End-to-End Label Uncertainty Modeling in Speech Emotion Recognition Using Bayesian Neural Networks and Label Distribution Learning

Navin Raj Prabhu, Student Member, IEEE, Nale Lehmann-Willenbrock, and Timo Gerkmann, Senior Member, IEEE

Abstract—To train machine learning algorithms to predict emotional expressions in terms of arousal and valence, annotated datasets are needed. However, as different people perceive others’ emotional expressions differently, their annotations are subjective. To account for this, annotations are typically collected from multiple annotators and averaged to obtain ground-truth labels. However, when exclusively trained on this averaged ground-truth, the model is agnostic to the inherent subjectivity in emotional expressions. In this work, we therefore propose an end-to-end Bayesian neural network capable of being trained on a distribution of annotations to also capture the subjectivity-based label uncertainty. Instead of a Gaussian, we model the annotation distribution using Student’s t-distribution, which also accounts for the number of annotations available. We derive the corresponding Kullback-Leibler divergence loss and use it to train an estimator for the annotation distribution, from which the mean and uncertainty can be inferred. We validate the proposed method using two in-the-wild datasets. We show that the proposed t-distribution based approach achieves state-of-the-art uncertainty modeling results in speech emotion recognition, and also consistent results in cross-corpora evaluations. Furthermore, analyses reveal that the advantage of a t-distribution over a Gaussian grows with increasing inter-annotator correlation and a decreasing number of annotations available.

Index Terms—t-distributions, annotations, Bayesian neural networks, emotional expressions, end-to-end, Kullback-Leibler divergence loss, label distribution learning, speech emotion recognition, subjectivity, uncertainty.

I. INTRODUCTION

EMOTIONS are typically studied as emotional expressions that others subjectively perceive and respond to [1], [2]. A standard theoretical backdrop for analyzing emotions is the two-dimensional pleasure and arousal framework [3], which describes emotional expressions along two continuous, bipolar, and orthogonal dimensions: pleasure-displeasure (valence) and activation-deactivation (arousal). One way emotions become expressed in social interactions, and therefore accessible for social signal processing (SSP), concerns speech signals. Speech emotion recognition (SER) research spans roughly two decades [2], with ever improving state-of-the-art techniques. As a consequence, research on SER has shown increasing prominence in highly-critical and socially relevant domains, such as health, security, and employee well-being [2], [4], [5], [23].

A crucial challenge when studying emotional expressions in terms of arousal and valence is that their annotations are per se subjective because different people perceive others’ emotional expressions differently [2], [5]. To address this, these annotations are typically collected by multiple annotators, and consensus on ground-truth is reached using techniques such as average scores [6], majority voting [6], or evaluator-weighted mean (EWE) [7]. These techniques in principle can lead to loss of valuable information on the inherently subjective nature of emotional expressions, and also tend to mask less prominent emotional traits [5]. In the context of reliability in real-world applications, SER systems not only need to model ground-truth labels but also account for the subjectivity inherent in these labels [2], [8]. Moreover, by also capturing subjectivity, SER systems can be efficiently deployed in human-in-the-loop solutions, and aid in the development of algorithms for active learning, co-training, and curriculum learning [5], [8].

In this work, we tackle the problem of recognising emotional expressions using speech signals, in terms of time- and value-continuous arousal and valence. To this, we adopt an end-to-end learning framework. Common SER approaches rely on hand-crafted features to model emotion labels [9], [10]. Recently, end-to-end architectures have been shown to deliver state-of-the-art emotion predictions [11], [12], [13], by learning features rather than relying on hand-crafted features. For modeling subjectivity in emotions, scholars have suggested that end-to-end learning also promotes learning subjectivity dependent representations [14].

Uncertainty in machine learning (ML) is generally investigated in terms of two broader categories. First, model uncertainty, or epistemic uncertainty, accounts for the uncertainty in model parameters, and the resulting uncertainty can be reduced given enough data-samples [15], [16], [17]. Second, label uncertainty, or aleatoric uncertainty, captures noise inherent in the data-samples, such as sensor noise or label noise [15], [16]. Label uncertainty cannot be reduced even if more data-samples...
are collected. Label uncertainty has been further categorized into **homo- psedastic** uncertainty, which remains constant across data-samples, and **heteroscedastic** uncertainty, whose uncertainty depends on the respective data-sample. This work specifically aims to model the heteroscedastic label uncertainty, henceforth simply mentioned as label uncertainty, that corresponds to the **inherent subjectivity in emotion annotations**.

We propose to use **Bayes by Backpropagation** (BBB), a Bayesian neural network (BNN) technique, in order to capture label uncertainty. In ML, stochastic and probabilistic models have mainly been used for uncertainty modeling, through ensemble learning [18], encoder-decoder architectures [19], neural processes [20], [21], and BNNs [22], [23], [24]. Among these, the Bayesian frameworks show improved performance over non-Bayesian baselines in previous works [15], [25], making BNNs such as Monte-Carlo dropout [22] and BBB [23] promising candidates for modeling label uncertainty in SER. BBB learns a distribution over weights to produce stochastic outputs, which makes it capable of being trained on a distribution of annotations.

With BBB capable of being trained on a distribution of annotations, we capture label uncertainty using the **label distribution learning** (LDL) technique [25], leveraging Kullback-Leibler (KL) divergence-based loss functions. Subjective annotations of emotion create a label distribution to represent the subjectivity in emotions [5]. For simplicity, histograms [5], [26], and Gaussians [27] have been employed to represent the label distributions. However, Gaussians and histograms with **limited** and **sparse** observations are sensitive to outliers thereby losing their robustness in this scenario [28], [29], [30]. Note here that publicly available SER datasets commonly comprise only limited annotations (e.g., 3 to 6) [31], [32], [33], [34], [35], and there is consensus in the literature that gaining more annotations is expensive and resource inefficient [5], [36]. At the same time, a significant degree of subjectivity in annotations is also well noted [10], [14], thereby leading to sparse annotations with outliers. To tackle this, in this work, we model emotion annotation distributions as a Student’s t-distribution, or simply t-distribution. Kotz et al. [29], and, Bishop [28], note that in scenarios of limited and sparse observations with outliers, the t-distribution becomes more robust over a Gaussian, by producing robust mean and standard deviation estimates of the distribution.

We derive a KL divergence loss for label uncertainty that quantifies distribution similarity between stochastic emotion predictions, modeled as a Gaussian distribution, and ground-truth emotion annotations, modeled as a t-distribution. Subsequently, we present analyses to reveal the benefits of using t-distribution over a Gaussian. We validate the proposed model in two in-the-wild datasets, AVEC’16 [37] and MSP-Conversation [31]. We show that the proposed model can aptly capture label uncertainty with state-of-the-art results for both datasets, along with a robust loss curve. To emphasize the benefits of the t-distribution, we present experiments studying the impact of the number of emotion annotations available. Finally, we perform an ablation study to understand specific benefits of the respective modules in the architecture.

This work is based on two prior conference contributions [38], [39], which to the best of our knowledge are the first in the literature to use BBB and LDL in SER. These works were also the first to tackle the problem of limited emotion annotations from an ML perspective. Previously, we only validated the method in one dataset, and with limited experiments [38], [39]. In this extension, we additionally validate the method in a larger and more complex dataset, the MSP-Conversation [31], along with cross-corpora evaluations. This extension is also the first in the literature to present SER results in this novel dataset [31]. Existing analyses and experiments from [38], [39] were also extended to MSP-Conversation. Moreover, we performed additional experiments that include an experiment to understand the impact of the number of annotations available, and an ablation study. Code for the proposed model and loss function is available online.1

## II. BACKGROUND AND RELATED WORK

### A. Ground-Truth Labels

To handle subjectivity in emotional expressions, annotations \{y_1, y_2, ..., y_a\} for emotions are typically collected from multiple annotators (a) [33], [35]. The **ground-truth label** is then obtained as the mean m across all annotations from a annotators [40]

\[
m = \frac{1}{a} \sum_{i=1}^{a} y_i.
\]

Alternatively, the EWE, which weights annotations with inter-annotator correlations, has been proposed as the **gold-standard \(\bar{m}\)** [7]. Both m and \(\bar{m}\) based approximation of ground-truth leads to loss of information on subjectivity [5].

Given a raw audio sequence of T frames \(\mathcal{X} = \{x_1, x_2, ..., x_T\}\), traditional SER approaches aim to estimate either the \(m_t\) or \(\bar{m}_t\) for each time frame \(t \in [1, T]\), referred to as \(\hat{m}_t\). The concordance correlation coefficient (CCC) [41] has been widely used as a loss function for this task [2]. For Pearson correlation r, the CCC between m and \(\hat{m}\), for T frames is

\[
\mathcal{L}_{\text{CCC}}(m) = \frac{\sigma_m^2 + \sigma_{\hat{m}}^2 - 2r \sigma_m \sigma_{\hat{m}}}{\sigma_m^2 + \sigma_{\hat{m}}^2 + (\mu_m - \mu_{\hat{m}})^2},
\]

where \(\mu_m = \frac{1}{T} \sum_{t=1}^{T} m_t\), \(\sigma_m^2 = \frac{1}{T} \sum_{t=1}^{T} (m_t - \mu_m)^2\), and \(\mu_{\hat{m}}, \sigma_{\hat{m}}^2\) are obtained similarly for \(\hat{m}\). The CCC metric measures the agreement between two variables, in our case the ground-truth m and its estimate \(\hat{m}\). It ranges from -1 to +1, with perfect agreement at +1. In contrast to Pearson’s correlation r, CCC takes both the linear correlation and the bias into account, which makes it preferable over Pearson’s correlation as the loss and evaluation metric in SER.

