Holistic Affect Recognition Using PaNDA: Paralinguistic Non-metric Dimensional Analysis

Holistic Affect Recognition using PaNDA: Paralinguistic Non-metric Dimensional Analysis

Yue Zhang, Felix Weninger, Björn Schuller, *Fellow, IEEE* and Rosalind W. Picard, *Fellow, IEEE*

Abstract—Humans perceive emotion from each other using a holistic perspective, accounting for diverse personal, non-emotional variables, such as age and personality, that shape expression. In contrast, today's algorithms are mainly designed to recognize emotion in isolation, and are usually demonstrated only within one relatively narrow database. In this work, we propose a multi-task learning approach to jointly learn the recognition of affective states from speech along with various speaker attributes. A problem with multi-task learning is that sometimes inductive transfer can negatively impact performance. To mitigate negative transfer, we introduce the Paralinguistic Non-metric Dimensional Analysis (PaNDA) method that systematically measures task relatedness and also enables visualizing the topology of affective phenomena as a whole. In addition, we present a generic framework that conflates the concepts of single-task and multi-task learning. Using this framework, we construct two models that demonstrate holistic affect recognition: one treats all tasks as equally related, whereas the other one incorporates the task correlations between a main task and its supporting tasks obtained from PaNDA. Both models employ a multi-task deep neural network, in which separate output layers are used to predict discrete and continuous attributes, while hidden layers are shared across different tasks. On average across 18 classification and regression tasks, the weighted multi-task learning with PaNDA significantly improves performance compared to single-task and unweighted multi-task learning.

Index Terms—Holistic context, speaker attributes, affective space, task relatedness, multi-task learning.

1 Introduction

> The whole is greater than the sum of its parts.
>
> ARISTOTLE

In Affective Computing, research has aimed at endowing machines with emotional intelligence, which should support collaboration and interaction with humans. Recent years have seen an upsurge of interest in affective technologies for a multitude of applications [1, 2], such as conversational interfaces, automotive assistants and smart solutions for human wellbeing.

Many state-of-the-art systems are able to recognize well-established emotion concepts such as valence, arousal and discrete emotion categories. Despite the achievements of today's systems, they are mostly designed to recognize emotion in isolation. However, studies in neuroscience and psychology, as well as social science have identified contextual cues as playing a central part in human perception of other people's emotion: People attend to individual differences in emotion expression, including differences attributed to personal factors and social influence, e. g., personality, gender and cultural background [3, 4]. Therefore, analyzing contextual information, e. g., demographic, personal, socio-cultural, psychophysiological, and environmental factors, helps improve affect recognition [5, 6].

Transient speaker states and permanent speaker traits fall under the umbrella of paralinguistic speech phenomena. In paralinguistic research, there are a few works on the joint learning of speaker attributes, such as deception and sincerity [7], inebriation and sleepiness [8], and native language and non-native English prosody [9]. However, the interrelations between manifold affective and other human phenomena remain hitherto under-explored. One reason for this shortcoming is the scarcity of multi-label databases (i. e., with labels along multiple target dimensions), which might be attributable to the traditional single-task learning (STL) paradigm. In contrast, Multi-task learning (MTL) is an approach to inductive transfer that improves generalization performance by sharing information between related tasks trained in parallel [10]. A significant problem arising with use of MTL is that knowledge transfer between unrelated tasks, with data drawn from different domains, may be counterproductive: It may cause performance loss [11]. For example, since feature relevance differs among tasks, it might not be efficient or even possible to learn consistent input feature weights for all tasks at the same time. To avoid negative transfer, it is thus important to define an appropriate notion of task relatedness, the fundamental component MTL is designed to exploit in order to successfully improve performance over STL.

In this work, we propose a novel approach that we call "holistic affect recognition", which refers to learning affective states together with any attributes that shape, influence, or interrelate with emotion. We posit that both affective and other non-affective characteristics that may contribute to affective communication or perception should be conceived as a whole, and in relation to each other, and that this holistic approach will lead to improvements in affect recognition systems. To implement a specific case of

- *Y. Zhang and R. W. Picard are with the Affective Computing Group, MIT Media Lab, Massachusetts Institute of Technology, Cambridge, MA 02139 USA. E-mail: {yuefw,picard}@mit.edu*
- *F. Weninger is with Nuance Communications, Burlington, MA 01803 USA. E-mail: felix@weninger.de*
- *B. Schuller is with Imperial College London, SW7 2AZ London, UK. E-mail: schuller@ieee.org*

Manuscript received date; revised date

holistic affect recognition, in this work we focus on affect recognition in speech, and demonstrate its improvement in performance by using non-affective attributes of speakers. We also expand upon the current state of the art by showing how to combine large numbers of single-task datasets to achieve holistic affect recognition, using MTL. Our main contributions are:

- This work presents the idea of holistic affect recognition, and illustrates it with a study combining 18 paralinguistic recognition tasks that have served as benchmarks for the research community, but have previously been treated as isolated tasks.
- We present the PaNDA method that is aimed at analyzing task relations in a data-driven way.
- We devise novel measures of task relatedness based on acoustic correlates and learned representations.
- We use a NMDS-based visualization to discern a big picture of the affective space and to make the task relations more human-interpretable.
- We propose a weighted MTL algorithm that exploits task correlations to reduce negative transfer and show that it is significantly more effective than STL across 18 recognition tasks.
- The new dynamic approach reconciles STL and MTL paradigms under one generic algorithm.

The remainder of the paper is structured as follows: Section 2 reviews prior work on learning multiple affective dimensions and measuring task relatedness. Section 3 lays out the conceptual foundations for holistic affect recognition. Section 4 describes the paralinguistic databases and the acoustic feature set. In Section 5, the PaNDA method for analyzing task relatedness is explained. Next, Section 6 elaborates on how the MTL framework is used for holistic affect recognition. The experimental results are discussed in Section 7 and a conclusion of the work is given in Section 8.

2 Related Work

2.1 Multi-Dimensional Affect Recognition

A wealth of research pertains to multi-label learning for affect recognition, using a single database labeled with a few affective dimensions, predominantly arousal/ valence/ dominance [12, 13, 14]. Xia and Liu [15] proposed a method to learn a main task (emotion classification) and a secondary task (arousal/ valence recognition). In their approach, the loss functions from these tasks are combined to train a deep belief network and subsequently the last hidden layer activations are used as features to train a support vector machine for the main task. Eyben et al. [16] used a long short-term memory recurrent neural network (LSTM-RNN) for modelling five affective dimensions, encompassing activation, expectation, intensity, power/ dominance, and valence. Chen [17] applied a similar model for the joint prediction of arousal/ valence and sentiment.

In contrast to multi-label learning [18], MTL is aimed at learning different tasks in cross-corpus settings, and thus deals with the problem of missing labels in disparate, task-specific datasets. One common technique of MTL using neural network models is the hard parameter sharing approach, using shared hidden layers and task-specific output layers [19]. Caruana [10] elucidated the underlying mechanisms, namely statistical data amplification, attribute selection, eavesdropping and representation bias. All these functionalities derive from the summing of error gradients from the task-specific output layers when calculating the gradient of the shared hidden layers, which is then propagated further down the network [10]. Effectively, the gradient of each task-specific loss is augmented by a regularization term, which ensures that no steps can be taken in a direction that would hurt the performance on the other tasks, and conversely, steps are encouraged in directions that improve the performance on more than one task. In our previous work [20], multi-task shared-hidden-layer DNNs were used to learn dimensional and categorical emotion representations using different emotion datasets, obviating the need for label mapping. Yet, task relations were not considered in this work.

Previous work on multi-dimensional affect recognition focused mostly on a few speaker attributes. Zhang et al. [7, 21] proposed a cross-task labelling (CTL) method to generate multi-label data by aggregating single-task datasets. Based on self-training, an ensemble of task-specific classifiers are iteratively trained to complete missing labels. The predicted labels are then used as auxiliary attributes in classifier chains to improve recognition performance on a specific task [22]. However, this method is susceptible to error accumulation in the iterative training process. Another approach to data aggregation was suggested in the work [23], which uses deep neural network (DNN) and long short-term memory (LSTM) to learn gender and naturalness as auxiliary tasks for emotion classification. Similar to the work [15], weighted loss functions are used to model task interdependencies. However, the weight parameters are empirically obtained, and the task relatedness itself remains elusive in these works.

Apart from recognizing emotion from speech, several other studies have been carried out for applying MTL to recognizing affective states from physiological and other behavioral data. Correa et al. [24] proposed a multi-task cascaded DNN, consisting of an affective convolutional neural network for predicting arousal/ valence from EEG data while reducing dimensionality, and a personal recurrent neural network for modelling personal factors, including the Big-five personality traits, mood (Positive Affect and Negative Affect Schedules) and social context (individual vs group). Jaques et al. [25] devised personalized models based on domain adaptation for predicting tomorrow's reported mood, stress, and physical health from wearable sensor and mobile phone data.

2.2 Studies on Task Relatedness

Numerous studies support the premise that task relatedness is key to inductive transfer learning (TL) [11, 26, 27], and multi-task learning [28, 29, 30]. In fact, learning unrelated tasks together can be counter-productive due to negative transfer, which implicates the performance on these tasks. Within the active research area of "learning to learn" [31], task relatedness has been extensively studied, with the aim to improve generalization by preventing negative transfer.

