MODELING UNCERTAINTY IN PREDICTING EMOTIONAL ATTRIBUTES FROM SPONTANEOUS SPEECH

Kusha Sridhar and Carlos Busso

Multimodal Signal Processing (MSP) laboratory, Department of Electrical and Computer Engineering
The University of Texas at Dallas, Richardson TX 75080, USA

Kusha.Sridhar@utdallas.edu, busso@utdallas.edu

ABSTRACT

A challenging task in affective computing is to build reliable *speech emotion recognition* (SER) systems that can accurately predict emotional attributes from spontaneous speech. To increase the trust in these SER systems, it is important to predict not only their accuracy, but also their confidence. An intriguing approach to predict uncertainty is *Monte Carlo* (MC) dropout, which obtains predictions from multiple feed-forward passes through a *deep neural network* (DNN) by using dropout regularization in both training and inference. This study evaluates this approach with regression models to predict emotional attribute scores for valence, arousal and dominance. The analysis illustrates that predicting uncertainty in this problem is possible, where the performance is higher for samples in the test set with lower uncertainty. The study evaluates uncertainty estimation as a function of the emotional attributes, showing that samples with extreme values have lower uncertainty. Finally, we demonstrate the benefits of uncertainty estimation with reject option, where a classifier can decline to give a prediction when its confidence is low. By rejecting only 25% of the test set with the highest uncertainty, we achieve relative performance gains of 7.34% for arousal, 13.73% for valence and 8.79% for dominance.

Index Terms— Speech Emotion Recognition, Monte Carlo dropout, activation functions, reject option.

1. INTRODUCTION

Emotion is externalized in speech affecting several acoustic properties [1]. We can easily decode emotional cues on speech during human interaction, which helps us to infer the emotional and cognitive state of others. This information shapes the way in which we communicate, influencing our decision-making process [2]. *Speech emotion recognition* (SER) systems aim to mimic our emotional skills to identify emotional cues that can be used to predict perceived emotions. Designing a robust SER system can have wide applications in healthcare, education, security and defense and *human computer interactions* (HCIs). A challenging problem in SER is the modeling of spontaneous human interactions in everyday life, which involve complex emotional behaviors [3–5]. Understanding and quantifying the degree of certainty in the predictions of a SER system is important to increase the trust in systems.

There are several advantages of knowing the uncertainty in SER predictions. Several applications and formulations in affective computing become possible when a SER model is aware of what it does not know. Such models are useful for critical applications such as security and healthcare, where incorrect predictions have important consequences. Uncertainty prediction facilitates human-in-the-loop solutions, where only uncertain cases are carefully reviewed by a person. Uncertainty prediction can also facilitate machine learning solutions for semi-supervised and unsupervised algorithms. For example, methods such as active learning [6–8] can use uncertainty prediction to identify unlabeled samples that maximize the performance of a classifier after obtaining their corresponding labels. Uncertainty prediction can also be useful for methods such as co-training [9], where unlabelled data with confident predictions from multi-view training are used to augment the labeled data. It is also useful for curriculum learning [10], where the training set is presented in order, starting from easy samples and ending with difficult samples. The difficulty of a sample can be inferred from uncertainty.

Ambiguous emotional content is common in spontaneous speech, where the performance of a SER system can be low [1]. It is difficult to predict uncertainty if the task leads to low performance. Our proposed approach to target this problem is based on a technique called *Monte Carlo* (MC) dropout, which explores the ability of *deep neural networks* (DNNs) to capture model uncertainty [11]. Gal et al. [11] formulated dropout regularization of DNNs as an approximation to Bayesian inference in deep Gaussian processes. The network is evaluated multiple times with different dropout configurations, creating a distribution of predictions for each sample. Multiple iterations through a network with dropout are analogous to obtaining predictions from an ensemble of thinner networks. This study investigates this technique in SER systems, applied to the prediction of emotional attributes. We observe that samples with higher uncertainty were predicted with valence, arousal and dominance scores in the middle of their scales. We also observe lower performance for samples with higher uncertainty, showing that MC dropout is useful for regression problems in SER.