### B. Label Uncertainty in SER

As an alternative to exclusively modeling \(m_t\) or \(\bar{m}_t\), previous research has attempted to model ground truth that also explains inter-annotator disagreement, for example by means of soft labels [5] and entropy of disagreement [42]. Sridhar et al. [5] proposed an auto-encoder technique that jointly models soft- and hard-labels of emotion annotations and subsequently estimates label uncertainty as the entropy on soft-labels. Fayek et al. [43] and Tarantino et al. [44] proposed to learn soft labels instead of \(m_t\) with improved performance. Steidl et al. [42] quantified

---

1https://github.com/sp-uhh/label-uncertainty-ser
Label uncertainty has also been approached as a prediction task by estimating the moments of a distribution [9], [45]. Han et al. [9], [45] used a multi-task learning (MTL) framework to model the unbiased standard deviation $s$ of annotators as an auxiliary task

$$s = \sqrt{\frac{1}{a-1} \sum_{i=1}^{a} (y_i - m)^2},$$

where $m$ is the mean and standard deviation

Similarly, Dan et al. [46] captured the temporal dependencies in the annotation signals, using multi-rater Gaussian mixture regression and Kalman filters. Sridhar et al. [10] introduced a Monte-Carlo (MC) dropout model to obtain uncertainty estimates from the distribution of stochastic outputs. However, their model was not explicitly trained on any label uncertainty estimate and hence could only capture the model uncertainty, but not the label uncertainty. A similar MC dropout was used by Rizos et al. [47], who proposed a meta-learning framework that uses uncertainty estimates to potentially detect highly-uncertain samples and perform soft data selection for the training process.

Research efforts have also been made to estimate emotion annotations as a distribution, using LDL [26], [27], [38], [39]. Fotinopoulos et al. [27] trained an MTL network using a KL divergence loss that models emotion annotations as a univariate Gaussian with mean $m$ and unknown variance. Chou et al. [26] used LDL to convert subjective into histogram-based distributional labels for training. In our preliminary work [38], we modeled emotion annotations as a Gaussian using BBK-based uncertainty modeling. Notwithstanding the improved performance of these approaches, a drawback concerns the limited annotations on which previous histogram assumptions were based [26], [27]. These assumptions are susceptible to unreliable $m$ and $s$ for lower values of $a$ and sparsely distributed annotations [28], [29]. In our subsequent work [39] and in this extension, we tackle this problem by modeling emotion annotation distribution as a $t$-distribution and show advantages over a Gaussian assumption.

C. On Distributions

A Gaussian distribution $Y \sim \mathcal{N}(\mu, \sigma^2)$ is a continuous probability distribution for a real-valued random variable $y$, with the general form of its probability density function [28]

$$p(y \mid \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}.$$  

The parameters $\mu$ and $\sigma$ are the mean and standard deviation of the distribution, respectively. Due to its simplicity, Gaussians are often used to model random variables whose distributional family are unknown [23], [38]. However, Gaussians are sensitive to outliers, especially in cases of limited and sparse observations of the random variable [28]. In this case, the $t$-distribution is noted to become more robust and realistic over a Gaussian [28], [29].

Student’s $t$-distribution, $Y_t \sim \mathcal{N}(\nu, \mu, \sigma)$, arises when estimating the moments of a normally distributed population in situations where the sample size is small [29], [48], with the probability density function given by [30], [49]

$$p(y \mid \nu, \mu, \sigma) = \frac{1}{B\left(\frac{\nu}{2}, \frac{\nu}{2}\right)} \frac{1}{\sqrt{\nu} \sigma^2} \left(1 + \frac{(y - \mu)^2}{\nu \sigma^2}\right)^{-\frac{\nu+1}{2}},$$

where $\nu$ denotes the degrees of freedom and $B(\cdot,\cdot)$ is the Beta function, for Gamma function $\Gamma$, formulated as

$$B(i, j) = \frac{\Gamma(i) \Gamma(j)}{\Gamma(i + j)}.$$  

The density function (5) is symmetric, and its overall shape resembles the bell shape of a normally distributed variable, except that it has heavier tails, meaning that it better captures values that fall far from its mean (i.e., outliers) [28], [29]. The degree of freedom $\nu$, also known as the normality parameter, controls the normality of the distribution and is correlated with the standard deviation of the distribution $\sigma$ [28], [29]. In (5), the standard deviation $\sigma$ takes the scaled form, where $\sigma$ is scaled using the normality parameter $\nu$

$$\sigma \sqrt{\nu \nu - 2} \text{ for } \nu > 2,$$

As $\nu$ increases, the $t$-distribution approaches the normal distribution [30]. The normality parameter $\nu$, in our case, allows the $t$-distribution to also account for the number of annotations available.

The robustness of the $t$-distribution, in cases of limited and sparse observations of the random variable, is associated with its ability to better capture the outliers by also accounting for the number of observations of the random variable [28]. This is the key motivation behind using the $t$-distribution to model emotion annotations, to produce robust mean and standard deviation estimates by also accounting for the number of annotations available.

III. PROPOSED LABEL UNCERTAINTY MODEL

In order to better represent subjectivity in emotional expressions, we estimate the emotion annotation distribution $Y_e$ for each time-frame $t$, given raw audio $x_t$. While the true distributional family of subjectively perceived emotions $Y_e$ is unknown, for simplicity, we can assume that it follows a Gaussian distribution

$$Y_e \sim \mathcal{N}(m_e, s_e^2).$$

However, with only a limited number of annotations, and in cases where the annotations are sparsely distributed with outliers, we argue that a Gaussian assumption is rather crude [28], [29]. Instead, we propose to model the emotion annotations as a $t$-distribution, with degrees of freedom $\nu$

$$Y_e \sim \mathcal{N}(\nu, m_e, s_e^2).$$

Thus, the goal is to obtain an estimate $\hat{Y}_e$ of $Y_e$ and infer both $\hat{m}_e$ and $\hat{s}_e$ from realizations of $\hat{Y}_e$.

A. End-to-End DNN Architecture

We propose an end-to-end architecture that uses a feature extractor to learn temporal-paralinguistic features from $x_t$, and an uncertainty layer to estimate $Y_e$ (see Fig. 1). The feature extractor, inspired by [11], consists of three Conv1D layers followed by two stacked long short term memory (LSTM) layers.
The uncertainty layer is devised using the BBB technique [23], comprising three BBB-based MLP.

B. Model Uncertainty Loss

Unlike a standard neuron which optimizes a deterministic weight \( w \), the BBB-based neuron learns a probability distribution on the weight \( w \) by calculating the variational posterior \( P(w|D) \) given the training data \( D \) [23]. Intuitively, this regularizes \( w \) to also capture the inherent uncertainty in \( D \). In contrast to learning a deterministic weight \( w \) to exclusively estimate \( m_t \), the BBB neuron learns a Gaussian weight distribution \( \mathcal{N}(\mu_w, \sigma_w) \), thereby allowing the model to not only estimate \( m_t \) but also \( s_t \). Estimation of \( s_t \) is achieved by calculating the standard deviation of the stochastic estimates obtained from stochastically sampled weights \( w^{(i)} \).

To obtain a non-negative estimate of the standard deviation of the standard deviation of \( w \), we re-parameterize the standard deviation as \( \sigma_w = \log (1 + \exp(\rho_w)) \) based on an initial estimate \( \rho_w \). This way \( \theta = (\mu_w, \rho_w) \) can be optimized using simple backpropagation and still ensure a non-negative \( \sigma_w \).

For an optimized \( \theta \), the predictive distribution \( \hat{Y}_t \) for \( x_t \), is given by \( P(\hat{y}_t|x_t) = \mathbb{E}_{P(w|D)}[P(\hat{y}_t|x_t, w)] \), where \( \hat{y}_t \) are realizations of \( \hat{Y}_t \). Unfortunately, the expectation under the posterior of weights is intractable. To tackle this, [23] proposed to learn \( \theta \) as a weight distribution \( q(w|\theta) \), the variational posterior, that minimizes the Kullback-Leibler (KL) divergence with the true Bayesian posterior, resulting in the negative evidence lower bound (ELBO).

\[
f(w, \theta)_{\text{BBB}} = \text{KL}[q(w|\theta)||P(w)] - \mathbb{E}_{q(w|\theta)} \left[ \log P(D|w) \right],
\]  

(10)

In BBB, stochastic outputs are achieved using multiple forward passes \( n \) with stochastically sampled weights \( w \), thereby modeling \( \hat{Y}_t \) using the \( n \) stochastic estimates. To account for the stochastic outputs, (10) is approximated as

\[
\mathcal{L}_{\text{BBB}} \approx \sum_{i=1}^{n} \log q(w^{(i)}|\theta) - \log P(w^{(i)}) - \log P(D|w^{(i)}),
\]  

(11)

where \( w^{(i)} \) denotes the \( i \)th weight drawn from \( q(w|\theta) \). The BBB window-size \( b \) controls how often new weights are sampled for time-continuous SER. The degree of uncertainty is assumed to be constant within this time period. During testing, the uncertainty estimate \( \hat{s}_t \) is the standard deviation of \( \hat{Y}_t \), and, \( \hat{m}_t \) is the realization \( \hat{y}_t \) obtained using the mean of the optimized weights \( \mu_w \). Obtaining \( \hat{m}_t \) using \( \mu_w \) helps overcome the randomization effect of sampling from \( q(w|\theta) \), which showed better performances in our case.

Note that variables \( n, a, \) and \( \nu \) are closely related to one another. The three variables all denote the number of samples used to model distribution, either \( \hat{Y}_t \) or \( Y_t \). Variable \( n \) represents the number of forward passes, thereby the number of stochastic estimates used to model the estimate distribution \( \hat{Y}_t \). Variable \( a \) represents the number of annotations used to model the ground-truth distribution \( Y_t \). In the probability density function of a \( t \)-distribution (5), \( \nu \) denotes the degree of freedom. In this work, the \( \nu \) of a \( t \)-distribution is set to \( a \) enabling the ground-truth distribution \( Y_t \) to also account for the number of annotations available.

