Estimating the Uncertainty in Emotion Attributes using Deep Evidential Regression

Anonymous ACL submission

Abstract

In automatic emotion recognition (AER), labels assigned by different human annotators to the same utterance are often inconsistent due to the inherent complexity of emotion and the subjectivity of perception. Though deterministic labels generated by averaging or voting are often used as the ground truth, it ignores the intrinsic uncertainty revealed by the inconsistent labels. This paper proposes a Bayesian approach, deep evidential emotion regression (DEER), to estimating the uncertainty in emotion attributes. Treating the emotion attribute labels of an utterance as samples drawn from an unknown Gaussian distribution, DEER places an utterance-specific normal-inverse gamma prior over the Gaussian likelihood and predicts its hyper-parameters using a deep neural network model. It enables a joint estimation of emotion attributes along with the aleatoric and epistemic uncertainties. AER experiments on the widely used MSP-Podcast and IEMOCAP datasets showed DEER produced state-of-the-art results for both the mean values and the distribution of emotion attributes[1].

1 Introduction

Automatic emotion recognition (AER) is the task that enables computers to predict human emotional states based on multimodal signals, such as audio, video and text. An emotional state is defined based on either categorical or dimensional theory. The categorical theory claims the existence of a small number of basic discrete emotions (*i.e.* anger and happy) that are inherent in our brain and universally recognised (Gunes et al., 2011; Plutchik, 2001). Dimensional emotion theory characterises emotional states by a small number of roughly orthogonal fundamental continuous-valued bipolar dimensions (Schlosberg, 1954; Nicolaou et al., 2011) such as valence-arousal and approach–avoidance (Russell and Mehrabian, 1977; Russell, 1980; Grimm et al., 2007). These dimensions are also known as emotion attributes, which allow us to model more subtle and complex emotions and are thus more common in psychological studies. As a result, AER includes a classification approach based on emotion-class-based labels and a regression approach based on attribute-based labels. This paper focuses on attribute-based AER with speech input.

Emotion annotation is challenging due to the inherent ambiguity of mixed emotion, the personal variations in emotion expression, the subjectivity in emotion perception, *etc.* Most AER datasets use multiple human annotators to label each utterance, which often results in inconsistent labels, either as emotion categories or attributes. This is also a typical manifestation of the intrinsic data uncertainty, also referred to as aleatoric uncertainty (Matthies, 2007; Der Kiureghian and Ditlevsen, 2009), that arises from the natural complexity of emotion data. It is common to replace such inconsistent labels with deterministic labels obtained by majority voting (Busso et al., 2008, 2017) or (weighted) averages (Ringeval et al., 2013; Lotfian and Busso, 2019; Kossaifi et al., 2019; Grimm and Kroschel, 2005). However, this causes a loss of data samples when a majority agreed emotion class doesn't exist (Majumder et al., 2018; Poria et al., 2018; Wu et al., 2021) and also ignores the discrepancies between annotators and the aleatoric uncertainty in emotion data.

In this paper, we propose to model the uncertainty in emotion attributes with a Bayesian approach based on deep evidential regression (Amini et al., 2020), denoted deep evidential emotion regression (DEER). In DEER, the inconsistent human labels of each utterance are considered as observations drawn independently from an unknown Gaussian distribution. To probabilistically estimate the mean and variance of the Gaussian distribution, a normal inverse-gamma (NIG) prior is introduced,

[1] Code will be publicly available upon acceptance.

which places a Gaussian prior over the mean and an inverse-gamma prior over the variance. The AER system is trained to predict the hyper-parameters of the NIG prior for each utterance by maximising the per-observation-based marginal likelihood of each observed label under this prior. As a result, DEER not only models the distribution of emotion attributes but also learns both the aleatoric uncertainty and the epistemic uncertainty (Der Kiureghian and Ditlevsen, 2009) without repeating the inference procedure for sampling. Epistemic uncertainty, also known as model uncertainty, is associated with uncertainty in model parameters that best explain the observed data. Aleatoric and epistemic uncertainty are combined to induce the total uncertainty, also called predictive uncertainty, that measures the confidence of attribute predictions. As a further improvement, a novel regulariser is proposed based on the mean and variance of the observed labels to better calibrate the uncertainty estimation. The proposed methods were evaluated on the MSP-Podcast and IEMOCAP datasets.

The rest of the paper is organised as follows. Section 2 summarises related work. Section 3 introduces the proposed DEER approach. Sections 4 and 5 present the experimental setup and results respectively, followed by the conclusion.

2 Related Work

There has been previous work by AER researchers to address the issue of inconsistent labels. For emotion categories, a single ground-truth label can be obtained as either a continuous-valued mean vector representing emotion intensities (Fayek et al., 2016; Ando et al., 2018), or as a multi-hot vector obtained based on the existence of emotions (Zhang et al., 2020; Ju et al., 2020). Recently, distribution-based approaches have been proposed, which consider the labels as samples drawn from emotion distributions (Chou et al., 2022; Wu et al., 2022b).

For emotion attributes, annotators often assign different values to the same attribute of each utterance. Davani et al. (2022) proposed a multi-annotator model which contains multiple heads to predict each annotator's judgement. This approach is computationally viable only when the number of annotators is relatively small. The method requires sufficient annotations from each annotator to be effective. Deng et al. (2012) derived confidence measures based on annotator agreement to build emotion-scoring models. Han et al. (2017, 2021) proposed predicting the standard deviation of the attribute label values as an extra task in the multi-task training framework. Dang et al. (2017, 2018) included annotator variability as a representation of uncertainty in a Gaussian mixture regression model. These techniques take the variance of human annotations either as an extra target or as an extra input. More recently, Bayesian deep learning has been introduced to the task, which models the uncertainty in emotion annotation without explicitly using the variance of human annotations. These include the use of Gaussian processes (Atcheson et al., 2018, 2019), variational auto-encoders (Sridhar et al., 2021), Bayesian neural networks (Prabhu et al., 2021), Monte-Carlo dropout (Sridhar and Busso, 2020b) and sequential Monte-Carlo methods (Markov et al., 2015; Wu et al., 2022a).

So far, these methods have not distinguished aleatoric uncertainty from epistemic uncertainty which are defined in the introduction. Our proposed DEER approach can simultaneously model these two uncertainties. In addition, our approach is more generic. It has no limits on the number of annotators, the number of annotators per utterance, and the number of annotations per annotator, and thus can cope with large crowd-sourced datasets.

3 Deep Evidential Emotion Regression

3.1 Problem setup

In contrast to Bayesian neural networks that place priors on model parameters (Blundell et al., 2015; Kendall and Gal, 2017), evidential deep learning (Sensoy et al., 2018; Malinin and Gales, 2018; Amini et al., 2020) places priors over the likelihood function. Every training sample adds support to a learned higher-order prior distribution called the evidential distribution. Sampling from this distribution gives instances of lower-order likelihood functions from which the data was drawn.

Consider an input utterance $\mathbf{x}$ with M emotion attribute labels $y^{(1)}, \ldots, y^{(M)}$ provided by multiple annotators. Assuming $y^{(1)}, \ldots, y^{(M)}$ are observations drawn *i.i.d.* from a Gaussian distribution with unknown mean μ and unknown variance σ^2, where μ is drawn from a Gaussian prior and σ^2 is drawn from an inverse-gamma prior:

$$y^{(1)}, \ldots, y^{(M)} \sim \mathcal{N}(\mu, \sigma^2)$$

$$\mu \sim \mathcal{N}(\gamma, \sigma^2 \upsilon^{-1}), \quad \sigma^2 \sim \Gamma^{-1}(\alpha, \beta)$$

where $\gamma \in \mathbb{R}$, $\upsilon > 0$, and $\Gamma(\cdot)$ is the gamma function with $\alpha > 1$ and $\beta > 0$.

