Handling Ambiguity in Emotion: From Out-of-Domain Detection to Distribution Estimation

Anonymous ACL submission

Abstract

The perception and interpretation of speech emotion are highly subjective, resulting in inconsistent labels from human annotators. Typically, only data with majority-agreed labels are used to train emotion classifiers, which results in the exclusion of data without majority-agreed labels and poses challenges to the model's generalisation ability when ambiguous emotional expressions are encountered in test. To handle ambiguous emotional speech, three methods are studied in this paper. First, an approach based on evidence theory is introduced to quantify the uncertainty in emotion class prediction and detect utterances with ambiguous emotions as out-of-domain samples using the uncertainty score. Second, to obtain fine-grained distinctions among ambiguous emotions, we propose re-framing emotion classification as a distribution estimation task, where every individual label is taken into account in training, not just the majority opinion. Finally, we extend the evidential uncertainty measure for classification to quantify the uncertainty in emotion distribution estimation. Experimental results on the IEMOCAP and CREMA-D datasets show that our method produces effective emotion representations with a reliable uncertainty measure[1].

1 Introduction

The inherent subjectivity of human emotion perception introduces complexity in annotating speech emotion recognition (SER) datasets. Multiple annotators are often involved in labelling each utterance and the majority-agreed (MA) class is usually used as the ground truth (Busso et al., 2008; Cao et al., 2014). Utterances that have no majority-agreed (NMA) labels (*i.e.*, with tied votes) are typically excluded during emotion classifier training (Kim et al., 2013; Poria et al., 2017; Yang et al., 2021), which may result in out-of-domain (OOD) issues in practical applications.

To handle ambiguous emotional data, a naive approach is to aggregate them into an extra OOD class in emotion classification (Wu et al., 2023). However, since such utterances contain a blend of different emotions, the model needs to classify the more complex and diverse NMA emotional expressions into one OOD class while distinguishing the rest of the data into their MA emotional classes.

In this paper, we first investigate if an emotion classifier is able to respond "I don't know" for the ambiguous emotional data. An evidential deep learning (EDL) approach (Sensoy et al., 2018) based on Dempster–Shafer belief theory (Dempster, 1968) is adapted to quantify the uncertainty in emotion classification. When a SER classifier trained on MA data encounters an NMA utterance during the test, the model should identify it as an OOD sample by providing a high uncertainty score, indicating its uncertainty about the specific emotion classes to which the NMA utterance may belong to. Assuming the probability assignment over the emotion classes as a multinomial distribution, this method places a Dirichlet distribution over the multinomial distributions to model their probabilities as second-order probabilities. The concentration parameters of the Dirichlet distribution for uncertainty estimation are predicted by a neural network model.

Consider the example shown in Fig. 1 with the annotations assigned to three utterances. For instance, in utterance (a), eight annotators interpret the speaker as angry while one interprets it as frustrated. Since the majority emotion classes are "angry" for both utterances (a) and (b), they will be assigned to the same ground-truth label "angry" in the aforementioned classification system, which implies that they convey the same emotional content and is evidently unsuitable. On the contrary, utterance (c), though being an NMA utterance, is

[1] Code will be available upon acceptance.

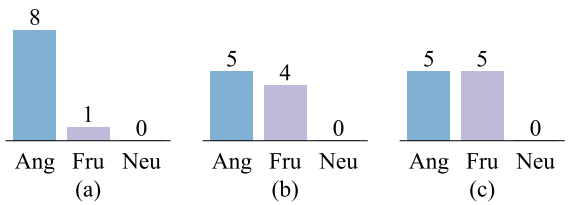


Figure 1: The bar chart shows the number of labels assigned by annotators to the emotion class "angry" (Ang), "frustrated" (Fru), and "neutral" (Neu) in an example.

more likely to share similar emotional content with utterance (b). To obtain more comprehensive representations of emotional content, we propose representing emotion as a distribution and re-framing emotion recognition as a density estimation problem rather than a classification problem. The objective is to estimate the underlying emotion distribution given observed human annotations. In this approach, the system is trained to maximise the marginal likelihood of observing all human annotations from a multinomial distribution under the Dirichlet prior. The EDL approach is then generalised to quantify the uncertainty in distribution estimation. Multiple evaluation metrics are adopted to evaluate the proposed system in terms of majority prediction, uncertainty measure, and distribution estimation. Rather than simply saying "I don't know", the proposed system demonstrates the ability to estimate the emotion distributions of the NMA utterances and also offer a reliable uncertainty measure for the distribution estimation.

The rest of the paper is organised as follows. Section 2 summarises related work. Sections 3 and 4 introduces the proposed approach of uncertainty quantification and distribution estimation. Evaluation metrics and experimental setup are presented in Sections 5 and 6 respectively. Experimental results are shown in Section 7, followed by the conclusions.

2 Related work

Human annotators often interpret the emotion of the same utterance differently due to their personal experiences and cultural backgrounds (Busso et al., 2008; Cowen and Keltner, 2017; Sethu et al., 2019). Instead of using the MA annotation as the ground truth label, some research suggests treating SER as a multi-label task (Mower et al., 2010; Zadeh et al., 2018; Chochlakis et al., 2023) where all emotion classes assigned by any annotator are considered as correct classes and the ground truth label is presented as a multi-hot vector. The SER model is trained to predict the presence of each emotion class for each utterance. An issue with this approach is that it ignores the differences in strengths of different emotion classes.

An alternative approach uses "soft labels" as the proxy of ground truth defined as the relative frequency of occurrence of each emotion class (Fayek et al., 2016; Han et al., 2017; Kim and Kim, 2018). The Kullback–Leibler (KL) divergence or distance metrics between the soft labels and model predictions are used to train the model. However, soft labels, being maximum likelihood estimates (MLE) of the underlying distribution based on observed samples, might not provide an accurate approximation to the unknown distribution when the number of observations (annotations) is limited.

So far, the calibration of SER models has not been extensively studied. In this study, we introduce a novel approach for SER combining Dempster–Shafer belief theory (Dempster, 1968) and evidential deep learning (Sensoy et al., 2018), which provides not only better emotion content estimation but also a reliable measure of the model's prediction confidence.

3 Detecting NMA as OOD

3.1 Limitation of modelling class probabilities with the softmax activation function

A neural network model classifier transforms the continuous logits at the output layer into class probabilities by a softmax function. The model prediction can thus be interpreted as a categorical distribution with the discrete class probabilities associated with the model outputs. The model is then optimised by maximising the categorical likelihood of the correct class, known as the cross-entropy loss.