C. Label Uncertainty Loss

While (11) exclusively captures model uncertainty, the aim of this work is to also capture label uncertainty. For this, using LDL, inspired by [16], we introduce a KL divergence-based loss to fit our model to the annotation distribution \( Y_t \), with either a Gaussian assumption (in Section III-C1) or a \( t \)-distribution assumption (in Section III-C2).

1) Gaussian \( Y_t \) KL Divergence: For a Gaussian assumption on \( Y_t \) (8), the label uncertainty loss, the KL divergence between two Gaussians \( Y_t \sim \mathcal{N}(m_t, s_t^2) \) and \( \hat{Y}_t \sim \mathcal{N}(\hat{m}_t, \hat{s}_t^2) \) is formulated as [28]

\[
\mathcal{L}_{KL}(Y_t||\hat{Y}_t) = \log \left( \frac{\hat{s}_t}{s_t} \right) + \frac{s_t^2 + (m_t - \hat{m}_t)^2}{2\hat{s}_t^2} - \frac{1}{2}.
\]  

(12)

The KL divergence is asymmetric, making the order of distributions crucial. In (12), we choose \( \hat{Y}_t \) to follow \( Y_t \), for a mean-seeking approximation, rather than a mode-seeking one, to capture the full distribution [50, p. 76]. See Supplementary Section 3, available online for further details on the choice between mean- and mode-seeking approximation using \( \mathcal{L}_{KL} \).

2) \( t \)-Distribution \( Y_t \) KL Divergence: For \( Y_t \) as a \( t \)-distribution (9), we derive the KL divergence between \( Y_t \sim \mathcal{N}(\nu, m_t, s_t^2) \) and the Gaussian outputs \( \hat{Y}_t \sim \mathcal{N}(\hat{m}_t, \hat{s}_t^2) \). Assuming a Gaussian on \( \hat{Y}_t \) is fair, as the number of stochastic outputs to model \( \hat{Y} \) can be controlled using \( n \) in (11). In this work, we intend to fix \( n \geq 30 \), as a \( t \)-distribution converges to a stable Gaussian with 30 samples [30], [49]. As a positive side effect, we result in deriving the KL divergence between a Gaussian and a \( t \)-distribution, in contrast to between two \( t \)-distributions, with the latter involving mathematical complexities in calculating intractable expectations for a loss function.

For a Gaussian \( \hat{Y} \) (see 4), and a \( t \)-distributed \( Y \) (see 5), the \( \mathcal{L}_{KL} \) is formulated as [51], [52]

\[
\mathcal{L}_{KL}(Y_t||\hat{Y}_t) = H(Y_t, \hat{Y}_t) - H(Y_t),
\]  

(13)

where \( H(\cdot, \cdot) \) is the cross-entropy between two distributions, and \( H(\cdot) \) is the entropy of a distribution. The cross-entropy term \( H(\cdot, \cdot) \) in (13), using (4), can be further formulated as

\[
H(Y_t, \hat{Y}_t) = -\int Y_t(y) \log \hat{Y}_t(y) \, dy
\]  

\[
= \frac{1}{2} \log(2\pi \hat{s}_t^2) + \int Y_t(y) \left( \frac{y - \hat{m}_t^2}{2\hat{s}_t^2} \right) \, dy
\]
= \frac{1}{2} \log(2\pi s_t^2) + \frac{1}{2s_t^2} \left[ \int Y_t(y)^2 dy \right] - 2\hat{m}_t \int Y_t(y) y dy + \hat{m}_t^2 \int Y_t(y) dy \right]. \quad (14)

Noting that \( \int Y_t(y)^2 dy = m_t^2 + s_t^2 \), \( \int Y_t(y) y dy = m_t \), and \( \int Y_t(y) dy = 1 \), where \( m_t \) and \( s_t \) are parameters of the \( t \)-distribution, \( p(y | \nu, m_t, s_t) \), the (14) becomes

\[
= \frac{1}{2} \log(2\pi s_t^2) + \frac{1}{2s_t^2} \left[ s_t^2 + m_t^2 - 2\hat{m}_t m_t + \hat{m}_t^2 \right] \\
= \frac{1}{2} \log(2\pi s_t^2) + \frac{s_t^2 + (m_t - \hat{m}_t)^2}{2s_t^2} \quad (15)
\]

Finally, using (15) in (13), our proposed KL divergence is

\[
\mathcal{L}_{KL} = \frac{1}{2} \log(2\pi s_t^2) + \frac{s_t^2 + (m_t - \hat{m}_t)^2}{2s_t^2} - H(Y_t). \quad (16)
\]

We implement (16) as a custom loss function by extending the \texttt{studentT} pytorch sub-package [53].

3) Comparing Gaussian and \( t \)-Distribution Loss: While the two loss-functions (12) and (16) have their second term in common, two differences can be noted. First, as (12) calculates the divergence between two similar distributions, \( Y_t \) and \( \hat{Y}_t \), it includes the logarithm of the ratio between the two Gaussian’s standard deviation in its formulation. However, in (16), the deviations of \( Y_t \) and \( \hat{Y}_t \) are separately quantified using terms \( \frac{1}{2} \log(2\pi s_t^2) \) and \( H(Y_t) \), respectively. Second, (16) is dependent on the number of annotations available through scaling \( s_t \) with the normality factor \( \nu \).

To further understand the advantages of the \( t \)-distribution \( \mathcal{L}_{KL} \) (16) over the Gaussian \( \mathcal{L}_{KL} \) (12), we plot the \( \mathcal{L}_{KL} \) values as a function of varying \( s_t \) for (16) and (12). We perform this analysis under four different scenarios, for different values of \( \hat{s}_t \) and \( \nu \), i) Fig. 2(a) for scenario \( \hat{s}_t = 0.5 \) and \( \nu = 6 \), ii) Fig. 2(b) for scenario \( \hat{s}_t = 1.0 \) and \( \nu = 6 \), iii) Fig. 2(c) for scenario \( \hat{s}_t = 1.0 \) and \( \nu = 12 \), and, iv) Fig. 2(d) for scenario \( \hat{s}_t = 1.0 \) and \( \nu = 30 \).

From Fig. 2, first, we see that \( \mathcal{L}_{KL} \) behaves differently when the ground-truth \( Y_t \) is modeled as a \( t \)-distribution (16), in comparison to the Gaussian assumption (12). Specifically, from Fig. 2(a), for \( \hat{s}_t = 0.5 \) and \( \nu = 6 \), we see that the minimum \( \mathcal{L}_{KL} \) (16) is achieved only at \( s_t = 0.61 \), in contrast to the Gaussian (12) \( s_t = 0.5 \). While the Gaussian attempts exactly fitting the model to the ground-truth \( s_t = 0.5 \), the \( t \)-distribution tries to fit on a more relaxed \( s_t = 0.61 \) by also considering the reduced degree of freedom \( \nu = 6 \). This behaviour is similar to the confidence intervals calculation using a \( t \)-distribution [54, Section 9.5], where relaxation on \( s_t \) is noted with respect to \( \nu \). Moreover, [28] associate this relaxed \( s_t \) towards the increased robustness of the \( t \)-distribution to sparse distributions with outliers.

Second, we note that the observed relaxation on \( s_t \) is dependent on two factors, 1) the standard deviation of the stochastic outputs \( \hat{s}_t \), and 2) the degree of freedom of the ground-truth \( \nu \). From Figs. 2(a) and (b), we see that, while \( \nu \) is constant, the relaxation on \( s_t \) increases along with an increase in \( \hat{s}_t \). At \( \hat{s}_t = 0.5 \), a relaxation of 0.11 is made by the \( t \)-distribution (16) from 0.5 to 0.61, while a larger relaxation of 0.22 is made for \( \hat{s}_t = 1.0 \). Similarly, from Figs. 2(c) and (d), we see that, while \( \hat{s}_t \) is constant, as \( \nu \) increases the relaxation on \( s_t \) decreases. That is, the \( t \)-distribution (16) starts behaving similar to a Gaussian, in line with literature that states that as the degree of freedom \( \nu \) of \( t \)-distribution increases, the distribution converges into a Gaussian [29, 30, 49]. This is also in line with our initial motivation behind using the \( t \)-distribution, which we expected to account for the number of annotations available while fitting on annotation distribution \( Y_t \).

From an ML and SER perspective, from Fig. 2, we note several benefits of \( t \)-distribution based loss term towards label uncertainty modeling. First, training on a \( t \)-distribution \( \mathcal{L}_{KL} \) (16) leads to training on a relaxed \( s_t \), and can lead to better capturing of the whole ground-truth label distribution. In other words, this can lead to the \( t \)-distribution better accounting for sparse annotations with outliers, where a relatively high likelihood is associated along the tails of the distribution, as noted by Bishop [28]. Second, we note that in all cases, the \( t \)-distribution \( \mathcal{L}_{KL} \) (16) values are always higher than Gaussian \( \mathcal{L}_{KL} \) for lower values of \( s_t \) and \( \hat{s}_t \). This might lead to larger penalization of the model through the \( \mathcal{L}_{KL} \) loss, and may thereby promote better and quicker convergence during training, in comparison to the Gaussian \( \mathcal{L}_{KL} \) (12). Finally, the \( t \)-distribution \( \mathcal{L}_{KL} \) (16) can also adapt to different datasets by also accounting for the number of annotations available during training.
D. Training Loss

The proposed end-to-end uncertainty loss is formulated as

$$\mathcal{L} = (1 - \mathcal{L}_{CCC}(m)) + \mathcal{L}_{BBB} + \alpha \mathcal{L}_{KL}.$$  (17)

Intuitively, $\mathcal{L}_{CCC}(m)$ optimizes for mean predictions $m$, $\mathcal{L}_{BBB}$ optimizes for BBB weight distributions, and $\mathcal{L}_{KL}$ optimizes for the label distribution $\gamma_t$. For $\alpha = 0$, the model only captures model uncertainty (MU). For $\alpha = 1$, the model also captures label uncertainty (MU + LU or t-LU). $\mathcal{L}_{CCC}(m)$ is used as part of $\mathcal{L}$ to achieve faster convergence and jointly optimize for mean predictions. Including $\mathcal{L}_{CCC}(m)$ might lead to better optimization of the feature extractor [11], [55].

