INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Ensemble of Students Taught by Probabilistic Teachers to Improve Speech Emotion Recognition

Kusha Sridhar, Carlos Busso

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

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

Abstract

Reliable and generalizable *speech emotion recognition* (SER) systems have wide applications in various fields including healthcare, customer service, and security and defense. Towards this goal, this study presents a novel *teacher-student* (T-S) framework for SER, relying on an ensemble of probabilistic predictions of teacher embeddings to train an ensemble of students. We use uncertainty modeling with *Monte-Carlo* (MC) dropout to create a distribution for the embeddings of an intermediate dense layer of the teacher. The embeddings guiding the student models are derived by sampling from this distribution. The final prediction combines the results obtained by the student ensemble. The proposed model not only increases the prediction performance over the teacher model, but also generates more consistent predictions. As a T-S formulation, the approach allows the use of unlabeled data to improve the performance of the students in a semi-supervised manner. An ablation analysis shows the importance of the MC-based ensemble and the use of unlabeled data. The results show relative improvements in *concordance correlation coefficient* (CCC) up to 4.25% for arousal, 2.67% for valence and 4.98% for dominance from their baseline results. The results also show that the student ensemble decreases the uncertainty in the predictions, leading to more consistent results.

Index Terms: Speech Emotion Recognition, Monte Carlo Dropout, Semi-Supervised learning, Teacher-Student network

1. Introduction

Human communication involves complex emotions that regulate our interaction. Therefore, designing socially smart systems that are aware of the emotional state of the user is a key challenge. *Speech emotion recognition* (SER) is particularly important, given the pervasiveness of speech-based devices [1]. For SER systems to be useful in practical application, the solutions have to generalize well to new conditions (e.g., different microphones, speakers, environments, or noises). For example, generalized models can minimize scalability issues in cross-modal learning, where domain adaptation is important. SER systems should also be consistent, where small perturbations should not affect the predictions. For example, psychometric analysis in behavioral studies require test-retest reliability to learn patient traits. Therefore, it is important that SER systems provide similar predictions in the presence of similar inputs.

Studies have proposed several formulations to increase the generalization of SER systems including active learning [2–4], use of ensemble [5, 6], domain adaptation [7, 8] and *multitask learning* (MTL) with supervised [9–11] or unsupervised [12,13] secondary tasks. Another flexible approach to construct models that generalize well to new conditions is by using the *teacher-student* (T-S) framework, where complex, well-trained models (i.e., teachers) can transfer knowledge to lighter, generalized models (i.e., students). Lighter models are preferred at inference time since they can adapt better to changes in the recording conditions by leveraging unlabeled data.

This study introduces a novel T-S formulation for SER to predict emotional attributes in a scalable and consistent manner. We train an ensemble of teachers by regularizing the *deep neural networks* (DNNs) with different dropout rates, increasing the diversity in the ensemble. Furthermore, we use *Monte-Carlo* (MC) dropout [14] for each teacher in the ensemble to create probabilistic predictions of its intermediate embeddings. MC dropout is a technique to approximate Bayesian inference in *deep neural networks* (DNNs) using dropout regularization while training and testing the models. By running the teacher multiple times with MC dropout, we obtain a probabilistic approximation of the posterior distribution of the emotional predictions. We infer the mean from the distribution of the intermediate embeddings of a teacher, which is used to train a student model. For each teacher, we train a student model in a semi-supervised manner using the predicted teacher embedding with unlabeled data. The proposed approach simultaneously captures the diversity in the emotional predictions by creating an ensemble of N teachers and N students using this MC dropout formulation. The final prediction of the proposed system is the average of the outputs of the student models.

The proposed T-S framework is evaluated with the MSP-Podcast corpus. We implement the approach as a regression problem, where the task is to predict the emotional attributes arousal, valence and dominance. The experimental results show an improvement in the performance of the student models over the teacher models, showing the success in transferring knowledge with this formulation. We evaluate the performance of the student models using *concordance correlation coefficient* (CCC). We achieve relative gains in CCC up to 4.25% for arousal, 2.67% for valence and 4.98% for dominance over a baseline network implemented with a similar architecture to the teacher model. Furthermore, the standard deviations in the predictions are lower for the students than the teacher models, showing improvements in the consistency of the proposed approach. These results show the benefits of our novel SER formulation using the T-S framework.