We demonstrate the use of uncertainty prediction in the context of reject option for SER problems, where the goal is to allow a SER system to decline a prediction when its confidence is low. Our previous study explores reject option for categorical emotion classification [12]. This study demonstrates this principle in regression problems in predicting valence, arousal and dominance scores. We accept or reject a sample based on the uncertainty predicted by the MC dropout method. We evaluate the performance of the model by studying the tradeoff between test coverage (i.e., number of accepted test samples) and performance, measured with *concordance correlation coefficient* (CCC). With 'tanh' as the activation function, we achieve relative gains in CCC up to 7.34% for arousal, 13.73% for valence and 8.79% for dominance for a 75% test coverage (i.e., reject 25% of the samples). These results show the benefits of estimating uncertainty, and its role in increasing the reliability in SER.

2. RELATED WORK

Prediction of emotional attributes with high precision is a challenging task. In recent years, previous studies have made important advances in this area using domain adversarial methods [13], multi-

This work was supported by NSF under Grant CNS-1823166 and CAREER Grant IIS-1453781.

task learning [14], and semi-supervised methods such as ladder networks [15, 16]. However, the accuracies of SER systems in spontaneous speech are still low for several applications, especially for valence where the accuracies of speech-based systems are particularly low [17, 18]. While we build the infrastructure and improve the SER models, it can be useful to formulate this problem with an alternative approach, where a SER system provides not only its predictions, but also its confidence.

Few studies in SER have predicted or modeled confidence measures. Deng et al. [19] derived confidence measures based on human labeler agreement to build emotion scoring models. They showed that the fusion of these scores correlate well with the unweighted average recall of the classifier for a predicted emotion state. Deng et al. [20] used a semi-supervised approach to include data from the target domain into training based on confidence levels obtained on the target data through multi-corpora training. Another approach that relies on uncertainty prediction is the reject option framework, where a classifier can decline a prediction when its confidence is low. Reject options have been used in machine learning, but mostly on classification problems [21–23]. Our previous work focused on applying reject options to emotion classification using DNNs [12]. We used two criteria to accept or reject a sample: (1) a threshold learned using the softmax response, and (2) a threshold on the difference between the two highest predictions of the softmax output under a risk minimization framework. Applying reject option to regression problems is less common, since measuring uncertainty is less intuitive.

Our study evaluates uncertainty prediction using MC dropout, which was designed by Gal et al. [11] to approximate a Gaussian process by placing a distribution over the weights of a DNN using dropout regularization. MC dropout has achieved competitive performance on image classification [24, 25] and simple regression tasks [25]. Dey et al. [26] used MC dropout to capture uncertainty in text transcriptions generated by an *automatic speech recognition* (ASR) system, selecting the best hypothesized outcome. Vyas et al. [27] used uncertainties on *word error rates* (WER) to perform ASR, using the uncertainty predictions to localize errors. Abdelwahab et al. [8] adopted the MC dropout technique in SER as a sampling method for active learning. They sampled unlabeled data based on their posterior probability estimates to train an autoencoder, using its bottleneck features along with a small amount of labeled data to predict the emotions. Our study explores the use of MC dropout in SER, focusing on regression problems. We illustrate the benefits of uncertainty prediction in SER by designing a regression model with reject option.

3. RESOURCES

3.1. The MSP-Podcast Database

This study uses the MSP-Podcast corpus [28], which consists of spontaneous speech from various audio-sharing websites collected using the strategy suggested in Mariooryad et al. [29]. The content of the recordings are diverse with broad topics discussing areas such as academics, sports, health, art, entertainment, and politics. This is a naturalistic speech database with rich emotional content. We use a diarization toolkit to identify distinct speaker segments from the recordings. A number of pre-processing steps are applied to these turns to obtain clean speech from a single speaker with no overlap or background music, with durations between 2.75 and 11s [28].