In (17), $\alpha$ is the tuning parameter that decides how much we want to regularize our model to also account for the label uncertainty. While the proposed models only use two values for the $\alpha$ (0 and 1), as an additional study, we also experimented with varying regularization on the label uncertainty loss term $\mathcal{L}_{KL}$ (see Supplementary Section 5, available online).

IV. EXPERIMENTAL SETUP

A. Dataset

To validate our proposed methodology, we use two publicly available in-the-wild datasets, with time- and value-continuous annotations for arousal and valence. First, the AVEC’16 [37] version of the RECOLA dataset [33], which has 2.15 hrs of annotated dyadic interactions. Second, the MSP-Conversation dataset, which has 15.15 hrs of annotated interactions with groups of 2-7 interlocutors.

1) AVEC’16 Dataset: The dataset consists of arousal and valence annotations by $a = 6$ annotators at 40 ms frame-rate, or 25 frames per second (fps). The arousal and valence annotations in the dataset are distributed on average with $\mu_{m} = 0.01$ and $\mu_{m} = 0.11$, and $\mu_{s} = 0.23$ and $\mu_{s} = 0.14$, respectively, where $\mu_{s} = \frac{1}{T} \sum_{t=1}^{T} s_{t}$. Further, in Fig. 3 the distribution of $s_{t}$ is illustrated. It can be noted from Fig. 3 that $s_{t}$ distributions are skewed towards high standard deviations $s_{t}$, thereby indicating the high-level of subjectivity present in the dataset. The high skewness is even more evident in the log-histogram plotted along in Fig. 3. The dataset is divided into speaker disjoint partitions for training, development, and testing, with nine 300 s recordings each. Results with respect to the AVEC’16 are presented only in terms of the development partition, as the annotations for the test partition are not publicly available. Similarly, the hyperparameters are fine-tuned on the train partition for this particular dataset. Note that the posterior distribution $P(w|D)$ and the time-shift for post-processing are the only parameters tuned using the partitions. See Supplementary Section 1 for the complete list of hyperparameters used, available online.

2) MSP-Conversation Dataset: The MSP-Conversation, or simply MSPConv, is approximately 7 times larger than AVEC’16, comprising of in-the-wild podcasts. The wide range of podcast recordings leads to high variability in terms of population size, group size, and more importantly its emotional content [31], [32], making the MSPConv a more complex dataset to model.

The dataset consists of time- and value-continuous annotations for arousal and valence, performed by at least $a = 6$ annotators at $\approx 16$ ms frame-rate, or 60 fps, however not uniform in all cases [31]. For uniform sampling rate, we perform median filtering with a window-size of 500 ms, as suggested in [31]. To keep the sampling rate consistent between the two datasets, for cross-corpora evaluations, we use a step-size of 1/25s in median filtering. A local normalization, i.e., for each annotated sequence and for each annotator, was performed using zero-mean unit-deviation normalization [33], similar to AVEC’16. As illustrated in Fig. 4, in the MSPConv dataset [31], arousal and valence annotations are distributed on average with $\mu_{m} = -0.01$ and $\mu_{m} = 0.00$, and $\mu_{s} = 0.18$ and $\mu_{s} = 0.21$, respectively. Further revealing the complexity of MSPConv, when comparing Figs. 3 and 4, we see that the level of subjectivity in MSPConv is higher than the AVEC’16 dataset, where in MSPConv the $s_{t}$ distribution tail is more skewed towards high subjectivity. The high skewness is more evident in the log-histogram plotted along in Fig. 4.

In preliminary experiments, we noted that the arousal and valence annotations were prone to periodic distortion noises, especially from particular annotators—001, 007, and 009. This could have originated from any technical error or from a human error by the annotator. Directly training on these noisy annotations degraded the performance of all models in comparison. Ignoring the noisy annotations might lead to a loss of information, and might also result in a reduced number of available annotations to derive ground-truth. To reduce these periodic distortions and still retain the inherent annotation information, we use a low-pass filter [56] with a cut-off frequency of 0.25 Hz. The cut-off frequency was tuned using a Fourier analysis [57] followed by a qualitative analysis of the filtered annotations. Filtering was performed only on annotations with periodic distortions, i.e., from the three annotators—001, 007, and 009.
B. Baselines and Proposed Model Versions

E2E Baselines: This baseline is a reimplementation of [11], with the same end-to-end framework as our proposed model but a multi-layer perceptron instead of the uncertainty layer. The model does not capture any form of uncertainty, and is exclusively trained on the $L_{\text{CCC}}(m)$ loss (2).

Time-continuous ground-truth annotations contain temporal dependencies [58], where an annotation at time $t$ can be expected to have a high correlation with annotations at time $t + 1$ and $t - 1$. Our proposed architecture accounts for this temporal dependency using two stacked LSTM layers. Moreover, temporal modeling is achieved by batching annotations into sequences of 12 s each (300 frames of 40 ms each). With this setup, the LSTM operation is performed over the sequence rather than over a single frame, thereby directly learning temporal dependencies.

To assess the impact of this temporal modeling, we use an additional baseline: E2E Baseline w/o Temp where the LSTM operation is performed on the feature dimension, in contrast to the E2E Baseline where the operation is performed on the temporal dimension. This way the number of parameters is kept the same for the two allowing for a fair comparison.

MTL Baselines: From [9], [45], as the baselines, we use the perception uncertainty (MTL PU) and single-task models (STL). The MTL PU is a label uncertainty model that also models $s_t$ as an auxiliary task. The STL does not capture uncertainty and is exclusively trained on $L_{\text{CCC}}(m)$ (2). For a fair comparison, we reimplemented these baselines. Crucially, the reimplementation also enables us to compare the models in terms of their standard deviation $s$ estimates, which were not presented in Han et al.’s work [9].

Proposed BBB-LDL Versions: We use three versions of the proposed label uncertainty model. First, the Model Uncertainty (MU) version, which shares the same DNN architecture as the other BBB version but is trained on (17) with $\alpha = 0$. Second, the Label Uncertainty (MU + LU) version also captures the label uncertainty and is trained on (17) with $\alpha = 1$. The MU + LU version however makes a Gaussian assumption on $\hat{y}_t$, thereby $L_{\text{KL}}$ follows (12). Finally, the t-distribution Label Uncertainty (t-LU) version, which is trained on the same loss function (17) but models $\hat{y}_t$ as a t-distribution, and $L_{\text{KL}}$ follows (16).

Finally, for all the models two post-processing techniques are applied, namely, median filtering [11] and time-shifting [59] (with shifts between 0.04 s and 10 s). See supplementary Section 2 for further detailed information, available online.

C. Validation Measures

To validate the proposed method’s mean and standard deviation estimates, we use $L_{\text{CCC}}(m)$ and $L_{\text{CCC}}(s)$ metrics, respectively, widely used in literature [9], [11], [55]. However, $L_{\text{CCC}}(m)$ and $L_{\text{CCC}}(s)$ validate mean and standard deviation estimates separately. To further jointly validate mean and standard deviation estimates, as label distribution $\hat{y}_t$, we use the $L_{\text{KL}}$ measure. For a fair comparison, we validate all the models in comparison using $L_{\text{KL}}$ based on their respective distribution assumptions on $\hat{y}_t$, as the models are trained in a similar fashion. The proposed t-LU version is validated and trained on the t-distribution $L_{\text{KL}}$ (16), and the baselines on the Gaussian $L_{\text{KL}}$ (12). Nevertheless, from the experiments, we also noted that the proposed t-LU performs better in terms of both (16) and (12). Finally, the statistical significance of results is estimated using a one-tailed t-test, asserting significance for $p$-values $\leq 0.05$.

V. RESULTS AND DISCUSSION

A. Quantitative Analysis of Estimates

Table I shows the average performance of the baselines and the proposed models, in terms of their mean $m$, standard deviation $s$, and distribution $\hat{y}_t$ estimations, $L_{\text{CCC}}(m)$, $L_{\text{CCC}}(s)$, and $L_{\text{KL}}$, respectively. Results are presented with respect to two datasets, for AVEC’16 in Table I a, and for MSPConv in Table I b. From the analysis presented in Section IV-A, we note that the MSPConv is a more complex dataset, in terms of modeling label uncertainty.

1) Comparison on Mean Estimates: In terms of arousal, Table I shows that the proposed t-LU model performs the best in comparison with the baselines, in both AVEC’16 (Table I a) and MSPConv (Table I b) datasets, with statistical significance. Four key takeaways can be noted from the $L_{\text{CCC}}(m)$ results for arousal. First, the proposed BBB-LDL versions (MU, MU + LU, and t-LU) achieve better $L_{\text{CCC}}(m)$ than the MTL baselines (STL, and MTL PU). In the more challenging MSPConv dataset, the performance improvement is even more evident, which highlights the robustness of the proposed approach. For example, while the t-LU improves over MTL PU by 0.042 in AVEC’16, a larger improvement of 0.093 $L_{\text{CCC}}(m)$ can be noted in the MSPConv. Second, between the BBB-LDL versions, the superiority of the proposed t-distribution $L_{\text{KL}}$ (16) over the Gaussian $L_{\text{KL}}$ (12) is noted, with t-LU outperforming MU + LU in both the datasets. Third, when incorporating uncertainty modeling in the E2E Baseline, a compromise on $L_{\text{CCC}}(m)$ is made with improving uncertainty estimates ($L_{\text{CCC}}(s)$ and $L_{\text{KL}}$). This can be noted when comparing the results of MU and MU + LU with that of the E2E Baseline. However, the proposed t-LU is free from this compromise, outperforming the E2E Baseline and other BBB-LDL versions. The t-LU achieves a $L_{\text{CCC}}(m)$ of 0.782 in AVEC’16 and 0.389 in MSPConv, with E2E Baseline achieving 0.770 and 0.373, respectively. Finally, the E2E Baseline w/o Temp performs the worst in comparison. The improved performance of E2E Baseline and the proposed models over the E2E Baseline w/o Temp emphasises the fact that temporal modeling exists in the proposed models and is achieved through the inclusion of the LSTM layers in their architecture.

