Generative Approach Using Soft-Labels to Learn Uncertainty in Predicting Emotional Attributes

Kusha Sridhar
Multimodal Signal Processing (MSP)
The University of Texas at Dallas
Richardson TX 75080, USA
Kusha.Sridhar@utdallas.edu

Wei-Cheng Lin
Multimodal Signal Processing (MSP)
The University of Texas at Dallas
Richardson TX 75080, USA
wei-cheng.lin@utdallas.edu

Carlos Busso
Multimodal Signal Processing (MSP)
The University of Texas at Dallas
Richardson TX 75080, USA
busso@utdallas.edu

***Abstract*—This paper presents a novel *speech emotion recognition* (SER) method to capture the uncertainty in predicting emotional attributes using the true distribution of scores provided by annotators as ground truth (i.e., soft-labels). Reliable, generalizable, and scalable SER systems are important in areas such as healthcare, customer service, security, and defense. A barrier to build these systems is the lack of quality labels due to the expensive annotation process, leading to poor generalization. To address this limitation, this study proposes a semi-supervised generative modeling approach using a *variational autoencoder* (VAE) with an emotional regressor at the bottleneck trained with soft-labels of emotional attributes. We demonstrate that estimating uncertainties in predicting emotional attribute scores is possible with soft-labels. We analyze the benefits of uncertainty estimation with a reject option formulation, where the model can abstain from predicting emotion when it is less confident. At 60% test coverage, we achieve relative improvements in *concordance correlation coefficient* (CCC) up to 16.85% for valence, 7.12% for arousal, and 8.01% for dominance. Furthermore, we propose an uncertainty transfer learning strategy where uncertainties learned from one attribute are used as a sample re-ordering criterion for another attribute, achieving additional improvements in prediction performance for valence. We also demonstrate the generalization power of our method in comparison to other uncertainty estimating methods using cross-corpus evaluations. Finally, we demonstrate that our method has lower computational complexity than alternative approaches.**

***Index Terms*—Generative Models, Variational Autoencoder, Speech Emotion Recognition, Emotional Attributes**

I. Introduction

With the growing popularity of *human computer interaction* (HCI) in areas such as healthcare and security, and the ubiquitous nature of speech based devices to perform HCI tasks, robust and reliable *speech emotion recognition* (SER) systems have enormous potential to improve the user experience. A major drawback is the time consuming nature of acquiring good quality emotional labels for the speech samples. Despite listening to the same audio clip, different annotators may disagree on a particular emotional label. These subjective labels create a perceptual distribution where the confusion between the annotators is evident. The conventional approach is to consolidate individual annotations with rules such as the majority vote for categorical emotions (e.g., happiness, anger or sadness) or simply the average of the scores in the case of emotional attributes (valence, arousal and dominance). These aggregation approaches reduce the confusion between the labelers, but they can lose some valuable information and mask less prominent emotional traits. Many studies have directly used soft-labels in emotion classification tasks, leveraging the labels provided by individual evaluators, even if they do not agree with the consensus-labels [1], [2]. We expect that the same idea can be explored for emotion regression tasks for emotional attributes. Soft-label probabilities of the emotional attribute scores can provide information about the disagreement between annotators. This problem can be formulated as a classification task by binning the attribute scores and using an appropriate loss to train a *deep neural network* (DNN) that directly compares the predicted and ground truth distributions.

Soft-labels encode the confusion between labelers. Studies have shown that ambiguous samples for human labelers are also ambiguous samples for SER systems [3]. Therefore, soft-labels can be leveraged to learn the uncertainty in a SER model to predict emotion, providing a reliability score associated with its predictions. Knowing the uncertainty in the predictions makes a SER system more versatile in mission critical applications with human-in-the-loop solutions, where only the uncertain cases are reviewed by humans. Knowledge about prediction uncertainties can also help in developing unsupervised and semi-supervised algorithms for machine learning problems such as active learning [4]–[6], co-training [7], and curriculum learning [8].

This study proposes a *semi-supervised learning* (SSL) method using a *variational autoencoder* (VAE) [9] to leverage soft-labels in the prediction of emotional attributes. The approach has an emotional regressor attached to its bottleneck representation layer and is trained with soft-labels of emotional attributes to perform SER. Soft-labels help in incorporating information from the label distributions into the latent space of the VAE. We employ *Monte Carlo* (MC) samplings from the latent space to obtain multiple predictions for the same input instance, from which we measure the uncertainty in the predictions using entropy. We demonstrate the use of uncertainty predictions using a reject option framework, where the model can selectively accept or reject a sample based on its confidence in the predictions.

We evaluate the SER performance of the proposed model by studying the tradeoff between test coverage (i.e., number of accepted test samples) and performance, measured with

This work was supported by NSF grants CNS-2016719 and IIS-1453781.

concordance correlation coefficient (CCC). At 60% test coverage, we achieve relative improvements in CCC up to 16.85% for valence, 7.12% for arousal, and 8.01% for dominance. We also show that the uncertainties learned when predicting one emotional attribute can be used as a sample re-ordering criterion to improve the prediction of another attribute. We refer to this effect as *uncertainty transfer learning* (UTL). We evaluate the generalization ability of our method with cross-corpus evaluations, where our proposed approach leads to better performance than alternative uncertainty prediction algorithms. We evaluate the computational efficiency of our method, showing a significant time reduction during inference compared to *Monte Carlo dropout* (MCD) based approaches.

