CONFIDENCE ESTIMATION FOR SPEECH EMOTION RECOGNITION BASED ON THE RELATIONSHIP BETWEEN EMOTION CATEGORIES AND PRIMITIVES

Yang Li[*†], *Constantinos Papayiannis*[†], *Viktor Rozgic*[†], *Elizabeth Shriberg*[†], *Chao Wang*[†]

[*]Dept. of Electrical and Computer Engineering, University of Rochester
yli131@ur.rochester.edu
[†] Amazon Alexa
{yllimm,papayiac,rozgicv,shriberg,wngcha}@amazon.com

ABSTRACT

Confidence estimation for Speech Emotion Recognition (SER) is instrumental in improving the reliability in the behavior of downstream applications. In this work we propose (1) a novel confidence metric for SER based on the relationship between emotion primitives: arousal, valence, and dominance (AVD) and emotion categories (ECs), (2) EmoConfidNet - a DNN trained alongside the EC recognizer to predict the proposed confidence metric, and (3) a data filtering technique used to enhance the training of EmoConfidNet and the EC recognizer. For each training sample, we calculate distances from corresponding AVD annotation vectors to centroids of each EC in the AVD space, and define EC confidences as functions of the evaluated distances. EmoConfidNet is trained to predict confidence from the same acoustic representations used to train the EC recognizer. EmoConfidNet outperforms state-of-the-art confidence estimation methods on the MSP-Podcast and IEMOCAP datasets. For a fixed EC recognizer, after we reject the same number of low confidence predictions using EmoConfidNet, we achieve a higher F1 and unweighted average recall (UAR) than when rejecting using other methods.

Index Terms— Confidence estimation, Speech emotion recognition, Emotion categories, Emotion primitives, Data preprocessing

1. INTRODUCTION

Speech Emotion Recognition (SER) aims to automatically recognize emotions and affective states from human speech/voice [1,2]. Many technologies in scientific and business focus, including conversational AI [3], human-to-human interaction understanding [4], and health & wellness assessment [5] require reliable SER that preferably knows how to distinguish high- from low-confidence predictions, i.e. SER models should be aware of boundaries of their knowledge, minimizing risks stemming from wrong predictions.

Deep learning and DNNs are widely used nowadays. Although DNNs often perform tasks with high classification accuracy, they are often poorly calibrated and unable to tell how confident their predictions are [6]. This issue motivated research on confidence estimation for DNNs [7, 8] where the goal is to have a model that outputs predictions only when it is sufficiently confident, or outputs predictions together with calibrated confidence scores. Confidence estimation is critical in areas where failures can cause damage or harm [9], such as autonomous driving [6, 10] and computer-aided diagnosis [5].

In this work, we study confidence estimation for SER with the objective to build a state-of-the-art SER model that predicts emotions and estimates a measure of confidence for each prediction. Such model enables us to discard low confidence predictions, and pass only high-confidence predictions to downstream tasks. We propose a novel confidence metric definition based on the relationship between two types of emotion representations, both typically annotated in emotion datasets: a) discrete emotion categories (EC) such as happy, neutral, sad and angry, and b) the continuous emotion primitives - arousal, valence, and dominance. While there exists an extensive research on EC-only and AVD-only predictions [11–13], less work has been dedicated to investigating how the EC-to-AVD relationship can be exploited for predictions of speech emotions [14, 15].

The relationship between EC and AVD annotations is important not only for prediction but also for confidence estimation. We estimate the distribution of primitives for each EC on annotated training data and calculate confidence vectors as a function of primitive likelihoods evaluated for each EC using these distributions. We teach a DNN, named EmoConfidNet, to predict the confidence vectors from acoustic embeddings. In training, we leverage the distributions of the primitives for different categories and the annotation of primitives for the given sample. EmoConfidNet is trained using a ranking loss to minimize the distance between the target and predicted confidence vectors. In Section 4 we present more details on the proposed confidence modeling approach. In Section 5, we evaluate the performance of EmoConfidNet by studying the trade-off between sample coverage and classification performance measured by the F1-score and Unweighted Average Recall (UAR). In comparison with the state-of-the-art method of [13], our method achieves a higher F1-score and UAR with higher sample coverage. We also explored data preprocessing by removing samples with conflicting ECs and primitives from the training. This preprocessing improves the performance of both the emotion classifier and EmoConfidNet. To the best of our knowledge, this is the first work to utilize the EC-to-AVD relationship to filter training data for SER and to estimate the confidence during inference.