In terms of valence, in the AVEC’16, the MTL PU baseline performs significantly better than the proposed models. However, in the larger and more complex MSPConv dataset, the proposed t-LU performs the best with statistical significance. The MTL PU requires dataset dependent tuning of the loss using the average correlation between $m_t$ and $s_t$. For example, $L = L_{\text{ccc}}(m) - L_{\text{ccc}}(s)$ for datasets with negative average correlation, and $L = L_{\text{ccc}}(m) + L_{\text{ccc}}(s)$ for a positive one [9]. In AVEC’16 the average correlation between $m_t$ and $s_t$ is +0.103 (a positive correlation exists). In MSPConv the average correlation between $m_t$ and $s_t$ is 0.002 where no correlation exists. With these statistics, we note that the MTL PU is robust only in datasets where a correlation between $m_t$ and $s_t$ exists, and not robust in cases of complex datasets like the MSPConv.
TABLE I
COMPARISON ON MEAN \( \bar{m} \), STANDARD DEV. \( s \), AND LABEL DISTRIBUTION ESTIMATES \( Y \), IN TERMS OF \( L_{\text{ccc}}(m) \), \( L_{\text{ccc}}(s) \), AND \( L_{\text{KL}} \), RESPECTIVELY

|        | Arousal |               | \( L_{\text{ccc}}(m) \) \( \uparrow \) | \( L_{\text{ccc}}(s) \) \( \uparrow \) | \( L_{\text{KL}} \) \( \downarrow \) | Valence |               | \( L_{\text{ccc}}(m) \) \( \uparrow \) | \( L_{\text{ccc}}(s) \) \( \uparrow \) | \( L_{\text{KL}} \) \( \downarrow \) |
|--------|---------|---------------|--------------------------------|--------------------------------|--------------------------------|---------|---------------|--------------------------------|--------------------------------|--------------------------------|
| E2E Baseline w/o Temp | 0.581 | -             | 0.129                         | -                             | -                             | E2E Baseline [11] | 0.770 | -             | 0.561                         | -                             | -                             |
| STL [9] | 0.727 | -             | 0.389                         | -                             | -                             | MTL PU [9] | 0.740 | 0.310         | 0.776                         | 0.420**                       | 0.032                        | 0.960 |
| MU [38] | 0.762 | 0.077         | 0.675                         | 0.332                        | 0.040                         | 0.631 |
| MU+LU [38] | 0.751 | 0.361         | 0.250                         | 0.301                        | 0.048                         | 0.405 |
| \( t-LU \) (proposed) | 0.782** | 0.381**       | 0.228**                       | 0.400*                       | 0.050*                        | 0.386** |

(a) Quantitative results on AVEC’16 dataset.

(b) Quantitative results on MSPConv dataset.

Larger CCC indicates improved performance as indicated by \( \uparrow \). Lower KL indicates improved performance as indicated by \( \downarrow \). ** indicates that the respective approach achieves statistically significant better results than all other approaches in comparison. * indicates that it achieves statistically significant better results over only some of the approaches in comparison. Results in brackets (.) are for the respective development partition of the dataset.

Moreover, with the dataset dependent tuning of the loss, MTL PU is also not robust in cross-corpora evaluation (see Section V-D).

2) Comparison on Uncertainty Estimates: Table I shows that the proposed \( t-LU \) achieves the best uncertainty estimates across datasets, in terms of both \( L_{\text{ccc}}(s) \) and \( L_{\text{KL}} \). In AVEC’16, the improvements are statistically significant over all baselines in comparison. In MSPConv, the improvements are statistically significant over all baselines only with respect to the \( L_{\text{KL}} \) measure. In terms of the \( L_{\text{ccc}}(s) \) measure, improvements are not statistically significant over the MU + LU baseline alone. For instance, in AVEC’16, \( t-LU \) achieves 0.381 \( L_{\text{ccc}}(s) \) and 0.228 \( L_{\text{KL}} \), improving with statistical significance. In MSPConv, \( t-LU \) achieves 0.118 \( L_{\text{ccc}}(s) \) and 0.357 \( L_{\text{KL}} \), where statistical significance over all other baselines exists only for \( L_{\text{KL}} \). The reason for this trend is that, first, the MSPConv is more complex with larger levels of subjectivity (see Section IV-A). Second, the model is exclusively trained on \( L_{\text{KL}} \), so direct improvements over \( L_{\text{KL}} \) is expected rather than on \( L_{\text{ccc}}(s) \).

For valence in AVEC’16, unlike the \( L_{\text{ccc}}(m) \) performances, Table I a shows that the proposed \( t-LU \) achieves improved uncertainty estimates, in terms of both the measures \( L_{\text{ccc}}(s) \) and \( L_{\text{KL}} \). Moreover, the improvements are statistically significant over all other baselines in terms of the \( L_{\text{KL}} \) measure, but only over the MTL-based baselines in terms of the \( L_{\text{ccc}}(s) \) measure. Similar improvement trends can also be noted in the more complex MSPConv dataset (from Table I b). This improved uncertainty estimates of the proposed \( t-LU \) across datasets emphasizes the advantage of using the \( t \)-distribution based \( L_{\text{KL}} \) loss (16) for label uncertainty modeling. The \( t \)-distribution, as seen in Fig. 2, promotes the model to fit on a more relaxed \( s_t \), thereby more robust in capturing the whole label distribution. The fitting on a relaxed \( s_t \) leads to increased robustness towards outliers, as noted in [28].

B. Qualitative Analysis of Estimates

For qualitative analyses, we plot the mean \( \bar{m}_t \) and standard deviation \( \tilde{s}_t \) estimates of \( \bar{Y}_t \) against the \( m_t \) and standard deviation \( s_t \) of ground-truth distribution \( Y_t \). Plots for a test subject from AVEC’16, in terms of arousal and valence, can be seen in Fig. 5(a) and (b), respectively, and, for MSPConv, in Fig. 5(c) and (d), respectively. Parts of the plots are boxed and numbered to note clear performance differences.

For arousal, in Fig. 5(a) and (c), further backing the results in Table I, the proposed \( t-LU \) model best captures \( m_t \) and \( s_t \) of the annotation distribution \( Y_t \), in comparison with MU and MU + LU. For example, in AVEC’16 (see Fig. 5(a)), in boxes 2 and 3, \( t-LU \) best captures the whole distribution \( Y_t \), where \( \tilde{s}_t \) best resembles \( s_t \). This further highlights the robustness of training on a relaxed \( s_t \) through a \( t \)-distribution. Backing the quantitative results in Table I, improvements are more evident in MSPConv, noted from boxes 2 and 3 in Fig. 5(c). Crucially, along with the \( \tilde{s}_t \) improvements by \( t-LU \), notable improvements are also seen on mean estimates \( m_t \).

For valence, Fig. 5(a) and (d) show that the proposed \( t-LU \) evidently improves on mean estimates \( \bar{m}_t \) on both datasets, with only small improvements on standard deviation estimates \( \tilde{s}_t \). This can be seen for instance in box 1 of Fig. 5(b). Hence, capturing \( s_t \) in valence by only relying on audio is a challenging task, and more complex in datasets such as the MSPConv where some frames have a very high subjectivity (see log histograms.
Fig. 5. Label distribution estimation results for a test subject.

Fig. 6. Loss curve comparison between Gaussian $L_{KL}$ (12) and proposed $t$-distribution $L_{KL}$ (16). The comparisons can be seen in Fig. 6. It is a common trend in the literature that the audio modality insufficiently explains ground-truth valence $m_t$ [13], [60], and this trend is even more challenging for modeling $s_t$ in valence.

C. Analysis on Training Loss Curve

To further study the advantages of the proposed $t$-distribution $L_{KL}$ (16) during the training phase, we compare the testing loss curve of (16) with the Gaussian $L_{KL}$ in MU + LU (12). The comparisons can be seen in Fig. 6.

Fig. 6 illustrates two crucial advantages of the proposed $t$-distribution $L_{KL}$ loss term (16) during training in both datasets. First, we see that in the initial epochs, before epoch 20, the proposed loss converges quicker than the Gaussian $L_{KL}$ (12). This is the result of the proposed $L_{KL}$ (16) loss term which penalizes more for lower $s_t$ values, in comparison to the Gaussian $L_{KL}$ (12) (see Section III-C3), thereby achieving faster convergence. Second, during the later epochs, after epoch 70, the Gaussian $L_{KL}$ (12) shows signs of overfitting, which is more evident in the MSPConv dataset. However, at the same time, the proposed $t$-distribution $L_{KL}$ (16) converges to the best minima during the later epochs. For instance, in MSPConv, the proposed (16) achieves minima $L_{KL}$ at epoch 86, with $L_{KL}$ of 0.357 for arousal and 0.373 for valence, while the Gaussian achieves a minima well before the later epochs, at epoch 35, with $L_{KL}$ of 0.369 for arousal and 0.411 for valence. The proposed $L_{KL}$ (16) is free from overfitting in the later stages of
training and also learns the optima at this stage, noticed across two datasets. This behaviour can be attributed to the nature of the proposed $L_{KL}$ (16) which promotes the model to learn a more relaxed $s_t$, thereby introducing more regularization to the model, preventing overfitting and converging on an improved $s_t$.