II. Related Work

A. Soft-labels for SER

The conventional way of aggregating emotional attribute annotations from several labelers is to average them, obtaining a consensus-label. This approach ignores disagreements across evaluators, removing information about inter-labeler variability. An alternative approach is to train DNN models using the true annotators' distribution, which is referred to as soft-labels. Studies have explored SER with soft-labels as ground truth labels for categorical classification tasks [1], [2], [10]–[12]. Fayek et al. [2] proposed a DNN to learn a mapping from spectrograms to emotion classes by using soft-labels to model the perceptual variations between multiple annotators. This approach improved the performance compared to methods trained with ground truth labels obtained by consensus-labels. Lotfian and Busso [1] formulated emotion perception as a probabilistic model, where each individual annotation corresponds to a realization sampled from an unknown multivariate Gaussian distribution representing the emotions of a sentence. Each labeler is modeled as a point in the distribution, reporting the emotional category that has the largest intensity measure. Tarantino et al. [11] implemented an emotion classifier using a transformer model with self-attention and global windowing. They compared the SER performance with soft-labels and hard-labels, showing that soft-labels are able to capture better features from raw audio inputs, outperforming the SER model trained with hard-labels. Kim et al. [12] proposed a soft-label classification technique to deal with samples with no agreement between annotators, reporting better performance over a method that disregarded these samples.

While the use of soft-labels has been popular in speech emotion classification tasks, fewer studies have explored its use for regression tasks on emotional attribute descriptors. Zhang et al. [13] proposed the *f-similarity preservation gain* (f-SPG) loss for SER using soft-labels of emotional attributes. This loss is added as an auxiliary loss in a *multi-task learning* (MTL) formulation, enforcing that the embedding learned by the model preserves label similarity between samples. Gideon et al. [14] used an adversarial training strategy using soft-labels of emotional attributes to learn a common representation between the source and target domains to improve cross-corpus SER. Cai et al. [15] used a SER system trained with soft-labels of emotional attributes to develop a *text-to-speech* (TTS) system that incorporates emotional expressiveness.

B. Uncertainty Prediction in SER

While soft-labels are capable of reflecting the diversity of human annotation, modeling the label ambiguity in SER tasks is a real challenge. Sethu et al. [16] developed a mathematical framework called *AMBiguous Emotion Representation* (AMBER) that takes different emotional representations into account (categorical, numerical and ordinal) to create an ambiguity function that reflects the perceptual uncertainties in emotional label spaces. Studies have also predicted or modeled confidence measures in SER. Deng et al. [17] used a SSL approach to include target domain data during the training of the models based on confidence scores obtained on the target data through multi-corpora training. Steidl et al. [18] used an entropy based measure to judge a classifier's output, evaluating the classifier by comparing the entropy in its predictions and the entropy calculated from the labelers' confusion. Studies such as Chou et al. [19] and Han et al. [20] have also utilized both hard and soft-labels to jointly model the emotional state and the perceptual uncertainty in the annotations.

A technique that relies on uncertainty prediction is the reject option formulation. A model with a reject option can decline a prediction when its confidence is low. Recent studies by Sridhar and Busso [21], [22] demonstrated the application of reject options in SER. In Sridhar et al. [21], they devised a reject option framework for emotion classification based on an empirical risk minimization framework. They also used MCD to model uncertainties in predicting emotional attributes [22].

This study evaluates uncertainty prediction using soft-labels with a generative modeling approach. We show the benefits of modeling uncertainty using soft-labels for predicting emotional attributes. Furthermore, we demonstrate that our approach achieves better generalization using less computational resources during inference when compared to alternative uncertainty prediction approaches such as MCD based methods.

III. Resources

A. Datasets

The primary corpus used in this study is the MSP-Podcast corpus [23], which is the largest naturalistic speech emotion database that is publicly available (release v1.6). The corpus consists of emotionally rich spontaneous speech recordings gathered from podcasts from various audio-sharing websites. The data collection protocol relied on retrieving emotionally rich segments with existing SER models to balance the emotional content of the corpus by following the strategy suggested in Mariooryad et al. [24]. The database is split into train, test and development sets with the goal of creating partitions with minimal speaker overlap. The test set has 10,124 samples from 50 speakers, the development set has 5,958 samples from 40 speakers, and the train set has 34,280 samples from the rest of the speakers. This study uses the emotional attributes valence (negative versus positive), arousal (calm versus active), and dominance (weak versus strong), which are annotated on a seven point Likert scale. The annotation process uses a crowdsourcing protocol similar to the one discussed in Burmania et al. [25], where at least five annotators assessed each sentence. There are around 600,000 speech segments that are not yet annotated with emotional labels. We use a portion of these recordings as the unlabeled set. This study uses the true annotator distribution (soft-labels) to represent the ground truth. The soft-labels for emotional attributes are constructed by dividing the attribute space into bins, counting the number of votes in each bin, creating a probability distribution for each sentence. We expect that this approach provides a better representation of the fuzzy labels, since this approach captures the true variability in the emotional content, as perceived by several annotators.

In addition to the MSP-Podcast corpus, we use the USC-IEMOCAP corpus [26] as an additional test set for cross-corpus evaluations. This is an audiovisual corpus, but this study only uses the audio modality. It contains dyadic interactions from 10 actors in improvised scenarios. The corpus contains 10,039 speaking turns annotated for arousal, valence, and dominance by at least two raters on a 5-Likert scale.

B. Acoustic Features

This paper uses the Interspeech 2013 computational paralinguistics challenge acoustic features [27], extracted using the OpenSmile toolkit [28]. This feature set consists of frame level features called *low level descriptors* (LLDs) such as energy, fundamental frequency and *Mel-frequency cepstral coefficients* (MFCCs). A set of sentence level statistics are calculated over these LLDs (e.g., mean and variance of energy), which are referred to as *high level descriptors* (HLDs). For each speech utterance this process generates a 6,373 dimensional feature vector, regardless of its duration.