Denote $\{\mu, \sigma^2\}$ and $\{\gamma, \upsilon, \alpha, \beta\}$ as $\mathbf{\Psi}$ and $\mathbf{\Omega}$. The posterior $\mathrm{p}(\mathbf{\Psi}|\mathbf{\Omega})$ is a NIG distribution, which is the Gaussian conjugate prior:

$$\begin{aligned}\mathrm{p}(\mathbf{\Psi}|\mathbf{\Omega}) &= \mathrm{p}(\mu|\sigma^2, \mathbf{\Omega})\, \mathrm{p}(\sigma^2|\mathbf{\Omega}) \\ &= \mathcal{N}(\gamma, \sigma^2\upsilon^{-1})\, \Gamma^{-1}(\alpha, \beta) \\ &= \frac{\beta^\alpha \sqrt{\upsilon}}{\Gamma(\alpha)\sqrt{2\pi\sigma^2}} \left(\frac{1}{\sigma^2}\right)^{\alpha+1} \\ &\cdot \exp\left\{-\frac{2\beta + \upsilon(\gamma-\mu)^2}{2\sigma^2}\right\}\end{aligned}$$

Drawing a sample $\mathbf{\Psi}_i$ from the NIG distribution yields a single instance of the likelihood function $\mathcal{N}(\mu_i, \sigma_i^2)$. The NIG distribution therefore serves as the higher-order, evidential distribution on top of the unknown lower-order likelihood distribution from which the observations are drawn. The NIG hyper-parameters $\mathbf{\Omega}$ determine not only the location but also the uncertainty, associated with the inferred likelihood function.

By training a deep neural network model to output the hyper-parameters of the evidential distribution, evidential deep learning allows the uncertainties to be found by analytic computation of the maximum likelihood Gaussian without the need for repeated inference for sampling (Amini et al., 2020). Furthermore, it also allows an effective estimate of the aleatoric uncertainty computed as the expectation of the variance of the Gaussian distribution, as well as the epistemic uncertainty defined as the variance of the predicted Gaussian mean. Given an NIG distribution, the prediction, aleatoric, and epistemic uncertainty can be computed as:

$$\text{Prediction: } \mathbb{E}[\mu] = \gamma$$

$$\text{Aleatoric: } \mathbb{E}[\sigma^2] = \frac{\beta}{\alpha - 1}, \quad \forall\, \alpha > 1$$

$$\text{Epistemic: } \mathrm{Var}[\mu] = \frac{\beta}{\upsilon(\alpha - 1)}, \quad \forall\, \alpha > 1$$

3.2 Training

The training of DEER is structured as fitting the model to the data while enforcing the prior to calibrate the uncertainty when the prediction is wrong.

3.2.1 Maximising the data fit

The likelihood of an observation y given the evidential distribution hyper-parameters $\mathbf{\Omega}$ is computed by marginalising over the likelihood parameters $\mathbf{\Psi}$:

$$\begin{aligned}\mathrm{p}(y|\mathbf{\Omega}) &= \int_{\mathbf{\Psi}} \mathrm{p}(y|\mathbf{\Psi})\mathrm{p}(\mathbf{\Psi}|\mathbf{\Omega})\ \mathrm{d}\mathbf{\Psi} \\ &= \mathbb{E}_{\mathrm{p}(\mathbf{\Psi}|\mathbf{\Omega})}[\mathrm{p}(y|\mathbf{\Psi})]\end{aligned} \tag{1}$$

An analytical solution exists in the case of placing an NIG prior on the Gaussian likelihood function:

$$\begin{aligned}\mathrm{p}(y|\mathbf{\Omega}) &= \frac{\Gamma(1/2+\alpha)}{\Gamma(\alpha)} \sqrt{\frac{\upsilon}{\pi}} \left(2\beta(1+\upsilon)\right)^\alpha \\ &\cdot \left(\upsilon(y-\gamma)^2 + 2\beta(1+\upsilon)\right)^{-(\frac{1}{2}+\alpha)} \\ &= \mathrm{St}_{2\alpha}\left(y|\gamma, \frac{\beta(1+\upsilon)}{\upsilon\,\alpha}\right)\end{aligned} \tag{2}$$

where $\mathrm{St}_\nu\,(t|r,s)$ is the Student's t-distribution evaluated at t with location parameter r, scale parameter s, and ν degrees of freedom. The predicted mean and variance can be computed analytically as

$$\mathbb{E}[y] = \gamma, \quad \mathrm{Var}[y] = \frac{\beta(1+\upsilon)}{\upsilon(\alpha-1)} \tag{3}$$

$\mathrm{Var}[y]$ represents the total uncertainty of model prediction, which is equal to the summation of the aleatoric uncertainty $\mathbb{E}[\sigma^2]$ and epistemic uncertainty $\mathrm{Var}[\mu]$ according to the law of total variance:

$$\begin{aligned}\mathrm{Var}[y] &= \mathbb{E}[\mathrm{Var}[y|\Psi]] + \mathrm{Var}[\mathbb{E}[y|\Psi]] \\ &= \mathbb{E}[\sigma^2] + \mathrm{Var}[\mu]\end{aligned}$$

To fit the NIG distribution, the model is trained by maximising the sum of the marginal likelihoods of each human label $y^{(m)}$. The negative log likelihood (NLL) loss can be computed as

$$\mathcal{L}^{\mathrm{NLL}}(\mathbf{x}; \mathbf{\Theta}) = -\frac{1}{M}\sum_{m=1}^{M} \log \mathrm{p}(y^{(m)}|\mathbf{\Omega}) \tag{4}$$

$$= -\frac{1}{M}\sum_{m=1}^{M} \log\left[\mathrm{St}_{2\alpha}\left(y^{(m)}|\gamma, \frac{\beta(1+\upsilon)}{\upsilon\,\alpha}\right)\right]$$

This is our proposed per-observation-based NLL loss, which takes each observed label into consideration for AER. This loss serves as the first part of the objective function for training a deep neural network model $\mathbf{\Theta}$ to predict the hyper-parameters $\{\gamma, \upsilon, \alpha, \beta\}$ to fit all observed labels of $\mathbf{x}$.

3.2.2 Calibrating the uncertainty on errors

The second part of the objective function regularises training by calibrating the uncertainty based on the incorrect predictions. A novel regulariser is formulated which contains two terms: $\mathcal{L}^\mu$ and $\mathcal{L}^\sigma$ that respectively regularises the errors on the estimation of the mean μ and the variance σ^2 of the Gaussian likelihood.

The first term $\mathcal{L}^\mu$ is proportional to the error between the model prediction and the average of the observations:

$$\mathcal{L}^\mu(\mathbf{x}; \mathbf{\Theta}) = \Phi\,|\bar{y} - \mathbb{E}[\mu]|$$

where $|\cdot|$ is L1 norm, $\bar{y} = \frac{1}{M}\sum_{m=1}^{M} y^{(m)}$ is the averaged label which is usually used as the ground truth in regression-based AER, and Φ is an uncertainty measure associated with the inferred posterior. The reciprocal of the total uncertainty is used as Φ in this paper, which can be calculated as

$$\Phi = \frac{1}{\text{Var}[y]} = \frac{\upsilon(\alpha-1)}{\beta(1+\upsilon)}$$

The regulariser imposes a penalty when there's an error in prediction and dynamically scales it by dividing by the total uncertainty of inferred posterior. It penalises the cases where the model produces an incorrect prediction with a small uncertainty, thus preventing the model from being over-confident. For instance, if the model produces an error with a small predicted variance, Φ is large, resulting in a large penalty. Minimising the regularisation term enforces the model to produce accurate prediction or increase uncertainty when the error is large.