However, the softmax activation function is known to have a tendency to inflate the probability of the predicted class due to the exponentiation applied to transform the logits, resulting in unreliable uncertainty estimations (Gal and Ghahramani, 2016; Guo et al., 2017). Furthermore, cross-entropy is essentially MLE, a frequentist technique lacking the capability to infer the variance of the predictive distribution. In this section, evidential deep learning (EDL) (Sensoy et al., 2018) is introduced to estimate the model uncertainty which places a second-order probability over the categorical distribution.

3.2 Quantify emotion classification uncertainty by evidential deep learning

Consider an emotion class label as a one-hot vector $\boldsymbol{y}$ where y_k is one if the emotion belongs to class k else zero. $\boldsymbol{y}$ is sampled from a categorical distribution $\boldsymbol{\eta}$ where each component η_k corresponds to the probability of sampling a label from class k:

$$\boldsymbol{y} \sim \mathrm{P}(\boldsymbol{y}|\boldsymbol{\eta}) = \mathrm{Cat}(\boldsymbol{\eta}) = \eta_k^{y_k}. \tag{1}$$

Assume the categorical distribution is sampled from a Dirichlet distribution:

$$\boldsymbol{\eta} \sim \mathrm{p}(\boldsymbol{\eta}|\boldsymbol{\alpha}) = \mathrm{Dir}(\boldsymbol{\eta}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} \eta_k^{\alpha_k - 1} \tag{2}$$

where $B(\cdot)$ is the Beta function, α_k is the hyperparameter of the Dirichlet distribution and $\alpha_0 = \sum_{k=1}^{K} \alpha_k$ is the Dirichlet strength. The output of a standard neural network classifier is a probability assignment over the possible classes and the Dirichlet distribution represents the density of each such probability assignment, hence modelling second-order probabilities and uncertainty.

Subjective logic (Jsang, 2018) establishes a connection between the Dirichlet distribution and the belief representation in Dempster–Shafer belief theory (Dempster, 1968), also known as evidence theory. Consider K classes each associated with a belief mass b_k and an overall uncertainty mass u, which satisfies $u + \sum_{k=1}^{K} b_k = 1$. The belief mass assignment corresponds to the Dirichlet hyperparameter α_k: $b_k = (\alpha_k - 1)/\alpha_0$, where $e_k = \alpha_k - 1$ is usually termed evidence (Sensoy et al., 2018). The overall uncertainty can then be computed as:

$$u = \frac{K}{\alpha_0}. \tag{3}$$

A neural network $\boldsymbol{f}_{\boldsymbol{\Lambda}}$ can be trained to predict $\mathrm{Dir}(\boldsymbol{\eta}^{(i)}|\boldsymbol{\alpha}^{(i)})$ for a given sample $\boldsymbol{x}^{(i)}$ where $\boldsymbol{\Lambda}$ is the model parameters. The network is similar to standard neural networks for classification except that the softmax output layer is replaced with a ReLU activation layer to assure non-negative outputs, which is taken as the evidence vector for the predicted Dirichlet distribution: $\boldsymbol{f}_{\boldsymbol{\Lambda}}(\boldsymbol{x}^{(i)}) = \boldsymbol{e}^{(i)}$. The concentration parameter of the Dirichlet distribution can be calculated as $\boldsymbol{\alpha}^{(i)} = \boldsymbol{f}_{\boldsymbol{\Lambda}}(\boldsymbol{x}^{(i)}) + 1$. Given $\mathrm{Dir}(\boldsymbol{\eta}^{(i)}|\boldsymbol{\alpha}^{(i)})$, the estimated probability of class k can be calculated by:

$$\mathbb{E}[\eta_k^{(i)}] = \frac{\alpha_k^{(i)}}{{\alpha_0}^{(i)}}. \tag{4}$$

3.2.1 Training

For brevity, superscript i is omitted in this section. Given one-hot label $\boldsymbol{y}$ and predicted Dirichlet $\mathrm{Dir}(\boldsymbol{\eta}|\boldsymbol{\alpha})$, the network can be trained by maximising the marginal likelihood of sampling $\boldsymbol{y}$ given the Dirichlet prior. Since the Dirichlet distribution is the conjugate prior of the categorical distribution, the marginal likelihood is tractable:

$$\begin{aligned} \mathrm{P}(\boldsymbol{y}|\boldsymbol{\alpha}) &= \int \mathrm{P}(\boldsymbol{y}|\boldsymbol{\eta})\mathrm{p}(\boldsymbol{\eta}|\boldsymbol{\alpha})d\boldsymbol{\eta} \\ &= \int \prod_k \eta_k^{y_k} \frac{1}{B(\boldsymbol{\alpha})} \prod_k \eta_k^{\alpha_k - 1} \\ &= \frac{B(\boldsymbol{\alpha} + \boldsymbol{y})}{B(\boldsymbol{\alpha})} = \frac{\prod_{k=1}^{K} \alpha_k^{y_k}}{\alpha_0^{\sum_{k=1}^{K} y_k}}. \end{aligned} \tag{5}$$

It is equivalent to training the model by minimising the negative log marginal likelihood:

$$\mathcal{L}^{\mathrm{NLL}} = \sum_{k=1}^{K} y_k(\log(\alpha_0) - \log(\alpha_k)). \tag{6}$$

Following (Sensoy et al., 2018), a regularisation term is added to penalise the misleading evidence:

$$\mathcal{L}^{\mathrm{R}} = \mathcal{KL}(\mathrm{Dir}(\boldsymbol{\eta}|\tilde{\boldsymbol{\alpha}})||\,\mathrm{Dir}(\boldsymbol{\eta}|\mathbf{1})), \tag{7}$$

where $\mathrm{Dir}(\boldsymbol{\eta}|\mathbf{1})$ denotes a Dirichlet distribution with zero total evidence and $\tilde{\boldsymbol{\alpha}} = \boldsymbol{y} + (1 - \boldsymbol{y}) \odot \boldsymbol{\alpha}$ is the Dirichlet parameters after removal of the non-misleading evidence from predicted $\boldsymbol{\alpha}$. This penalty explicitly enforces the total evidence to shrink to zero for a sample if it cannot be correctly classified. The overall loss is $\mathcal{L} = \mathcal{L}^{\mathrm{NLL}} + \lambda\mathcal{L}^{\mathrm{R}}$ where λ is the regularisation coefficient.