IV. Proposed Method

This study explores the use of soft-labels of emotional attributes in uncertainty predictions for SER problems. The proposed method relies on a generative model with VAE, trained in a SSL manner.

A. Generative Approach for SER Using Soft-labels

Figure 1 shows an overview of the proposed approach, which consists of a VAE with an emotional regressor (ER) attached to its bottleneck layer. VAEs [9] are probabilistic graphical models with a network structure similar to an *autoencoder* (AE), where an encoder (ϕ) compresses the high-dimensional input features into a lower dimensional latent representation and a decoder (θ) decompresses the latent representations to reconstruct the input. The difference from an AE is that the latent space has a distributional constraint in the form of a prior, forcing it to learn the parameters of the probability distribution of the inputs. The benefits of this formulation are many, such as the inclusion of domain knowledge about the inputs through the prior, learning a continuous latent space that enables easy random samplings and interpolations, and estimating uncertainty in predictions through MC samplings in the latent space.

If z represents a latent vector that can generate an observation x, then the characteristics of z can be inferred

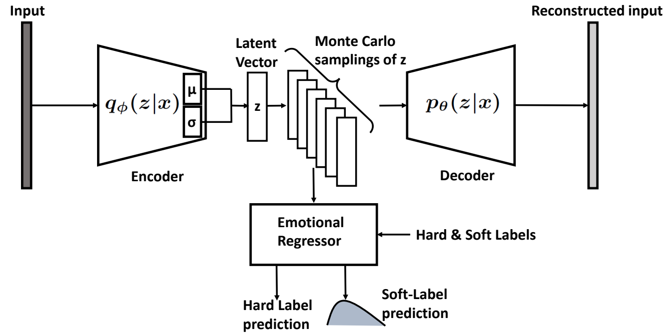


Fig. 1. Proposed approach, which has a VAE and an emotional regressor trained with soft and hard labels.

by computing the posterior probability $p_\theta(z|x)$. Due to the intractable nature of $p_\theta(x)$, it is approximated using variational inference. Here, $p_\theta(z|x)$ is approximated by a tractable distribution $q_\phi(z|x)$ where its parameters can be optimized such that $p_\theta(z|x) \approx q_\phi(z|x)$. This step can be achieved by minimizing the *Kullback-Leibler divergence* (KLD) between $p_\theta(z|x)$ and $q_\phi(z|x)$, leading to the formulation of the variational lower bound as shown in Equation 1,

$$KLD(q_\phi(z|x)||p_\theta(z|x)) - log(p_\theta(x)) = \\ - E_{q_\phi(z|x)} log(p_\theta(x|z)) + KLD(q_\phi(z|x)||p_\theta(z)) \quad (1)$$

where $p_\theta(x|z)$ is the likelihood of the generated data and $p_\theta(z)$ is the prior distribution over the latent variable z. The maximization of the left hand side of Equation 1 (lower bound of the probability of generating data) can be achieved by minimizing $[-E_{q_\phi(z|x)} log(p_\theta(x|z)) + KLD(q_\phi(z|x)||p_\theta(z))]$. The first term denotes the reconstruction likelihood and the second term ensures that the learned distribution $q_\phi(z|x)$ is similar to the true prior distribution $p_\theta(z)$. In this study, we use a VAE with an ER attached to the latent representation layer such that the input to the ER is the sampled vector from the latent space (the same input that the decoder receives). We take the SSL approach to train our model. First, we train the VAE on unlabelled data. Unsupervised pre-training enables the model to learn prior information about the target domain, which helps generalize the model as demonstrated on cross-corpus evaluations. Pre-training the model also provides a better initialization for the model parameters. The pre-trained VAE and ER are jointly trained on the labelled data in a MTL fashion. We use both the consensus and soft-labels for the emotional attributes at the ER, weighing more the loss on the soft-labels (Fig. 1). The information about the confusion between annotators can be better leveraged with soft-label probabilities. Additionally, joint training with soft-labels enables the latent space of the VAE to jointly model the label distribution and the input distribution while backpropagating the gradients. We use consensus-labels to reinforce the gradients and put a stronger constraint on the latent space to model the true label distributions. Equations 2 and 3 show the cost function used for training our model,

$$\mathcal{L}_{unlabelled} = -E_{q_\phi(z|x)} log(p_\theta(x|z)) \quad (2)$$

$$\mathcal{L}_{labelled} = \alpha \cdot (NLL)_{VAE} + \beta \cdot (1 - CCC)_{consensus} + \gamma \cdot (KLD)_{soft} \quad (3)$$

where the Equation 2 corresponds to the *negative log-likelihood* (NLL) loss used for the unsupervised pre-training phase, and the Equation 3 corresponds to the MTL loss function used for the joint training phase. We minimize the (1 - CCC) loss for the consensus-labels and the KLD loss for the soft-labels. We weight the hyperparameters (α, β, γ) such that the losses are in the same scale. We implement a grid-search by varying the values of the hyper-parameters ($(\alpha, \beta, \gamma) \in \{0.1, 0.2, 0.3, \ldots, 2.5\}$). The combination that maximizes the classification loss on the development set is $\alpha = 0.5$, $\beta = 0.8$ and $\gamma = 2$. These weights give more importance to the classification loss on the soft-label (i.e., KLD), since our proposed method derives its main power by learning uncertainties from the soft-labels.