2. Related Work

Emotion recognition becomes more challenging when the input modality is spontaneous speech. Recent studies have proposed many novel DNN approaches such as *multi-task learning* (MTL) [9–11], domain adversarial methods [7, 8] and ladder networks [12, 13] to improve SER. However, the naturalness in spontaneous speech makes SER even more complex, often leading to predictions that are spurious or unreliable. Alternative paradigms based on Bayesian learning may be beneficial.

Some of the deep learning strategies inspired by Bayesian theory are ensemble of model predictions [15], regularization by penalizing output distributions [16], and *prior networks* (PN)

http://dx.doi.org/10.21437/Interspeech.2020-2694

to model predictive uncertainty [17]. Approximating Bayesian inference using DNNs can be used to learn well calibrated probabilistic predictions leading to construction of reliable and generalized models. *Knowledge distillation* (KD) [18] via T-S framework provides a flexible way to achieve this goal by training guided models that estimate better prediction uncertainties and generalize better to new conditions by diversifying the models. A study by Gurau et al. [19] proposed a *distilled dropout network* (DDN) using a T-S paradigm on image classification tasks. They transfer knowledge from multiple MC samples of the soft targets generated by the teacher along with the ground truth labels of the training samples to build a student model. For *automatic speech recognition* (ASR) tasks, Wong et al. [20] proposed an ensemble T-S framework using MTL on context dependent phonetic targets. They improved model diversity by using multi-task ensembles for the teacher, which led to lower *word error rate* (WER) on conversational telephone speech tasks. In *natural language processing* (NLP), Sun et al. [21] implemented a multi-layer patient KD scheme by training a student with features extracted from multiple intermediate layers of the teacher. They used *Bidirectional Encoder Representations from Transformers* (BERT) as the pre-trained teacher model, showing improvements in the student performance. Similarly, there are several feature-based ensemble KD methods using some form of feature map or non-linear transformations to match representation layers of teacher and student to achieve knowledge distillation [22, 23].

In SER, Albanie et al. [24] used cross-modal distillation to train a T-S model for solving an audiovisual SER problem. The teacher was trained on a facial emotion recognition task. The representations from the teacher were used to train a student along with unlabeled speech embeddings to predict emotional scores. They showed that transferring knowledge across modalities from faces to speech can reduce noise in the labels, increasing the robustness towards ambiguous annotations. Mower Provost et al. [25] used emotion distillation as a pre-processing stage to detect emotionally salient regions in audio-visual inputs. Our study explores the use of uncertainty modeling and unlabeled data in a T-S framework to improve SER. We show the benefits of using MC dropout and ensemble of feature representations to improve prediction of emotional attributes.

3. Resources

3.1. The MSP-Podcast Database

This study uses the MSP-Podcast corpus [26] which is a collection of emotionally rich spontaneous speech recordings from various audio-sharing websites. We use a set of pre-processing steps to clean and segment the podcast recordings into speaking turns without background music, noise or overlapped speech. We select single speaker segments with duration between 2.75s and 11s. To balance the emotional content of the corpus, we retrieve emotionally rich samples based on the strategy suggested in Mariooryad et al. [27]. The approach relies on using alternative machine learning formulations to identify segments expected to have emotional content. This study uses version 1.6 of the corpus, which consists of 50,362 speaking turns (83h 29m). The emotional evaluations are conducted with a crowdsourcing protocol similar to the one discussed in Burmania et al. [28]. This study uses the emotional attributes valence (negative versus positive), arousal (calm versus active), and dominance (weak versus strong) annotated with *self-assessment manikins* (SAMs) on a seven point Likert scale. The ground truth labels for the attributes of each speech segment are obtained by averaging the scores of five or more annotators. More details on this corpus is provided in Lotfian and Busso [26].

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. There are around 400,000 speech segments that have not yet been retrieved by our algorithms, so they have not been annotated.

3.2. Acoustic Features

This study uses the Interspeech 2013 computational paralinguistics challenge acoustic features [29], extracted using the OpenSmile toolkit [30]. The feature set consists of *low level descriptors* (LLDs) such as energy, fundamental frequency and *Mel-frequency cepstral coefficients* (MFCCs), extracted at the frame level using 20ms windows. A set of utterance level statistics are calculated over these LLDs (e.g., mean of the energy), which are referred to as *high level descriptors* (HLDs). This process generates a 6,373 dimensional feature vector for each speech segment, regardless of its duration.