Thrun and O'Sullivan [32] described a task clustering algorithm that selectively transfers knowledge from the most similar task cluster to new tasks. A similar Bayesian task clustering technique based on prior distributions was developed by Bakker and Heskes [33]. Lee et al. [34] designed an algorithm for the joint learning of informative priors on feature relevance from an ensemble of related prediction tasks that share a similar relevance structure. Using a data generation model, Ben and Schuller [29] defined a notion of task relatedness and derived generalization error bounds on the information complexity. Evgeniou and Pontil [35] suggested a kernel function that uses a task-coupling parameter to model task relations, minimizing regularization functionals. Kumar and Daume [30] introduced an approach to modeling task grouping and overlap structure in order to selectively share the information among related tasks. Recently, Liu et al. [36] tackled the problem of learning asymmetric task relations by constructing a weighted directed regularization graph between multiple tasks. To our knowledge, task relatedness has not been systematically analyzed for paralinguistic speech phenomena.

In this work, we propose a principled and data-driven method to handle the large variety of affective states and speaker attributes. Based on the obtained task correlations and distance measures, we project various speaker attributes as embeddings in a 2D space. In contrast to the work [37] that shows the principal components of emotion recognizers trained on animated GIFs, the scope of our study goes beyond discrete emotions, encompassing a wide range of affect-centered human phenomena.

3 Holistic Affect Recognition

The concept of holistic affect recognition is based on two fundamental assumptions: (1) Affective and non-affective human characteristics interrelate within a holistic context; (2) Speaker attributes and their interrelations externalize through speech, and hence can be discerned from vocal cues. We now explain and substantiate the rationale behind the speaker attributes considered in this work, focusing on the properties they have in common with affective states.

From both an encoding and decoding perspective, as argued by others [38, 39], emotion manifests and is perceived in complex relationships with a multitude of speaker characteristics. We construe the holistic context as a general frame for *any influencing, dependent, or confounding variable of affect*, such as a given person's demographic traits, personal characteristics, sociocultural background, and psychophysiological states, as well as environmental factors (Fig. 1).

Demographic factors refer to a person's inherent traits, such as age, gender, and ethnicity. A number of studies have ascertained age differences in emotional experience, expression, and regulation, ascribing less negative affectivity and greater emotional control to elderly people [40]. Over a lifespan, the voice production characteristics undergo substantial changes, including decreases in fundamental frequency (F0), speaking rate, and voice quality [41].

In the literature, it is widely acknowledged that social, cultural, and interpersonal contexts exert a profound influence on emotion [6]. Intuitively, interpersonal conflict and emotions are natural concomitants in social interactions,

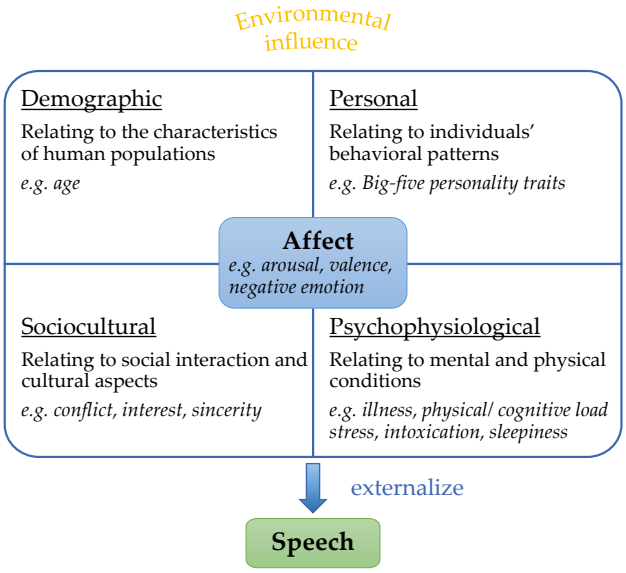


Fig. 1: Hypothetical framework illustrating the holistic context stemming from affect-related speaker attributes.

performing mutually initiating, amplifying and modulating functions [42]. Another social signal is interest that connects to active emotions and openness, and manifests in a faster speaking rate and greater range in vocal frequency [43]. In addition, we examine the perceived sincerity of apologies that is often driven and reinforced by social emotions [44]. On an illocutionary level, it has been shown that a high pitch accent and a low boundary tone reflect sincere apologies, whereas a double pitch accent and a high boundary tone are associated with ostensible apologies [45].

Personal characteristics are presumably the most important factors in shaping emotion. Early studies in psychology identified several vocal parameters common to both personality and emotion. For example, high pitch and speaking rate are associated both with a nervous personality type and also with 'active' emotions such as anger and fear [46].

Last but not least, we consider mental and physical states in the holistic context based on amassed evidence of psychophysiological effects on affect. In physiological analysis, bodily responses, e. g., higher skin conductance and heart rate, are coupled with emotional states [47], but they can also be induced by physical activities and health changes. We investigate a condition of illness with an upper respiratory tract infection, i. e., cold, which impairs voice quality, as well as influencing affective states [48]. Studies of alcohol's effects on the emotional trajectory have produced discrepant results, evidencing dampened emotional reactivity on the one hand, and elevated levels of arousal on the other hand [49, 50]. In speech analysis, most studies found that the speaking rate and the overall amplitude decrease after alcohol consumption, whereas the pitch variability, the F0, and the sentence duration increase [51]. Furthermore, consistent findings emerge from numerous studies that sleep deprivation adversely impacts emotional functioning and affects vocal expression of emotion in terms of decreases in pitch, intensity in certain high frequency bands, and vocal sharpness [52]. Psychological modulators of affect included

in this study are cognitive load and stress that alter speaking rate, energy contour, F0, and spectral parameters [53]. Besides internal (e. g., emotional, cognitive) stressors, there are also external stressors due to environmental factors [54]. For example, people tend to raise their vocal effort under noisy conditions, a.k.a., the Lombard effect.

Overall, contemporary research provides converging evidence for many complex relationships between affective states and speaker attributes. By analyzing their acoustic correlates, we aim to shed light on their interplay, and examine if the resulting insights and holistic combination can be used to provide more robust affect recognition.

4 Machine Analysis of Speaker Attributes

A speaker's voice conveys a wealth of information about the person behind it. The INTERSPEECH Computational Paralinguistic Challenges (ComParE) provide a variety of machine analysis tasks for recognizing speaker characteristics from vocal cues. The speech databases we use in this work come from the ComParE challenges; thus, they permit benchmark comparisons across the work of a large research community and allow for comparability of the results we present here. We further aim to practice the highest quality in this work by following the strict guidelines of the ComParE Challenges (speaker-independent partitioning so that test data contains different speakers than training data, stratification according to meta data, reproducibility of results, etc.). That said, the last 10 consecutive ComParE Challenges have treated paralinguistic tasks as isolated phenomena. In contrast, our work aims to aggregate all of the single-task databases for holistic affect recognition. This has never been done before in paralinguistic research. For the purpose of this study, we processed the data of 18 paralinguistic tasks that represent the speaker attributes motivated in Section 3. Note that we exclude speaker identity as an attribute, because we aim at speaker-independent recognition.

4.1 Datasets and Tasks

In this section, we briefly describe the benchmark datasets used for the recognition of speaker attributes. Further details can be found in Table 1, as well as in the work [67]. Note that in Table 1, the tasks are sorted according to the labelling scheme of the paralinguistic databases, i. e., binary, ordinal, continuous, as this label conversion plays an important role in understanding the task relatedness in subsequent sections.

The **Geneva Multimodal Emotion Portrayals (GEMEP)** corpus [55] contains emotional speech featuring 18 different emotional expressions portrayed by professional actors, including admiration, amusement, anxiety, cold anger, contempt, despair, disgust, elation, hot anger, interest, panic fear, pleasure, pride, relief, sadness, shame, surprise, tenderness. For the classification task, the multi-class labels were mapped to positive/ negative arousal/ valence.

The **Aibo Emotion Corpus (AEC)** [56] contains spontaneous, emotionally colored speech. The recordings were collected from 51 children, interacting with Sony's pet robot Aibo. To elicit negative emotions, the children were led to believe that Aibo was following their orders, where the robot was actually controlled by an unseen human operator and sometimes disobeyed them on purpose.

The **Speaker Personality Corpus (SPC)** [57] contains speech data collected from the Swiss national broadcast. The clips were annotated regarding the speaker's personality traits using the Big Five Inventory (BFI-10) [68, 69]. The personality scores were dichotomized to above or below the average of their ratings for the respective trait.

The **Upper Respiratory Tract Infection Corpus (URTIC)** [58, 70] contains read, prompted and free speech. To check their health condition, the subjects underwent a self-assessment using the Wisconsin upper respiratory symptom survey [71], based on an illness severity scale from 0 (not sick) to 7 (severely sick). The ratings were binarized to 'non-cold' vs. 'cold'.

The **Munich Bio-voice Corpus (MBC)** [59] contains read speech as well as physiological data (heart rate and skin conductivity) while speaking. Based on the sensory measurements, two levels of physical load ('low' vs 'high') were evoked, i. e., before and after exercises such as fast stair-climbing and running.

The **Cognitive Load with Speech and EGG (CLSE)** database [60] is used for studying the impact of cognitive load, caused by a Stroop test and a reading span task, on speech production. Here, cognitive load refers to the working memory, i. e., the brain's limited capacity for storing and processing temporary information [72].

The **Speech Under Simulated and Actual Stress (SUSAS)** database [54] contains recordings in noisy environments, where different types of stress were exerted on the subjects, such as cognitive load, noise and motion fear. The speech tests were e. g., calibrated work load tracking task, acquisition and compensatory tracking task, and amusement park roller-coaster.