B. Uncertainty Estimation from Soft-labels

The advantage of using soft-labels is that they provide prior knowledge about the label distribution to be fitted, which reduces the search space of the VAE and leads to better generalization. For every input sample, we generate a distribution of soft-label predictions to compute uncertainty in the predictions. We use multiple MC samplings from the bottleneck layer of the VAE to get multiple predictions for a single input sample. With this approach, we approximate the expected log likelihood ($[E_{q_\phi(z|x)} log(p_\theta(x|z))]$) multiple times for each data point in a batch both during training and inference. We average the generated distributions to obtain a single soft-label prediction. We calculate the entropy of the predicted distribution to quantify its uncertainty. We operate on the hypothesis that confidence of the model increases as the entropy in the prediction decreases (e.g., sharper distribution).

C. Implementation Details

The regression of emotional attributes using consensus-labels is used as an auxiliary task. We construct the encoder and decoder of the VAE using three dense layers with 512 nodes per layer. The encoder and decoder networks mirror each other. The purpose of the encoder is to approximate the true posterior distribution $p_\theta(z|x)$. A proxy distribution $q_\phi(z|x)$ is used with its local variational parameters, which is approximated using a multivariate Gaussian distribution with a diagonal covariance matrix parametrized by (μ_ϕ, Σ_ϕ). The mean and log variance of this distribution is specified as the output of the encoder, represented by two fully connected output layers with 256 nodes. We use *rectified linear unit* (ReLU) activations at the encoder, tanh activations at the hidden layers of the decoder, and sigmoid at the output layer. The ER is constructed using four dense layers with 128 nodes per layer and tanh activations at the hidden layers. The prediction layer for the consensus-label is a linear activation layer with a single node. The classification layer for the soft-label is a softmax layer with seven nodes, matching the ground truth emotional attribute scores in the MSP-Podcast corpus that range from 1 to 7. We use dropout in the hidden layers of the VAE and ER. We use a higher dropout rate of p=0.7 for valence and p=0.5 for arousal and dominance, since studies have showed that valence requires higher regularization [29].

An important part of the VAE implementation is the calculation of gradients with respect to the model parameters θ and the variational parameters ϕ (see Sec. IV-A). Due to the intractable nature of these gradients, a MC estimate of the gradients is computed using a reparametrization trick. The random variable representing the latent space of the VAE, $z \sim q_\phi(z|x)$, is expressed as a deterministic transformation of another random variable ϵ and input x such that $z = f_\phi(x, \epsilon)$, where $\epsilon \sim p(\epsilon)$. Hence, the Gaussian approximation of $q_\phi(z|x)$ is achieved using Equation 4.

$$z \sim f_\phi(x, \epsilon) = \mu_\phi(x) + \sigma_\phi(x) \odot \epsilon; \epsilon \sim \mathcal{N}(0, \mathcal{I}) \quad (4)$$

This reparametrization allows for smooth computation of the stochastic gradients by drawing noise samples ϵ from $p(\epsilon)$. For every input sample, we draw 100 MC samples from the latent space for all our experiments.

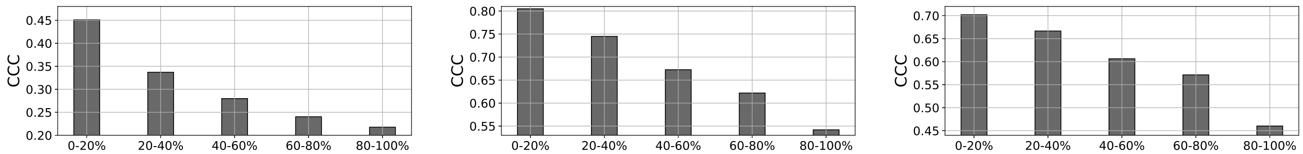
The unlabeled data for the VAE pre-training phase is selected by randomly sampling 10,124 samples (matching the size of the test set) from the unlabeled set of the MSP-Podcast corpus (Sec III-A). We use *adaptive moment estimation* (ADAM) optimizer with a learning rate of 5e-5 for the unsupervised training phase. Then, we attach the ER to the bottleneck representation layer of the VAE and jointly train the VAE-ER model by increasing the learning rate to 3e-4. The input to the network is a 6,373D feature vector (Sec. III-B). We use the KLD loss for the soft-label classification, where we directly compare the predicted distributions to the ground truth soft-labels. We separately train SSL models for arousal, valence and dominance and save the best models based on their performances on the development set. We train all our models using the Keras deep learning library with a Tensorflow backend on a single NVIDIA Quadro P4000 8GB GPU.

V. Experimental Results and Analysis

This section analyzes the results obtained with our proposed method to estimate uncertainty. The SER network and uncertainty prediction networks are separate, but share the same network structure. We create *single-task learning* (STL) SER models by training separate regression models for arousal, valence, and dominance. These models are constructed using a DNN architecture similar to the ER model in our proposed model. Since they have the same structure, we expect that the predicted uncertainties reflect the uncertainties of the SER models. We train the models for 200 epochs, optimizing their performance on the development set.