The data collection is an ongoing effort in our laboratory, where this study uses version 1.4 of the corpus. This version consists of about 56 hours of speech (33,262 speaking turns). We annotate the corpus with emotional labels using categorical and attribute-based descriptors using a crowdsourcing protocol. This study uses the emotional attributes valence (negative versus positive), arousal (calm versus active), and dominance (weak versus strong) annotated with a seven-Likert-type scales. The perceptual evaluation uses a modified version of the crowdsourcing protocol presented in Burmania et al. [30] to track the performance of the annotators in real-time. Each speaking turn is annotated by at least 5 annotators. In the database, we manually identified 703 speakers. The dataset is partitioned into train, validation, and test sets. The test set consists of 9,255 sentences from 50 speakers. The validation set has 4,300 sentences from 30 speakers. The training set has 19,707 sentences from the rest of the speakers, including the sentences without speaker information. This dataset partition aims to create speaker-independent splits, where data from one speaker is exclusively contained in only one of the sets.

3.2. Acoustic Features

This study uses the Interspeech 2013 *computational paralinguistics challenge* (ComParE) [31] feature set, extracted with the OpenSmile toolkit [32]. For each speaking turn, the toolkit extracts *low level descriptors* (LLDs) using 20ms windows. The LLDs consist of frame level features such as energy, fundamental frequency and *Mel-frequency cepstral coefficients* (MFCCs). Then, sentence-level statistics are calculated over the LLDs (e.g., mean and standard deviation of the energy). These statistics are called *high level descriptors* (HLDs). There are a total of 6,373 HLDs extracted using this approach, regardless of the duration of the speaking turn.

4. METHODOLOGY

This study explores uncertainty prediction for SER problems. We focus our study on regression models, where the goal is to predict emotional attributes (valence, arousal and dominance). The proposed approach relies on MC dropout, which this section describes.

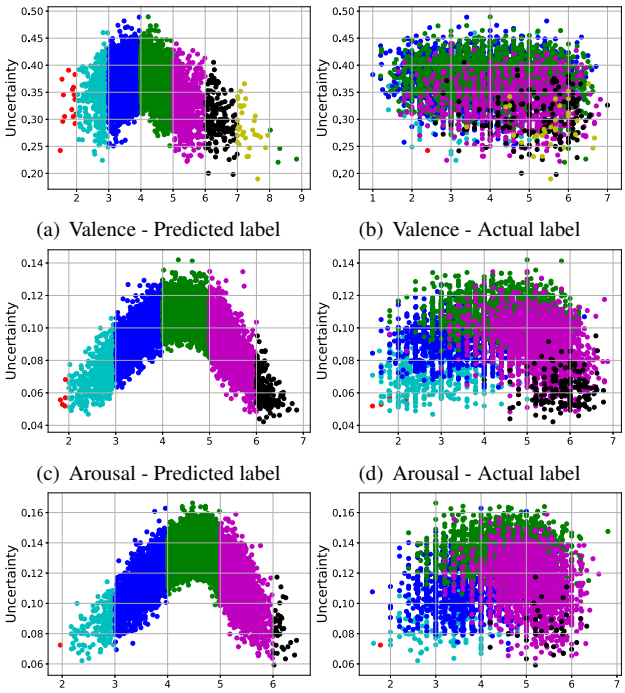
4.1. Monte Carlo Dropout

If X represents the training data and ω the parameters of a model, then the posterior predictive distribution for a test sample x_{test}, is given by

$$p(x_{test}|X) \approx \int p(x_{test}|\omega)p(\omega|X)d\omega \quad (1)$$

The integral in Equation 1 is intractable, since we do not know the posterior probability $p(\omega|X)$. Therefore, this expression can be approximated with sampling methods such as *Markov chain monte carlo* (MCMC) and *variational inference* (VI). A specific type of VI for DNNs consists of placing a distribution over the weights. We obtain different predictions by stochastically changing the weights during training, which can be used to estimate uncertainty. MC dropout [11] is an appealing approach to implement VI for DNNs without having to increase the number of hyper-parameters, changing the simplifying assumptions and structure of the network, or increasing the computational cost. We obtain a distribution of the estimations for each new sample by sampling from the predictive posterior distribution.