4 Emotion distribution estimation

In order to obtain a fine-grained emotion representation, we then describe emotion by a distribution rather than a single class label. Consider an input utterance $\boldsymbol{x}^{(i)}$ associated with M_i labels from human annotators $\{\boldsymbol{y}_m^{(i)}\}_{m=1}^{M_i}$ where $\boldsymbol{y}_m = [y_{m1}, \ldots, y_{mK}]$ is a one-hot vector. Instead of representing the emotional content by the majority vote class, we propose estimating the underlying emotion distribution $\boldsymbol{\eta}$ based on the observations $\{\boldsymbol{y}_m^{(i)}\}_{m=1}^{M_i}$. The emotion classification problem is thus re-framed as a distribution estimation problem. In contrast to the "soft label" method in Section 2 which approximates the emotion distribution of each $\boldsymbol{x}^{(i)}$ solely based on $\mathcal{D}^{(i)} = \{\boldsymbol{y}_m^{(i)}\}_{m=1}^{M_i}$ by

MLE and trains the model to learn this proxy in a supervised manner, the proposed approach meta-learns a distribution estimator $\boldsymbol{f}_{\boldsymbol{\Lambda}}$ across all data points $\mathcal{D}_{\text{meta}} = \{\mathcal{D}^{(i)}\}_{i=1}^{N}$ where N is the number of utterances in training. This uses the knowledge about the emotion expression and annotation variability across different utterances.

For brevity, superscript i is omitted thereafter. Assume $\{\boldsymbol{y}_m\}_{m=1}^{M}$ are samples drawn from a multinomial distribution. Let $\hat{\boldsymbol{y}} = \sum_{m=1}^{M} \boldsymbol{y}_m$ represent the counts of each emotion class:

$$\{\boldsymbol{y}_m\}_{m=1}^{M} \sim \mathrm{P}(\boldsymbol{y}|\boldsymbol{\eta}) = \mathrm{Mult}(\boldsymbol{\eta}, M) \quad (8)$$

$$\mathrm{Mult}(\boldsymbol{\eta}, M) = \frac{\Gamma(M+1)}{\prod_{k=1}^{K}\Gamma(\hat{y}_k+1)}\eta_k^{\hat{y}_k}. \quad (9)$$

The categorical distribution in Eqn. (1) is the special case when $M = 1$.

The network is trained by maximising the marginal likelihood of sampling $\{\boldsymbol{y}_m\}_{m=1}^{M}$ given the predicted Dirichlet prior $\mathrm{Dir}(\boldsymbol{\eta}|\boldsymbol{\alpha})$:

$$\mathrm{P}(\{\boldsymbol{y}_m\}_{m=1}^{M}|\boldsymbol{\alpha}) = \int \mathrm{P}(\{\boldsymbol{y}_m\}_{m=1}^{M}|\boldsymbol{\eta})\mathrm{p}(\boldsymbol{\eta}|\boldsymbol{\alpha})d\boldsymbol{\eta}$$

$$= \frac{\Gamma(M+1)}{\prod_{k=1}^{K}\Gamma(\hat{y}_k+1)}\frac{\prod_{k=1}^{K}\alpha_k^{\hat{y}_k}}{\alpha_0^{\sum_{k=1}^{K}\hat{y}_k}}. \quad (10)$$

The multinomial coefficient is independent of $\boldsymbol{\alpha}$, we thus verify that $\mathcal{L}^{\text{NLL}}$ in Eqn. (11) can be generalised to the distribution estimation framework by replacing one-hot majority label $\boldsymbol{y}$ with $\hat{\boldsymbol{y}}$:

$$\mathcal{L}^{\text{NLL}^*} = \sum_{k=1}^{K}\hat{y}_k(\log(\alpha_0) - \log(\alpha_k)). \quad (11)$$

The regulariser in Eqn. (7) is then modified as:

$$\mathcal{L}^{\text{R1}} = \mathcal{KL}(\mathrm{Dir}(\boldsymbol{\eta}|\hat{\boldsymbol{\alpha}})||\,\mathrm{Dir}(\boldsymbol{\eta}|\mathbf{1})) \quad (12)$$

where $\hat{\boldsymbol{\alpha}} = \bar{\boldsymbol{y}} + (1-\bar{\boldsymbol{y}}) \odot \boldsymbol{\alpha}$ and $\bar{\boldsymbol{y}} = \frac{1}{M}\sum_{m=1}^{M}\boldsymbol{y}_m$ is the soft label. An alternative regulariser is proposed in order to explicitly regularise the predicted multinomial distribution:

$$\mathcal{L}^{\text{R2}} = \mathcal{KL}(\bar{\boldsymbol{y}}||\mathbb{E}[\boldsymbol{\eta}]). \quad (13)$$

Hence, we have extend the EDL method described in Section 3.2 for classification to quantify the uncertainty in distribution estimation, with the original method (Sensoy et al., 2018) being a special case when $M = 1$ and $\hat{\boldsymbol{y}}$ becomes the one-hot majority label $\boldsymbol{y}$. In addition, it's worth noting that the proposed approach does not require a fixed number of annotators for every utterance and can easily generalise to a large number of annotators (*i.e.*, for crowd-sourced datasets).

5 Evaluation metrics

The proposed method is evaluated in terms of majority prediction, uncertainty estimation, OOD detection, and distribution estimation.

Majority prediction. Majority prediction for MA utterances is evaluated by classification accuracy (ACC) and unweighted average recall (UAR) which is the sum of class-wise accuracy divided by the number of classes.

Uncertainty estimation. Model calibration is evaluated by expected calibration error (ECE) (Naeini et al., 2015) and maximum calibration error (MCE) (Naeini et al., 2015). ECE measures model calibration by computing the difference in expectation between confidence and accuracy. Predictions are partitioned into Q bins equally spaced in the [0,1] range and ECE can be computed as follows:

$$\mathrm{ECE} = \sum_{q=1}^{Q}\frac{|B_q|}{n}\left|\mathrm{Acc}(B_q) - \mathrm{Conf}(B_q)\right|. \quad (14)$$

MCE is a variation of ECE which measures the largest calibration gap:

$$\mathrm{MCE} = \max_{q\in\{1,\ldots,Q\}}\left|\mathrm{Acc}(B_q) - \mathrm{Conf}(B_q)\right|. \quad (15)$$

OOD detection. The area under the receiver operating characteristic (AUROC) and the area under the precision-recall curve (AUPRC) are used to evaluate the performance of OOD detection. The estimated uncertainty is used as a decision threshold for both AUROC and AUPRC. The baseline is 50% for AUROC and is the fraction of positives for AUPRC. NMA utterances are set as the positive class to detect.