2. RELATED WORK

Existing confidence estimation research in the SER domain has focused either on predictions of emotion primitives [12] or emotion classification [13]. For emotion primitives, [12] proposed predicting uncertainty via Monte Carlo dropout, a general approach proposed by [16]. For emotion classification, [17] introduced emotion classification using emotion profiles, or soft labels rather than raw acoustic features or categorical labels to capture the nature of emotion expressions. [18] used multiple classifiers to assess prediction reliability by introducing confidence ratios from confusion matrices. These ratios are then combined into a single confidence measure. [19] proposed a new evaluation methodology that utilized the underlying distribution of the emotion labels rather than the majority votes. More recently, [13] applied the rejection options for SER where a classifier only provides predictions over samples with reliability above a given threshold. Despite the extensive research work done in this area, to the best of our knowledge, no previous work has been done to combine EC and primitives for confidence estimation in SER.

Besides hand-crafted rules and ensemble-based methods, a parameterized model can be trained to predict a selected confidence metric from acoustic embeddings. In [20], a confidence-network was trained to output predictions and confidence scores simultaneously. Approaches, including output calibration [6] and loss function regularization [7], aim to adapt native outputs to specific output-to-confidence metric mappings. [9] defined confidence with a specific learning scheme in which a DNN is trained to learn confidence scores of training samples and then applied on test data.

Our work is the closest to [9] with a similar procedure used to train a confidence neural network. In [9], the confidence targets are derived solely from the classifier outputs, while we utilized both annotations of emotion primitives and ECs to define confidence targets by utilizing additional information embedded in the EC-to-AVD relationship.

3. RESOURCES

3.1. Datasets

Two datasets are used in this work. The first, the MSP-Podcast corpus [21], is a naturalistic emotional database relying on spontaneous recordings obtained from audio-sharing websites. The version used in this work contains 10,124 test samples, a development set of 5,958 samples, and a training set of 34,280 samples. The partitioning avoids speaker overlap across train, test and development sets. The emotion labels are provided by annotators, each one of them being asked to rate the emotion primitives (arousal, valence, dominance) from 1 to 7, and also to pick a primary EC. The final labels for the speech samples are the mean values of the labeled primitives, and the result of the majority votes on the primary ECs selected by the annotators. We limit the ECs to 5 (“Happy”, “Sad”, “Disgusted”, “Neutral”, “Angry”) as in [13] for emotion classification and confidence estimation.

The second dataset used is the IEMOCAP dataset [22], which contains recorded dialogues from improvisations of affective scenarios and performances of theatrical scripts. Ten actors were involved in the recording of 5 dyadic sessions with 2 in each session. The annotations of the utterances follow a similar procedure as for MSP-Podcast, while the major difference is the 1 to 5 range of emotion primitive scores. In this work, we evaluated the performance of emotion recognition and confidence estimation on samples that belong to the selected 5 ECs: “Happy”, “Frustrated”, “Angry”, “Sad” and “Neutral”.

3.2. Acoustic Features

For both datasets we use the same acoustic features as in [11–13,23]. The feature set was introduced in Interspeech 2016 [24] and includes 6,373 static features from the computation of various functionals over low-level descriptors (LLD) contours such as fundamental frequency and Mel-frequency cepstral coefficients (MFCCs). A more thorough description of the feature set can be found in [25]. The OpenSmile toolkit [26] is used in this work to extract a vector with 6,373 acoustic features for each sample, which in our case is the input to the DNN emotion classifier.

3.3. Data Preprocessing

We conducted an analysis of the annotation data for MSP-Podcast and IEMOCAP, which revealed samples with conflicting primitives and EC annotations. This contradicts the established and assumed relationship between them [14]. Examples of such conflicts are samples labeled as “Happy” but with a low valence annotation. A more thorough summary of all the possible conflicts is given in Table 1. We hypothesize that conflicts arise when we apply mean and majority vote operators on annotations with a high disagreement levels. Conflicts suggest that corresponding samples may have lower quality or ambiguous patterns that hinder annotators in terms of providing consistent labels. With indications that the eventual labels for such samples might be of lower quality, we stipulate that inclusion of such samples in the model training is counterproductive and we process the dataset by removing them during training. Despite the fact that we believe that these rules enable us to reject low-quality annotations, we acknowledge that they are based on deviation from prototypical relationships between emotion categories and emotion primitives. Therefore, a trade-off exists between the induced accuracy gains for samples similar to the prototypical cases and other rarer cases.