Starting with a simple DNN, we only have to add a prior distribution (l2 regularization term) over its weights, and use dropout as regularization to stochastically change the weight during the training process. Dropout has to be used both during training and testing the models. This procedure is mathematically equivalent to solving the intractable integral in Equation 1 using MC integration. A detailed explanation of the method can be found in Gal et al. [11]. For a given sample in the test set, the MC dropout approach computes unbiased

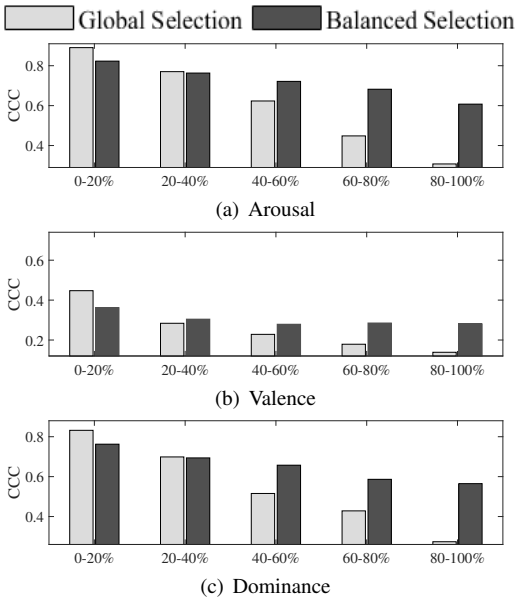


(a) Valence - Predicted label (b) Valence - Actual label

(c) Arousal - Predicted label (d) Arousal - Actual label

(e) Dominance - Predicted label (f) Dominance - Actual label

Fig. 1. Scatter plot showing the uncertainty as a function of the emotional attributes. The results are provided in terms of predicted and actual scores. We assign colors to the samples in bins with similar predicted scores for better visualization (best viewed in color).



(a) Arousal

(b) Valence

(c) Dominance

Fig. 2. Performance of regression models on sets with different uncertainty. The first set (0% - 20%) includes the samples with the lowest uncertainty, and the fifth set (80%-100%) the samples with the most uncertainty. The results are presented with global and balanced selection (Sec. 5).

estimates for the mean and variance of its predictive posterior probability. In the derivation in Gal et al. [11], the only additional parameter is λ, which is a regularization term for the weights. Equation 2 gives the formula for λ, which depends on the dropout probability p, the prior length-scale l usually set to 0.01, the number of data samples in the training set N, and the model precision τ.

$$\lambda = \frac{(1-p)l^2}{2N\tau} \tag{2}$$

We obtain the optimal values for p and τ using a grid-search approach over a list of values ($p \in \{0.05, 0.1, 0.2, \ldots, 0.9\}$; $\tau \in \{0.025, 0.05, 0.1, 0.2, \ldots, 0.5\}$), choosing the configuration with the highest log-likelihood value over the validation set.

4.2. Implementation Details

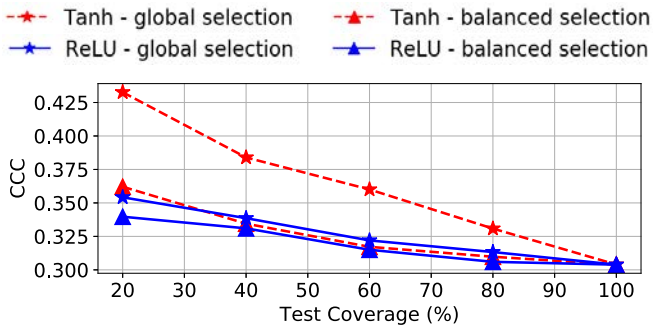
The prediction of emotional attributes is formulated as a regression problem implemented with DNNs. We use three dense layers with 512 nodes per layer (adding more layers does not necessarily lead to better SER performance [33]). We use *stochastic gradient descent* (SGD) with a learning rate of 0.001 to optimize the parameters of the network. We use the cost function $\mathcal{L} = (1 - CCC)$, where the goal is to minimize its value. The input to the network is a 6,373D feature vector (Sec. 3.2). We use a linear activation at the output layer with a single node for regression predictions. We normalize the features using the mean and standard deviation values estimated over the training samples.