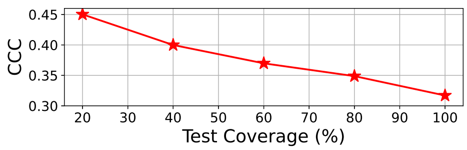
A. Analysis of Uncertainty Predictions

Evaluating uncertainty prediction is not straightforward. This study evaluates the CCC performance in predicting emotional attributes after grouping the samples in the test set according to their uncertainty. We expect that the uncertainty value dictates the performance of the model, where better CCC is achieved for sets with lower uncertainty. We split the test set into five sets with equal number of sentences by sorting the test

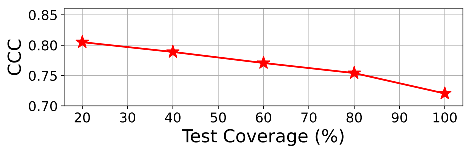


(a) Valence (b) Arousal (c) Dominance

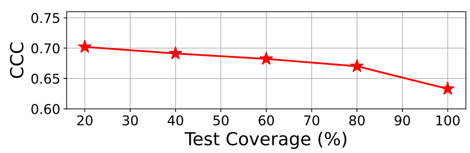
Fig. 2. Performance of regression models on sets with different uncertainty. The first set (0% - 20%) includes samples with the lowest uncertainty, and the fifth set (80%-100%) includes the samples with the highest uncertainty.



(a) Valence



(b) Arousal



(c) Dominance

Fig. 3. Regression performance in CCC using a reject option for emotional attributes. The figures show the tradeoff between coverage and performance.

samples according to their uncertainty scores. The first set has 20% of the samples with the lowest uncertainty scores. The fifth set has 20% of the samples with the highest uncertainty scores (e.g., 80 to 100 percentile in the list). Figure 2 shows the results for this evaluation. We observe improvements in CCC values as the uncertainty of the samples decreases, following our expectation. The ranges of performance are broad, creating important performance gap across sets. The trend is observed in valence, arousal, and dominance. The result shows that our proposed approach to estimate uncertainty is effective for SER problems implemented with regression models.

B. Application in Reject Options

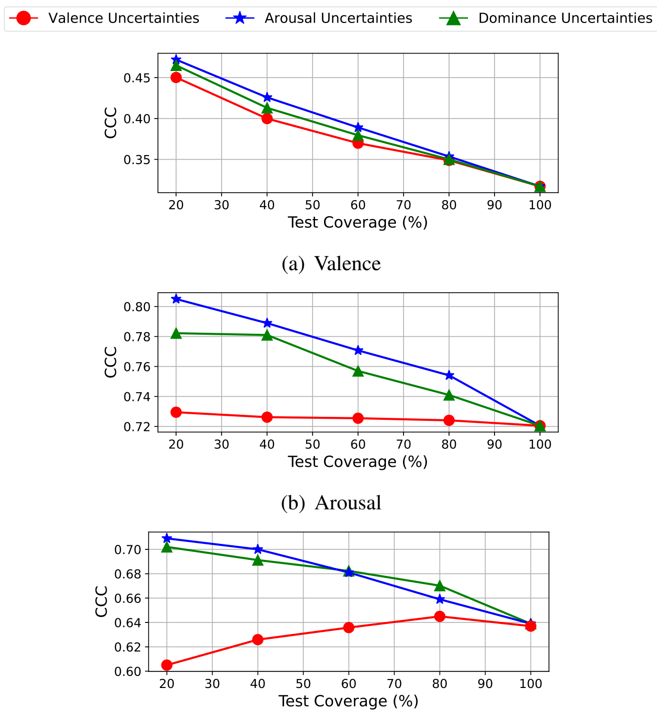
An application area for uncertainty predictions is in selective prediction. We demonstrate the benefits of our model in SER by training regression models with a reject option. With this formulation, the SER system can selectively accept or reject a test sample based on prediction uncertainties. As more uncertain samples are rejected, the performance is expected to improve. However, the coverage decreases as the system provides predictions on fewer samples (i.e., tradeoff between coverage and performance). The goal of a system with a reject option is to obtain the best possible performance while keeping the coverage (accepted samples) as high as possible. Our approach rejects samples with more uncertainty.

The DNN models trained to create the STL baselines are used here for inference. During inference, we accept or reject a test sample based on the uncertainty prediction obtained with our proposed method. Figure 3 shows the tradeoff between the test coverage and CCC performance as we progressively reject samples based on their predicted uncertainties. The results at 100% coverage (baseline performance) correspond to the CCC values achieved on the entire test data. The baseline performances on valence, arousal, and dominance are $CCC_{val} = 0.3170$, $CCC_{aro} = 0.7205$, and $CCC_{dom} = 0.6329$, respectively. We observe clear improvements in CCC as we progressively reject uncertain samples. At 60% test coverage, we achieve relative gains in CCC up to 16.85% for valence, 7.12% for arousal, and 8.01% for dominance.

C. Uncertainty Transfer Learning (UTL)

The baseline CCC values of arousal and dominance are higher than the CCC value for valence. Although we use a higher regularization (dropout rate of $p = 0.7$) to train a DNN for valence prediction [29], estimating valence from speech is inherently difficult [30]. Since the prediction of valence is lower, we expect that the prediction of uncertainty may also be less accurate. Can uncertainty predictions from one emotional attribute (e.g., arousal) be transferred to another emotional attribute (e.g., valence)? We refer to this approach as *uncertainty transfer learning* (UTL). The prediction performance on each attribute is evaluated using a reject option formulation with two different scenarios: (a) self-learned uncertainty, where the prediction uncertainty of an emotional attribute is used for the same attribute (e.g., uncertainty prediction of valence used for reject option on valence), and (b) transferred uncertainty, where the prediction uncertainty of an emotional attribute is used for a different attribute (e.g., uncertainty prediction of arousal used for reject option on valence). The second scenario corresponds to the UTL case. The UTL approach changes the order of the samples to be rejected, directly affecting the performance of the system.