Emotion Category	Conflicts
Happy	Low arousal or low valence
Sad/Disgusted/Frustrated	High valence
Angry	Low arousal or high valence

Table 1: Types of conflicts between ECs and primitives.

To demonstrate the effectiveness of the proposed data preprocessing, we used two versions of the training set: one with the removal of samples with conflicts listed in Table 1 and the other with raw data. The emotion classifier and the confidence estimation model described in the next section were trained twice on both training set versions. To ensure a fair comparison with previous work using the same datasets, the data preprocessing is only applied on the training set while the validation set and test set remain unchanged from the original distributions. We expect that with data preprocessing: a) the emotion classifier will receive higher quality samples during training, leading to a more accurate classifier; and b) our confidence estimation model (which relies on the EC and primitives relationship) will produce more reliable confidence estimation scores. This is because fundamental assumptions between their relationship are not contradicted during the training process.

4. CONFIDENCE ESTIMATION

4.1. Linking Sample and EC-Centroid Distance to Confidence

Analyzing IEMOCAP and MSP-Podcast annotations of the emotion primitives AVD, after the above pre-processing is applied, shows a compact clustering behavior for samples of the same EC. For each EC, we calculate the center of the cluster in this 3D space, the EC centroid. Given this centroid, a simple way to define confidence would be to compute the distance between the primitives' annotation vector and the centroid of the predicted EC. If this distance is large, the sample is less likely to belong to the predicted EC.

While this confidence definition can be effective [27], in the context of SER it can prove problematic. For example, a sample with extremely high valence and arousal and labeled as “Happy“ may appear far from the centroid of the “Happy” EC when measuring the scalar Euclidean distance. The above naive distance-based definition of confidence will assign a low confidence score to this sample, but it is actually a very strong expression of happy. To address this, we do not use the distance to the specific EC centroid in isolation, but in relation to distances to centroids of other ECs. Back to the example above, the sample's distance to the centroid of the “Happy” EC, compared with its distances to the centroids of other ECs, is smaller, indicating that this sample is more likely to be an expression of “Happy” than any other EC considered.

Based on this distance-based confidence definition, we proposed a DNN named EmoConfidNet that operates in parallel with the EC classifier, and maps acoustic embeddings to a confidence-score vector. The training procedure considers the predicted EC, the primitives' annotations and the ensemble of category centroids.

4.2. EmoConfidNet

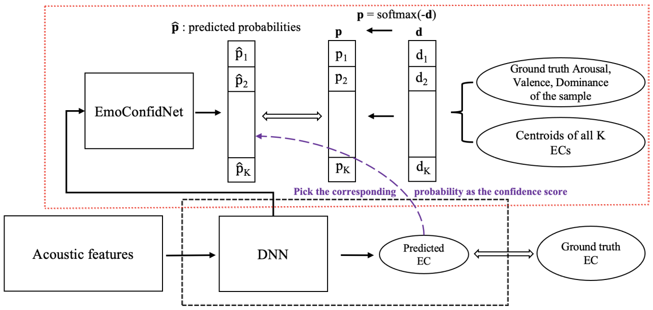

$\hat{\mathbf{p}}$: predicted probabilities
$\mathbf{p}$ = softmax(-$\mathbf{d}$)
$\mathbf{p}$
$\mathbf{d}$
$\hat{p}_1$
$\hat{p}_2$
$\hat{p}_K$
p_1
p_2
p_K
d_1
d_2
d_K
EmoConfidNet
Ground truth Arousal, Valence, Dominance of the sample
Centroids of all K ECs
Pick the corresponding probability as the confidence score
Acoustic features
DNN
Predicted EC
Ground truth EC

Fig. 1: Scheme of the complete pipeline, including the emotion classifier (black dashed box) and the proposed EmoConfidNet (red dashed box) for confidence estimation.