Distribution estimation. Emotion distribution estimation performance is measured by the negative log-likelihood (NLL) of sampling human annotations from the predicted multinomial distribution.

6 Experimental setup

6.1 Baselines

The proposed EDL-based method is compared to baselines including a deterministic classification network with softmax activation trained by the cross-entropy loss between the majority vote label and model predictions (**MLE**), a Monte-Carlo dropout (Gal and Ghahramani, 2016) model with a dropout rate of 0.5 (**MCDP**) which is forwarded 100 times to obtain 100 samples during inference,

an ensemble (Lakshminarayanan et al., 2017) of 10 models with the same structure trained by bagging (**Ensemble**), and a MLE model with NMA as an extra class (**MLE+**). An additional baseline is designed for distribution estimation: a deterministic model with softmax activation trained by minimising KL divergence between the soft label $\bar{\boldsymbol{y}}$ and predictions as defined in Eqn. (13) (**MLE***), which is an extension of the MLE system from one-hot majority vote labels to soft labels.

The system described in Section 3.2 is denoted as **"EDL"**. **"EDL*(R1)"** and **"EDL*(R2)"** refer to the systems proposed in in Section 4 using regularisation terms defined in Eqn. (12) and Eqn. (13) respectively. Uncertainty estimation of EDL models are computed by Eqn. (3) while max probability is used as confidence measure for other methods.

6.2 Datasets

Two publicly available datasets are used in the experiments: IEMOCAP (Busso et al., 2008) and CREMA-D (Cao et al., 2014).

The IEMOCAP corpus is one of the most widely used SER datasets. It consists of 10,039 English utterances from 5 dyadic conversational sessions. Each utterance is evaluated by a minimum of three human annotators for 10 emotion categories, resulting in an average of 3.42 labels per utterance. Only 16.1% of utterances have an all-annotators-agreed emotion label. The emotion distribution is represented using a five-dimensional categorical distribution, including happy (merged with excited), sad, neutral, angry, and others. The "others" category includes all emotions not covered in the previous four categories, primarily dominated by frustration, which accounts for over 92% of this category. After the grouping, 1429 (14.2%) utterances don't have a majority agreed emotion class label.

The CREMA-D corpus contains 7,442 English utterances from 91 actors with a variety of backgrounds. Actors spoke from a selection of 12 sentences using one of six different emotions (anger, disgust, fear, happy, neutral and sad). The dataset is annotated by crowd-sourcing. Participants rated the emotion based on the combined audiovisual presentation, the video alone, and the audio alone. Ratings based on audio alone are used in this work. 95% of the clips have more than 7 ratings and utterances have 9.21 ratings on average. 644 (8.7%) utterances don't have a majority agreed emotion class label.

Both datasets are divided into an MA subset and an NMA subset. Except for MLE+, all other methods are trained on MA data only. For MLE+, NMA is split into 75% train and 25% test. The NMA (train) data is included in MLE+ training. Therefore, OOD detection is evaluated only on NMA (test) data for MLE+ while on the whole NMA subset for all other methods. All other methods are also evaluated on the NMA (test) subset for comparison. For IEMOCAP, unless otherwise stated, models are trained on MA data from Sessions 1-4 and MA data from Session 5 is held out as the MA test set. For the CREMA-D dataset, the MA subset is split into train, validation, test in the ratio 70 : 15 : 15 following prior work (Ristea and Ionescu, 2021).

6.3 Model structure

The backbone structure used in this paper follows an upstream-downstream paradigm (Bommasani et al., 2021). The upstream model uses the universal speech model (USM) (Zhang et al., 2023) with 300M parameters which contains a CNN-based feature extractor and 12 Conformer (Gulati et al., 2020) encoder blocks of dimension 1024 with 8 attention heads. The USM is pre-trained by BEST-RQ (Chiu et al., 2022) which uses a BERT-style training task for the audio input to predict masked speech features. The structure of the downstream model follows SUPERB (Yang et al., 2021), a benchmark for evaluating pre-trained upstream models, which performs utterance-level mean-pooling followed by a fully-connected layer. The pre-trained upstream USM model is frozen. The downstream model computes the weighted sum of the hidden states extracted from each layer of the upstream model.

6.4 Implementation details

The model is implemented using Pax[2]. The batch size is set to 256, The coefficient λ is set to 0.8 for IEMOCAP and 0.2 for CREMA-D. The Adafactor optimiser and Noam learning rate scheduler are used with 200 warm up steps and a peak learning rate of 8.84×10^{-4}. Since the CREMA-D dataset is extremely imbalanced (*i.e.*, neutral accounts for over 50%), a balanced sampler is applied during training which makes sure samples in each training batch are roughly balanced. The model is trained for 20k steps which takes $\sim$ 5 hours on 8 TPU v4s.

[2]https://github.com/google/paxml

	Classify MA				Detect NMA (all)		Detect NMA (test)	
	ACC ↑	UAR ↑	ECE ↓	MCE ↓	AUROC ↑	AUPRC ↑	AUROC ↑	AUPRC ↑
MLE+	0.447	0.438	0.303	0.383	/	/	0.461	0.139
MLE	0.582	0.577	0.206	0.239	0.550	0.471	0.549	0.177
MCDP	0.584	0.572	0.128	0.184	0.566	0.491	0.568	0.203
Ensemble	0.593	0.595	0.439	0.594	0.567	0.491	0.563	0.192
EDL	**0.611**	**0.596**	**0.103**	**0.145**	**0.610**	**0.530**	**0.620**	**0.227**

Table 1: Results of quantifying uncertainty in emotion classification on the IEMOCAP dataset. The baseline for AUPRC is 0.433 for the entire NMA set and 0.160 for the NMA test subset. The best value in each column is indicated in bold, and the second-best value is underlined.