Figure 4 shows the results obtained with the UTL strategy. Uncertainties learned from arousal and dominance improve the recognition of valence, where this trend becomes more prominent as the coverage decreases. The uncertainty information from arousal and dominance help valence more than the uncertainties learned from valence itself, showing the benefits of the UTL approach. At 60% test coverage we achieve relative gains in CCC up to 19.55% for valence using arousal uncertainties. However, UTL is not as useful when the accuracy in predicting the attribute is high. In these cases, the self-learned uncertainties are also expected to be good. Figures 4(b) and 4(c) show that the UTL approach does not



(a) Valence

(b) Arousal

(c) Dominance

Fig. 4. Results with UTL using a reject option. The order of the samples to be rejected is determined with self-learned or transferred uncertainties.

help much with the arousal and dominance performances. The only slight exception is the case of dominance, where arousal uncertainties help in improving dominance predictions after the 60% coverage point (Fig. 4(c)).

D. Model Generalization

This section analyzes the generalization of our proposed method using a reject option formulation with and without UTL on a different speech emotional database. We use the proposed VAE-ER model exclusively trained on the MSP-Podcast corpus, performing inference on the IEMOCAP corpus (i.e., uncertainty estimation). In contrast, the STL SER models are trained with the IEMOCAP data (emotion prediction). The IEMOCAP corpus with acted and improvised speech recordings serves as a good example of domain mismatch between the train and test conditions. The IEMOCAP database has ground truth attribute labels using a 5-point Likert scale and our proposed model has a softmax classification layer with 7 nodes. For the prediction of uncertainty, we do not modify the width of the classification layer to fit the experiments on the IEMOCAP database, since we only do inference on this corpus and its labels are not used for our proposed model. We compare the generalization ability of our method with the following MCD based methods:

- **MCD**: Uses the same model and training procedure proposed in Sridhar and Busso [22]. The model uses the consensus-labels of the emotional attributes.
- **MCD soft**: A DNN implemented using MCD, following the implementation procedure presented in Sridhar and Busso [22], but only trained with soft-labels using the KLD loss.
- **AE-MCD**: An AE model with a MCD decoder and an ER attached to the bottleneck layer. This model is trained in a MTL manner using both consensus and soft-labels. The key differences with our proposed model are: (a) the bottleneck layer of the AE has no distributional constraint, and (b) the decoder network is implemented with MCD. The loss function is the *mean-squared error* (MSE) for the AE and [(1 - CCC) + KLD] for the ER. We pre-train the AE with unlabelled data and jointly train the entire model using the labeled data.
- **MCD MTL**: A DNN implemented using MCD, following the implementation procedure presented in Sridhar and Busso [22]. This model is trained in a MTL manner using both consensus (1 - CCC) and soft-label (KLD) annotations.

TABLE I
CROSS-CORPUS EVALUATION RESULTS ON THE IEMOCAP DATABASE USING SELF-LEARNED UNCERTAINTIES. THE RESULTS SHOW THE CCC VALUES ACHIEVED AT DIFFERENT COVERAGE LEVELS WITH A REJECT OPTION IMPLEMENTED USING OUR PROPOSED UNCERTAINTY PREDICTION APPROACH AND OTHER MCD BASED BASELINES. † INDICATES THAT ONE METHOD LEADS TO SIGNIFICANTLY BETTER RESULTS THAN THE OTHERS.

Attribute	Approach	CCC at different coverage (%)			
		80%	60%	40%	20%
Valence	Proposed	**0.5489**†	**0.5710**†	**0.6122**†	0.6433
	AE-MCD	0.5380	0.5561	0.5918	**0.6450**
	MCD MTL	0.5400	0.5650	0.5888	**0.6492**†
	MCD soft	0.5391	0.5691	0.5820	0.6000
	MCD	0.5340	0.5620	0.5725	0.6181
Arousal	Proposed	**0.7859**†	**0.8022**†	0.8108	0.8180
	AE-MCD	0.7750	0.7920	0.8001	0.7590
	MCD MTL	0.7480	0.7172	0.6509	0.6460
	MCD soft	0.7500	0.7105	0.6823	0.6588
	MCD	0.7830	0.7990	**0.8482**†	**0.8791**†
Dominance	Proposed	**0.6795**†	**0.7201**†	**0.7505**†	0.7653
	AE-MCD	0.6450	0.6409	0.6621	0.6098
	MCD MTL	0.6789	0.7170	0.7412	**0.7808**†
	MCD soft	0.6205	0.6054	0.6011	0.5923
	MCD	0.6790	0.7055	0.7415	0.7487

We calculate the entropy of the predicted distribution to quantify the uncertainty in predictions whenever soft-labels are used to train the different models (proposed approach, AE-MCD, MCD MTL, and MCD soft). For the MCD model, we quantify uncertainty using the standard deviations of the different dropout results, following the uncertainty estimation procedure presented in Sridhar and Busso [22]. We implement the reject options based on the predicted uncertainties achieved from different methods and compare their performances. The SER model is a DNN constructed with a similar architecture as our ER. We create five sets of train-test partitions of the IEMOCAP dataset, using a five-fold cross-validation strategy to train the STL SER model. In each fold, we consider two speakers as the test speakers and the rest as train speakers. The final results are averaged across the 5-folds. We estimate the significance of the results using one-tailed t-test over 10 trials, asserting significance when p-value$\leq$ 0.05.

Table I shows the results obtained on the IEMOCAP database using the self-learned uncertainties for reject options. The baseline CCC values achieved on all the test

TABLE II

CROSS-CORPUS EVALUATION RESULTS ON THE IEMOCAP DATABASE USING THE UTL STRATEGY. † INDICATES THAT ONE APPROACH LEADS TO SIGNIFICANTLY BETTER RESULTS THAN THE OTHERS.