In addition to imposing a penalty on the mean prediction as in (Amini et al., 2020), a second term $\mathcal{L}^{\sigma}$ is proposed in order to calibrate the estimation of the aleatoric uncertainty. As discussed in the introduction, aleatoric uncertainty in AER is shown by the different emotional labels given to the same utterance by different human annotators. This paper uses the variance of the observations to describe the aleatoric uncertainty in the emotion data. The second regularising term is defined as:

$$\mathcal{L}^{\sigma}(\mathbf{x};\boldsymbol{\Theta}) = \Phi\,|\bar{\sigma}^2 - \mathbb{E}[\sigma^2]|$$

where $\bar{\sigma}^2 = \frac{1}{M}\sum_{m=1}^{M}(y^{(m)} - \bar{y})^2$.

3.3 Summary and implementation details

For an AER task that consists of N emotion attributes, DEER trains a deep neural network model to simultaneously predict the hyper-parameters $\{\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_N\}$ associated with the N attribute-specific NIG distributions, where $\boldsymbol{\Omega}_n = \{\gamma_n, \upsilon_n, \alpha_n, \beta_n\}$. A DEER model thus has $4N$ output units. The system is trained by minimising the total loss *w.r.t.* $\boldsymbol{\Theta}$ as:

$$\mathcal{L}_{\text{total}}(\boldsymbol{x};\boldsymbol{\Theta}) = \sum_{n=1}^{N} \epsilon_n \mathcal{L}_n(\boldsymbol{x};\boldsymbol{\Theta}) \quad (5)$$

$$\mathcal{L}_n(\boldsymbol{x};\boldsymbol{\Theta}) = \mathcal{L}_n^{\text{NLL}}(\boldsymbol{x};\boldsymbol{\Theta}) + \lambda_n\left[\mathcal{L}_n^{\mu}(\boldsymbol{x};\boldsymbol{\Theta}) + \mathcal{L}_n^{\sigma}(\boldsymbol{x};\boldsymbol{\Theta})\right] \quad (6)$$

where ϵ_n is the weight satisfying $\sum_{n=1}^{N}\epsilon_n = 1$, λ_n is the scale coefficient that trades off the training between data fit and uncertainty regulation.

At test-time, the predictive posteriors are N separate Student's t-distributions $\text{p}(y|\boldsymbol{\Omega}_1), \text{p}(y|\boldsymbol{\Omega}_2), \ldots, \text{p}(y|\boldsymbol{\Omega}_N)$, each of the same form as derived in Eqn. (2)[2]. Apart from obtaining a distribution over the emotion attribute of the speaker, DEER also allows analytic computation of the uncertainty terms, as summarised in Table 1.

Term	Expression
Predicted mean	$\mathbb{E}[y] = \mathbb{E}[\mu] = \gamma$
Predicted variance (Total uncertainty)	$\text{Var}[y] = \frac{\beta(1+\upsilon)}{\upsilon(\alpha-1)}$
Aleatoric uncertainty	$\mathbb{E}[\sigma^2] = \frac{\beta}{\alpha-1}$
Epistemic uncertainty	$\text{Var}[\mu] = \frac{\beta}{\upsilon(\alpha-1)}$

Table 1: Summary of the uncertainty terms.

4 Experimental Setup

4.1 Dataset

The MSP-Podcast (Lotfian and Busso, 2019) and IEMOCAP datasets (Busso et al., 2008) were used in this paper. The annotations of both datasets use $N = 3$ with valence, arousal (also called activation), and dominance as the emotion attributes. MSP-Podcast contains natural English speech from podcast recordings and is one of the largest publicly available datasets in speech emotion recognition. A seven-point Likert scale was used to evaluate valence (1-negative vs 7-positive), arousal (1-calm vs 7-active), and dominance (1-weak vs 7-strong). The corpus was annotated using crowd-sourcing. Each utterance was labelled by at least 5 human annotators and has an average of 6.7 annotations per utterance. Ground-truth labels were defined by the average value. Release 1.8 was used in the experiments, which contains 73,042 utterances from 1,285 speakers amounting to more than 110 hours of speech. The average variance of the labels assigned to each sentence is 0.975, 1.122, 0.889 for valence, arousal, and dominance respectively. The standard splits for training (44,879 segments), validation (7,800 segments) and testing (15,326 segments) were used in the experiments.

The IEMOCAP corpus is one of the most widely used AER datasets. It consists of approximately 12 hours of English speech including 5 dyadic conversational sessions performed by 10 professional

[2]Since NIG is the Gaussian conjugate prior, the posterior is in the same parametric family as the prior. Therefore, the predictive posterior has the same form as the marginal likelihood. Detailed derivations see Appendix A.

actors with a session being a conversation between two speakers. There are in total 151 dialogues including 10,039 utterances. Each utterance was annotated by three human annotators using a five-point Likert scale. Again, ground-truth labels were determined by taking the average. The average variance of the labels assigned to each sentence is 0.130, 0.225, 0.300 for valence, arousal, and dominance respectively. Unless otherwise mentioned, systems on IEMOCAP were evaluated by training on Session 1-4 and testing on Session 5.

4.2 Model structure

The model structure used in this paper follows the upstream-downstream framework (Yang et al., 2021), as illustrated in Figure 1. WavLM (Chen et al., 2022) was used as the upstream model, which is a speech foundation model pre-trained by self-supervised learning. The BASE+ version[3] of the model was used in this paper which has 12 Transformer encoder blocks with 768-dimensional hidden states and 8 attention heads. The parameters of the pre-trained model were frozen and the weighted sum of the outputs of the 12 Transformer encoder blocks was used as the speech embeddings and fed into the downstream model.

The downstream model consists of two 128-dimensional Transformer encoder blocks with 4-head self-attention, followed by an evidential layer that contains four output units for each of the three attributes, which has a total of 12 output units. The model contains 0.3M trainable parameters. A Softplus activaton[4] was applied to $\{\upsilon, \alpha, \beta\}$ to ensure $\upsilon, \alpha, \beta > 0$ with an additional +1 added to α to ensure $\alpha > 1$. A linear activation was used for $\gamma \in \mathbb{R}$. The proposed DEER model was trained to simultaneously learn three evidential distributions for the three attributes. The weights in Eqn. (5) were set as $\epsilon_v = \epsilon_a = \epsilon_d = 1/3$. The scale coefficients were set to $\lambda_v = \lambda_a = \lambda_d = 0.1$ for Eqn. (6)[5].

A dropout rate of 0.3 was applied to the transformer parameters. The system was implemented using PyTorch and the SpeechBrain toolkit (Ravanelli et al., 2021). The Adam optimizer was used with an initial learning rate set to 0.001. Training took $\sim$ 8 hours on an NVIDIA A100 GPU.

[3]https://huggingface.co/microsoft/wavlm-base-plus

[4]$\text{Softplus}(x) = \ln(1 + \exp(x))$

[5]The values were manually selected from a small number of candidates.