**D. Cross-Corpora Evaluation**

To validate the robustness and generalisation capabilities of the models, we performed cross-corpora evaluations. In Fig. 7, results are presented in terms of $L_{ccc}(m)$ and $L_{KL}$, under two conditions. The Match condition where the train and the test partitions come from the same dataset, and the Unmatch condition where the train partition is from a different dataset. Apart from the dataset size, other dataset-specific factors, such as population demographics and social context, severely challenge the cross-corpora performances because human behaviour varies across group-sizes [36], [61] and social contexts [32]. Crucial differences exist between the AVEC’16 and MSPConv datasets. While the social context of AVEC’16 is a dyadic interaction in a virtual setting, MSPConv comprises of larger groups in a face-to-face setting. Moreover, AVEC’16 was collected from French-speaking persons, while MSPConv from English-speaking persons.

Fig. 7 illustrates that the proposed $t$-LU achieves the best cross-corpora performances on both datasets, and MU with the second best performances. Under the Unmatch condition, for arousal in AVEC’16 (see Fig. 7(a)), $t$-LU achieves 0.421 $L_{ccc}(m)$ and 0.409 $L_{KL}$, while MU achieves 0.342 and 0.490, respectively. Similarly, in MSPConv (see Fig. 7(c)), $t$-LU achieves 0.260 $L_{ccc}(m)$ and 0.600 $L_{KL}$, while MU achieves 0.216 and 0.655, respectively.

All models degrade in performance from the Match to Unmatch conditions. For both arousal and valence, across datasets and metrics, $t$-LU achieves the least degrade percentage while the MTL PU results in the highest degrade. For instance, in AVEC’16, in terms of arousal mean-estimates $L_{ccc}(m)$ (see Fig. 7(a)), $t$-LU achieves the least degradation of 46% and MTL PU degrades the most with 61%. Similarly, for valence (see Fig. 7(b)), $t$-LU degrades least with 53%, and MTL PU degrades the most with 62%. This further emphasises on the robustness of the proposed $t$-LU model and clearly highlights the lack of robustness of the MTL PU baseline. The MTL PU which achieves the best $L_{ccc}(m)$ for valence on the AVEC’16 (see Table I(a), degrades the most on cross-corpora evaluations. This drawback of the MTL PU baseline stems from the dataset-dependent tuning of loss function that it relies on. The proposed $t$-LU is free from such dataset-dependent tuning and hence more robust. The degrade percentage in $L_{KL}$ is not comparable as the scale of the measure is not linear (depicted in Fig. 2). Also notable is that, for all models, the degrade percentage is larger for valence than for arousal.

**E. Impact of Number of Annotations Available**

In Section V-A, we noted the benefits of modeling $Y_t$ as a $t$-distribution, with six available annotations. To capitalise on the benefits of the $t$-distribution $t$-LU over the Gaussian MU + LU, especially when fewer annotations are available, we performed experiments by varying $a$ and thereby the degrees of freedom $v$. The results are presented in Fig. 8, under 4 settings, $a = 3$, $a = 4$, $a = 5$, and, $a = 6$. Annotations were ignored to achieve conditions of $a \leq 5$. The order of annotation to be ignored was handled based on Pearson’s correlations measure. For instance, under setting $a = 4$, annotations from two annotators, with the
least inter-annotator correlation, for the whole audio file were ignored to model ground-truth annotation distribution $\mathcal{Y}_t$.

Fig. 8 shows that, especially when $a \geq 4$, the t-distribution based t-LU shows superior performance over the Gaussian MU + LU on both datasets. Crucially, the improvements are larger and more evident when $a = 4$ and $a = 5$ than when $a = 6$. In the case of $\mathcal{L}_{\text{CCC}}(m_t)$, a non-monotonic behavior with the available number of annotations is notable; $\mathcal{L}_{\text{CCC}}(m_t)$ initially increases from $a = 6$ to $a = 5$ and subsequently decreases with reducing annotations $a \leq 4$. The initial increase is noticed as annotations are ignored in the order of reducing Pearson’s correlation, hence we can expect better consensus in $m_t$ for $a = 5$ than when $a = 6$. The subsequent decrease can be associated with the reduced number of annotations to model a stable distribution $\mathcal{Y}_t$. This emphasises the advantage of t-distribution over the Gaussian with increasing inter-annotator correlation and reducing number of available annotations. In the case of $a = 3$, the performance of t-LU drops below that of the Gaussian MU + LU, as t-LU becomes highly uncertain with only 3 annotations because of the large relaxation on $s_t$ introduced by the scaling in (7) (see Supplementary Section 2 for theoretical analysis, available online). This behaviour is similar to the t-test calculation, where models become more uncertain with reducing $\nu$. For modeling emotion annotations as a distribution and uncertainty modeling, we therefore recommend the t-distribution over the Gaussian when more than 3 annotations are available. Noting that both t-distribution and Gaussian drop in performances with only 3 annotations, we suggest collecting at least 4 annotations to obtain a reliable annotation distribution and its ground-truth consensus.

F. Ablation Study

The proposed end-to-end label uncertainty model has three essential modules, namely 1) feature extractor, 2) uncertainty layer, and 3) label uncertainty loss. To understand the modules’ specific contributions, we perform an ablation study and present its results in Table II. In case of the feature extractor, E2E, $\checkmark$ denotes usage of an end-to-end feature extractor and $\times$ the hand-crafted features [20], [62]. In case of the uncertainty layer, BBB, $\checkmark$ denotes the usage of the BBB-based uncertainty layer and $\times$ the MTL-based $s_t$ estimator. For label uncertainty loss, KL, $\checkmark$ denotes using $\mathcal{L}_{KL}$ loss and $\times$ denotes usage of $\mathcal{L}_{\text{CCC}}(s)$ loss.

Table II first shows that end-to-end models achieve better uncertainty estimates than hand-crafted feature models. For instance, in AVEC’16, E2E based BBB-KL model achieves 0.381 $\mathcal{L}_{\text{CCC}}(s)$ and 0.228 $\mathcal{L}_{KL}$, improving over hand-crafted features based BBB-KL model which achieves 0.330 and 0.276, respectively. Similarly, in the larger and more complex MSPConv, the E2E based BBB-KL model achieves the best uncertainty estimate performances, against all other models in comparison, with 0.118 $\mathcal{L}_{\text{CCC}}(s)$ and 0.357 $\mathcal{L}_{KL}$. This trend is inline with literature that suggests end-to-end learning, that learns emotion representations in a data-driven manner, for uncertainty modeling [14]. Second, the combination of BBB-based uncertainty layer and KL-based loss term (BBB + KL) results in improved performances in both mean and uncertainty estimates, recommending the combination of BBB-layer and KL-loss for label uncertainty modeling in SER. The performance of BBB-layer with a $\mathcal{L}_{\text{CCC}}(s)$ loss term degrades performance across metrics. An intuition behind this is that KL-based distribution loss is apt for optimizing the weight distributions $P(w|D)$, rather than a loss with only optimizes for $s_t$ of label distribution. Third, across datasets, for both arousal and valence, the KL-based loss term contributes to the improvement of both uncertainty and mean estimates, as the KL loss jointly optimizes for $m_t$ and $s_t$. For instance, in terms of arousal, the inclusion of KL loss to the E2E+BBB architecture results in a 5% improvement on mean estimates $\mathcal{L}_{\text{CCC}}(m)$ in AVEC’16 and 26% in MSPConv.
TABLE II
ABSTRACTION RESULTS OF THE T-LU MODEL, ON THE AVEC’16 [37] AND MSP-CONVERSATION [31] DATASETS

| Modules | Arousal | Valence |
|---------|---------|---------|
|         | L_{cc}(m) ↑ | L_{cc}(s) ↑ | L_{KL} ↓ | L_{cc}(m) ↑ | L_{cc}(s) ↑ | L_{KL} ↓ |
| E2E     | ✓ ✓ ✓    | 0.782* | 0.381* | 0.225* | 0.400 | 0.050 | 0.390* |
|         | ✓ ✓ x    | 0.743 | 0.356 | 0.412 | 0.329 | 0.054 | 0.594 |
|         | ✓ x ✓    | 0.704 | 0.315 | 0.299 | 0.373 | 0.039 | 0.426 |
|         | ✓ x x    | 0.772 | 0.395* | 0.512 | 0.401 | 0.064 | 0.863 |
| AVEC’16 | x ✓ ✓    | 0.772* | 0.330 | 0.276* | 0.366 | 0.033 | 0.413 |
|         | x x ✓    | 0.798 | 0.329 | 0.446 | 0.330 | 0.050 | 0.601 |
|         | x x x    | 0.716 | 0.329 | 0.318 | 0.381 | 0.039 | 0.446 |
|         | ✓ x x    | 0.740 | 0.310 | 0.776 | 0.420* | 0.032 | 0.960 |
|         | ✓ ✓ x    | 0.786* | 0.118* | 0.357* | 0.213* | 0.032 | 0.373* |
|         | ✓ ✓ x    | 0.266 | 0.097 | 0.412 | 0.180 | 0.029 | 0.495 |
|         | ✓ x x    | 0.163 | 0.051 | 0.515 | 0.122 | 0.009 | 0.537 |
|         | ✓ x x    | 0.271 | 0.100 | 0.489 | 0.174 | 0.012 | 0.593 |
|         | x ✓ ✓    | 0.401* | 0.056 | 0.392* | 0.230* | 0.017 | 0.391* |
|         | x x x    | 0.308 | 0.078 | 0.551 | 0.181 | 0.026 | 0.416 |
|         | x x x    | 0.247 | 0.040 | 0.490 | 0.140 | 0.005 | 0.549 |
|         | x x x    | 0.296 | 0.107 | 0.527 | 0.181 | 0.030 | 0.560 |

At the same time, improvements on uncertainty estimates are also noted, 7% improvement of $L_{cc}(s)$ in AVEC’16 and 18% in MSPConv.