Attribute-Uncertainty Pairs	Approach	CCC at different coverage (%)			
		80%	60%	40%	20%
Val-Aro	Proposed	**0.5500**†	**0.5929**†	**0.6292**†	**0.6558**†
	AE-MCD	0.5420	0.5650	0.5900	0.6329
	MCD MTL	0.5322	0.5690	0.5899	0.5980
	MCD soft	0.5480	0.5710	0.5902	0.5981
	MCD	0.5280	0.5315	0.5428	0.5750
Val-Dom	Proposed	**0.5498**†	**0.5717**†	**0.6073**†	**0.6490**†
	AE-MCD	0.5411	0.5650	0.5881	0.6126
	MCD MTL	0.5325	0.5350	0.5510	0.5738
	MCD soft	0.5450	0.5680	0.5838	0.5811
	MCD	0.5255	0.5364	0.5427	0.5833

TABLE III

ABLATION STUDY: A, B AND C REPRESENT DIFFERENT KEY COMPONENTS OF THE PROPOSED MODEL. A: VAE, B: USE OF SOFT-LABELS AT THE ER, AND C: NEED FOR MC SAMPLING AT THE LATENT SPACE OF VAE.

				CCC at different coverage (%)				
Attributes	A	B	C	100%	80%	60%	40%	20%
Aro	✓	✓	100	0.720	**0.754**	**0.770**	**0.788**	**0.805**
	✓	-	100	0.720	0.741	0.758	0.770	0.791
	✓	✓	1	0.720	0.731	0.739	0.745	0.750
Val	✓	✓	100	0.317	**0.348**	**0.369**	**0.399**	**0.450**
	✓	-	100	0.317	0.325	0.348	0.366	0.372
	✓	✓	1	0.317	0.319	0.325	0.333	0.348
Dom	✓	✓	100	0.632	**0.670**	**0.682**	**0.699**	**0.702**
	✓	-	100	0.632	0.652	0.655	0.659	0.675
	✓	✓	1	0.632	0.639	0.643	0.651	0.658

samples (100% coverage) are $CCC_{val} = 0.5224$ for valence, $CCC_{aro} = 0.7661$ for arousal, and $CCC_{dom} = 0.6439$ for dominance. Our proposed method performs significantly better than the MCD-based approaches for the coverage at 80% and 60%, which are the most important cases for a reject option formulation, improving the performance without compromising too much coverage. For the 40% coverage, only the MCD approach is able to achieve better performance than our approach, but only for arousal.

Based on results shown in Section V-C, we also evaluate if the UTL approach leads to better performance for valence in cross-corpus evaluations. We report the results on valence alone, since we do not observe improvements with UTL for arousal and dominance. Table II shows the results obtained on the IEMOCAP database using the UTL approach. The first column indicates the attribute used for the transferred uncertainties. For example, *Val-Aro* indicates the prediction performances on valence using uncertainties learned from arousal. We see that the UTL approach works best with our proposed method, achieving significantly higher CCC scores as we reject more ambiguous samples. The results in Table I and II show the generalizability of our proposed method and the advantages of the UTL approach.

E. Ablation Study

We evaluate different contributing factors of the model architecture with an ablation study for within corpus experiments (on the MSP-Podcast corpus) with self-learned uncertainties using a reject option formulation. We evaluate the performance after removing: (A) the VAE (removing the distributional constraint at the bottleneck, replacing it with a simple AE), (B) the soft-labels to train the ER, and (C) the need for MC sampling at the latent space of the VAE. Table III shows the results, which indicate that removing one of these components lead to performance drops between 1% and 5% (absolute) compared to the full system.

F. Computational Resources During Inference

A SER model should be efficient at inference time, not requiring enormous computational memory or time to provide a solution. To evaluate the complexity of our proposed method, we calculate the average time that our proposed model takes for inference on the entire test set of the MSP-Podcast corpus, comparing it with that of the MCD model trained with consensus-labels [22]. We take the average over 10 trials with different random initializations of the parameters of the models. On average, our method takes 11.39s for inference whereas the MCD model takes 19.86s. This indicates that our method is 74.36% faster than the MCD approach. Even though our proposed generative model (7,519,341 parameters) has more parameters than the MCD model (865,537 parameters), this result shows that it is more computationally effective than the MCD method during inference. The inference time for other approaches used in Table I and II are 21.26s for MCD soft, 41.38s for AE-MCD, and 26.32s for MCD MT.

VI. CONCLUSIONS

This study presented a novel approach for using soft-labels of emotional attributes in uncertainty prediction. We use soft-labels of emotional attributes to incorporate the confusion between annotators and estimate uncertainty using a VAE-ER model trained using SSL. We evaluate the application of uncertainty prediction in SER as a reject option problem, reporting the tradeoff between performance and test coverage. At a test coverage of 60%, we achieve relative gains in prediction performance in terms of CCC values up to 16.85% for valence, 7.12% for arousal, and 8.01% for dominance. We proposed a UTL strategy where prediction uncertainties are transferred across emotional attributes. This approach is particularly useful for improving valence uncertainty, where SER models achieve lower prediction performance than arousal or dominance. The analysis demonstrated that transfer learning with uncertainties is feasible in SER problems. We improved the performance gains up to 19.55% on valence at 60% test coverage. The results on cross-corpus evaluations showed that our proposed method generalizes better than other MCD approaches to estimate uncertainty. We also discussed the efficiency of our method in terms of computational complexity during inference, showing that our approach is significantly faster than other MCD approaches. As a future work, we will explore using these ideas in curriculum learning [8]. We expect improvements by defining the curriculum according to the predicted uncertainties of the training samples.