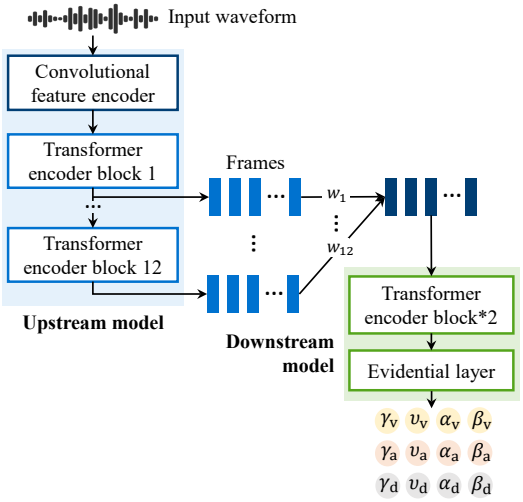


Figure 1: Illustration of the model structure. Weights $w_1, \ldots, w_{12}$ for the weighted sum of the 12 Transformer encoder outputs are trainable and satisfy $\sum_{i=1}^{12} w_i = 1$.

4.3 Evaluation metrics

4.3.1 Mean prediction

Following prior work in continuous emotion recognition (Ringeval et al., 2015, 2017; Sridhar and Busso, 2020a; Leem et al., 2022), the concordance correlation coefficient (CCC) was used to evaluate the predicted mean. CCC combines the Pearson's correlation coefficient with the square difference between the mean of the two compared sequences:

$$\rho_{\text{ccc}} = \frac{2\rho\,\sigma_{\text{ref}}\sigma_{\text{hyp}}}{\sigma_{\text{ref}}^2 + \sigma_{\text{hyp}}^2 + \left(\mu_{\text{ref}} - \mu_{\text{hyp}}\right)^2},$$

where ρ is the Pearson correlation coefficient between a hypothesis sequence (system predictions) and a reference sequence, where μ_{hyp} and μ_{ref} are the mean values, and σ_{hyp}^2 and σ_{ref}^2 are the variance values of the two sequences. Hypotheses that are well correlated with the reference but shifted in value are penalised in proportion to the deviation. The value of CCC ranges from -1 (perfect disagreement) to 1 (perfect agreement).

The root mean square error (RMSE) averaged over the test set is also reported. Since the average of the human labels, $\bar{y}$, is defined as the ground truth in both datasets, $\bar{y}$ were used as the reference in computing the CCC and RMSE. However, using $\bar{y}$ also indicates that these metrics are less informative when the aleatoric uncertainty is large.

4.3.2 Uncertainty estimation

It is common to use NLL to measure the uncertainty estimation ability (Gal and Ghahramani, 2016; Amini et al., 2020). NLL is computed by fitting data to the predictive posterior $\text{q}(y)$.

MSP-Podcast	CCC ↑			RMSE ↓			NLL(avg) ↓			NLL(all) ↓		
	v	**a**	**d**	**v**	**a**	**d**	**v**	**a**	**d**	**v**	**a**	**d**
$\mathcal{L}$ in Eqn. (6)	0.506	0.698	0.613	0.772	0.680	0.576	1.334	1.285	1.156	1.696	1.692	1.577
$\mathcal{L}^{\sigma}=0$	0.451	0.687	0.607	0.784	0.679	0.580	1.345	1.277	1.159	1.706	1.705	1.586
$\mathcal{L}^{\text{NLL}}=\bar{\mathcal{L}}^{\text{NLL}}$	0.473	0.682	0.609	0.808	0.673	0.566	1.290	1.060	0.899	2.027	2.089	1.969
IEMOCAP	**v**	**a**	**d**	**v**	**a**	**d**	**v**	**a**	**d**	**v**	**a**	**d**
$\mathcal{L}$ in Eqn. (6)	0.596	0.755	0.569	0.755	0.457	0.638	1.070	0.795	1.035	1.275	1.053	1.283
$\mathcal{L}^{\sigma}=0$	0.582	0.752	0.553	0.772	0.466	0.655	1.180	0.773	1.061	1.408	1.069	1.294
$\mathcal{L}^{\text{NLL}}=\bar{\mathcal{L}}^{\text{NLL}}$	0.585	0.759	0.555	0.786	0.444	0.633	1.001	0.727	1.036	1.627	1.329	1.441

Table 2: DEER results variations of the loss in Eqn. (6). 'v' , 'a', 'd' stands for valence, arousal, dominance. '↑' denotes the higher the better, '↓' denotes the lower the better. The '$\mathcal{L}$ in Eqn. (6)' row systems used the complete total loss of DEER. The '$\mathcal{L}^{\sigma}=0$' row systems had no $\mathcal{L}^{\sigma}$ regularisation term in the total loss. The '$\mathcal{L}^{\text{NLL}}=\bar{\mathcal{L}}^{\text{NLL}}$' row systems replaced the individual human labels with $\bar{\mathcal{L}}^{\text{NLL}}$ in the total loss.

In this paper, NLL(avg) defined as $-\log \mathrm{q}(\bar{y})$ and NLL(all) defined as $-\frac{1}{M}\sum_{m=1}^{M}\log \mathrm{q}(y^{(m)})$ are both used. NLL(avg) measures how much the averaged label $\bar{y}$ fits into the predicted posterior distribution, and NLL(all) measures how much how much every single human label $y^{(m)}$ fits into the predicted posterior. A lower NLL indicates better uncertainty estimation.

5 Experiments and Results

5.1 Effect of the aleatoric regulariser $\mathcal{L}^{\sigma}$

First, by setting $\mathcal{L}^{\sigma}=0$ in the total loss, an ablation study of the effect of the proposed extra regularising term $\mathcal{L}^{\sigma}$ is performed. The results are given in the '$\mathcal{L}^{\sigma}=0$' rows in Table 2. In this case, only $\mathcal{L}^{\mu}$ is used to regularise $\mathcal{L}^{\text{NLL}}$ and the results are compared to those trained using the complete loss defined in Eqn. (6), which are shown in the '$\mathcal{L}$ in Eqn. (6)' rows. From the results, $\mathcal{L}^{\sigma}$ improves the performance in CCC and NLL(all), but not in NLL(ref), as expected.

5.2 Effect of the per-observation-based $\mathcal{L}^{\text{NLL}}$

Next, the effect of our proposed per-observation-based NLL loss defined in Eqn. (4), $\mathcal{L}^{\text{NLL}}$, is compared to an alternative. Instead of using $\mathcal{L}^{\text{NLL}}$,

$$\bar{\mathcal{L}}^{\text{NLL}} = -\log \mathrm{p}(\bar{y}|\boldsymbol{\Omega})$$

is used to compute the total loss during training, and the results are given in the '$\mathcal{L}^{\text{NLL}}=\bar{\mathcal{L}}^{\text{NLL}}$' rows in Table 2. While $\mathcal{L}^{\text{NLL}}$ considers the likelihood of fitting each individual observation into the predicted posterior, $\bar{\mathcal{L}}^{\text{NLL}}$ only considers the averaged observation. Therefore, it is expected that using $\bar{\mathcal{L}}^{\text{NLL}}$ instead of $\mathcal{L}^{\text{NLL}}$ yields a smaller NLL(avg) but larger NLL(all), which have been validated by the results in the table.

5.3 Baseline comparisons

Three baseline systems were built:

- A Gaussian Process (GP) with a radial basis function kernel, trained by maximising the per-observation-based marginal likelihood.
- A Monte Carlo dropout (MCdp) system with a dropout rate of 0.4. During inference, the system was forwarded 50 times with different dropout random seeds to obtain 50 samples.
- An ensemble of 10 systems initialised and trained with 10 different random seeds.