4. Proposed Method

Figure 1 shows the conceptual idea of our proposed ensemble model using the T-S framework. The motivation for using a T-S formulation is that we need deep, complex models during training process with more capacity to incorporate knowledge from large amounts of training data. During inference, however, it may be better to have a lighter model that generalizes well to a target domain. The T-S framework is suitable for this kind of problem where the teacher is the deep, complex model and the student is the shallow, lighter model. Another motivation in our formulation is to leverage the benefits observed in previous studies by using an ensemble of classifiers for SER tasks [6]. Our proposed approach uses an ensemble of N teachers and N students to build a model that generalizes well on unseen data. A key step in improving robustness and generalization is through regularization, which can be achieved with various approaches including data augmentation, early stopping and dropout. This study uses dropout as a regularization, which has been successfully used in SER [31]. The third motivation in our formulation is to capture the model uncertainty while building our teacher and student models. A robust system that generalizes well should be able to handle out-of-distribution samples or inputs sampled from sparse regions in the in-domain data. The knowledge about such uncertain conditions can be captured by estimating model uncertainty (epistemic uncertainty) [32]. One approach to estimate model uncertainty is by using MC dropout [14]. We rely on MC ensembles of the predicted teacher embeddings to distill knowledge to the student ensemble.

4.1. The Teacher Model

The first part of our formulation is the teacher model, which is implemented as an ensemble of N networks. While building an ensemble, it is important that the teachers are diverse, providing complementary information. We achieve this goal by using different dropout rates, which introduces model diversity for the same input data points (e.g., one teacher is trained with $p = 0.45$ while another teacher is trained with $p = 0.6$). In addition, we use MC dropout for the N teacher models to construct a strong teacher network that preserves model uncertainty. We use the method followed in our previous study [33] to implement the MC dropout. We extract 100 MC samples of each prediction and average them to get a single prediction for each teacher in the ensemble. This process not only captures

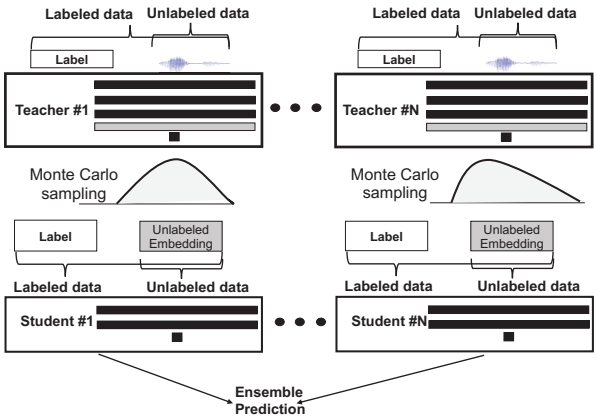


Figure 1: *Proposed T-S framework using a teacher ensemble with MC dropout to distill knowledge to a student ensemble.*

the mean of the ensemble, but also preserves the model's uncertainty in its predictions. Therefore, we reduce the risk of making falsely confident predictions. Furthermore, we expect that MC dropout will produce better results than a one-time inference with maximum likelihood prediction.

For each of the N teachers, we use the embeddings or feature representations, extracted from the penultimate layer (dense layer) of the teacher models to train a student network.

4.2. The Student Model

The student models are lighter (e.g., reduced number of layers) because the redundancy of the knowledge from the teachers is reduced after distillation. In addition to the labeled data, the student models are trained by the feature representations learned by the teachers. The goal of the student is to learn simpler feature representations that mirror the responses of the teacher.

We train a student model for each of the N teacher models. The final prediction is the average of the student model predictions. We use labeled and unlabeled data during training to guide the students to learn generalizable representations. The use of unlabeled data along with the supervision from the teacher representations, boosts the performance and the robustness of the students. Training students on MC dropout ensembles of N teacher predictions helps the students to benefit from the signal in the uncertainties even when a teacher makes mistakes. We train by alternating between batches of unlabeled and labeled data during each epoch. The use of unlabeled data in an iterative manner also has a regularization effect while training students, resulting in better generalization. We do not use dropout layers in training the student models.

4.3. Implementation Details

The prediction of emotional attributes is formulated as a regression problem implemented with DNNs. Our proposed approach is implemented with an ensemble of five teachers and five students (N=5). We implement the DNNs for the teacher with four dense layers, with 512 nodes each. We use *rectified linear unit* (ReLU) as the activation function for the hidden layers and a linear activation for the output layer.