Finally, MTL-based $s_t$ estimating model achieves the best $L_{cc}(m)$ performance for valence, but only in AVEC’16 (see last row in Table II). However, in MPSConv, the proposed BBB+KL based models achieve better results. This improvement, noted only for valence in the AVEC’16, again stems from the dataset-dependent tuning of the loss that is required by MTL-based $s_t$ estimating models (see Section V-A1). However, this tuning also results in MTL-based $s_t$ estimating models losing their robustness and generalisation capabilities, as shown in cross-corpora evaluations (see Section V-D). Moreover, the MTL-based uncertainty models collapse when not trained on $L_{cc}(s)$ loss, and are not capable of distribution learning using the $L_{KL}$ loss. Overall, these trends suggest that BBB-based $Y_t$ learning models are to be preferred over MTL-based $s_t$ estimating models for label uncertainty modeling in SER.

VI. CONCLUSION

We introduced an end-to-end BNN capable of modeling emotion annotations as a label distribution, thereby accounting for the inherent subjectivity-based label uncertainty. In the literature, emotion annotations are commonly modeled using a Gaussian distribution or a histogram representation, however with assumptions based on only limited annotations. In contrast, in this work, we modeled ground-truth emotion annotations as a Student’s $t$-distribution, which also accounts for the number of annotations available. Specifically, we derived a $t$-distribution based KL divergence loss that, for limited and sparse annotations, produces robust mean estimates and standard deviation estimates that well capture the outliers. We showed that the proposed $t$-distribution loss term leads to training on a relaxed standard deviation, which is adaptable with respect to the number of annotations available. We validated our approach on two publicly available in-the-wild datasets. Quantitative analysis of the results showed that our proposed approach achieved state-of-the-art results for mean and uncertainty estimations, in terms of both CCC and KL divergence measures, which were also consistent for cross-corpora evaluations. By analysing the loss curves, we showed that the proposed loss term yields faster and improved convergence, and is less prone to overfitting than the Gaussian loss term. Our results further revealed that the advantage of $t$-distribution over the Gaussian grows with increasing inter-annotator correlation and decreasing numbers of available annotations. Finally, our ablation study suggests that, for modeling label uncertainty in SER, BBB-based label distribution learning models are to be preferred over estimating standard deviation as an auxiliary task.

A. Limitations and Future Avenues

In our work, we modeled the emotion annotations as a label distribution using a BNN. However, the BNN introduced here, both MU+LU and $t$-LU, jointly captures the two types of uncertainty—model and label uncertainty. In future work, it would be interesting to focus on disentangling the two types of uncertainty for reliable label uncertainty aware SER systems. One possible way to achieve this concerns Prior Networks (PNs) [63], a variant of BNNs, which could be employed to exclusively capture the label uncertainty. PNs do this by parameterizing a prior distribution over predictive label distributions.

This work specifically focused on modeling emotion annotations in a time- and value-continuous manner. In future work, the proposed methodology can be directly extended to model emotion annotations at the utterance-level, as opposed to time-continuous annotations, by simply adding a pooling layer to the feature extractor. However, the method cannot be directly

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
extended to modeling discrete emotion annotations (e.g., classification tasks). Note that the model architecture introduced here (Fig. 1) can be modified to classify discrete emotion labels, but the introduced label uncertainty loss (16) operates only on value-continuous annotations samples. To further extend the introduced loss function for classification tasks, future work may focus on the discrete variant of t-distributions. In that case, similar to the loss function derivation in Section III-C2, KL divergence loss for discrete t-distributions would need to be derived.

While the proposed model achieved significantly better state-of-the-art performances in terms of the arousal dimension of emotion across datasets, in one of the datasets (AVEC’16) it did not achieve state-of-the-art performance in terms of valence. Note however that state-of-the-art valence performance was achieved in the more complicated MSPConv dataset. It is well documented in the literature that the audio modality insufficiently explains the valence dimension of emotions [55]. This is likely also the reason why the best performing t-LU model, in terms of valence in Tables I a, I b, and II, did not achieve statistical significance in some of the metrics despite its improved performance. To overcome this drawback, future work could also include the video and semantic modalities in the feature extractor module, thereby achieving multimodality.

REFERENCES

[1] G. A. Van Kleef, “How emotions regulate social life: The emotions as social information (EASI) model,” Curr. Directions Psychol. Sci., vol. 18, no. 3, pp. 184–188, 2009.

[2] B. W. Schuller, “Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends,” Commun. ACM, vol. 61, no. 5, pp. 90–99, Apr. 2018.

[3] J. A. Russell, “A circumplex model of affect,” J. Pers. Soc. Psychol., vol. 39, no. 6, 1980, Art. no. 1161.

[4] D. Dukes et al., “The rise of affectivism,” Nature Hum. Behav., vol. 5, no. 7, pp. 816–820, 2021.

[5] K. Srithar, W.-C. Lin, and C. Busso, “Generative approach using soft-labels to learn uncertainty in predicting emotional attributes,” in Proc. 9th Int. Conf. Affect. Comput. Intell. Interact., Oct. 2021, pp. 1–8.

[6] C. Busso et al., “IEMOCAP: Interactive emotional dyadic motion capture database,” Lang. Resour. Eval., vol. 42, no. 4, pp. 335–359, 2008.

[7] M. Grimm and K. Kroschel, “Evaluation of natural emotions using self-assessment manikins,” in Proc. Workshop Autom. Speech Recognit. Understanding, 2005, pp. 381–385.

[8] H. Gunes and B. Schuller, “Categorical and dimensional affect analysis in continuous input: Current trends and future directions,” Image Vis. Comput., vol. 31, pp. 120–136, 2013.

[9] J. Han, Z. Zhang, Z. Ren, and B. Schuller, “Exploring perception uncertainty for emotion recognition in dyadic conversation and music listening,” Cogn. Comput., vol. 13, pp. 231–240, Mar. 2021.

[10] K. Srithar and C. Busso, “Modeling uncertainty in predicting emotional attributes from spontaneous speech,” in Proc. Int. Conf. Acoust., Speech Signal Process., Barcelona, Spain, May 2020, pp. 8384–8388.

[11] P. Tzirakis, J. Zhang, and B. W. Schuller, “End-to-end speech emotion recognition using deep neural networks,” in Proc. Int. Conf. Acoust., Speech, Signal Process., Calgary, Canada, Apr. 2018.

[12] J. Huang, Y. Li, J. Tao, Z. Lian, and J. Yi, “End-to-end continuous emotion recognition from video using 3D ConvLSTM networks,” in Proc. Int. Conf. Acoust., Speech, Signal Process., Apr. 2018, pp. 6837–6841.

[13] P. Tzirakis, J. Chen, S. Vafeiadis, and B. Schuller, “End-to-end multimodal affect recognition in real-world environments,” Inf. Fusion, vol. 68, pp. 46–53, 2021.

[14] S. Alisamir and F. Ringeval, “On the evolution of speech representations for affective computing: A brief history and critical overview,” IEEE Signal Process. Mag., vol. 38, pp. 12–21, 2021.

[15] A. Kendall and Y. Gal, “What uncertainties do we need in Bayesian deep learning for computer vision?,” in Proc. Int. Conf. Neural Inf. Process. Syst., Dec. 2017, pp. 5574–5584.

[16] R. Zheng, S. Zhang, L. Liu, Y. Luo, and M. Sun, “Uncertainty in Bayesian deep label distribution learning,” Appl. Soft Comput., vol. 101, Mar. 2021, Art. no. 107046.

[17] M. K. Tellamekala, T. Giesbrecht, and M. Valstar, “Dimensional affect uncertainty modeling for apparent personality recognition,” IEEE Trans. Affect. Comput., vol. 13, no. 4, pp. 2144–2155, Fourth Quarter 2022.

[18] J. Liu, J. Paisley, M.-A. Kioumourtzoglou, and B. Couill, “Accurate uncertainty estimation and decomposition in ensemble learning,” in Proc. Int. Conf. Neural Inf. Process. Syst., Vancouver, Dec. 2019, pp. 8950–8961.

[19] S. Kohl et al., “A probabilistic U-Net for segmentation of ambiguous images,” in Proc. Int. Conf. Neural Inf. Process. Syst., Montreal, Canada, Dec. 2018, pp. 6965–6975.

[20] M. K. Tellamekala, E. Sanchez, G. Trimitropoulos, T. Giesbrecht, and M. Valstar, “Stochastic process regression for cross-cultural speech emotion recognition,” in Proc. 22nd Annul. Int. Conf. Speech Commun. Assoc., Brno, Sep. 2021, pp. 3390–3394.

[21] M. Garcelo et al., “Conditional neural processes,” in Proc. Int. Conf. Mach. Learn., Stockholm, Sweden, Jul. 2018, pp. 1690–1699.

[22] Y. Gal and Z. Ghahramani, “Dropout as a Bayesian approximation: Representing model uncertainty in deep learning,” in Proc. Int. Conf. Mach. Learn., New York City, NY, USA, Jun. 2016, pp. 1050–1059.

[23] C. Blundell, J. Cornebee, K. Kavukcuoglu, and D. Wierstra, “Weight uncertainty in neural networks,” in Proc. Int. Conf. Mach. Learn., Lille, France, Jul. 2015, pp. 1613–1622.

[24] H. Fang, T. Peer, S. Wernter, and T. Gerkmann, “Integrating statistical uncertainty into neural network-based speech enhancement,” in Proc. Int. Conf. Acoust., Speech, Signal Process., Singapore, Jan. 2022, pp. 386–390.

[25] B.-B. Gao, C. Xing, C.-W. Xie, J. Wu, and X. Geng, “Deep label distribution learning with level ambiguity,” IEEE Trans. Image Process., vol. 26, no. 6, pp. 2825–2838, Jun. 2017.

[26] H.-C. Chou, W.-C. Lin, C.-C. Lee, and C. Busso, “Exploiting annotators’ typed description of emotion perception to maximize utilization of ratings for speech emotion recognition,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., Singapore, Jan. 2022, pp. 7717–7721.

[27] N. M. Fotinopoulos, C. Tzelepis, and I. Patras, “Estimating continuous affect with label uncertainty,” in Proc. IEEE Int. Conf. Affect. Comput. Intell. Interact., Virtual Event, Oct. 2021, pp. 1–8.