Fig. 1 presents the complete framework proposed in this work, which includes two sequential training stages. The first stage is to train an emotion classifier with acoustic features as the input that predicts the EC. The second stage is the training of EmoConfidNet with the following details:

- **Input:** The input of EmoConfidNet is the same as the input to the last layer of the emotion classifier that was trained in the first step.
- **Output:** The output of EmoConfidNet is a probability array $\hat{\mathbf{p}}$ that contains the estimated probability that the sample belongs to each EC. For each training sample, the Manhattan distances between its ground truth arousal, valence and dominance, and the centroids of each EC are computed. With K ECs, at this step for each training sample, there will be K distances, denoted by the vector $\mathbf{d}$ in Fig. 1. The next step is to translate the distances to probabilities by applying a softmax activation function over the negatives of the distances, resulting in a vector $\mathbf{p}$ and each element in this array relating to the probability that the sample belongs to a certain EC. The reason behind the distance-to-probability transformation is two-fold. First, directly predicting the distances is equal to solving the problem of primitives estimation which is hard and less related to our objective. Second, such transformation will ensure that the training targets for all the samples are in the same scale. The latter is important in interpreting the values during inference.
- **Training:** The training objective of EmoConfidNet is to produce an output $\hat{\mathbf{p}}$ that preserves the pairwise ranking in the target vector $\mathbf{p}$. For example, suppose in $\mathbf{p}$ we have $p_{happy} > p_{sad}$, meaning that the sample is more likely to be "Happy" than "Sad". Ideally, in $\hat{\mathbf{p}}$ we also want $\hat{p}_{happy} > \hat{p}_{sad}$. For this purpose, the mean pairwise squared loss function in (1) is used which preserves not only the pairwise ranking but also the margin between probabilities. Consequently, we expect to generate $\hat{\mathbf{p}}$ whose elements will approximate the pre-computed probabilities of the sample belonging to each EC in the training set.

$$Loss = \frac{\sum_{i \neq j}\left(\left(\hat{p}_i - \hat{p}_j\right) - \left(p_i - p_j\right)\right)^2}{C_K^2}, i=\{1,\ldots,K\}, j=\{1,\ldots K\} \quad (1)$$

- **Confidence score in testing:** Given a new input sample, the whole scheme will generate a predicted EC $\hat{c}$ and a probability vector $\hat{\mathbf{p}}$. The final confidence score of this prediction is set to be $\hat{\mathbf{p}}_{\hat{c}}$, which is the probability in $\hat{\mathbf{p}}$ that corresponds to the predicted EC $\hat{c}$.

4.3. Model Structure

The emotion classifier built in this work consists of two dense layers with 256 nodes per layer. We used Rectified Linear Unit (ReLU) activation [28] for intermediate layers and softmax activation on the output layer. Batch normalization is applied on the hidden layers and the classifier is trained with a dropout rate of 0.5 and cross-entropy loss. Early-stopping with a patience equal to 10 is applied to monitor the loss on the validation set to avoid overfitting. The ADAM optimizer [29] was used with a learning rate 10^{-4}. The input of the classifier is the extracted acoustic features from each audio sample standardized using the mean and standard deviation computed over the samples in the training set. EmoConfidNet is trained separately after the classifier is trained and the input to EmoConfidNet is the output of the last dense layer of the classifier (see Fig. 1). All parameters such as the number of layers, early-stopping patience etc. are the same as the classifier described above.

5. RESULT AND ANALYSIS

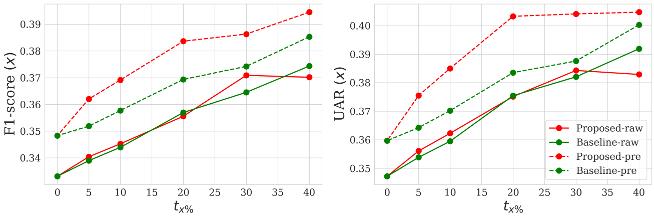
In this section, the performance of the proposed EmoConfidNet is evaluated and compared with the sample rejection criteria proposed in [13] where the confidence is defined as the difference between the top-two probabilities from the softmax output of the classifier.