The **Alcohol Language Corpus (ALC)** [61] contains genuine intoxicated speech in automotive environment. For data collection, the subjects underwent a systematic intoxication test within a certain range of blood alcohol concentration (BAC). The required amount of alcohol was calculated from the individuals' biometric data (e. g., body mass, body fat percentage, gender) using the Watson- and Widmark formula [73]. The database also contains control recordings from the subjects in sober condition.

The **Sleepy Language Corpus (SLC)** [62] was collected in sleep deprivation studies, including sustained vowels, read stories, commands and control in a driver assistance system etc. The annotation based on the Karolinska sleepiness scale (KSS) was completed by the subjects (self-assessment) and additionally by two assessors (observer-assessment). The scores range from 1 (extremely alert) to 10 (cannot stay awake).

The **SSPNet Conflict Corpus (SCC)** [63] was extracted from the Canal 9 Corpus [74], a collection of Swiss political debates. Each clip was rated in terms of the intensity of conflict on an interval scale $[-10, +10]$via Amazon Mechanical Turk, using a questionnaire including physical (objective observation) and inferential (subjective interpretation) questions [75].

The **Audiovisual Interest Corpus (AVIC)** [64] contains sessions in which a salesperson advertised a product to the subjects. The speech segments were annotated in terms

TABLE 1: Overview of paralinguistic speech databases used in the proposed approach to holistic affect recognition.

Task	Dataset	Task Description	# Inst	hh:mm	# Subj	Label conversion
Arousal Valence	GEMEP [55]	Portrayed emotion	1 260	00:52	5 m 5 f	Binary: Low (0), high (1) Binary: Negative (0), positive (1)
Negative Emotion	AEC [56]	Elicited emotion	18 216	09:12	21 m 30 f	Binary: Negative (0) or idle (1)
Openness Conscientiousness Extroversion Agreeableness Neuroticism	SPC [57]	Big-five personality traits	640	01:43	263 m 59 f	Binary: absence (0) or presence (1) of a trait
Cold (illness)	URTIC [58]	Upper respiration tract infection	28 652	44:24	382 m 248 f	Binary: Non-cold (0), cold (1)
Physical Load	MBC [59]	Speaking before and after exercise	1 088	00:22	15 m 4 f	Binary: Low (0), high (1)
Cognitive Load	CSLE [60]	Working memory	2 418	05:34	20 m 6 f	Ordinal: Low (0), medium (1), high (2)
Stress	SUSAS [54]	Level of stress	3 593	01:01	4 m 3 f	Ordinal: Low (0), medium (1), high (2)
Intoxication	ALC [61]	Blood alcohol concentration (BAC)	12 360	39:05	84 m 78 f	Continuous, range [0, 1.75] in per mille
Sleepiness	SLC [62]	Karolinska sleepiness scale (KSS)	9 089	21:16	43 m 56 f	Continuous, range [1, 10]
Conflict	SCC [63]	Conflict in dyadic political debates	1 430	11:55	92 m 18 f	Continuous, range [-10,10]
Interest	AVIC [64]	Degree of interest in conversations	3 880	02:17	11 m 10 f	Continuous, range [-1, 1]
Sincerity	SSC [65]	Degree of sincerity when apologizing	911	02:20	15 m 17 f	Continuous, range [-2.5, 1.7]
Age	aGender [66]	Age in years	65 364	47:00	404 m 410 f 131 x	Continuous, range [7–80]

Abbreviations: Inst: instances; Subj: subjects (m: male, f: female).

of the degree of interest (LoI) shown by the subjects. The five-point ordinal scale designates disinterest, indifference, neutrality, interest, and curiosity. The gold-standard labels were obtained by averaging the ratings from four expert labelers and mapping the mean values to the interval [-1,1].

The **Sincerity Speech Corpus (SSC)** [65] contains sentences of apologies in different prosodic styles. The utterances were annotated in terms of perceived sincerity on a 5-point Likert scale, ranging from 0 (not sincere at all) to 4 (extremely sincere). To eliminate individual biases, the ratings of each annotator were standardized to zero mean and unit variance. The gold-standard labels were obtained by averaging the standardized ratings from the annotators.

The **aGender** Corpus [66] contains telephone speech in mixed environment. The subjects were prompted by an automated interactive voice response system to repeat sentences or speak freely.

4.2 Acoustic Features

The ComParE set of supra-segmental acoustic features [76] serves as the standard feature set in the ComParE Challenges. It contains 6 373 acoustic features obtained from the computation of various functionals over low-level descriptor (LLD) contours. The features are extracted with openSMILE [76]. Important subgroups of the ComParE feature set comprise prosodic, cepstral, spectral, and voice quality features.

Since the ComParE feature set contains many redundant features in practice (e. g., various types of means), we employ correlation-based feature selection (CFS) [77]. The CFS algorithm searches for features which are highly predictive for the task labels, yet uncorrelated among each other [78]. Typically, CFS discards well over half of the features without sacrificing performance [79]. The reduced feature set is composed of the relevant features selected for each task.

5 Paralinguistic Non-metric Dimensional Analysis (PaNDA)

To systematically assess task similarities, we propose the PaNDA method that helps identify related tasks. Fig. 2 depicts the workflow of the proposed method, which is described in the following sections.

5.1 Measure of Task Relatedness

The proposed measure of task relatedness is established in the acoustic feature space $\mathcal{X}$. Let us denote the total number of recognition tasks by S (cf. Section 4.1). For each task $s \in \{1, \ldots, S\}$ and acoustic feature $f \in \{1, \ldots, F\}$, the correlation between the feature values and the corresponding labels is computed on the training instances of task s.

To this end, the Pearson product-moment correlation coefficient (CC) is used. The rationale is that the Pearson's r is suitable for measuring the strength and direction of association between a continuous variable (feature f) and a continuous/ binary/ ordinal variable (the target label) [80]. For binary tasks, the nominal labels are mapped to 1 and 0, indicating 'presence' or 'absence' of the dichotomous attribute. Here, it can be shown that the point bi-serial CC, which is used for measuring the relationship between a continuous and a dichotomous variable [81], is equivalent to the Pearson's r and similar to a two-sample t-test. Accordingly, ordinal scales (e. g., low, medium, high) are

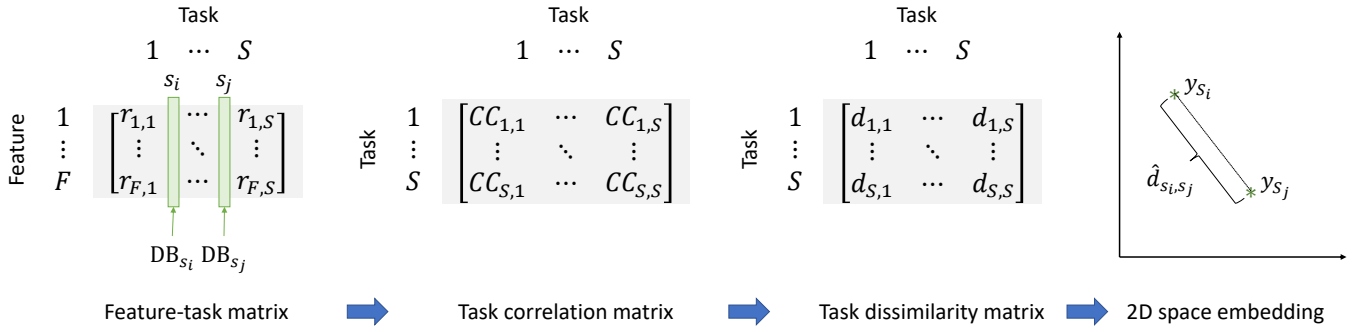


Fig. 2: Schematic of the Paralinguistic Non-metric Dimensional Analysis (PaNDA) method for visualizing task relatedness.

represented in ranks. The conversion of nominal attributes to numbers for computing the feature-label correlations is shown in Table 1. To account for non-linear relationships, we can also compute the feature-label correlations on the acoustic features transformed through the hidden layers of a DNN. Given an F-dimensional feature space, we obtain an $F \times S$-dimensional feature-task matrix $\mathbf{R}$ (cf. Fig. 2), of which the s-th column is an F-dimensional vector $\mathbf{r}_s$ describing task s in terms of feature-label correlations.

Based on the feature-task matrix, we compute the task correlation matrix $\mathbf{C} = (CC_{s_i,s_j})$, where CC_{s_i,s_j} is the correlation coefficient of $\mathbf{r}_{s_i}$ and $\mathbf{r}_{s_j}$. Consequently, we consider tasks s_i and s_j to be highly related if the value of CC_{s_i,s_j} is near 1. Conversely, if CC_{s_i,s_j} is near -1, the tasks can be considered as 'antipodes'. Note that the opposite tasks can still be jointly learned in neural networks since the output layers can be trained to swap binary labels or reverse the the sign of continuous-valued labels.

Based on the task correlation matrix, we define the task dissimilarity matrix as $\mathbf{D} = \mathbf{1} - \mathbf{C}$, where $\mathbf{1}$ is a matrix of ones. It follows that the task dissimilarity d_{s_i,s_j} is in the interval $[0, 2]$, and the highest dissimilarity (2) is measured for opposite tasks, while medium dissimilarity (1) indicates unrelated tasks.