[28] C. M. Bishop, Pattern Recognition and Machine Learning, Berlin, Germany: Springer, 2006.

[29] S. Kotz and S. Nadarajah, Multivariate t-Distributions and Their Applications. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[30] C. Villa and F. J. Rubio, “Objective priors for the number of degrees of freedom of a multivariate t distribution and the t-copula,” Comput. Statist. Data Anal., vol. 124, pp. 197–219, 2018.

[31] L. Martinez-Lucas, M. Abdelwahab, and C. Busso, “The MSP-conversation corpus,” in Proc. Annu. Conf. Int. Speech Commun. Assoc., Shanghai, China, Oct. 2020, pp. 1823–1827.

[32] R. Lofman and C. Busso, “Building naturalistic emotionally balanced speech corpora by retrieving emotional speech from existing podcast recordings,” IEEE Trans. Autom. Control, vol. 10, no. 4, pp. 471–483, Fourth Quarter 2019.

[33] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, “Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions,” in Proc. IEEE Int. Conf. Workshops Autom. Face Gesture Recognit., Shanghai, China, Apr. 2015, pp. 1–8.

[34] J. Kossaifi et al., “SEWA DB: A rich database for audio-visual emotion and sentiment research in the wild,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 3, pp. 1022–1040, 2019.

[35] N. Raj Prabhu, C. Raman, and H. Hung, “Defining and quantifying conversation quality in spontaneous interactions,” in Proc. Companion Work. Conf. Int. Conf. Multimodal Interact., Sep. 2020, pp. 196–205.

[36] D. Gatica-Perez, “Automatic nonverbal analysis of social interaction in small groups: A review,” in Proc. Companion Work. Conf. Int. Conf. Multimodal Interact., Shanghai, China, Oct. 2020, pp. 1823–1827.

[37] N. Raj Prabhu, G. Caruajal, N. Lehmann-Willebrock, and T. Gerkmann, “End-to-end label uncertainty modeling for speech-based arousal recognition using Bayesian neural networks,” in Proc. 23rd Annul. Int. Conf. Speech Commun. Assoc., Incheon, Korea, Sep. 2022, pp. 151–155.
[39] N. Raj Prabhu, N. Lehmann-Willenbrock, and T. Gerkmann, “Label uncertainty modeling and prediction for speech emotion recognition using t-distributions,” in Proc. IEEE Int. Conf. Affect. Comput. Intell. Interact., Nara, Japan, Oct. 2022, pp. 1–8.

[40] M. Abdelwahab and C. Busso, “Active learning for speech emotion recognition using deep neural network,” in Proc. IEEE Int. Conf. Affect. Comput. Intell. Interact., Cambridge, U.K., Sep. 2019, pp. 1–7.

[41] I. Lawrence and K. Lin, “A concordance correlation coefficient to evaluate reproducibility,” Biometrics, vol. 45, no. 1, pp. 255–268, 1989.

[42] S. Steidl, M. Levit, A. Battiner, E. Noth, and H. Niemann, “Of all things the measure is man” automatic classification of emotions and inter-labeler consistency,” in Proc. Int. Conf. Acoust., Speech, Signal Process., Philadelphia, USA, 2005, pp. 317–320.

[43] H. M. Fayek, M. Lech, and L. Cavedon, “Modeling subjectiveness in emotion recognition with deep neural networks: Ensembles vs soft labels,” in IEEE Int., Joint Conf., Neural Netw., Vancouver, Canada, Jul. 2016, pp. 566–570.

[44] L. Tarantino, P. N. Garner, and A. Lazaridis, “Self-attention for speech emotion recognition,” in Proc. 20th Annu. Conf. Int. Speech Commun. Assoc., Graz, Sep. 2019, pp. 2578–2582.

[45] J. Han, Z. Zhang, M. Schmitt, M. Pantic, and B. Schuller, “From hard to soft: Towards more human-like emotion recognition by modelling the perception uncertainty,” in Proc., 25th ACM Int. Conf. Multimedia, Mountain View, USA, Oct. 2017, pp. 890–897.

[46] T. Dang, V. Sethu, and E. Ambikairajah, “Dynamic multi-rater gaussian mixture regression incorporating temporal dependencies of emotion uncertainty using kalman filters,” in Proc. Int. Conf. Acoust., Speech, Signal Process., Calgary, Canada, Apr. 2018, pp. 4929–4933.

[47] G. Rizos and B. Schuller, “Modelling sample informativeness for deep affective computing,” in Proc. Int. Conf. Acoust., Speech, Signal Process., Brighton, U.K., May 2019, pp. 3482–3486.

[48] R. E. Walpole, R. H. Myers, S. L. Myers, and K. Ye, Probability & Statistics for Engineers and Scientists, London, U.K.: Pearson Education, 2007.

[49] C. Villa and S. G. Walker, “Objective prior for the number of degrees of freedom of at distribution,” Bayesian Anal., vol. 9, no. 1, pp. 197–220, 2014.

[50] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, 3rd ed., vol. 4. Cambridge, MA, USA: MIT, Jul. 2016, pp. 51–77.

[51] S. Kullback and R. A. Leibler, “On information and sufficiency,” Ann. Math. Statist., vol. 22, no. 1, pp. 79–86, 1951.

[52] K. P. Murphy, Machine Learning: A Probabilistic Perspective. Cambridge, MA, USA: MIT, 2012.

[53] A. Paszke et al., “PyTorch: An imperative style, high-performance deep learning library,” in Proc. Int. Conf. Neural Inf. Process. Syst., Vancouver, Dec. 2019, pp. 8026–8037.

[54] D. Rees, “Essential statistics,” in American Statistician, vol. 55, London, U.K.: Chapman & Hall, 2001.

[55] P. Tzirakis, A. Nguyen, S. Zafeiriou, and B. W. Schuller, “Speech emotion recognition using semantic information,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., Toronto, Jun. 2021, pp. 6279–6283.

[56] S. Butterworth et al., “On the theory of filter amplifiers,” Wireless Engineer, vol. 7, no. 6, pp. 536–541, 1930.

[57] J. W. Cooley and J. W. Tukey, “An algorithm for the machine calculation of complex fourier series,” Math. Comput., vol. 19, no. 90, pp. 297–301, 1965.

[58] R. Gupta, K. Audhkhasi, Z. Jacokes, A. Rozga, and S. Narayanan, “Modeling multiple time series annotations as noisy distortions of the ground truth: An expectation-maximization approach,” IEEE Trans. Autom. Control, vol. 9, no. 1, pp. 76–89, First Quarter 2018.

[59] S. Marioyyad and C. Busso, “Correcting time-continuous emotional labels by modeling the reaction lag of evaluators,” IEEE Trans. Autom. Control, vol. 6, no. 2, pp. 97–108, Second Quarter 2014.

[60] K. Sridhar and C. Busso, “Unsupervised personalization of an emotion recognition system: The unique properties of the externalization of valence in speech,” IEEE Trans. Affect. Comput., vol. 13, no. 4, pp. 1959–1972, Fourth Quarter 2022.

[61] C. Ramana, N. R. Prabhu, and H. Hung, “Perceived conversation quality in spontaneous interactions,” IEEE Trans. Autom. Control, to be published, doi: 10.1109/TAFFC.2023.3233950.

[62] F. Eyben et al., “The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing,” IEEE Trans. Autom. Control, vol. 7, no. 2, pp. 190–202, Second Quarter 2016.

[63] A. Malinin and M. Gales, “Predictive uncertainty estimation via prior networks,” in Proc. Int. Conf. Neural Inf. Process. Syst., 2018, pp. 7047–7058.

Navel Raj Prabhu (Student Member, IEEE) received the BTech degree in computer science from SRM University, India, in 2015, and the MS degree in computer science from the Intelligent Systems Department, Delft University of Technology, Delft, The Netherlands, in 2020. Currently, he is working toward the PhD degree with the Signal Processing Lab and Organisation Psychology Lab, University of Hamburg, Hamburg, Germany. His research interests include affective computing, social signal processing, deep learning, uncertainty modelling, speech signal processing, and group affect.

Nale Lehmann-Willenbrock received the PhD degree in psychology from Technische Universität Braunschweig, in 2012. She studied psychology with the University of Goettingen and University of California, Irvine. After several years working as an Assistant Professor with Vrije Universiteit Amsterdam and associate professor with the University of Amsterdam, she joined Universität Hamburg as a full professor and chair of Industrial/Organizational Psychology, in 2018, where she also leads the Center for Better Work. She studies emergent behavioral patterns in organizational teams, social dynamics among leaders and followers, and meetings with the core of organizations. Her research program blends organizational psychology, management, communication, and social signal processing. She serves as associate editor for the Journal of Business and Psychology as well as for Small Group Research.

Timo Gerkmann (Senior Member, IEEE) received the Dipl.-Ing and Dr.-Ing degrees in electrical engineering and information sciences from the Ruhr-Universität Bochum, Bochum, Germany, in 2004 and 2010, respectively. He studied electrical engineering and information sciences with the Universität Bremen and the Ruhr-Universität Bochum in Germany. In 2005, he spent six months with Siemens Corporate Research in Princeton, NJ, USA. During 2010 to 2011 he was a postdoctoral researcher with the Sound and Image Processing Lab, Royal Institute of Technology (KTH), Stockholm, Sweden. From 2011 to 2015 he was a professor for speech signal processing with the Universität Oldenburg, Oldenburg, Germany. During 2015 to 2016 he was a principal scientist for audio & acoustics with Technicolor Research & Innovation in Hanover, Germany. Since 2016 he is a professor for signal processing with the Universität Hamburg, Germany. His main research interests are on statistical signal processing and machine learning for speech and audio applied to communication devices, hearing instruments, audio–visual media, and human-machine interfaces. Timo Gerkmann serves as an elected member of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing and as an associate editor of the IEEE/ACM Transactions on Audio, Speech and Language Processing. He received the VDE IFG award 2022.