5.1. Evaluation Methodology

To evaluate the proposed confidence estimation method, we studied the trade-off between the classification performance and the sample coverage level determined by thresholds computed on the validation set. After training the emotion classifier and EmoConfidNet on the training set, they were used to produce EC predictions and confidence scores respectively for each sample of the validation set. To determine the threshold associated with a level of coverage, we sorted the confidence scores of the validation samples in descending order and picked the score that marked the desired coverage level. For example, for a hypothetical validation set of 100 samples, the 80th confidence score (after sorting) is 0.4. Then 0.4 is set to be the 20%-threshold corresponding to an 80% coverage level and all predictions with confidence scores lower than 0.4 are removed at this level of coverage. We denote such thresholds as $t_{20\%}$, where in the example above $t_{20\%} = 0.4$ and in all cases $t_{0\%} = 0$. The computed thresholds were then applied on the test set, where predictions with confidence scores lower than a threshold would be removed with respect to the corresponding threshold level. F1-score and UAR were computed on the remaining predictions at each threshold level, both expected to increase as low-confidence samples are being removed. To compute the F1-score, the precision and recall were first computed for each EC. The average precision and average recall were used to compute the F1-score, giving the same weights per EC [13].

5.2. Experiment results

Fig. 2 shows the evaluation of the proposed method and the baseline [13] on the MSP-Podcast dataset. As shown in the figure, with data preprocessing applied to the training data, the performance of both the emotion classifier and the confidence-estimation method improve with respect to the F1-score and UAR. The starting points of the dashed lines are higher than those of the solid lines in both plots, indicating that the classifier trained using the preprocessed training set offers better performance on the test set, which is unchanged between the two cases. As preprocessing leads to better performance on the same test set, the following analyses will focus on the comparison between the EmoConfidNet and the baseline, both trained using the preprocessed training set.



(a) (b)

Fig. 2: (a) F1-score and (b) UAR against threshold level of EmoConfidNet (red) and baseline (green). $t_{x\%}$ represents the $x\%$-threshold determined on the validation set; '-raw' refers to training with raw training set and '-pre' refers to training with preprocessed training set.

As shown in Fig. 2, the increase in the F1 score and the UAR when shifting the threshold from $t_{20\%}$ to $t_{30\%}$ is not significant, thus we choose to conduct a thorough performance analysis at $t_{20\%}$, which provides a reasonable trade-off between coverage and accuracy gains. Compared with the $t_{0\%}$ case (no predictions removed), at $t_{20\%}$ the relative gain in F1 for EmoConfidNet is 10.1% and 12.1% in UAR, while the baseline offers a 6% relative gain in F1 and 6.6% in UAR. Therefore, the proposed method provides higher relative gains in F1 and UAR at $t_{20\%}$ than the baseline. It is worth noting that in order to reach the same F1 or UAR, EmoConfidNet provides higher coverage by rejecting fewer samples, which is very important in real-world applications of these methods.

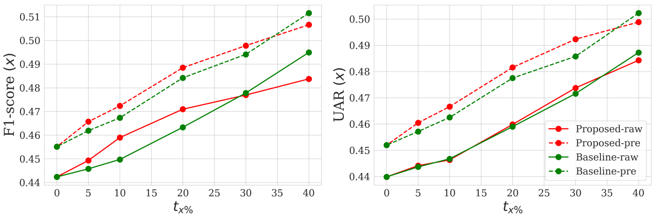
Metric	Rejection Method	Angry	Disgusted	Happy	Neutral	Sad
Precision	No rejection	0.166	0.123	0.615	0.675	0.109
	Baseline [13]	0.185	0.131	**0.657**	**0.689**	0.119
	EmoConfidNet	**0.252**	**0.166**	0.616	0.675	**0.120**
Recall	No rejection	0.329	0.372	0.511	0.289	0.297
	Baseline [13]	0.354	**0.431**	0.551	0.273	0.308
	EmoConfidNet	0.354	0.300	**0.623**	**0.385**	**0.355**
F1-score	No rejection	0.221	0.185	0.558	0.405	0.159
	Baseline [13]	0.243	0.201	0.599	0.391	0.172
	EmoConfidNet	**0.295**	**0.214**	**0.620**	**0.490**	**0.179**

Table 2: F1, precision and recall for each EC at $t_{20\%}$ for the baseline [13] and EmoConfidNet. Results also shown for $t_{0\%}$ (no rejection).

Table 2 shows the performance of each EC at $t_{20\%}$ and at $t_{0\%}$ (no predictions removed). Both the baseline and EmoConfidNet improve the performance by removing low-confidence predictions, and EmoConfidNet has higher F1-scores for all 5 ECs. However, lower recall for "Disgusted" is observed. Further investigation reveals that incorrect predictions ("Disgusted" predicted to be "Happy" or "Neutral") are more frequent when rejecting at $t_{20\%}$ by EmoConfidNet, leading to lower recall for "Disgusted" and slightly lower precision in "Happy" and "Neutral". The reason may be that these samples have high arousal, which brings them close to clusters of samples of the "Happy" and "Neutral" categories. Consequently, they were assigned relatively higher confidence scores that exceed $t_{20\%}$.