5.2 Non-Metric Dimensional Scaling

Non-metric dimensional scaling (NMDS) [82, 83] is used for visualizing the task dissimilarity matrix, projecting task attributes as embeddings in a 2D space. This helps humans interpret the learned task correlations as a sanity check. NMDS is based on classical multidimensional scaling (MDS). In MDS, given a distance matrix, a corresponding set of points in an Euclidean space is analytically obtained by centering the distance matrix and subsequent eigenvalue decomposition. In NMDS, one operates on a matrix $\mathbf{D}$ of general dissimilarities instead of distances. Thus, in addition to the optimal configuration of points one also looks for a (potentially non-linear) transformation of the dissimilarities into distances. To this end, a loss function called *stress* is minimized w.r.t. the distances $\hat{d}_{s_i,s_j} = |\mathbf{y}_{s_i} - \mathbf{y}_{s_j}|^2$ (cf. Fig. 2) to obtain the configuration of points $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_S)$. The variant of NMDS used in this work is Sammon's non-linear mapping [83]. In accordance with this work, the stress is defined as

$$\text{Stress}(\mathbf{Y}) = \frac{1}{\sum_{s_i<s_j} d_{s_i,s_j}} \sum_{s_i<s_j} \frac{(d_{s_i,s_j} - \hat{d}_{s_i,s_j})^2}{d_{s_i,s_j}}. \quad (1)$$

This stress function is minimized by steepest descent [83]. In order to improve convergence speed, the initial configuration of $\mathbf{Y}$ is set to the classical MDS solution where the dissimilarities are interpreted as distances.

5.3 Visualizations

In Fig. 3, the two-dimensional map obtained by PaNDA is shown. Notably, phenomena concomitant with 'activation', such as stress, conflict, neuroticism, extroversion, interest, and intoxication cluster around (high) arousal. Due to the emotion elicitation scenario (cf. Section 4.1), negative emotions (e. g., angry, reprimanding) mainly manifest in high arousal and negative valence, which also explains the far distance to the (positive) valence task (cf. Table 1). In close proximity to the arousal cluster, one can find openness and conscientiousness as neighboring tasks of extroversion, as well as physical load and cold. In contrast, tasks that cannot be directly related to activation (cognitive load, valence, agreeableness, sincerity, sleepiness, age) are allocated in the right hemisphere, and are much more scattered. It is noted that 'agreeableness' is found to be opposed to 'extroversion' despite the fact that the data of these tasks were recorded under the exact same conditions. This demonstrates that our notion of task relatedness abstracts away from pure acoustic similarity, so is not simply a function of recording conditions within a dataset. Moreover, the tasks of sleepiness and intoxication are located far away from each other, indicating their dissimilarity that was also found in the work [8].

This pattern is also reflected in Fig. 4, in which the task correlation matrix $\mathbf{C}$ is visualized as a heatmap (blue: strong negative correlation, purple: strong positive correlation), using hierarchical clustering with the Euclidean distance. In line with our previous findings, we observe a clear grouping into activation-related tasks and others, with conscientiousness, extroversion and openness showing the strongest association, and agreeableness and conflict being furthest apart.

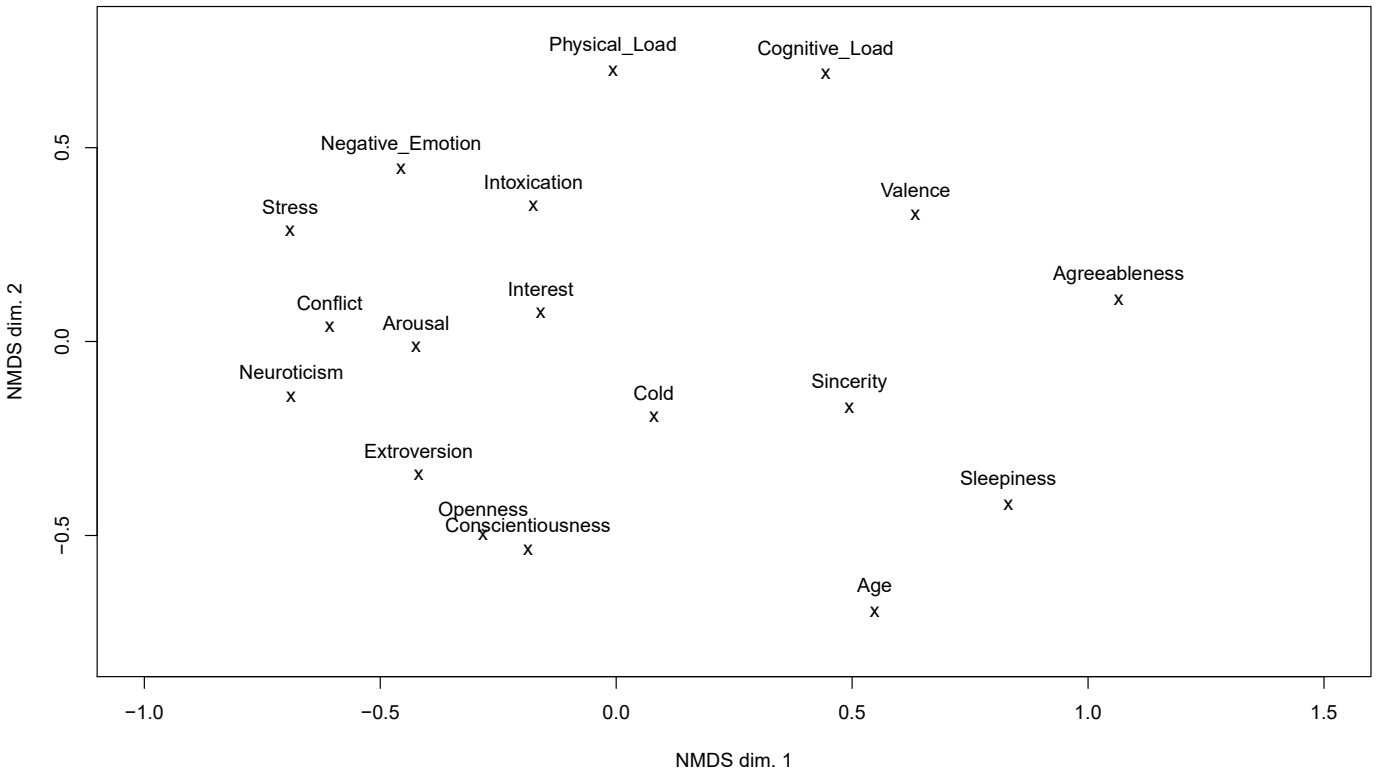


Fig. 3: Visualization of task relations obtained by the PaNDA method based on non-metric dimensional scaling (NMDS): the task embeddings indicate 'presence' of the respective speaker attributes (cf. label conversion in Table 1).

6 Multi-Task Learning Framework

In this study, we compare two MTL models: one treats all tasks as equally related, whereas the other one incorporates the task correlations between a main task and its supporting tasks obtained from PaNDA. To this end, we apply a multi-task DNN with a shared representation across tasks. The advantages of this multitasking model are manifold: first, it enables an effective data aggregation scheme in joining diverse datasets; second, the neural network topology is highly compact in comparison with a set of single-task DNNs; third, the model is trained to be predictive for a broad array of discrete and continuous speaker attributes.

Mathematically, the output $\hat{\mathbf{y}}$ of the multi-task DNN on an acoustic feature vector $\mathbf{x}$ is composed of sub-vectors for each task $1, \ldots, S$:

$$\hat{\mathbf{y}} = \left[\hat{\mathbf{y}}^{(1)}; \hat{\mathbf{y}}^{(2)}; \ldots; \hat{\mathbf{y}}^{(S)}\right], \tag{2}$$

where each $\hat{\mathbf{y}}^{(s)}$, $s = 1, \ldots, S$ corresponds to a transformation of the last hidden layer activation with a task-specific weight matrix $\mathbf{W}^{(s)}$:

$$\hat{\mathbf{y}}^{(s)} = u^{(s)}(\mathbf{W}^{(s)}\mathbf{h}) = u^{(s)}(\mathbf{W}_H^{(s)}u'(\mathbf{W}_{H-1}(\cdots u'(\mathbf{W}_1\mathbf{x})))), \tag{3}$$

with layer-specific output activation functions $u^{(s)}$ and a hidden layer activation function u' for H hidden layers. A common choice for u' is the rectified linear activation function. Biases are omitted in the above equation for ease of exposition, but are used in the experiments. Optimization of the parameters $\mathcal{W} = \{\mathbf{W}_1, \ldots, \mathbf{W}_H, \mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(S)}\}$ of the shared-hidden-layer DNN is done via error back-propagation and stochastic gradient descent (SGD). Assuming that each acoustic feature vector $\mathbf{x}_k$ in the training set belongs to exactly one task s_k with label y_k, the multi-task loss function to be minimized is:

$$J^{\mathrm{MT}} = \sum_k J(\hat{\mathbf{y}}^{(s_k)}, y_k). \tag{4}$$

Hence, the forward and backward propagation mechanism needs to evaluate only one output layer per input vector. A main improvement in comparison with the prior work [20] is that the output layers here are trained to predict both discrete and continuous labels. To this end, different activation functions are used in the output layers depending on the type of task. Specifically, we use the sigmoid function with cross-entropy loss in binary classification and the linear activation function with mean squared error loss in regression. Optionally, the softmax function with cross-entropy can be used to generalize to classification tasks. Note that ordinal scales are treated as continuous labels (cf. Table 1).

