 5, no. 1, pp. 3–55, 2001.

[37] L. R. James, R. G. Demaree, and G. Wolf, "Estimating within-group interrater reliability with and without response bias," *Journal of Applied Psychology*, vol. 69, no. 1, p. 85 pages, 1984.

[38] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, "A survey on perceived speaker traits: Personality, likability, pathology, and the first Challenge," *Computer Speech and Language, Special Issue on Next Generation Computational Paralinguistics*, vol. 29, no. 1, pp. 100–131, 2015.

[39] F. Hönig, T. Bocklet, K. Riedhammer, A. Batliner, and E. Nöth, "The automatic assessment of non-native prosody: Combining classical prosodic analysis with acoustic modelling," in *Proc. of INTERSPEECH*. Portland, Oregon: ISCA, 2012, pp. 823–826.

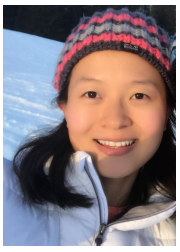
[40] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang, and F. Weninger, "The INTERSPEECH 2015 Computational Paralinguistics Challenge: Degree of Nativeness, Parkinson's & Eating Condition," in *Proc. of INTERSPEECH*. Dresden, Germany: ISCA, 2015, pp. 478–482.

[41] E. Coutinho, F. Hönig, Y. Zhang, S. Hantke, A. Batliner, E. Nöth, and B. Schuller, "Assessing the prosody of non-native speakers of English: Measures and feature sets," in *Proc. of Language Resources and Evaluation Conference (LREC)*. Portoroz, Slovenia: ELRA, 2016, pp. 1328–1332.

[42] A. Maier, T. Haderlein, U. Eysholdt, F. Rosanowski, A. Batliner, M. Schuster, and E. Nöth, "PEAKS–a system for the automatic evaluation of voice and speech disorders," *Speech Communication*, vol. 51, no. 5, pp. 425–437, 2009.

[43] F. Hönig, A. Batliner, and E. Nöth, "Automatic assessment of non-native prosody – annotation, modelling and evaluation," in *Proc. of International Symposium on Automatic Detection of Errors in Pronunciation*

Training (IS ADEPT). Stockholm, Sweden: KTH, Computer Science and Communication, 2012, pp. 21–30.
[44] F. Hönig, A. Batliner, K. Weilhammer, and E. Nöth, "Automatic assessment of non-native prosody for English as L2," in *Proc. of Speech Prosody*. Chicago, IL: ISCA, 2010, p. 4 pages.
[45] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in *Proc. of ACM Multimedia*, Barcelona, Spain, 2013, pp. 835–838.
[46] F. Weninger, F. Eyben, B. Schuller, M. Mortillaro, and K. R. Scherer, "On the acoustics of emotion in audio: What speech, music and sound have in common," *Frontiers in Emotion Science*, vol. 4, no. 292, pp. 1–12, 2013.
[47] Y. Zhang, F. Weninger, A. Batliner, F. Hönig, and B. Schuller, "Language proficiency assessment of English L2 speakers based on joint analysis of prosody and native language," in *Proc. of ICMI*. Tokyo, Japan: ACM, 2016, pp. 274–278.
[48] M. Hall, "Correlation-based feature selection for discrete and numeric class machine learning," in *Proc. of the 17th International Conference on Machine Learning (ICML)*. Stanford University, CA: Morgan Kaufmann Publishers, 2000, pp. 359–366.
[49] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*. San Francisco, CA: Morgan Kaufmann Publishers Inc., 2005.
[50] F. Chollet, "Keras," https://github.com/fchollet/keras, 2015.
[51] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio, "Theano: A CPU and GPU math compiler in Python," in *Proc. of Python in Science Conference (SciPy)*. Austin, TX: SciPy.org, 2010, pp. 3–10.
[52] M. Abadi, A. Agarwal, P. Barham *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: http://tensorflow.org/
[53] Y. Huang, W. Wang, and L. Wang, "Unconstrained multimodal multi-label learning," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 1923–1935, 2015.
[54] Z. Yu, Y. Zhang, C. P. Chen, J. You, H.-S. Wong, D. Dai, S. Wu, and J. Zhang, "Multiobjective semisupervised classifier ensemble," *IEEE Transactions on Cybernetics*, pp. 1–14, 2018.
[55] T. Baur, I. Damian, F. Lingenfelser, J. Wagner, and E. André, "NovA: Automated analysis of nonverbal signals in social interactions," in *Proc. of ACM Multimedia, Workshop on Human Behavior Understanding*. Barcelona, Spain: ACM, 2013, pp. 160–171.



Yue Zhang received her master's degree in Electrical Engineering and Information Technology (M. Sc.) from Technische Universität München (TUM) in 2013. In 2018, she received her PhD degree in Computing at Imperial College London, U.K. Currently, she is a Marie Curie fellow in the Affective Computing Group at the Massachusetts Institute of Technology. Her research interests lie in holistic machine perception of non-verbal cues, including affective and paralinguistic phenomena, as well as social signals.

Andrea Michi received his diploma (2016) in Computing from Imperial College London, U.K. He currently works as a software engineer at Google DeepMind, London, U.K.

Johannes Wagner received his master's degree in in Informatics and Multimedia in 2007 and his doctoral degree for his study on Online Systems for Multimodal Behaviour Analysis in 2016. He is now a research associate in the lab of Human Centered Multimedia (HCM) and has been working on several European projects (Humaine, Callas, Ilhaire, CEEDs, Aria-Valuspa). His main research area is the integration of Social Signal Processing (SSP) into real-life applications. He is the founding developer of the Social Signal Interpretation (SSI) framework, a framework for integrating multiple sensors into multimedia applications.

Elisabeth André is a full professor of Computer Science and Founding Chair of Human-Centered Multimedia at Augsburg University in Germany where she has been since 2001. She has multiple degrees in computer science from Saarland University, including a doctorate. Previously, she was a principal researcher at the German Research Center for Artificial Intelligence (DFKI GmbH) in Saarbrücken. Elisabeth André has a long track record in multimodal human-machine interaction, embodied conversational agents, social robotics, affective computing and social signal processing. In 2017, she was elected to the CHI Academy, an honorary group of leaders in the field of Human-Computer Interaction. Since January 2019, she is serving as the editor-in-chief of the IEEE Transactions on Affective Computing.

Björn Schuller received his diploma, doctoral degree, habilitation, and Adjunct Teaching Professorship all in EE/IT from TUM in Munich, Germany. He is currently a professor in Machine Learning at Imperial College London, U.K, full professor and chair of Embedded Intelligence for Health Care and Wellbeing at the University of Augsburg, Germany, and the co-founding CEO of audEERING GmbH. Previously, he headed the Machine Intelligence and Signal Processing Group at TUM from 2006 to 2014. He co-authored more than 600 technical contributions (more than 22 k citations). He is a Fellow of the IEEE and the former Editor-in-Chief of the IEEE Transactions on Affective Computing.

Felix Weninger received his diploma (2009) and his PhD degree (2015), both in computer science, from TUM, Munich, Germany. He is currently a tech lead at Nuance Communications, Burlington, MA, USA. From 2010–2014, he was a research assistant in the Machine Intelligence and Signal Processing Group at TUM's Institute for Human-Machine Communication, focusing on new machine learning techniques for noise-robust automatic speech recognition and related tasks. In 2013/14, he interned at Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA. His research interests are in the area of deep learning applied to speech and audio processing. He has published more than 90 peer-reviewed papers (4.1 k citations) in books, journals and conference proceedings.