We further study the samples removed at $t_{20\%}$. Altogether 1,559 predictions were removed by the baseline due to low confidence, among which 477 are actually correct predictions. EmoConfidNet removed 1,589 predictions and only 124 of them were correct. Compared with the baseline, our method removes far fewer correct predictions, indicating that the confidence scores computed via EmoConfidNet are better rejection indicators. In-depth analysis reveals that while the baseline removes both wrong and correct predictions across all 5 ECs, EmoConfidNet removes a lot of wrong predictions that predict "Happy", "Neutral", or "Sad" as "Angry" or "Disgusted", leading to the increased recall of "Happy", "Neutral" and "Sad" and increased precision of "Angry" and "Disgusted". In addition, EmoConfidNet brings an increase in precision and recall for the "Sad" EC. As "Sad" is different than other ECs, with samples typically annotated with low arousal and valence, we attribute these gains to the leveraging of the relationship between ECs and emotion primitives by EmoConfidNet.

Experiments on the IEMOCAP dataset followed the same procedure as with MSP-Podcast and the results are shown in Fig. 3. A 5-fold cross validation is carried out, with each model training process utilizing data from 3 dyadic sessions. The other two sessions were used as a validation and test set, respectively. This ensures that each session is used as a test set once and is unseen during training.



(a) (b)

Fig. 3: (a) F1-score and (b) UAR against threshold level of EmoConfidNet (red) and baseline (green) for IEMOCAP. Each dot stands for the mean value across sessions. Legend notions are the same as in Fig. 2.

Results on IEMOCAP are consistent with those for MSP-Podcast, indicating the robustness of the proposed EmoConfidNet model across different datasets. In addition, we notice that the performance gains for IEMOCAP are relatively smaller than for MSP-Podcast. We attribute this to the following: a) IEMOCAP is a smaller dataset; b) each session in IEMOCAP contains 2 subjects, leading to large variance between sessions and unstable classification performance; c) the rating of IEMOCAP is from 1 to 5, not 7, therefore, the granularity in the primitives' space is smaller, which means that irreducible errors from training of EmoConfidNet may be more likely to impact the performance. This finding also reveals that EmoConfidNet relies on well-annotated emotion primitives. Furthermore, as a DNN model, EmoConfidNet benefits from large amounts of data for training to ensure generalization.

6. CONCLUSION

In this work, a DNN model, EmoConfidNet, was proposed in order to leverage the relationship between ECs and primitives for SER confidence estimation, which remains relatively unexplored in the literature. Compared to the method of [13], inclusion of primitives better captures the notion of confidence in emotion recognition, leading to improved SER performance with a reduced level of compromise on sample coverage. In-depth study also showed that EmoConfidNet performs better across all ECs with fewer errors, such as in removals of correct predictions. We also studied and performed data preprocessing on the training set by removing samples with conflicting ECs and primitives' annotation. This results in higher-quality data for model training, increasing the efficacy of learning the relationship between acoustic signals and emotion attributes, evident by the improved performance of the classifier and EmoConfidNet. Future work will focus on model evaluation across corpora, and ensembles of multiple confidence measurements into one for potential gains in performance in terms of confidence estimation for SER.

7. ACKNOWLEDGMENT

The authors would like to thank Bo Yang for the discussions, suggestions and insights.

8. REFERENCES

[1] B. Schuller, "Speech emotion recognition," *Commun. ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[2] M. Li, B. Yang, J. Levy, A. Stolcke, V. Rozgic, S. Matsoukas, C. Papayiannis, D. Bone, and C. Wang, "Contrastive unsupervised learning for speech emotion recognition," in *IEEE Intl. Conf. Acoust., Speech and Signal Process. (ICASSP)*. IEEE, 2021, pp. 6329–6333.

[3] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, "Emotion recognition in conversation: Research challenges, datasets, and recent advances," *IEEE Access*, vol. 7, pp. 100 943–100 953, 2019.

[4] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," *Intl. J. Speech Technol.*, vol. 15, no. 2, pp. 99–117, 2012.

[5] M. Li and Z.-H. Zhou, "Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples," *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 37, no. 6, pp. 1088–1098, 2007.