	Classify MA				Detect NMA (all)		Detect NMA (test)	
	ACC ↑	UAR ↑	ECE ↓	MCE ↓	AUROC ↑	AUPRC ↑	AUROC ↑	AUPRC ↑
MLE+	0.568	0.540	0.216	0.476	/	/	0.552	0.156
MLE	0.714	0.672	0.150	0.156	0.578	0.467	0.571	0.179
MCDP	0.717	0.687	0.102	0.109	0.619	0.481	0.614	0.201
Ensemble	**0.731**	0.674	0.362	0.496	0.598	0.481	0.605	0.198
EDL	0.711	**0.714**	**0.057**	**0.080**	**0.645**	**0.506**	**0.657**	**0.234**

Table 2: Results of quantifying uncertainty in emotion classification on the CREMA-D dataset. The baseline for AUPRC is 0.387 for the entire NMA set and 0.097 for the NMA test subset.

Model	# Param	ACC (%)
Wav2vec 2.0 large (Baevski et al., 2020)	317M	65.64
Data2vec large (Baevski et al., 2022)	314M	66.31
HuBERT large (Hsu et al., 2021)	317M	67.62
WavLM large (Chen et al., 2022)	317M	70.62
USM-300M (Zhang et al., 2023)	290M	**71.06**

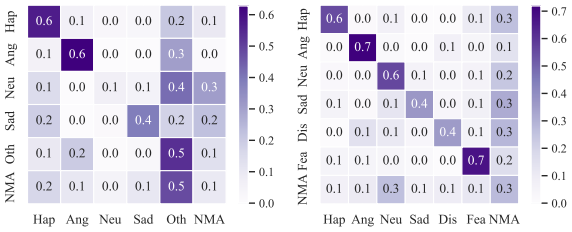
Table 3: Four-way classification results IEMOCAP following the SUPERB-ER benchmark setup.

7 Results

The USM-based backbone structure is first evaluated following the setup of the emotion recognition[3] task of the SUPERB benchmark (Yang et al., 2021). As shown in Table 3, the backbone structure outperforms state-of-the-art methods[4].

[3]SUPERB-ER setup: four-way emotion classification (happy, sad, angry, neutral) on IEMOCAP dataset with leave-one-session-out five-fold cross validation.

[4]https://superbbenchmark.org/leaderboard



(a) IEMOCAP (b) CREMA-D

Figure 2: Confusion matrix of the MLE+ system on IEMOCAP and CREMA-D.

7.1 Emotion classification uncertainty

The proposed EDL-based method is compared to baselines in Table 1 and 2 on the IEMOCAP and CREMA-D dataset respectively. First, the proposed method demonstrates comparable classification performance to the baselines, suggesting that the extension for uncertainty estimation does not undermine the model's capabilities. In addition, the proposed method offers superior model calibration, as evidenced by the lowest values of ECE and MCE. It also outperforms the baselines in effectively identifying NMA as OOD samples.

7.1.1 Including NMA as an additional category degrades the performance

The MLE+ results reveal that the addition of the NMA class has a detrimental impact on the clas-

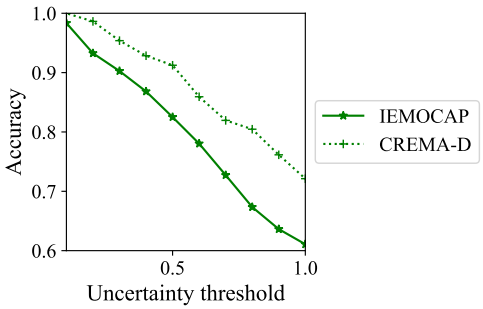


Figure 3: The change of accuracy with respect to the uncertainty threshold for EDL-based methods on IEMOCAP and CREMA-D.

sification performance of the original emotional classes. Figure 2 shows the confusion matrix of the MLE+ model. NMA itself is challenging to predict possibly because it essentially contains a mix of different emotion content. Grouping these utterances into one class can confuse the model, particularly for the classes neutral, sad, frustrated, and disgust.

7.1.2 Reject option for accuracy

Figure 3 shows the change of accuracy when samples with uncertainty larger than a threshold are excluded. The model tends to provide more accurate predictions when it is more confident about its prediction, which demonstrates the effectiveness of uncertainty prediction.

7.2 Emotion distribution estimation

The result of distribution-based methods on classification of MA data are shown in Table 4. Compared to the classification-based methods in Table 1 and Table 2, it can be seen that distribution-based

	IEMOCAP			
	ACC	**UAR**	**ECE**	**MCE**
MLE*	0.564	0.562	0.151	0.279
EDL*(R1)	0.623	0.612	0.081	0.208
EDL*(R2)	**0.624**	**0.616**	**0.025**	**0.201**
	CREMA-D			
	ACC	**UAR**	**ECE**	**MCE**
MLE*	0.693	0.621	0.109	0.115
EDL*(R1)	**0.740**	0.694	**0.029**	**0.095**
EDL*(R2)	0.718	**0.722**	0.084	0.107

Table 4: Performance of distribution-based methods on MA data.

	NLL$^{\text{MA}}$ ↓	**NLL$^{\text{NMA (all)}}$** ↓
MLE	1.310	1.924
MCDP	0.972	1.266
Ensemble	2.572	2.055
EDL	0.958	1.019
MLE*	0.941	1.137
EDL*(R1)	0.861	**0.951**
EDL*(R2)	**0.833**	0.953

Table 5: Distribution estimation results on IEMOCAP.

	NLL$^{\text{MA}}$ ↓	**NLL$^{\text{NMA (all)}}$** ↓
MLE	1.532	2.054
MCDP	0.965	1.292
Ensemble	2.285	2.089
EDL	0.757	1.021
MLE*	0.648	0.774
EDL*(R1)	0.614	0.722
EDL*(R2)	**0.606**	**0.698**

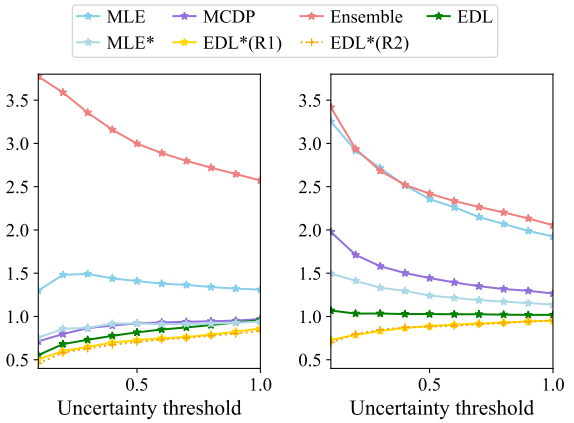
Table 6: Distribution estimation results on CREMA-D.

methods do not reduce the performance of emotion classification and model calibration on MA data.