5. ANALYSIS OF UNCERTAINTY PREDICTION

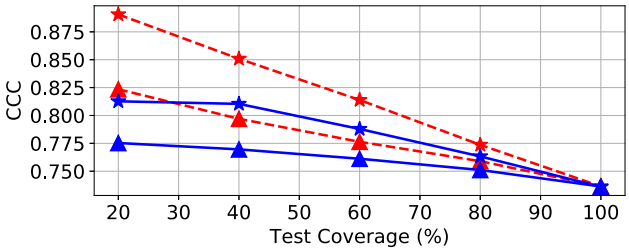
This section analyzes the uncertainty prediction results obtained with MC dropout. First, we train the models with dropout and weight regularization (Sec. 4.1), obtaining predictions on the test samples with their corresponding uncertainties. Based on the optimal parameters obtained from the grid search (Sec. 4.1), we select a dropout of $p = 0.4$ for arousal and dominance, and $p = 0.6$ for valence. These values are used for both training and inference. We also use L2 regularization on the weights of the hidden layers. The experiments are implemented with 'tanh' as the activation function.

After estimating the uncertainty of each sentence in the test set, we create a scatter plot showing uncertainty as a function of the emotional attribute. Figure 1 shows the results when we consider the predicted and actual labels. For better visualization, we create uniform bins using the predicted scores, assigning consistent colors for each bin. The first observation from Figure 1 is that samples predicted with emotional attribute scores in the middle have higher uncertainty. This result is consistently observed for valence (Fig. 1(a)), arousal (Fig. 1(c)), and dominance (Fig. 1(e)). Samples in the extreme have a stronger emotional content that our regression models can reliably predict. Samples with more neutral scores (i.e., values in the middle) include sentences with more ambiguous emotional content, increasing its uncertainty. For arousal and dominance, the trends for the predicted labels are similar to the trends in the scatter plots displaying the actual labels. The membership of the samples in each bin between predicted and actual labels are fairly consistent as the same sequence of colors is observed. This result shows that the performance for our regression models is high for arousal ($CCC_{aro} = 0.736$) and dominance ($CCC_{dom} = 0.668$). For valence, we observe that the scatter plots with the predicted (Fig. 1(b)) and real (Fig. 1(a) labels do not have the same trend, showing the challenges in predicting valence with acoustic features [17, 18] ($CCC_{val} = 0.304$).

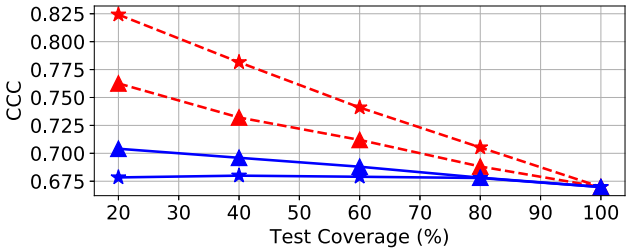
To understand whether the MC dropout approach is useful in regression problems for SER, we evaluate the CCC values for datasets with different uncertainty values. We expect that the performance of the model on these sets depends on their uncertainty value, where better CCC is achieved for sets with lower uncertainty. We split the test set into five sets following two alternative sample selection rules.



(a) Valence



(b) Arousal



(c) Dominance

Fig. 3. Regression models with reject option. The figures show the tradeoff between coverage and performance. At 100% coverage, the results are obtained using the entire test set.