Building upon this network topology, we propose a MTL technique that incorporates task correlations. The novelty of our MTL technique compared to prior works using weighted losses [15, 23] is that the weighting scheme we use is based on task relations measured on the training data. To this end, we define the weighted (task correlation based) loss function as:

$$J_{s^*}^{\mathrm{TC}} = \sum_k |CC_{s^*,s_k}|^{\alpha} J(\hat{\mathbf{y}}^{(s_k)}, y_k), \tag{5}$$

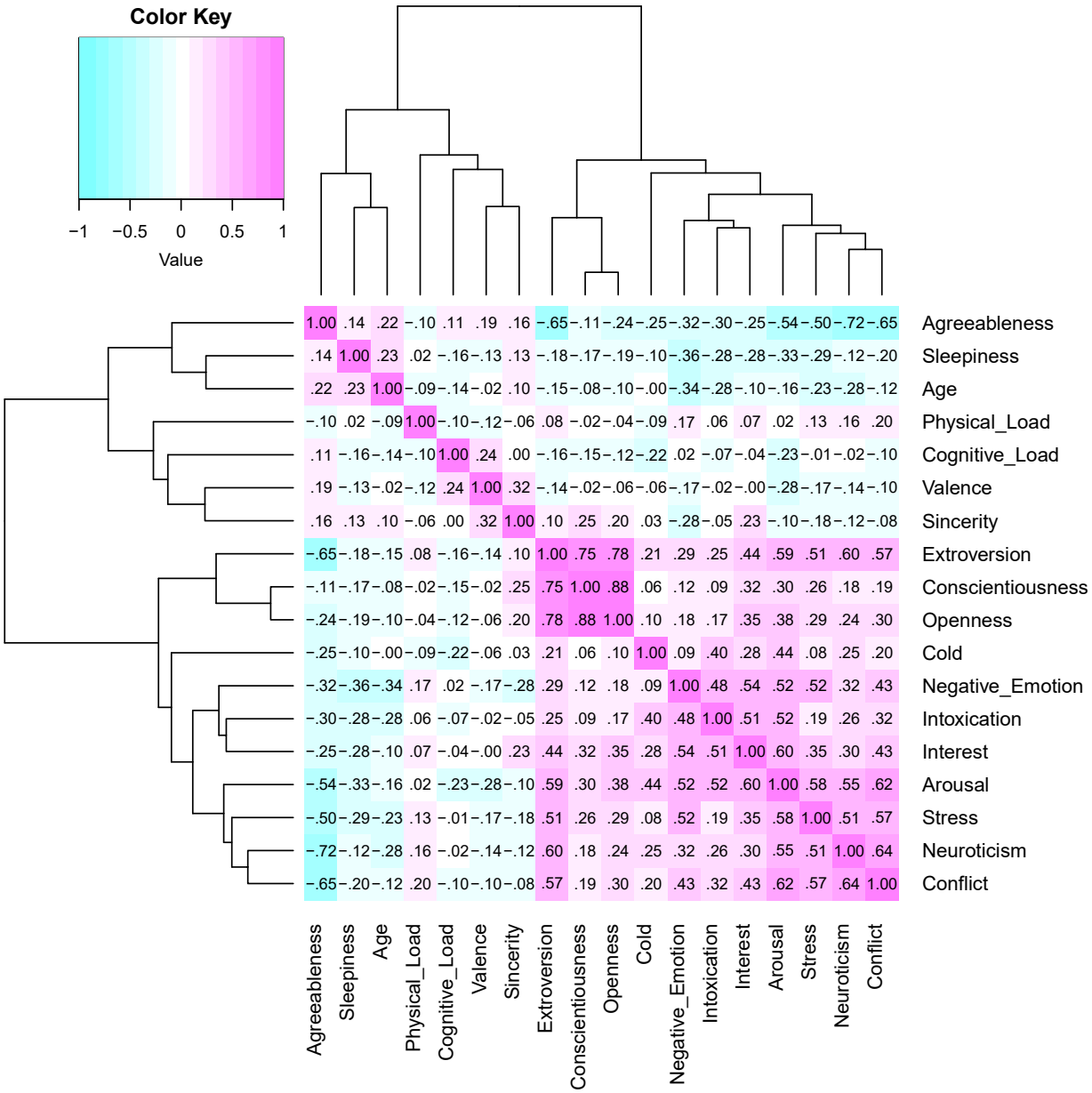


Fig. 4: Task correlation matrix $\mathbf{C}$ visualized as a heatmap with hierarchical clustering.

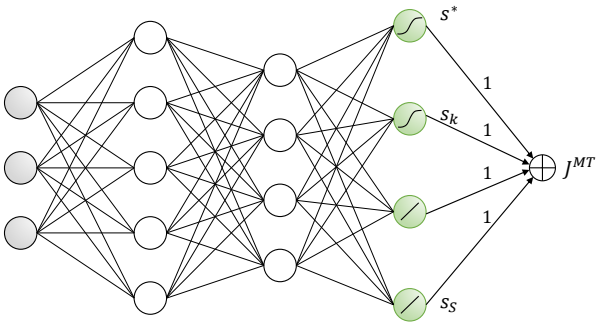
where s^* signifies a task as the main task. For $s_k = s^*$, the correlation CC_{s^*,s^*} is one. The exponent α parameterizes the power of the weights. Figuratively speaking, for a given main task in the map (Fig. 3), α influences the range of nearby tasks considered for joint learning. However, we avoid a hard decision which tasks are considered as related and which are not. For $\alpha \to \infty$, only the instances of the main task are taken into account by the loss $J^{\text{TC}}_{s^*}$, which thus resembles STL. For $\alpha = 0$, the loss is equivalent to the MTL loss. Thus, the proposed method unifies STL, MTL, and weighted MTL under one generic model.

For each main task s^*, a multi-task DNN $\mathcal{W}$ is pre-trained on all datasets. Subsequently, the parameters $\mathcal{W}^{(s^*)} = \{\mathbf{W}_1, \ldots, \mathbf{W}_H, \mathbf{W}^{(s^*)}\}$ are fine-tuned on the instances of task s^*. This approach is similar to transfer learning that deals with the problem of transferring knowledge from one or more source tasks to a target task especially when the latter has few training data. However, to exploit task correlations, the main task has to be included in the network training. Fig. 5 illustrates the architecture of the multi-task shared-hidden-layer (MT-SHL) DNNs.

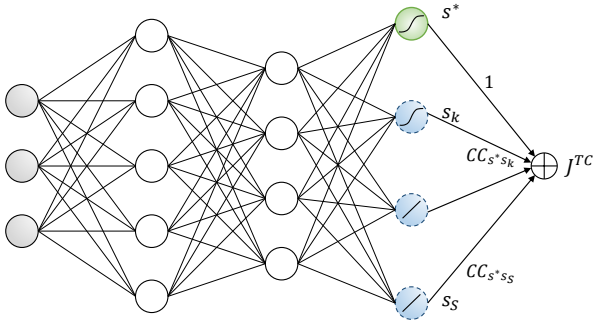
7 Experiments and Results

In this work, we compare using single-task (ST) DNNs and the proposed MTL models. In particular, we seek to examine the effectiveness of α in regulating the influence of supporting tasks.

The multi-tasking framework is implemented using the Python deep learning library Keras and TensorFlow [84]. For the ST and MT experiments, we train DNNs with 3 hidden layers (1 024, 512, 256 units) for 10 epochs. The batch size is set to 32 using stratified sampling. In training, we use a L2 regularizer and 10 % dropout [85] to prevent



(a) MT-SHL



(b) MT-SHL + PaNDA

Fig. 5: Multi-task shared-hidden-layer (MT-SHL) network architectures: (a) All tasks are trained simultaneously and the losses are weighted equally; (b) Each task s^* is trained as a main task while the other tasks s_k function as supporting tasks. The losses are weighted with the task correlations obtained from the PaNDA method.

overfitting. Optimization is done via SGD with 0.01 learning rate and harmonic decay. To minimize oscillations of the loss between minibatches, we add a Nesterov momentum to the parameter update. The number of hidden layers and the hidden layer sizes were tuned for the ST baseline so as to achieve competitive performance compared to the original Challenge baselines. These settings were also used for the MT models. The training hyperparameters were optimized for the ST and MT experiments separately in preliminary experiments. In the experiments using task correlations, the exponent α in Eq. 5 is set to the values 0, 0.5 or 1. The features of each training set are standardized to zero mean and unit variance; the corresponding test set is standardized using the same scales and offsets. Considering the imbalance in terms of dataset sizes, we tested up-/ down-sampling to an equal number of instances, as well as homogeneous minibatches as in the work [20]. However, we found that while these techniques helped the 'smaller' tasks, they decreased the performance on 'larger' tasks, and hence yielded similar results on average.

The evaluation metrics are unweighted average recall (UAR) for classification, and Pearson's CC for regression. To ensure reproducibility and comparability of the results, we set the random seeds to 42 for starting generated random numbers in a well-defined initial state. For each task, we compute the mean and standard deviation of the UAR/ CC values resulting from 10 trials.

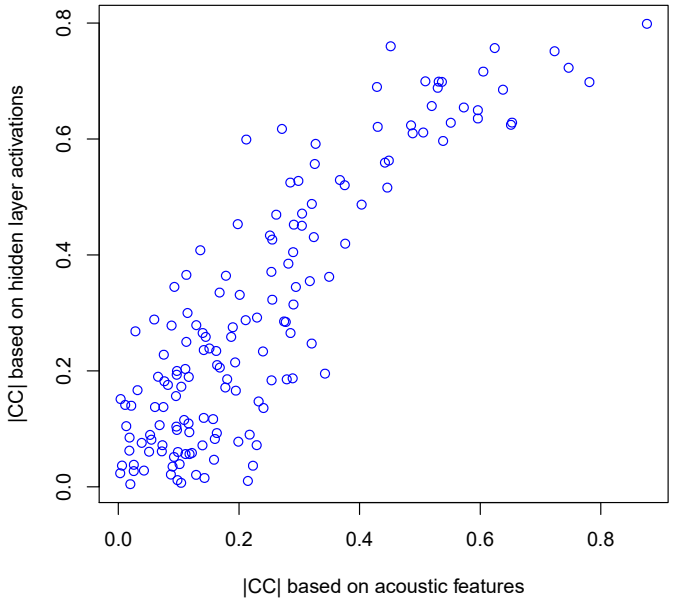


Fig. 6: Scatterplot of task correlations obtained from acoustic features and hidden layer activations of the MT-SHL-DNN.