The proposed EDL* methods are compared to the baselines in terms of the negative log likelihood of sampling target labels from the predicted emotion distribution. Results on IEMOCAP and CREMA-D are shown in Table 5 and Table 6. EDL* methods produce improved distribution estimation, achieving smaller NLL values on both MA and NMA data. Among the two EDL* methods employing different regularisation terms, EDL* with R2 (defined in Eqn. (13)), which directly applies regularisation to the predicted distribution, exhibits better distribution estimation without sacrificing the calibration capability of the model.

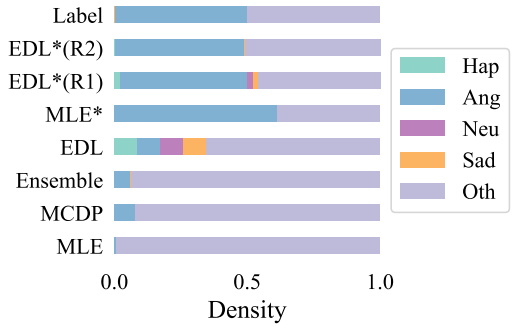
7.2.1 Reject option for NLL

A reject option is then evaluated for NLL instead of accuracy to examine the model calibration. For a well-calibrated model, a decrease in the NLL value, which is associated with improved distribution estimation, is expected when the model becomes more confident. Figure 4 visualises the change of NLL for MA data and NMA data when uncertainty increases. For MA, the type of data that has been seen by the models during training, most methods can successfully reject uncertain samples except for MLE and Ensemble, as shown by an increase

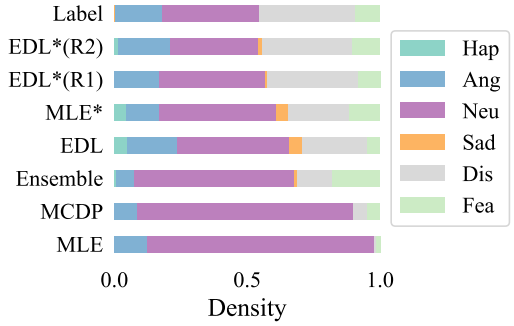


(a) MA data (b) NMA data

Figure 4: Reject option for NLL on IEMOCAP. Results on CREMA-D show similar trend which can be found in Appendix B.



(a) Utterance "Ses04M_impro02_F024"



(b) Utterance "1084_TSI_ANG_XX"

Figure 5: Visualisation of emotion distribution for case study. Utterance (a) is selected from IEMOCAP. Utterance (b) is selected from CREMA-D.

in NLL values when the uncertainty threshold increases. However, for NMA data which the model hasn't seen in training, only the EDL* methods exhibit the ability to demonstrate an increasing trend in NLL values.

The proficiency of the proposed EDL* methods in estimating the emotion distribution and providing reliable confidence predictions, demonstrate the method's capacity to estimate both aleatoric uncertainty (Matthies, 2007; Der Kiureghian and Ditlevsen, 2009), arising from data complexity (*i.e.*, the ambiguity of emotion expression), and epistemic uncertainty, which corresponds to the amount of uncommitted belief in subjective logic (Jsang, 2018).

7.2.2 Case study

Emotion distributions estimated by different methods are visualised against the label distributions for two representative examples in Figure 5. In general, distribution-based methods show superior performance in distribution estimation than classification-based methods. In the case of an utterance (a) which receives two "angry" labels and two "frustrated" labels, the proposed EDL* methods stands out by effectively capturing the tie between the emotions, whereas the predictions of classification-based methods tend to be predominantly skewed towards "frustrated". As for utterance (b), where both "disgust" and "neutral" receive four votes, along with two votes for "angry" and one for "fear", the emotion distributions predicted by the EDL* methods also show a similar pattern. Additional examples can be found in Appendix C and Appendix D.

8 Conclusion

This work re-examines the emotion classification problem, starting with an exploration of ways to handle data with ambiguous emotions. We first adopt evidence theory to allow the emotion classifier to output "I don't know" when it encounters utterances with ambiguous emotion. The model is trained to predict the hyperparameters of a Dirichlet distribution, which represents the second-order probability of the probability assignment over emotion classes. In order to capture more fine-grained emotion differences, the emotion classification problem is transformed into emotion distribution estimation where each annotation is taken into account rather than only the majority opinion. The EDL-based uncertainty measure is extended to quantify uncertainty in emotion distribution estimation. Results on the IEMOCAP and CREMA-D datasets show that given an utterance with ambiguous emotion which the model hasn't seen during training, the proposed approach is able to estimate its emotion distribution rather than just returning "I don't know".

Ethics Statement

In this work, all human annotations used for training were taken from existing publicly available corpora. No new human annotations were collected.

In subjective tasks like emotion recognition, there is usually no single "correct" answer. The conventional approach of imposing a single ground truth through majority voting may overlook valuable nuances within each annotator's evaluation and the disagreements between them, potentially resulting in the under-representation of minority views. This study, instead of exclusively relying on the majority vote, integrates emotion annotations from all annotators for each utterance during model training. It is hoped that this work could contribute to a more inclusive representation of human opinions.

Limitations

The proposed approach requires the raw labels from different human annotators for each sentence to be provided by the datasets. Although validated only for emotion recognition, the proposed method could also be applied to other tasks with disagreements in subjective annotations, which will be investigated in future work.

References

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *Proc. ICML*, Baltimore, USA.

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*, Vancouver, Canada.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359.

Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. 2014. CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390.

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*, 16(6):1505–1518.

Chung-Cheng Chiu, James Qin, Yu Zhang, Jiahui Yu, and Yonghui Wu. 2022. Self-supervised learning with random-projection quantizer for speech recognition. In *Proc. ICML*, Baltimore, USA.

Georgios Chochlakis, Gireesh Mahajan, Sabyasachee Baruah, Keith Burghardt, Kristina Lerman, and Shrikanth Narayanan. 2023. Leveraging label correlations in a multi-label setting: A case study in emotion. In *Proc. ICASSP*, Rhodes, Greece.

Alan S Cowen and Dacher Keltner. 2017. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the national academy of sciences*, 114(38):E7900–E7909.

Arthur P Dempster. 1968. A generalization of Bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(2):205–232.