The first rule is the *global selection*, where the sets are directly created by sorting the test data according to their uncertainty score. The first set has 20% of the samples with the lowest uncertainty scores. The fifth set has all the samples between the 80 and 100 percentile (i.e., most uncertain samples). The second rule is *balanced selection*, where we attempt to balance the sets in terms of emotional attributes. We rely on the predicted emotional scores, since the true labels in the test set are assumed to be hidden. We use the bins defined in Figures 1(a) (valence), 1(c) (arousal), and 1(e) (dominance). The first set includes the top 20% of the samples with the lowest uncertainty on each bin. We continue this approach until the fifth set, which includes 20% of the most uncertain samples on each bin. By selecting the data per bin, the sets are emotionally balanced replicating the overall distribution of the data. Figure 2 shows the results for global and balanced selections. We observe improved CCC values as the uncertainty of the samples decreases, following our expectation. This result is consistent with the two sample selection rules. The ranges of performance are broader for global selection, creating an important performance gap across sets. The trend is also observed for balanced selection, especially for arousal and dominance. The result shows that the MC dropout approach is effective for SER problems implemented with regression models.

6. APPLICATION IN REJECT OPTION FOR SER

An important area where uncertainty prediction can be used is in training a regression model with reject option. In this formulation, the SER system exercises the option to accept or reject a test sample based on uncertainty prediction. Rejecting ambiguous samples reduces the coverage in the test set, but improves performance of the system (i.e., the tradeoff between coverage and performance). This approach is ideal for human-in-the-loop applications.

We train a DNN model for 200 epochs, optimizing its performance on the validation set. The regression models are implemented with a dropout rate of $p = 0.5$ for all the emotional attributes, without weight regularization. We train the models with dropout, but we do not use it during inference. During inference, we accept or reject a test sample based on the uncertainty prediction obtained with MC dropout. We consider the global and balanced selection criteria (Sec. 5). We implement the regression models using different activation functions including tanh, sigmoid, *rectified linear unit* (ReLU), Leaky ReLU and *exponential linear unit* (ELU). However, we observed that not all the activation functions were as effective in capturing uncertainty using MC dropout. Our preliminary analysis showed that tanh and ReLU provided the best results, so this section only reports results with these activation functions.

Figure 3 shows the performance of our regression models with a reject option. The figures show the tradeoff between coverage and performance. The results at 100% coverage correspond to the CCC achieved on the entire test data. These values are the baseline results. As we reject uncertain samples, following the results of the MC dropout, we observe clear improvements in CCC. We observe that the reject option implemented with balanced selection leads to lower performance compared to global selection (the exception is dominance implemented with ReLU). For this application, it is better to reject samples without attempting to balance their emotional content. When we compare the performance obtained with tanh and ReLU, we observe that tanh always leads to better performance. At 75% test coverage and using tanh as the activation function, we achieve relative gains in CCC up to 7.34% for arousal, 13.73% for valence and 8.79% for dominance. These gains are possible by only rejecting 25% of the most uncertain samples, increasing the confidence of the system for samples that the regression model decides to accept.

7. CONCLUSIONS

This study explored uncertainty prediction in regression problems for attribute-based descriptors. We estimated uncertainty using MC dropout, analyzing the results as a function of the emotional attribute scores. We observed that the confidence of SER models for samples with emotional values in the middle (e.g., more neutral emotions) is lower than samples with extreme values. The analysis demonstrated that uncertainty prediction is feasible in SER problems, showing that regression results are better for samples with higher confidence. We observed a monotonic decrease in regression performance as the uncertainty increases, suggesting that the proposed criteria were effective to quantify uncertainty. The study evaluated the use of MC dropout in the context of reject options to demonstrate the benefits of estimating uncertainty. The results show that we can improve the regression predictions without compromising much the test coverage. When the coverage is 75% of the test set, the relative gains in CCC were up to 7.34% for arousal, 13.73% for valence and 8.79% for dominance. Our future work includes understanding better the impact of different activation functions in the estimation of uncertainty. We also want to explore the use of uncertainty prediction for curriculum learning, semi-supervised learning, and active learning.