Table 2 shows the results on 18 tasks as described in Section 4.1, using the ST and MT DNNs. On 10 out of 18 tasks, the MT-SHL model performs better than the ST baseline, especially for predicting cognitive load, intoxication, sleepiness, interest, sincerity and age.

Moreover, the correlation-based MTL method (MT-SHL + PaNDA) considerably improves upon the MT-SHL results, particularly for predicting cold, physical and cognitive load. In particular, the integration of task correlations makes up for the performance drop on cold and physical load. This is promising and underpins the benefit of task relations to reduce negative transfer. On the tasks where positive knowledge transfer is achieved by MT-SHL (valence, cognitive load, interest, sleepiness, age, and sincerity), the additional gain from task correlations is smaller. From this, we construe that task relatedness is more effective in suppressing negative transfer than fostering positive transfer. Juxtaposing the results obtained by PaNDA using different α values, we find that the performance measures with $\alpha = 0.5$ interpolate between the ones with $\alpha = 0$ and $\alpha = 1$.

In additional experiments, we evaluate the performance of using task correlations based on the learned representations of the DNN (PaNDA-hidden). To this end, we compute the activations of the last hidden layer of the trained MT-SHL-DNN for every training instance. From the hidden activations, we compute feature-label correlations and task correlations according to Section 5. The task correlations are then used in the loss (5). The values of the task correlations based on hidden activations are plotted against the task correlations based on acoustic features (cf. Fig. 4) in Fig. 6. We observe a Spearman's ρ of 0.80, demonstrating that our approach to measure task relatedness is not only applicable to acoustic features, but also to non-linear feature transformations. Moreover, we obtain similar recognition performance with PaNDA-hidden and PaNDA (60.7% on average).

TABLE 2: Performance measures on 18 tasks using single-task DNNs (ST), multi-task shared-hidden-layer DNN (MT-SHL), and MT-SHL with weighted loss functions based on task correlations (+ PaNDA). The parameter α regulates the weights of the supporting tasks according to Eq. 5. Evaluation metrics are UAR for classification tasks and CC for regression tasks (*).

Task	ST $\alpha \to \infty$	MT-SHL $\alpha = 0$	+ PaNDA $\alpha = 0.5$	+ PaNDA $\alpha = 1$
Arousal	**72.7 ± 0.7**	72.2 ± 0.6	72.2 ± 0.3	72.1 ± 0.3
Valence	62.4 ± 1.2	63.9 ± 1.6	**64.0 ± 1.2**	63.9 ± 1.1
Negative Emotion	69.0 ± 0.1	68.9 ± 0.3	69.1 ± 0.3	**69.4 ± 0.3**
Openness	**57.9 ± 1.5**	56.6 ± 1.1	55.9 ± 2.4	56.2 ± 2.1
Conscientiousness	**79.1 ± 1.2**	78.5 ± 1.1	78.2 ± 1.3	78.1 ± 1.2
Extroversion	**76.7 ± 1.2**	75.6 ± 1.1	75.1 ± 1.3	75.6 ± 1.4
Agreeableness	57.9 ± 1.2	58.6 ± 1.3	59.2 ± 1.0	**60.2 ± 1.6**
Neuroticism	62.7 ± 1.4	63.8 ± 1.2	63.8 ± 1.8	**64.2 ± 1.2**
Cold	68.8 ± 0.3	64.5 ± 0.9	67.2 ± 0.7	**69.7 ± 0.8**
Physical Load	**72.0 ± 1.9**	67.7 ± 1.4	69.5 ± 2.0	69.5 ± 1.5
Cognitive Load*	27.6 ± 0.6	32.8 ± 2.1	36.8 ± 0.7	**38.4 ± 1.3**
Stress*	**60.8 ± 1.4**	56.4 ± 1.8	56.7 ± 1.5	57.9 ± 1.3
Intoxication*	30.7 ± 0.8	35.0 ± 0.9	34.7 ± 1.5	**36.0 ± 1.0**
Sleepiness*	31.0 ± 0.8	41.7 ± 1.2	41.0 ± 0.9	**43.6 ± 0.5**
Conflict*	83.7 ± 0.5	84.0 ± 0.5	83.9 ± 0.9	**84.4 ± 0.4**
Interest*	33.1 ± 1.7	38.9 ± 0.7	**39.6 ± 1.1**	**39.6 ± 1.3**
Sincerity*	55.0 ± 0.9	57.6 ± 1.8	57.6 ± 2.8	**59.2 ± 1.1**
Age*	46.5 ± 0.3	52.4 ± 0.2	52.6 ± 0.3	**53.8 ± 0.5**
Mean	58.2 ± 1.0	59.4 ± 1.1	59.8 ± 1.2	**60.7 ± 1.0**

On average across 18 tasks and 10 trials, the MT-SHL + PaNDA method performs better than MT+SHL. The gains are statistically significant according to a one-sided Wilcoxon signed rank test ($p < .01$). In turn, the MT-SHL significantly outperforms the ST baseline ($p < .01$). These tests, spanning 18 datasets from different recording conditions, collected by different teams, and using 153 different task correlation pairs, show that the PaNDA method generalizes across a great variety of diverse test conditions.

Nevertheless, STL proves to be a strong baseline, as data distributions can vary dramatically across different tasks. Finally, despite relatively small data sizes for most tasks, the DNNs we use here yield competitive results compared to the SVM performance in the ComParE series (for a comprehensive overview of the Challenge baselines the reader is referred to the work [67]).

A limitation of the MT-SHL + PaNDA method is that a multi-task network as well as a task-specific network for each main task have to be trained, creating a large parameter space to learn, and incurring high computational cost. Nevertheless, we argue that the benefits of the proposed method outweigh this limitation since (1) it allows the focused training of a target task while including its related tasks, (2) it helps reduce negative transfer that may hurt performance, and (3) it enables a soft decision between ST (learning only one task) and unweighted MTL (learning all tasks equally), thereby unifying ST and MT algorithms under one generic learning paradigm.

8 Conclusion

Summing up, we presented a novel approach to holistic affect recognition by jointly predicting affect and non-affective speaker attributes, and evaluated its performance using diverse datasets collected under different conditions. The holistic approach aims to model all contextual factors contributing to communication of and perception of affective phenomena, including demographic, sociocultural, personal, psychophysiological, and environmental factors that influence human emotion – including its production, expression, and perception.

We presented a method to support holistic recognition, the PaNDA method, in which we derived measures of task similarities from the bi-variate (Pearson) correlations between acoustic features and labels of the input vectors. Using non-metric dimensional scaling, a big picture of inter-related patterns was revealed, displaying a prevalent cluster of 'active' states and traits.

To facilitate joint classification and regression, we used an MT-SHL DNN that employs different activation and loss functions in separate output layers to predict multiple discrete and continuous attributes at the same time. Moreover, we introduced a generic algorithm (MT-SHL + PaNDA) that unifies STL, MTL and task-correlation based MTL under one learning paradigm, using a hyperparameter α to regulate the influence of the supporting tasks.

On 18 exemplary tasks, our results demonstrate that both of the MTL methods significantly improve performance compared to STL. Importantly, it has been shown that, where MT-SHL suffers a performance drop, incorporating task correlations helps mitigate the effects of negative transfer. Thus, our findings corroborate the importance of task relatedness for inductive transfer learning, dovetailing with previous work [7, 8, 9]. For example, in prior work [8], it was hypothesized that sleepiness and alcohol intoxication are related, yet, transfer learning did not improve the performance. Our findings from PaNDA indicate that these tasks are actually dissimilar.

A major advantage of holistic modeling is that it bridges gaps that have made it a challenge to combine datasets with different labeling schemes. For example, there exist numerous emotion datasets containing different labels (categorical, dimensional, continuous etc.), some with overlapping emotion concepts. The standard approach to jointly use these datasets is label discretization, e.g., mapping discrete emotion classes into an arousal/ valence space. This, however, comes with considerable information loss. What we propose instead is a versatile model that can be applied to combine any labelled datasets, regardless of whether their labels are binary, ordinal, or continuous, and without having to impose a lossy conversion to a common label set.

In conclusion, we put forward a new approach to holistic affect recognition that leverages information across diverse but related concepts, while mitigating the impact of negative information transfer. We recognize that this work has been limited to the speech modality, and a certain selection of datasets. In future work, we will apply the PaNDA method to other modalities and cross-modal features. We will also explore expansions to new datasets and assess the robustness of the method to the selection of features and instances. We hope that the proposed holistic approach will be examined in many areas, with the potential to provide not only more robust affect recognition, but also greater insights into what factors contribute to the rich, diverse means in which human affect is produced, communicated,

and understood.

ACKNOWLEDGEMENT

This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement HOL-DEEP-SENSE No 797323. We thank the data providers of the INTERSPEECH Computational Paralinguistic Challenges.

REFERENCES

[1] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic, "A survey of multimodal sentiment analysis," *Image and Vision Computing*, vol. 65, pp. 3–14, 2017.