The MCdp and ensemble baselines used the same model structure as the DEER system, expect that the evidential output layer was replaced by a standard fully-connected output layer with three output units to predict the values of valence, arousal and dominance respectively. Following prior work (Al-Badawy and Kim, 2018; Atmaja and Akagi, 2020b; Sridhar and Busso, 2020b), the CCC loss,

$$\mathcal{L}_{\text{ccc}} = 1 - \rho_{\text{ccc}}$$

was used for training the MCdp and ensemble baselines. The CCC loss was computed based on the sequence within each mini-batch of training data. The CCC loss has been shown by previous studies to improve the continuous emotion predictions compared to the RMSE loss (Povolny et al., 2016; Trigeorgis et al., 2016; Le et al., 2017). For MCdp and ensemble, the predicted distribution of the emotion attributes were estimated based on the obtained samples by kernel density estimation.

The results are listed in Table 3. The proposed DEER system outperforms the baselines on most of the attributes and the overall values. In particular, DEER outperforms all baselines consistently in the NLL(all) metric.

MSP-Podcast	CCC ↑			RMSE ↓			NLL_ref ↓			NLL_all ↓		
	v	a	d	v	a	d	v	a	d	v	a	d
DEER	0.506	**0.698**	**0.613**	**0.772**	0.680	**0.576**	**1.334**	**1.285**	1.156	**1.696**	**1.692**	**1.577**
GP	0.342	0.595	0.486	0.811	**0.673**	0.566	1.447	1.408	1.297	1.727	1.808	1.592
MCdp	0.476	0.667	0.594	0.874	0.702	0.623	1.680	1.300	**1.071**	2.050	2.027	1.776
Ensemble	**0.511**	0.679	0.608	0.855	0.692	0.615	1.864	1.384	1.112	2.096	2.066	1.795
IEMOCAP	**v**	**a**	**d**	**v**	**a**	**d**	**v**	**a**	**d**	**v**	**a**	**d**
DEER	**0.596**	**0.756**	**0.569**	**0.755**	**0.457**	**0.638**	**1.070**	0.795	**1.035**	**1.275**	**1.053**	**1.283**
GP	0.535	0.717	0.512	0.763	0.479	0.657	1.209	**0.791**	1.047	1.295	1.205	1.380
MCdp	0.539	0.724	0.568	0.786	0.561	0.702	1.291	0.849	1.133	1.549	1.325	1.747
Ensemble	0.580	0.754	0.560	0.778	0.476	0.686	1.296	0.864	1.110	1.584	1.218	1.749

Table 3: Comparison with the baselines. 'v', 'a', 'd' stands for valence, arousal, dominance. '↑' denotes the higher the better, '↓' denotes the lower the better. Best results in each column shown in bold.

5.4 Cross comparison of mean prediction

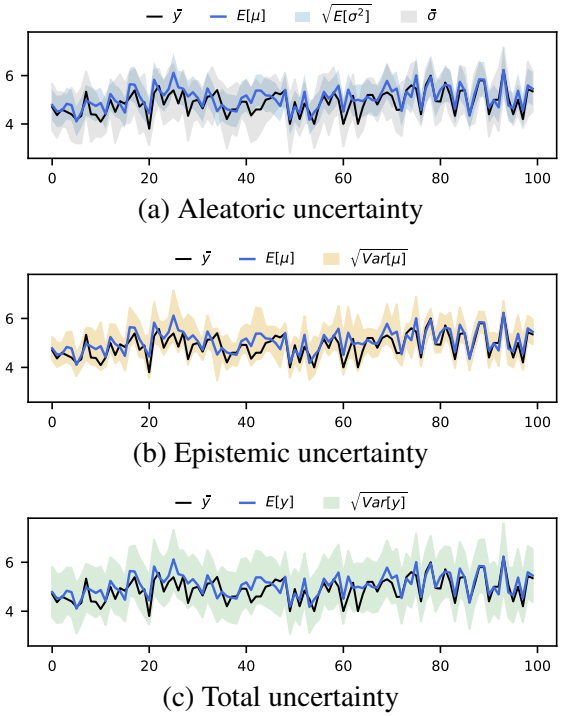
Table 4 compares results obtained with those previously published in terms of the CCC value. Previous papers have reported results on both version 1.6 and 1.8 of the MSP-Podcast dataset. For comparison, we also conducted experiments on version 1.6 for comparison. Version 1.6 of MSP-Podcast database is a subset of version 1.8 and contains 34,280 segments for training, 5,958 segments for validation and 10,124 segments for testing. For IEMOCAP, apart from training on Session 1-4 and testing on Session 5 (Ses05), we also evaluated the proposed system by a 5-fold cross-validation (5CV) based on a "leave-one-session-out" strategy. In each fold, one session was left out for testing and the others were used for training. The configuration is speaker-exclusive for both settings. As shown in Table 4, our DEER systems achieved state-of-the-art (SOTA) results both versions of MSP-Podcast and both test setting of IEMOCAP.

5.5 Analysis of uncertainty estimation

5.5.1 Visualisation

Based on a randomly selected subset test set of MSP-Podcast version 1.8, the aleatoric, epistemic and total uncertainty of the dominance attribute predicted by our proposed DEER system are shown in Figure 2.

Figure 2 (a) shows the predicted mean ± square root of the predicted aleatoric uncertainty ($\mathbb{E}[\mu] \pm \sqrt{\mathbb{E}[\sigma^2]}$) and the average label ± the standard deviation of the human labels ($\bar{y} \pm \bar{\sigma}$). It can be seen that the predicted aleatoric uncertainty (blue) overlaps with the label standard deviation (grey) and the overlapping is more evident when the mean predictions are accurate (*i.e.* samples around index 80-100).



(a) Aleatoric uncertainty

(b) Epistemic uncertainty

(c) Total uncertainty

Figure 2: Visualisation of (a) aleatoric (b) epistemic (c) total uncertainty of dominance for MSP-Podcast. x-asix is the test utterance index.

Figure 2 (b) shows the predicted mean ± square root of the predicted epistemic uncertainty ($\mathbb{E}[\mu] \pm \sqrt{\mathrm{Var}[\mu]}$). The epistemic uncertainty is high when the predicted mean deviates from the target (*i.e.* samples around index 40-50) while low then the predicted mean matches the target (*i.e.* samples around index 80-100).

Figure 2 (c) shows the predicted mean ± square root of the total epistemic uncertainty ($\mathbb{E}[y] \pm \sqrt{\mathrm{Var}[y]}$) which combines the aleatoric and epistemic uncertainty. The total uncertainty is high

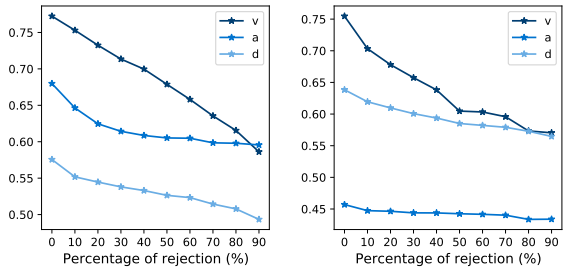
	Paper	Version	v	a	d	Average
MSP-podcast	Ghriss et al. (2022)	1.6	0.412	0.679	0.564	0.552
	Mitra et al. (2022)	1.6	0.57	0.75	0.67	0.663
	Srinivasan et al. (2022)	1.6	0.627	0.757	0.671	0.685
	DEER	1.6	**0.629**	**0.777**	**0.684**	**0.697**
	Leem et al. (2022)	1.8	0.212	0.572	0.505	0.430
	DEER	1.8	**0.506**	**0.698**	**0.613**	**0.606**
	Paper	**Setting**	**v**	**a**	**d**	**Average**
IEMOCAP	Atmaja and Akagi (2020a)	Ses05	0.421	0.590	0.484	0.498
	Atmaja and Akagi (2021)	Ses05	0.553	0.579	0.465	0.532
	DEER	Ses05	**0.596**	**0.756**	**0.569**	**0.640**
	Srinivasan et al. (2022)	5CV	0.582	0.667	0.545	0.598
	DEER	5CV	**0.625**	**0.720**	**0.548**	**0.631**

Table 4: Cross comparison of the CCC value on MSP-Podcast and IEMOCAP. 'v', 'a', 'd' stands for valence, arousal, dominance. 'Version' of MSP-Podcast denotes the release version of the dataset., and only the results from the same dateset version are comparable. 'Test set' of IEMOCAP denotes the train/set split. 'Ses05' denotes training on Session 1-4 and tested on Session 5. '5CV' denotes leave-one-session-out 5-fold cross validation.

either when the input utterance is complex or the model is not confident.