8. REFERENCES

[1] C. Busso, M. Bulut, and S.S. Narayanan, "Toward effective automatic recognition systems of emotion in speech," in *Social emotions in nature and artifact: emotions in human and human-computer interaction*, J. Gratch and S. Marsella, Eds., pp. 110–127. Oxford University Press, New York, NY, USA, November 2013.

[2] R. Picard, *Affective Computing*, MIT Press, Cambridge, MA, USA, 1997.

[3] E. Mower, A. Metallinou, C.-C. Lee, A. Kazemzadeh, C. Busso, S. Lee, and S.S. Narayanan, "Interpreting ambiguous emotional expressions," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2009)*, Amsterdam, The Netherlands, September 2009, pp. 1–8.

[4] R. Cowie and R.R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech Communication*, vol. 40, no. 1-2, pp. 5–32, April 2003.

[5] L. Devillers, L. Vidrascu, and L. Lamel, "Challenges in real-life emotion annotation and machine learning based detection," *Neural Networks*, vol. 18, no. 4, pp. 407–422, May 2005.

[6] 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.

[7] M. Abdelwahab and C. Busso, "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.

[8] M. Abdelwahab and C. Busso, "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.

[9] 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.

[10] 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.

[11] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *International Conference on Machine Learning (ICML 2016)*, New York, NY, USA, June 2016, pp. 1050–1059.

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

[13] M. Abdelwahab and C. Busso, "Domain adversarial for acoustic emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2423–2435, December 2018.

[14] S. Parthasarathy and C. Busso, "Jointly predicting arousal, valence and dominance with multi-task learning," in *Interspeech 2017*, Stockholm, Sweden, August 2017, pp. 1103–1107.

[15] S. Parthasarathy and C. Busso, "Ladder networks for emotion recognition: Using unsupervised auxiliary tasks to improve predictions of emotional attributes," in *Interspeech 2018*, Hyderabad, India, September 2018, pp. 3698–3702.

[16] S. Parthasarathy and C. Busso, "Semi-supervised speech emotion recognition with ladder networks," *ArXiv e-prints (arXiv:1905.02921)*, pp. 1–13, May 2019.

[17] 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.

[18] 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.

[19] J. Deng, W. Han, and B. Schuller, "Confidence measures for speech emotion recognition: A start," in *ITG Conference on Speech Communication*, Braunschweig, Germany, September 2012, pp. 1–4.

[20] 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.

[21] S. Kang, S. Cho, S.-J Rhee, and K.-S. Yu, "Reliable prediction of anti-diabetic drug failure using a reject option," *Pattern Analysis and Applications*, vol. 20, no. 3, pp. 883–891, August 2017.

[22] N. Hatami and C. Chira, "Classifiers with a reject option for early time-series classification," in *IEEE Symposium on Computational Intelligence and Ensemble Learning (CIEL 2013)*, Singapore, September 2013, pp. 9–16.

[23] R. El-Yaniv and Y. Wiener, "On the foundations of noise-free selective classification," *Journal of Machine Learning Research*, vol. 11, pp. 1605–1641, May 2010.

[24] Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks," in *In Advances in Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, December 2017, pp. 4878–4887.

[25] Y. Geifman and R. El-Yaniv, "SelectiveNet: A deep neural network with an integrated reject option," in *International Conference on Machine Learning (ICML 2019)*, Long Beach, CA, USA, June 2019.

[26] S. Dey, P. Motlicek, T. Bui, and F. Dernoncourt, "Exploiting semi-supervised training through a dropout regularization in end-to-end speech recognition," in *Interspeech 2019*, Graz, Austria, September 2019, pp. 734–738.

[27] A. Vyas, P. Dighe, S. Tong, and H. Bourlard, "Analyzing uncertainties in speech recognition using dropout," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019)*, Brighton, UK, May 2019, pp. 6730–6734.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] M. Abdelwahab and C. Busso, "Study of dense network approaches for speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, Calgary, AB, Canada, April 2018, pp. 5084–5088.