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, USA.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented Transformer for speech recognition. In *Proc. Interspeech*, Shanghai, China.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *Proc. ICML*, Sydney, Australia.

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, USA.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Audun Jsang. 2018. *Subjective Logic: A formalism for reasoning under uncertainty*. Springer Publishing Company, Incorporated.

Y. Kim, H. Lee, and E. M. Provost. 2013. Deep learning for robust feature generation in audiovisual emotion recognition. In *Proc. ICASSP*, Vancouver, Canada.

Yelin Kim and Jeesun Kim. 2018. Human-like emotion recognition: Multi-label learning from noisy labeled audio-visual expressive speech. In *Proc. ICASSP*, Calgary, Canada.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proc. NeurIPS*, Long Beach, USA.

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.

Emily Mower, Maja J Matarić, and Shrikanth Narayanan. 2010. A framework for automatic human emotion classification using emotion profiles. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(5):1057–1070.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proc. AAAI*, Austin, USA.

S. Poria, E. Cambria, R. Bajpai, and A. Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125.

Nicolae-Cătălin Ristea and Radu Tudor Ionescu. 2021. Self-paced ensemble learning for speech and audio classification. In *Proc. Interspeech*, Brno, Czech.

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

Vidhyasaharan Sethu, Emily Mower Provost, Julien Epps, Carlos Busso, Nicholas Cummins, and Shrikanth S. Narayanan. 2019. The ambiguous world of emotion representation. *arXiv preprint arXiv:1909.00360*.

Wen Wu, Chao Zhang, and Philip C. Woodland. 2023. Integrating Emotion Recognition with Speech Recognition and Speaker Diarisation for Conversations. In *Proc. Interspeech*, Dublin, Ireland.

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, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee. 2021. SUPERB: Speech Processing Universal PERformance Benchmark. In *Proc. Interspeech*, Brno, Czech.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proc. ACL*, Melbourne, Australia.

Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, et al. 2023. Google USM: Scaling automatic speech recognition beyond 100 languages. *arXiv preprint arXiv:2303.01037*.

A Alternative activation functions

As described in Section 3.2, ReLU is used as the output activation function to make sure the evidence is non-negative. This section investigates the use of softplus and exponential as alternative activation functions. As shown in Table 7, using exponential function tends to result in less effective model calibration. As shown by the empirical cumulative distribution function (ECDF) of uncertainty and entropy in Figure 8, exponential activation leads to smaller uncertainty and entropy, which echos the statement in Section 3.1 that exponential activation tends to inflate the probability of the correct class.

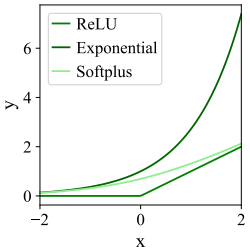
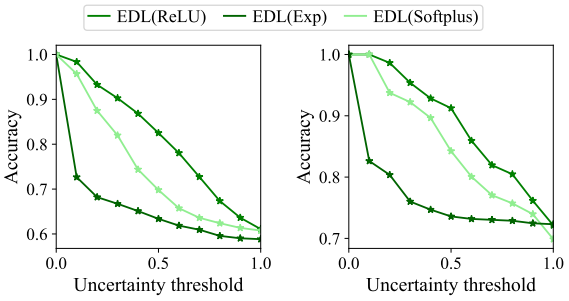


Figure 6: Illustration of the activation functions.



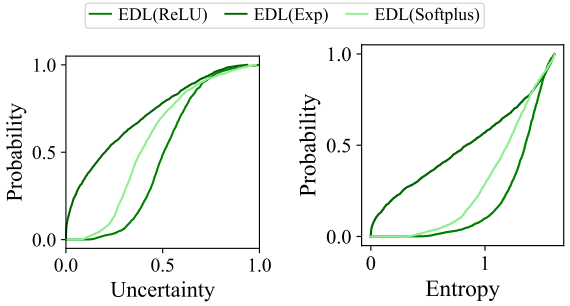
(a) IEMOCAP (b) CREMA-D

Figure 7: Reject option for accuracy for EDL methods with different activation functions.

IEMOCAP	Classify MA				Detect NMA (all)		Detect NMA (test)	
	ACC	UAR	ECE	MCE	AUROC	AUPRC	AUROC	AUPRC
EDL (ReLU)	**0.611**	0.596	0.103	**0.145**	0.610	0.530	0.620	0.227
EDL (Softplus)	0.608	0.574	**0.035**	0.173	**0.617**	**0.534**	**0.639**	**0.251**
EDL (Exponential)	0.588	**0.601**	0.167	0.230	0.593	0.502	0.619	0.225

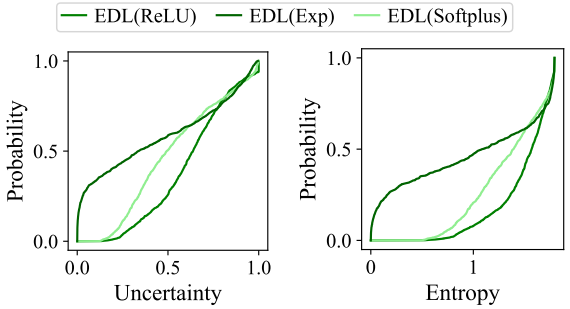
CREMA-D	Classify MA				Detect NMA (all)		Detect NMA (test)	
	ACC	UAR	ECE	MCE	AUROC	AUPRC	AUROC	AUPRC
EDL (ReLU)	0.701	**0.714**	**0.057**	**0.080**	**0.645**	**0.506**	**0.657**	**0.234**
EDL (Softplus)	0.692	0.696	0.113	0.309	0.640	**0.506**	0.633	0.230
EDL (Exponential)	**0.723**	0.602	0.277	0.277	0.623	0.495	0.626	0.197

Table 7: Comparison of EDL methods with different activation functions on IEMOCAP and CREMA-D.



(a) ECDF of uncertainty (b) ECDF of entropy

Figure 8: Empirical CDF of uncertainty (left) and entropy (right) on IEMOCAP for EDL method with different activation functions.



(a) ECDF of uncertainty (b) ECDF of entropy

Figure 9: Empirical CDF of uncertainty (left) and entropy (right) on CREMA-D for EDL method with different activation functions.

B Reject option for NLL on CREMA-D

This section shows the reject option for NLL on CREMA-D dataset. Similar to the findings in Section 7.2.1, most methods are effective for rejecting uncertain samples in the MA data, while only the EDL* methods are successful for NMA.