References

[1] R. Lotfian and C. Busso, "Formulating emotion perception as a probabilistic model with application to categorical emotion classification," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2017)*, San Antonio, TX, USA, October 2017, pp. 415–420.

[2] H. M. Fayek, M. Lech, and L. Cavedon, "Modeling subjectiveness in emotion recognition with deep neural networks: Ensembles vs soft labels," in *International Joint Conference on Neural Networks (IJCNN 2016)*, Vancouver, BC, Canada, July 2016, pp. 566–570.

[3] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. Mower Provost, "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, January-March 2017.

[4] M. Abdelwahab and C. Busso, "Incremental adaptation using active learning for acoustic emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, New Orleans, LA, USA, March 2017, pp. 5160–5164.

[5] ——, "Ensemble feature selection for domain adaptation in speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, New Orleans, LA, USA, March 2017, pp. 5000–5004.

[6] ——, "Active learning for speech emotion recognition using deep neural network," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2019)*, Cambridge, UK, September 2019, pp. 441–447.

[7] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the eleventh annual conference on Computational learning theory (COLT 1998)*, Madison, WI, USA, July 1998, pp. 92–100.

[8] R. Lotfian and C. Busso, "Curriculum learning for speech emotion recognition from crowdsourced labels," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 4, pp. 815–826, April 2019.

[9] D. Kingma and M. Welling, "Auto-encoding variational bayes," *ArXiv e-prints (arXiv:1312.6114)*, pp. 1–14, December 2013.

[10] E. Mower, M. Mataric, and S. Narayanan, "A framework for automatic human emotion classification using emotional profiles," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 5, pp. 1057–1070, May 2011.

[11] L. Tarantino, P. Garner, and A. Lazaridis, "Self-attention for speech emotion recognition," in *Interspeech 2019*, Graz, Austria, September 2019, pp. 2578–2582.

[12] Y. Kim and J. Kim, "Human-like emotion recognition: Multi-label learning from noisy labeled audio-visual expressive speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, Calgary, AB, Canada, April 2018, pp. 5104–5108.

[13] B. Zhang, Y. Kong, G. Essl, and E. Mower Provost, "f-similarity preservation loss for soft labels: A demonstration on cross-corpus speech emotion recognition," in *AAAI Conference on Artificial Intelligence (AAAI 2019)*, vol. 33, Honolulu, HI, USA, January-February 2019, pp. 5725–5732.

[14] J. Gideon, M. McInnis, and E. Mower Provost, "Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (ADDoG)," *IEEE Transactions on Affective Computing*, 2020.

[15] X. Cai, D. Dai, Z. Wu, X. Li, J. Li, and H. Meng, "Emotion controllable speech synthesis using emotion-unlabeled dataset with the assistance of cross-domain speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2021)*, Toronto, ON, Canada, June 2021, pp. 5734–5738.

[16] V. Sethu, E. M. Provost, J. Epps, C. Busso, N. Cummins, and S. Narayanan, "The ambiguous world of emotion representation," *ArXiv e-prints (arXiv:1909.00360)*, pp. 1–19, May 2019.

[17] J. Deng and B. Schuller, "Confidence measures in speech emotion recognition based on semi-supervised learning," in *Interspeech 2012*, Portland, OR, USA, September 2012, pp. 2226–2229.

[18] S. Steidl, M. Levit, A. Batliner, E. Nöth, and H. Niemann, ""Of all things the measure is man" automatic classification of emotions and inter-labeler consistency," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)*, vol. 1, Philadelphia, PA, USA, March 2005, pp. 317–320.

[19] H.-C. Chou and C.-C. Lee, "Every rating matters: Joint learning of subjective labels and individual annotators for speech emotion classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019)*, Brighton, UK, May 2019, pp. 5886–5890.

[20] J. Han, Z. Zhang, M. Schmitt, M. Pantic, and B. Schuller, "From hard to soft: Towards more human-like emotion recognition by modelling the perception uncertainty," in *ACM international conference on Multimedia (MM 2017)*, Mountain View, CA, USA, October 2017, pp. 890–897.

[21] K. Sridhar and C. Busso, "Speech emotion recognition with a reject option," in *Interspeech 2019*, Graz, Austria, September 2019, pp. 3272–3276.

[22] ——, "Modeling uncertainty in predicting emotional attributes from spontaneous speech," in *IEEE international conference on acoustics, speech and signal processing (ICASSP 2020)*, Barcelona, Spain, May 2020, pp. 8384–8388.

[23] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, October-December 2019.

[24] S. Mariooryad, R. Lotfian, and C. Busso, "Building a naturalistic emotional speech corpus by retrieving expressive behaviors from existing speech corpora," in *Interspeech 2014*, Singapore, September 2014, pp. 238–242.

[25] A. Burmania, S. Parthasarathy, and C. Busso, "Increasing the reliability of crowdsourcing evaluations using online quality assessment," *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 374–388, October-December 2016.

[26] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, December 2008.

[27] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Interspeech 2013*, Lyon, France, August 2013, pp. 148–152.

[28] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: the Munich versatile and fast open-source audio feature extractor," in *ACM International conference on Multimedia (MM 2010)*, Florence, Italy, October 2010, pp. 1459–1462.

[29] K. Sridhar, S. Parthasarathy, and C. Busso, "Role of regularization in the prediction of valence from speech," in *Interspeech 2018*, Hyderabad, India, September 2018, pp. 941–945.

[30] C. Busso and T. Rahman, "Unveiling the acoustic properties that describe the valence dimension," in *Interspeech 2012*, Portland, OR, USA, September 2012, pp. 1179–1182.