5.5.2 Reject option

A reject option was applied to analyse the uncertainty estimation performance, where the system has the option to accept or decline a test sample based on the uncertainty prediction. Since the evaluation of CCC is based on the whole sequence rather than individual samples, its computation would be affected when the sequence is modified by rejection (Wu et al., 2022a). Therefore, the reject option is performed based on RMSE.



(a) MSP-Podcast (b) IEMOCAP

Figure 3: Reject Option of RMSE based on predicted variance for (a) MSP-Podcast and (b) IEMOCAP.

Confidence is measured by the total uncertainty given in Eqn. (3). Figure 3 shows the performance of the proposed DEER system with a reject option on MSP-Podcast and IEMOCAP. A percentage of utterances with the largest predicted variance were rejected. The results at 0% rejection corresponds to the RMSE achieved on the entire test data. As the percentage of rejection increases, test coverage decreases and the average RMSE decreases showing the predicted variance succeeded in confidence estimation. The system then trades off between the test coverage and performance.

6 Conclusions

Two types of uncertainty exist in AER: (i) aleatoric uncertainty arising from the inherent ambiguity of emotion and personal variations in emotion expression; (ii) epistemic uncertainty associated with the estimated network parameters given the observed data. This paper proposes DEER for estimating those uncertainties in emotion attributes. Treating observed attribute-based annotations as samples drawn from a Gaussian distribution, DEER places a normal-inverse gamma (NIG) prior over the Gaussian likelihood. A novel training loss is proposed which combines a per-observation-based NLL loss with a regulariser on both the mean and the variance of the Gaussian likelihood. Experiments on the MSP-Podcast and IEMOCAP datasets show that DEER can produce SOTA results in estimating both the mean value and the distribution of emotion attributes. The use of NIG, the conjugate prior to the Gaussian distribution, leads to tractable analytic computation of the marginal likelihood as well as aleatoric and epistemic uncertainty associated with attribute prediction. Uncertainty estimation is analysed by visualisation and a reject option. Beyond the scope of AER, DEER could also be applied to other tasks with subjective evaluations yielding inconsistent labels.

Limitations

The proposed approach (along with other methods for estimating uncertainty in inconsistent annotations) is only viable when the raw labels from different human annotators for each sentence are provided by the datasets. However, some multiple-annotated datasets only released the majority vote or averaged label for each sentence (*i.e.* Poria et al., 2019).

The proposed method made a Gaussian assumption on the likelihood function for the analytic computation of the uncertainties. The results show that this modelling approach is effective. Despite the effectiveness of the proposed method, other distributions could also be considered.

Apart from aleatoric and epistemic uncertainty, distributional uncertainty can also occur when testing on unseen data. Distributional uncertainty (also called dataset shift) results from the mismatch between the training and test distributions. Aleatoric uncertainty is sometimes considered as "known-unknown" – the knowledge the model learnt from the training set can be well-generalised to the test data and the model knows how difficult it is to predict from the data. Distributional uncertainty is an "unknown-unknown" – the model is unfamiliar with the test data and cannot make predictions confidently. Unseen data is not discussed in this paper but can be an interesting extension to investigate and the DEER method should have the potential to address this issue.

Data collection processes for AER datasets vary in terms of recording conditions, emotional elicitation scheme, and annotation procedure *etc*. This work was tested on two typical datasets: IEMOCAP and MSP-Podcast. The two datasets are both publicly available and differ in various aspects:

- IEMOCAP contains emotion acted by professional actors while MSP-Podcast contains natural emotion.
- IEMOCAP contains dyadic conversations while MSP-Podcast contains Podcast recordings.
- IEMOCAP contains 10 speakers and MSP-Podcast contains 1285 speakers.
- IEMOCAP contains about 12 hours of speech and MSP-Podcast contains more than 110 hours of speech.
- IEMOCAP was annotated by six professional evaluators with each sentence being annotated by three evaluators. MSP-Podcast was annotated by crowd-sourcing where a total of 11,799 workers were involved and each work annotated 41.5 sentences on average.

The proposed approach has been shown effective over both datasets. We believe the proposed technique should be generic. Furthermore, although validated only for AER, the proposed method could also be applied to other tasks with disagreements in subjective annotations such as hate speech detection and language assessment.

Ethics Statement

For subjective tasks such as emotion recognition, it is common to employ multiple human annotators to give multiple annotations to each data instance. In face of annotator disagreements, majority voting and averaging are commonly used to derive single ground truth labels for training supervised machine learning models. However, in many subjective tasks, there is usually no single "correct" answer. And a potential risk of enforcing a single ground truth is ignoring the valuable nuance in each annotator's evaluation and their disagreements. This can cause minority views to be under-represented. The DEER approach proposed in this work could be beneficial to this concern as it models uncertainty in annotator disagreements and provides some explainability of the predictions.

While our method helps preserve minority perspectives, misuse of this technique might lead to ethical concerns. Emotion recognition is at risk of exposing a person's inner state to others and this information could be abused. Furthermore, since the proposed approach takes each annotation into consideration, it is important to protect the anonymity of annotators.

References

Ehab A AlBadawy and Yelin Kim. 2018. Joint discrete and continuous emotion prediction using ensemble and end-to-end approaches. In *Proc. ICMI*, Boulder.

Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. 2020. Deep evidential regression. In *Proc. NeurIPS*, Vancouver.

Atsushi Ando, Satoshi Kobashikawa, Hosana Kamiyama, Ryo Masumura, Yusuke Ijima, and Yushi Aono. 2018. Soft-target training with ambiguous emotional utterances for DNN-based speech emotion classification. In *Proc. ICASSP*, Brighton.

Mia Atcheson, Vidhyasaharan Sethu, and Julien Epps. 2018. Demonstrating and modelling systematic time-varying annotator disagreement in continuous emotion annotation. In *Proc. Interspeech*, Hyderabad.

Mia Atcheson, Vidhyasaharan Sethu, and Julien Epps. 2019. Using Gaussian processes with LSTM neural networks to predict continuous-time, dimensional emotion in ambiguous speech. In *Proc. ACII*, Cambridge.

Bagus Tris Atmaja and Masato Akagi. 2020a. Improving valence prediction in dimensional speech emotion recognition using linguistic information. In *Proc. O-COCOSDA*, Yangon.

Bagus Tris Atmaja and Masato Akagi. 2020b. Multitask learning and multistage fusion for dimensional audiovisual emotion recognition. In *Proc. ICASSP*, Conference held virtually.