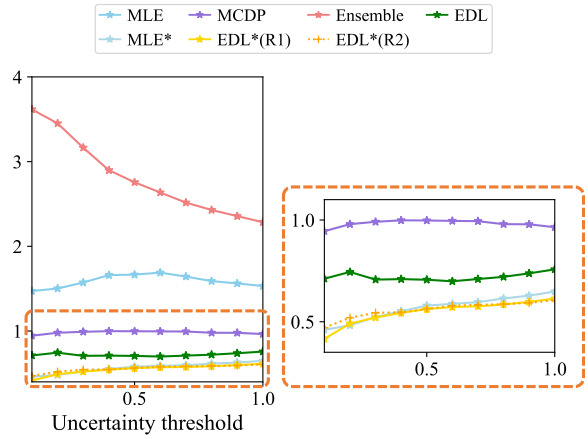


Figure 10: Reject option for NLL on MA data of CREMA-D.

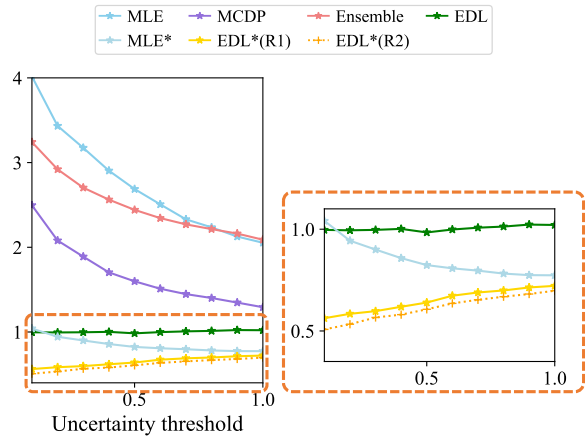
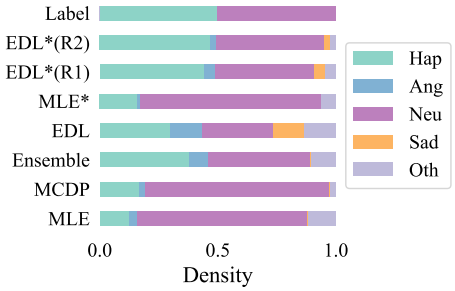


Figure 11: Reject option for NLL on NMA data of CREMA-D.

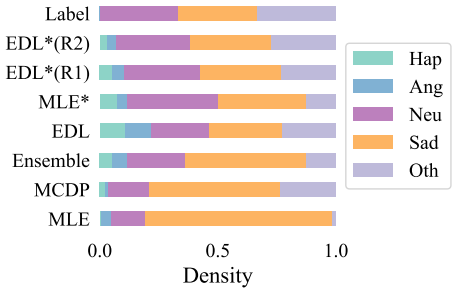
C More visualised examples on IEMOCAP

This section shows more examples on IEMOCAP. EDL* methods show better estimation of emotion

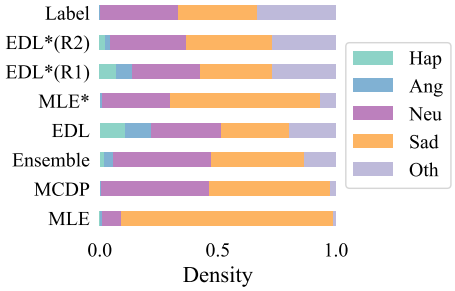
distribution.



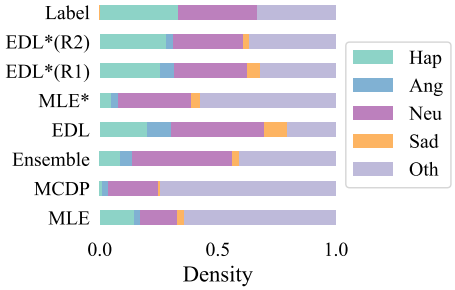
(a) Ses01M_impro07_M025



(b) Ses03F_script01_1_F016



(c) Ses04F_script01_1_M033

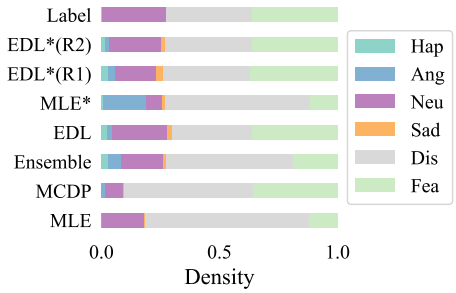


(d) Ses04M_script01_3_M013

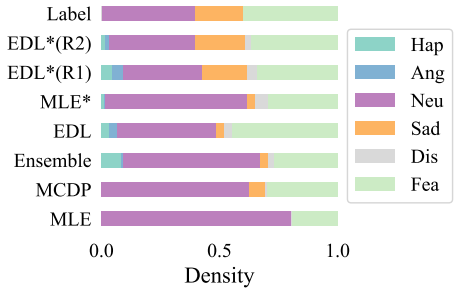
Figure 12: Case study on IEMOCAP.

D More visualised examples on CREMA-D

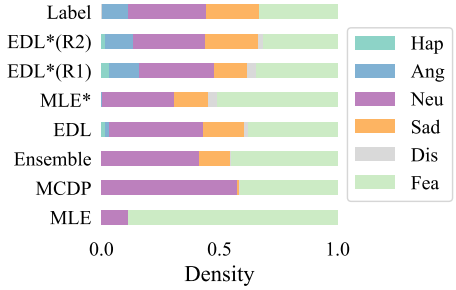
This section shows more examples on CREMA-D. As can be seen, EDL* methods can better approximate the distribution of emotional content of an utterance.



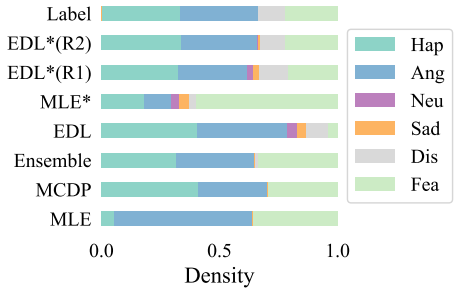
(a) 1033_IWW_DIS_XX



(b) 1052_ITH_FEA_XX



(c) 1068_ITH_SAD_XX



(d) 1009_IWL_FEA_XX

Figure 13: Case study on CREMA-D.