[6] L. Neumann, A. Zisserman, and A. Vedaldi, "Relaxed softmax: Efficient confidence auto-calibration for safe pedestrian detection," *Adv. Neural Inf. Process. Syst.*, 2018.

[7] J. Moon, J. Kim, Y. Shin, and S. Hwang, "Confidence-aware learning for deep neural networks," in *Intl. Conf. Mach. Learn.*, 2020, pp. 7034–7044.

[8] A. Subramanya, S. Srinivas, and R. V. Babu, "Confidence estimation in deep neural networks via density modelling," *arXiv preprint arXiv:1707.07013*, 2017.

[9] C. Corbière, N. Thome, A. Bar-Hen, M. Cord, and P. Pérez, "Addressing failure prediction by learning model confidence," in *Adv. Neural Inf. Process. Syst.*, 2019, pp. 2902–2913.

[10] N. Djuric, V. Radosavljevic, H. Cui, T. Nguyen, F.-C. Chou, T.-H. Lin, N. Singh, and J. Schneider, "Uncertainty-aware short-term motion prediction of traffic actors for autonomous driving," in *IEEE Winter Conf. Appl. Comput. Vis.*, 2020, pp. 2095–2104.

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

[12] K. Sridhar and C. Busso, "Modeling uncertainty in predicting emotional attributes from spontaneous speech," in *IEEE Intl. Conf. Acoust., Speech and Signal Process. (ICASSP)*. IEEE, 2020, pp. 8384–8388.

[13] ——, "Speech emotion recognition with a reject option," in *Interspeech*, 2019, pp. 3272–3276.

[14] A. Toisoul, J. Kossaifi, A. Bulat, G. Tzimiropoulos, and M. Pantic, "Estimation of continuous valence and arousal levels from faces in naturalistic conditions," *Nat. Mach. Intell.*, vol. 3, no. 1, pp. 42–50, 2021.

[15] V. Kowtha, V. Mitra, C. Bartels, E. Marchi, S. Booker, W. Caruso, S. Kajarekar, and D. Naik, "Detecting emotion primitives from speech and their use in discerning categorical emotions," in *IEEE Intl. Conf. Acoust., Speech and Signal Process. (ICASSP)*. IEEE, 2020, pp. 7164–7168.

[16] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *Intl. Conf. Mach. Learn.* PMLR, 2016, pp. 1050–1059.

[17] E. Mower, A. Metallinou, C.-C. Lee, A. Kazemzadeh, C. Busso, S. Lee, and S. Narayanan, "Interpreting ambiguous emotional expressions," in *Intl. Conf. Affect. Comput. Intell. Interact.* IEEE, 2009, pp. 1–8.

[18] J. Deng and B. Schuller, "Confidence measures in speech emotion recognition based on semi-supervised learning," in *Interspeech*, 2012.

[19] P. Meyer and T. Fingscheidt, "A new evaluation methodology for speech emotion recognition with confidence output," in *Speech Commun.*, 2014, pp. 1–4.

[20] S. Wan, T.-Y. Wu, W. H. Wong, and C.-Y. Lee, "Confnet: predict with confidence," in *IEEE Intl. Conf. Acoust., Speech and Signal Process. (ICASSP)*. IEEE, 2018, pp. 2921–2925.

[21] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Trans. on Affect. Comput.*, vol. 10, no. 4, pp. 471–483, 2017.

[22] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008.

[23] G. Aguilar, V. Rozgic, W. Wang, and C. Wang, "Multimodal and multi-view models for emotion recognition," *arXiv preprint arXiv:1906.10198*, 2019.

[24] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, K. Evanini *et al.*, "The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language," in *Interspeech*, 2016, pp. 2001–2005.

[25] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer, "On the acoustics of emotion in audio: What speech, music, and sound have in common," *Front. Psychol.*, vol. 4, p. 292, 2013.

[26] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the Munich versatile and fast open-source audio feature extractor," in *Proc. ACM Intl. Conf. Multimed.*, 2010, pp. 1459–1462.

[27] A. S. Shirkhorshidi, S. Aghabozorgi, and T. Y. Wah, "A comparison study on similarity and dissimilarity measures in clustering continuous data," *PloS one*, vol. 10, no. 12, p. e0144059, 2015.

[28] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Intl. Conf. Mach. Learn.*, 2010, pp. 807–814.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.