Bagus Tris Atmaja and Masato Akagi. 2021. Two-stage dimensional emotion recognition by fusing predictions of acoustic and text networks using svm. *Speech Communication*, 126:9–21.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. Wav2Vec 2.0: A framework for self-supervised learning of speech representations. In *Proc. NeurIPS*, Conference held virtually.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. 2015. Weight uncertainty in neural network. In *Proc. ICML*, Lille.

C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E.M. Provost, S. Kim, J.N. Chang, S. Lee, and S.S. Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359.

Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohammed AbdelWahab, Najmeh Sadoughi, and Emily Mower Provost. 2017. MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing*, 8(1):67–80.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*.

Huang-Cheng Chou, Wei-Cheng Lin, Chi-Chun Lee, and Carlos Busso. 2022. Exploiting annotators' typed description of emotion perception to maximize utilization of ratings for speech emotion recognition. In *Proc. ICASSP*, Singapore.

Ting Dang, Vidhyasaharan Sethu, and Eliathamby Ambikairajah. 2018. Dynamic multi-rater Gaussian mixture regression incorporating temporal dependencies of emotion uncertainty using Kalman filters. In *Proc. ICASSP*, Calgary.

Ting Dang, Vidhyasaharan Sethu, Julien Epps, and Eliathamby Ambikairajah. 2017. An investigation of emotion prediction uncertainty using gaussian mixture regression. In *Proc. Interspeech*, Stockholm.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Jun Deng, Wenjing Han, and Björn Schuller. 2012. Confidence measures for speech emotion recognition: A start. In *Speech Communication; 10. ITG Symposium*, pages 1–4. VDE.

Armen Der Kiureghian and Ove Ditlevsen. 2009. Aleatory or epistemic? does it matter? *Structural Safety*, 31(2):105–112.

H.M. Fayek, M. Lech, and L. Cavedon. 2016. Modeling subjectiveness in emotion recognition with deep neural networks: Ensembles vs soft labels. In *Proc. IJCNN*, Vancouver.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proc. ICML*, New York City.

Ayoub Ghriss, Bo Yang, Viktor Rozgic, Elizabeth Shriberg, and Chao Wang. 2022. Sentiment-aware automatic speech recognition pre-training for enhanced speech emotion recognition. In *Proc. ICASSP*, Singapore.

Michael Grimm and Kristian Kroschel. 2005. Evaluation of natural emotions using self assessment manikins. In *Proc. ASRU*, Cancun.

Michael Grimm, Kristian Kroschel, Emily Mower, and Shrikanth Narayanan. 2007. Primitives-based evaluation and estimation of emotions in speech. *Speech Communication*, 49(10-11):787–800.

Hatice Gunes, Björn Schuller, Maja Pantic, and Roddy Cowie. 2011. Emotion representation, analysis and synthesis in continuous space: A survey. In *Proc. FG*, Santa Barbara.

Jing Han, Zixing Zhang, Zhao Ren, and Björn Schuller. 2021. Exploring perception uncertainty for emotion recognition in dyadic conversation and music listening. *Cognitive Computation*, 13(2):231–240.

Jing Han, Zixing Zhang, Maximilian Schmitt, Maja Pantic, and Björn Schuller. 2017. From hard to soft: Towards more human-like emotion recognition by modelling the perception uncertainty. In *Proc. ACM MM*, Mountain View.

Xincheng Ju, Dong Zhang, Junhui Li, and Guodong Zhou. 2020. Transformer-based label set generation for multi-modal multi-label emotion detection. In *Proc. ACM MM*, Seattle.

Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? In *Proc. NeurIPS*, Long Beach.

Jean Kossaifi, Robert Walecki, Yannis Panagakis, Jie Shen, Maximilian Schmitt, Fabien Ringeval, Jing Han, Vedhas Pandit, Antoine Toisoul, Björn Schuller, et al. 2019. SEWA DB: A rich database for audio-visual emotion and sentiment research in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(3):1022–1040.

Duc Le, Zakaria Aldeneh, and Emily Mower Provost. 2017. Discretized continuous speech emotion recognition with multi-task deep recurrent neural network. In *Proc. Interspeech*, Stockholm.

Seong-Gyun Leem, Daniel Fulford, Jukka-Pekka Onnela, David Gard, and Carlos Busso. 2022. Not all features are equal: Selection of robust features for speech emotion recognition in noisy environments. In *Proc. ICASSP*, Singapore.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

R. Lotfian and C. Busso. 2019. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 10(4):471–483.

N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria. 2018. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowledge-Based Systems*, 161:124–133.

Andrey Malinin and Mark Gales. 2018. Predictive uncertainty estimation via prior networks. In *Proc. NeurIPS*, Montréal.

Konstantin Markov, Tomoko Matsui, Francois Septier, and Gareth Peters. 2015. Dynamic speech emotion recognition with state-space models. In *Proc. EUSIPCO*, Nice.

Hermann G Matthies. 2007. Quantifying uncertainty: Modern computational representation of probability and applications. In *Extreme man-made and natural hazards in dynamics of structures*, pages 105–135. Springer.

Vikramjit Mitra, Hsiang-Yun Sherry Chien, Vasudha Kowtha, Joseph Yitan Cheng, and Erdrin Azemi. 2022. Speech Emotion: Investigating Model Representations, Multi-Task Learning and Knowledge Distillation. In *Proc. Interspeech 2022*, pages 4715–4719.

Mihalis A Nicolaou, Hatice Gunes, and Maja Pantic. 2011. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing*, 2(2):92–105.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *Proc. ICASSP*, South Brisbane.

Robert Plutchik. 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89(4):344–350.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proc. ACL*, Florence.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Erik Cambria, Alexander Gelbukh, and Amir Hussain. 2018. Multimodal sentiment analysis: Addressing key issues and setting up the baselines. *IEEE Intelligent Systems*, 33(6):17–25.

Filip Povolny, Pavel Matejka, Michal Hradis, Anna Popková, Lubomír Otrusina, Pavel Smrz, Ian Wood, Cecile Robin, and Lori Lamel. 2016. Multimodal emotion recognition for avec 2016 challenge. In *Proc. ACM MM*, Amsterdam.

Navin Raj Prabhu, Guillaume Carbajal, Nale Lehmann-Willenbrock, and Timo Gerkmann. 2021. End-to-end label uncertainty modeling for speech emotion recognition using bayesian neural networks. *arXiv preprint arXiv:2110.03299*.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. 2021. SpeechBrain: A general-purpose speech toolkit. ArXiv:2106.04624.

Fabien Ringeval, Björn Schuller, Michel Valstar, Roddy Cowie, and Maja Pantic. 2015. AVEC 2015: The 5th international audio/visual emotion challenge and workshop. In *Proc. ACM MM*, Brisbane.

Fabien Ringeval, Björn Schuller, Michel Valstar, Jonathan Gratch, Roddy Cowie, Stefan Scherer, Sharon Mozgai, Nicholas Cummins, Maximilian Schmitt, and Maja Pantic. 2017. AVEC 2017: Real-life depression, and affect recognition workshop and challenge. In *Proc. ACM MM*, Mountain View.

Fabien Ringeval, Andreas Sonderegger, Jürgen Sauer, and Denis Lalanne. 2013. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *Proc. FG*, Shanghai.

James A Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161.

James A Russell and Albert Mehrabian. 1977. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11(3):273–294.

Harold Schlosberg. 1954. Three dimensions of emotion. *Psychological Review*, 61(2):81.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. 2018. Evidential deep learning to quantify classification uncertainty. In *Proc. NeurIPS*, Montréal.

Kusha Sridhar and Carlos Busso. 2020a. Ensemble of students taught by probabilistic teachers to improve speech emotion recognition. In *Proc. Interspeech*, Shanghai.