[2] S. Scherer, M. Glodek, G. Layher, M. Schels, M. Schmidt, T. Brosch, S. Tschechne, F. Schwenker, H. Neumann, and G. Palm, "A generic framework for the inference of user states in human computer interaction," *Journal on Multimodal User Interfaces*, vol. 6, no. 3-4, pp. 117–141, 2012.

[3] L. F. Barrett, B. Mesquita, and M. Gendron, "Context in emotion perception," *Current Directions in Psychological Science*, vol. 20, no. 5, pp. 286–290, 2011.

[4] J. Zaki, "Cue integration: A common framework for social cognition and physical perception," *Perspectives on Psychological Science*, vol. 8, no. 3, pp. 296–312, 2013.

[5] A. Göker and H. Myrhaug, "User context and personalisation," in *Proc. of 6th European Conference on Case Based Reasoning*, 2002, pp. 1–34.

[6] S. E. Bibri, *The Human Face of Ambient Intelligence*. Springer, 2015, 515 pages.

[7] Y. Zhang, F. Weninger, Z. Ren, and B. Schuller, "Sincerity and deception in speech: Two sides of the same coin? A transfer- and multi-task learning perspective," in *Proc. of 17th Conference of the International Speech Communication Association (INTERSPEECH)*. San Francisco, CA: ISCA, 2016, pp. 2041–2045.

[8] Y. Zhang, F. Weninger, and B. Schuller, "Cross-domain classification of drowsiness in speech: The case of alcohol intoxication and sleep deprivation," in *Proc. of 18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Stockholm, Sweden: ISCA, 2017, pp. 3152–3156.

[9] 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 18th ACM International Conference on Multimodal Interaction (ICMI)*. Tokyo, Japan: ACM, 2016, pp. 274–278.

[10] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, no. 1, pp. 41–75, 1997.

[11] M. Rosenstein, Z. Marx, L. P. Kaelbling, and T. Dietterich, "To transfer or not to transfer," in *Proc. of 18th Advances in Neural Information Processing Systems (NIPS), Workshop on Inductive Transfer*, vol. 898. Vancouver, Canada: MIT Press, 2005, pp. 1–4.

[12] 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, 2015.

[13] S. Parthasarathy and C. Busso, "Jointly predicting arousal, valence and dominance with multi-task learning," in *Proc. of 18th International Speech Communication Association (INTERSPEECH)*, Stockholm, Sweden, 2017, pp. 1103–1107.

[14] D. Le, Z. Aldeneh, and E. M. Provost, "Discretized continuous speech emotion recognition with multi-task deep recurrent neural network," in *Proc. of 18th Conference of the International Speech Communication Association (INTERSPEECH)*, Stockholm, Sweden, 2017, pp. 1108–1112.

[15] R. Xia and Y. Liu, "A multi-task learning framework for emotion recognition using 2d continuous space," *IEEE Transactions on Affective Computing*, no. 1, pp. 3–14, 2017.

[16] F. Eyben, M. Wöllmer, and B. Schuller, "A multitask approach to continuous five-dimensional affect sensing in natural speech," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 2, no. 1, p. 6, 2012.

[17] S. Chen, Q. Jin, J. Zhao, and S. Wang, "Multimodal multi-task learning for dimensional and continuous emotion recognition," in *Proc. of 7th Workshop on Audio/Visual Emotion Challenge*. Mountain View, CA: ACM, 2017, pp. 19–26.

[18] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 8, pp. 1819–1837, 2013.

[19] S. Ruder, "An overview of multi-task learning in deep neural networks," *CoRR*, vol. abs/1706.05098, 2017. [Online]. Available: http://arxiv.org/abs/1706.05098

[20] Y. Zhang, Y. Liu, F. Weninger, and B. Schuller, "Multi-task deep neural network with shared hidden layers: Breaking down the wall between emotion representations," in *Proc. of 42nd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. New Orleans, LA: IEEE, 2017, pp. 4990–4994.

[21] Y. Zhang, Y. Zhou, J. Shen, and B. Schuller, "Semi-autonomous data enrichment based on cross-task labelling of missing targets for holistic speech analysis," in *Proc. of 41st IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Shanghai, P. R. China: IEEE, 2016, pp. 6090–6094.

[22] B. Schuller, Y. Zhang, F. Eyben, and F. Weninger, "Intelligent systems' holistic evolving analysis of real-life universal speaker characteristics," in *Proc. of 5th International Workshop on Emotion Social Signals, Sentiment & Linked Open Data (ES^{3}LOD), satellite of the 9th Language Resources and Evaluation Conference (LREC)*. Reykjavik, Iceland: ELRA, 2014, pp. 14–20.

[23] J. Kim, G. Englebienne, K. Truong, and V. Evers, "Towards speech emotion recognition "in the wild" using aggregated corpora and deep multi-task learning," in *Proc. of 18th Conference of the International Speech Communication Association (INTERSPEECH)*. Stockholm, Sweden: ISCA, 2017, pp. 1113–1117.

[24] J. A. Mioranda-Correa and I. Patras, "A multi-task cascaded network for prediction of affect, personality, mood and social context using eeg signals," in *Proc. of 13th IEEE International Conference on Automatic Face &*

Gesture Recognition (FG 2018). IEEE, 2018, pp. 373–380.

[25] N. Jaques, S. Taylor, E. Nosakhare, A. Sano, and R. Picard, "Multi-task learning for predicting health, stress, and happiness," in *Proc. of 10th Conference on Neural Information Processing Systems (NIPS), Workshop on Machine Learning for Healthcare*. Barcelona, Spain: Curran Associates, 2016.

[26] E. Eaton, M. desJardins, and T. Lane, "Modeling transfer relationships between learning tasks for improved inductive transfer," in *Proc. of Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Antwerp, Belgium: Springer, 2008, pp. 317–332.

[27] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[28] J. Baxter, "A model of inductive bias learning," *Journal of Artificial Intelligence Research*, vol. 12, pp. 149–198, 2000.

[29] S. Ben-David and R. Schuller, "Exploiting task relatedness for multiple task learning," in *Learning Theory and Kernel Machines*. Springer, 2003, pp. 567–580.

[30] A. Kumar and H. Daume, "Learning task grouping and overlap in multi-task learning," in *Proc. of 29th International Conference on Machine Learning*. Edinburgh, UK: Omnipress, 2012, pp. 1383–1390.

[31] S. Thrun and L. Pratt, *Learning to Learn*. Springer Science & Business Media, 2012.

[32] S. Thrun and J. O'Sullivan, "Discovering structure in multiple learning tasks: The tc algorithm," in *Proc. of 13th International Conference on Machine Learning (ICML)*, vol. 96. Bari, Italy: Morgan Kaufmann Publishers, 1996, pp. 489–497.

[33] B. Bakker and T. Heskes, "Task clustering and gating for bayesian multitask learning," *Journal of Machine Learning Research*, vol. 4, pp. 83–99, 2003.

[34] S.-I. Lee, V. Chatalbashev, D. Vickrey, and D. Koller, "Learning a meta-level prior for feature relevance from multiple related tasks," in *Proc. of 24th International Conference on Machine Learning (ICML)*. Corvallis, OR: ACM, 2007, pp. 489–496.

[35] T. Evgeniou and M. Pontil, "Regularized multi-task learning," in *Proc. of 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Seattle, WA: ACM, 2004, pp. 109–117.

[36] S. Liu, Y. Liang, and A. Gitter, "Loss-balanced task weighting to reduce negative transfer in multi-task learning," in *Proc. of 33rd AAAI Conference on Artificial Intelligence (AAAI)*, vol. 793, Honolulu, Hawaii, 2019, p. 802.

[37] W. Chen and R. W. Picard, "Predicting perceived emotions in animated gifs with 3d convolutional neural networks," in *Proc. of IEEE International Symposium on Multimedia (ISM)*. San Jose, CA: IEEE, 2016, pp. 367–368.

[38] A. Kappas, U. Hess, and K. Scherer, "Voice and emotion," *Fundamentals of Nonverbal Behavior*, pp. 201–238, 1991.

[39] U. Hess, "Nonverbal communication," in *Encyclopedia of Mental Health (Second Edition)*. Oxford, UK: Academic Press, 2016, pp. 208–218.

[40] J. Gross, L. Carstensen, M. Pasupathi, J. Tsai, C. Götestam Skorpen, and A. Hsu, "Emotion and aging: Experience, expression, and control," *Psychology and Aging*, vol. 12, no. 4, p. 590, 1997.

[41] E. Stathopoulos, J. Huber, and J. Sussman, "Changes in acoustic characteristics of the voice across the life span: Measures from individuals 4–93 years of age," *Journal of Speech, Language, and Hearing Research*, pp. 1011–1021, 2011.

[42] W. Arsenio and M. Killen, "Conflict-related emotions during peer disputes," *Early Education and Development*, vol. 7, no. 1, pp. 43–57, 2010.

[43] P. Silvia, "Interest—the curious emotion," *Current Directions in Psychological Science*, vol. 17, no. 1, pp. 57–60, 2008.

[44] S. Hareli and Z. Eisikovits, "The role of communicating social emotions accompanying apologies in forgiveness," *Motivation and Emotion*, vol. 30, no. 3, pp. 189–197, 2006.

[45] S. Alexander, "Sincerity, intonation, and apologies: A case study of Thai EFL and ESL learners," Ph.D. dissertation, Indiana University, Bloomington, IN, 2011.