Kusha Sridhar and Carlos Busso. 2020b. Modeling uncertainty in predicting emotional attributes from spontaneous speech. In *Proc. ICASSP*, Conference held virtually.

Kusha Sridhar, Wei-Cheng Lin, and Carlos Busso. 2021. Generative approach using soft-labels to learn uncertainty in predicting emotional attributes. In *Proc. ACII*, Chicago.

Sundararajan Srinivasan, Zhaocheng Huang, and Katrin Kirchhoff. 2022. Representation learning through cross-modal conditional teacher-student training for speech emotion recognition. In *Proc. ICASSP*, Singapore.

Andreas Triantafyllopoulos, Johannes Wagner, Hagen Wierstorf, Maximilian Schmitt, Uwe Reichel, Florian Eyben, Felix Burkhardt, and Björn W. Schuller. 2022. Probing speech emotion recognition transformers for linguistic knowledge. In *Proc. Interspeech 2022*, pages 146–150.

George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Björn Schuller, and Stefanos Zafeiriou. 2016. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *Proc. ICASSP*, Shanghai.

Jingyao Wu, Ting Dang, Vidhyasaharan Sethu, and Eliathamby Ambikairajah. 2022a. A novel sequential Monte Carlo framework for predicting ambiguous emotion states. In *Proc. ICASSP*, Singapore.

Wen Wu, Chao Zhang, and Philip C. Woodland. 2021. Emotion recognition by fusing time synchronous and time asynchronous representations. In *Proc. ICASSP*, Toronto.

Wen Wu, Chao Zhang, Xixin Wu, and Philip C Woodland. 2022b. Estimating the uncertainty in emotion class labels with utterance-specific dirichlet priors. *arXiv preprint arXiv:2203.04443*.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al. 2021. SUPERB: Speech processing universal performance benchmark. *arXiv preprint arXiv:2105.01051*.

Dong Zhang, Xincheng Ju, Junhui Li, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2020. Multi-modal multi-label emotion detection with modality and label dependence. In *Proc. EMNLP*, Conference held virtually.

A Derivation of the predictive posterior

Since NIG is the Gaussian conjugate prior,

$$\begin{aligned}\mathrm{p}(\boldsymbol{\Psi}|\boldsymbol{\Omega}) &= \mathcal{N}(\gamma, \sigma^2\upsilon^{-1})\,\Gamma^{-1}(\alpha,\beta)\\ &= \frac{\beta^\alpha\sqrt{\upsilon}}{\Gamma(\alpha)\sqrt{2\pi\sigma^2}}\left(\frac{1}{\sigma^2}\right)^{\alpha+1}\\ &\cdot\exp\left\{-\frac{2\beta+\upsilon(\gamma-\mu)^2}{2\sigma^2}\right\}\end{aligned}$$

its posterior $\mathrm{p}(\boldsymbol{\Psi}|\mathcal{D})$ is in the same parametric family as the prior $\mathrm{p}(\boldsymbol{\Psi}|\boldsymbol{\Omega})$. Therefore, given a test utterance $\mathbf{x}_*$, the predictive posterior $\mathrm{p}(y_*|\mathcal{D})$ has the same form as the marginal likelihood $\mathrm{p}(y|\boldsymbol{\Omega})$, where $\mathcal{D}$ denotes the training set.

$$\mathrm{p}(y_*|\mathcal{D}) = \int \mathrm{p}(y_*|\boldsymbol{\Psi})\mathrm{p}(\boldsymbol{\Psi}|\mathcal{D})\,\mathrm{d}\boldsymbol{\Psi} \quad (7)$$

$$\mathrm{p}(y|\boldsymbol{\Omega}) = \int \mathrm{p}(y|\boldsymbol{\Psi})\mathrm{p}(\boldsymbol{\Psi}|\boldsymbol{\Omega})\,\mathrm{d}\boldsymbol{\Psi} \quad (8)$$

In DEER, the predictive posterior and posterior are both conditioned on $\boldsymbol{\Omega}$, written as $\mathrm{p}(y_*|\mathcal{D},\boldsymbol{\Omega})$ and $\mathrm{p}(\boldsymbol{\Psi}|\mathcal{D},\boldsymbol{\Omega})$ to be precise. Also, the information of $\mathcal{D}$ is contained in $\boldsymbol{\Omega}_*$ since $\boldsymbol{\Omega}_* = f_{\hat{\boldsymbol{\Theta}}}(\mathbf{x}_*)$ and $\hat{\boldsymbol{\Theta}}$ is the optimal model parameters obtained by training on $\mathcal{D}$. Then the predictive posterior can be written as $\mathrm{p}(y_*|\boldsymbol{\Omega}_*)$. Given the conjugate prior, the predictive posterior in DEER can be computed by directly substituting the predicted $\boldsymbol{\Omega}_*$ into the expression of marginal likelihood derived in Eqn. (2), skipping the step of calculating the posterior.

B Fusion with text modality

This appendix presents bi-modal experiments that incorporate text information into the DEER model. Transcriptions were obtained from a publicly available automatic speech recognition (ASR) model "wav2vec2-base-960h" [6] which fine-tuned the wav2vec 2.0 (Baevski et al., 2020) model on 960 hours Librispeech data (Panayotov et al., 2015). Transcriptions were first encoded by a RoBERTa model (Liu et al., 2019) and fed into another two-layer Transformer encoder. As shown in Figure 4, outputs from the text Transformer were concatenated with the outputs from the audio Transformer

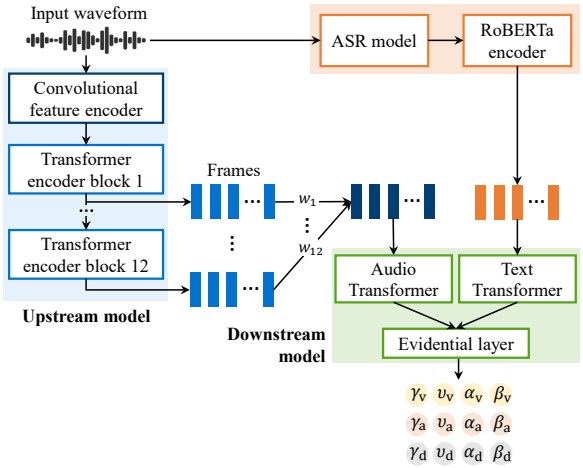


Figure 4: Model structure for bi-modal experiments.

	Modality	v	a	d
MSP-podcast	A	0.506	0.698	0.613
	A+T	0.559	0.699	0.614
	Modality	**v**	**a**	**d**
IEMOCAP	A	0.596	0.756	0.569
	A+T	0.609	0.754	0.575

Table 5: CCC value for bi-modal experiments. 'A' and 'T' stands for audio and text. 'v', 'a', and 'd' stand for valence, arousal, and dominance. Release 1.8 is used for MSP-Podcast. 'Ses05' setup used for IEMOCAP that trains on Session 1-4 and tested on Session 5.

encoder and fed into the evidential output layer. Results are shown in Table 5. Incorporating text information improves the estimation of valence but not necessarily for arousal and dominance. Similar phenomena were observed by (Triantafyllopoulos et al., 2022). A possible explanation is that text is effective for sentiment analysis (positive or negative) but may not be as informative as audio to determine a speaker's level of excitement. CCC for dominance improves more for IEMOCAP than MSP-Podcast possibly because IEMOCAP is an acted dataset and the emotion may be exaggerated compared with MSP-Podcast which contains natural emotion.

[6]Available at: https://huggingface.co/facebook/wav2vec2-base-960h