[46] W. Apple, L. Streeter, and R. Krauss, "Effects of pitch and speech rate on personal attributions," *Journal of Personality and Social Psychology*, vol. 37, no. 5, p. 715, 1979.

[47] R. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 10, pp. 1175–1191, 2001.

[48] R. Bucks, Y. Gidron, P. Harris, J. Teeling, K. Wesnes, and H. Perry, "Selective effects of upper respiratory tract infection on cognition, mood and emotion processing: a prospective study," *Brain, Behavior, and Immunity*, vol. 22, no. 3, pp. 399–407, 2008.

[49] J. Curtin, C. Patrick, A. Lang, J. Cacioppo, and N. Birbaumer, "Alcohol affects emotion through cognition," *Psychological Science*, vol. 12, no. 6, pp. 527–531, 2001.

[50] P. Sutker, A. Allain, P. Brantley, and C. Randall, "Acute alcohol intoxication, negative affect, and autonomic arousal in women and men," *Addictive Behaviors*, vol. 7, no. 1, pp. 17–25, 1982.

[51] O. Cooney, K. McGuigan, P. Murphy, and R. Conroy, "Acoustic analysis of the effects of alcohol on the human voice," *Journal of the Acoustical Society of America*, vol. 103, no. 5, p. 2895, 1998.

[52] E. McGlinchey, L. Talbot, K.-h. Chang, K. Kaplan, R. Dahl, and A. Harvey, "The effect of sleep deprivation on vocal expression of emotion in adolescents and adults," *Sleep*, vol. 34, no. 9, pp. 1233–1241, 2011.

[53] K. Scherer, D. Grandjean, T. Johnstone, G. Klasmeyer, and T. Bänziger, "Acoustic correlates of task load and stress," in *Proc. of 7th International Conference on Spoken Language Processing*. ISCA, 2002, pp. 2017–2020.

[54] J. Hansen, S. Bou-Ghazale, R. Sarikaya, and B. Pellom, "Getting started with SUSAS: a speech under simulated and actual stress database," in *Proc. of 5th European Conference on Speech Communication and Technology (EUROSPEECH)*, vol. 4, Rhodes, Greece, 1997, pp. 1743–1746.

[55] T. Bänziger, M. Mortillaro, and K. Scherer, "Introducing

the Geneva Multimodal Expression Corpus for experimental research on emotion perception," *Emotion*, vol. 12, pp. 1161–1179, 2012.

[56] S. Steidl, "Automatic classification of emotion-related user states in spontaneous children's speech," Ph.D. dissertation, University of Erlangen-Nuremberg, 2009.

[57] G. Mohammadi and A. Vinciarelli, "Automatic personality perception: Prediction of trait attribution based on prosodic features," *IEEE Transactions on Affective Computing*, vol. 3, no. 3, pp. 273–284, 2012.

[58] N. Cummins, M. Schmitt, S. Amiriparian, J. Krajewski, and B. Schuller, ""You sound ill, take the day off": Automatic recognition of speech affected by upper respiratory tract infection," in *Proc. of 39th International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. Jeju Island, Korea: IEEE, 2017, pp. 3806–3809.

[59] B. Schuller, F. Friedmann, and F. Eyben, "The Munich BioVoice Corpus: Effects of physical exercising, heart rate, and skin conductance on human speech production," in *Proc. of 9th Language Resources and Evaluation Conference (LREC)*. Reykjavik, Iceland: ELRA, 2014, pp. 1506–1510.

[60] T. F. Yap, J. Epps, E. Ambikairajah, and E. Choi, "Voice source under cognitive load: Effects and classification," *Speech Communication*, vol. 72, pp. 74–95, 2015.

[61] F. Schiel, C. Heinrich, and S. Barfüsser, "Alcohol Language Corpus: The first public corpus of alcoholized German speech," *Language Resources and Evaluation*, vol. 46, no. 3, pp. 503–521, 2011.

[62] J. Krajewski, A. Batliner, and M. Golz, "Acoustic sleepiness detection: Framework and validation of a speech-adapted pattern recognition approach," *Behavior Research Methods*, vol. 41, no. 3, pp. 795–804, 2009.

[63] S. Kim, M. Filippone, F. Valente, and A. Vinciarelli, "Predicting the conflict level in television political debates: An approach based on crowdsourcing, nonverbal communication and Gaussian processes," in *Proc. of 20th ACM International Conference on Multimedia*. Nara, Japan: ACM, 2012, pp. 793–796.

[64] B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker, and H. Konosu, "Being bored? Recognising natural interest by extensive audiovisual integration for real-life application," *Image and Vision Computing*, vol. 27, no. 12, pp. 1760–1774, 2009.

[65] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception & Sincerity," in *Proc. of 17th Conference of the International Speech Communication Association (INTERSPEECH)*. San Francsico, CA: ISCA, 2016, pp. 2001–2005.

[66] F. Burkhardt, M. Eckert, W. Johannsen, and J. Stegmann, "A database of age and gender annotated telephone speech," in *Proc. of 7th International Conference on Language Resources and Evaluation (LREC)*. Valletta, Malta: ELRA, 2010.

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

[68] M. Barrick and M. Mount, "The big five personality dimensions and job performance: A meta-analysis," *Personnel Psychology*, vol. 44, no. 1, pp. 1–26, 1991.

[69] B. Rammstedt and O. John, "Measuring personality in one minute or less: A 10-item short version of the big five inventory in English and German," *Journal of Research in Personality*, vol. 41, no. 1, pp. 203–212, 2007.

[70] B. Schuller, S. Steidl, A. Batliner *et al.*, "The INTERSPEECH 2017 Computational Paralinguistics Challenge: Addressee, Cold & Snoring," in *Proc. of 18th Conference of the International Speech Communication Association (INTERSPEECH)*. Stockholm, Sweden: ISCA, 2017, pp. 3442–3446.

[71] B. Barrett, R. Brown, M. Mundt, G. Thomas, S. Barlow, A. Highstrom, and M. Bahrainian, "Validation of a short form Wisconsin upper respiratory symptom survey," *Health and Quality of Life Outcomes*, vol. 7, no. 1, p. 76, 2009.

[72] A. Baddeley, "Working memory," *Science*, vol. 255, no. 5044, pp. 556–559, 1992.

[73] P. Watson, I. Watson, and R. Batt, "Prediction of blood alcohol concentrations in human subjects. Updating the Widmark equation," *Journal of Studies on Alcohol*, vol. 42, no. 7, pp. 547–556, 1981.

[74] A. Vinciarelli, A. Dielmann, S. Favre, and H. Salamin, "Canal9: A database of political debates for analysis of social interactions," in *Proc. of 3rd International Conference on Affective Computing and Intelligent Interaction (ACII)*. Amsterdam, Netherlands: IEEE, 2009, pp. 1–4.

[75] A. Vinciarelli, S. Kim, F. Valente, and H. Salamin, "Collecting data for socially intelligent surveillance and monitoring approaches: The case of conflict in competitive conversations," in *Proc. of 5th International Symposium on Communications Control and Signal Processing (ISCCSP)*. IEEE, 2012, pp. 1–4.

[76] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in *Proc. of 21st ACM International Conference on Multimedia*. Barcelona, Spain: ACM, 2013, pp. 835–838.

[77] M. Hall, "Correlation-based feature selection for discrete and numeric class machine learning," in *Proc. of 17th International Conference on Machine Learning (ICML)*. Stanford University, CA: Morgan Kaufmann Publishers, 2000, pp. 359–366.

[78] E. E. Ghiselli, *Theory of psychological measurement*. McGraw-Hill, 1964.

[79] M. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, University of Waikato, 1999.

[80] D. Hahs-Vaughn and R. Lomax, *An introduction to statistical concepts*. Routledge, 2013, 840 pages.

[81] R. Tate, "Correlation between a discrete and a continuous variable. point-biserial correlation," *The Annals of Mathematical Statistics*, vol. 25, no. 3, pp. 603–607, 1954.

[82] J. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika*, vol. 29, no. 1, pp. 1–27, 1964.

[83] J. Sammon, "A nonlinear mapping for data structure analysis," *IEEE Transactions on Computers*, vol. 100, no. 5, pp. 401–409, 1969.

[84] M. Abadi, A. Agarwal, P. Barham *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," pp. 1–21, 2015, software available from tensorflow.org.
[85] 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, no. 1, pp. 1929–1958, 2014.

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 are holistic machine perception of human phenomena, including affective states, social signals and speaker attributes.

Felix Weninger received his diploma (2009) and his PhD degree (2015), both in computer science, from TUM, Munich, Germany. He is currently a Principal Research Scientist 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 machine learning techniques for noise-robust automatic speech recognition. 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 (5 k citations) in books, journals and conference proceedings.

Björn Schuller, FIEEE, received his diploma, doctoral degree, habilitation, and Adjunct Teaching Professorship all in EE/IT from TUM in Munich, Germany. He is currently a professor of Artificial Intelligence 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.

Rosalind W. Picard, ScD, FIEEE, is Professor of Media Arts and Sciences at MIT, founder and director of the Affective Computing Research Group at the MIT Media Lab, and co-founder of the startups Affectiva and Empatica. She has a BS in Electrical Engineering from the Georgia Institute of Techology and an SM and ScD in Electrical Engineering and Computer Science from MIT. In 2019, she received one of the highest professional honors accorded an engineer, election to the National Academy of Engineering for her contributions on affective computing and wearable computing. Picard is credited with starting the branch of computer science known as affective computing with her 1997 book of the same name. She is author of the book Affective Computing, and author or co-author of over three hundred scientific articles (more than 49 k citations).