As described in Section 4.1, we build different teacher models by using different dropout rates (p). In particular, we use $p \in \{0.45, 0.5, 0.55, 0.6, 0.65\}$ to train five separate teacher models. The network parameters are optimized on the development set, using *stochastic gradient descent* (SGD) with a learning rate of $r = 0.001$. We train the network to minimize the loss function $\mathcal{L}_{teacher} = (1 - CCC)$. The input to the network is the 6,373 dimensional feature vector described in Section 3.2. The features are normalized using the mean and standard deviation values estimated over the training samples. All teacher models are trained for 150 epochs, and inference is drawn over 100 MC dropout samplings. The output is a separate predicted feature embedding for arousal, valence and dominance.

For the student, we use a DNN with two dense layers, each of them implemented with 512 nodes. We optimize the network using NADAM optimizer with a learning rate of $r = 0.0001$. Equation 1 shows the objective function for training students, which consists of the supervised loss $(1 - CCC)$ for labeled data and unsupervised loss for the unlabeled data, which is implemented with the *mean squared error* (MSE).

$$\mathcal{L}_{student} = \alpha \cdot (1 - CCC) + \beta \cdot (MSE) \quad (1)$$

where α and β are hyperparameters $\in \{0.1, 0.2, \ldots, 0.9\}$ optimized on the development set. The output of each student is a prediction score for arousal, valence or dominance.

Table 1: *Performance of the baselines and T-S systems in terms of CCC.* † *indicates that one method leads to significantly better results than other methods (the differences between the T-S framework (test) and the T-S framework (unlabeled) are not statistically significant so we add* † *for both settings).*

Methods	Arousal	Valence	Dominance
Baseline	0.7045	0.3146	0.6336
Teachers' MC ensemble	0.7217	0.3184	0.6480
T-S framework (test)	**0.7345**†	**0.3230**†	**0.6652**†
T-S framework (unlabeled)	0.7322†	0.3219†	0.6625†
T-S framework (Pseudo-Label)	0.7290	0.3213	0.6558
T-S framework (top 75%)	0.7279	0.3205	0.6508

5. Results and Analysis

This section evaluates the use of our proposed method in SER. We create *single-task learning* (STL) baselines for arousal, valence and dominance using a similar DNN architecture to the teacher model (Sec.4.3). The baseline models are created by training the models for 200 epochs, optimizing their performance on the development set.

5.1. Performance of Teacher-Student Models

We evaluate and compare the performance of our proposed T-S framework trained under different settings. The reported results are the average over 10 trials with different random initializations. Table 1 shows the results, where our proposed approach, referred to as *T-S framework (test)*, is given in the third row. This model directly uses the test set as the unlabeled data. Notice that the emotional labels are not used during training. This approach is significantly better than any of the other methods reported in this study, with the exception of *T-S framework (unlabeled)* (one-tailed t-test over 10 trials, asserting significance when p-value≤ 0.01). The table shows that our approach is significantly better than the baseline, showing relative improvements of 4.25% for arousal, 2.67% for valence and 4.98% for dominance. The second row of the table presents the performance achieved by the ensemble of teachers ($N = 5$). This approach does not benefit from the unlabeled data, since it does not use the student models. Transferring knowledge to the students leads to relative improvements of 1.77% for arousal, 1.44% for valence and 2.65% for dominance. These results demonstrate the benefits of using our proposed T-S framework.

Our proposed formulation uses the test set as the unlabeled data. Alternatively, we can also use unlabeled data from the segments of the podcasts that have not been annotated with

Table 2: *Ablation study. A, B and C represent different techniques used in training our system. A: unlabeled data, B: MC dropout, and C: number of teachers and students in the ensembles).* * *indicates that the full model (first row) is significantly better than all other settings.*

A	B	C	Arousal	Valence	Dominance
✓	✓	5	**0.7345***	**0.3230***	**0.6652***
-	✓	5	0.7300	0.3211	0.6585
✓	-	5	0.7205	0.3154	0.6480
✓	✓	1	0.7240	0.3172	0.6512
-	✓	1	0.7219	0.3166	0.6556
✓	-	1	0.6873	0.2673	0.6198

emotional labels. We evaluate this setting by randomly picking 10,124 unlabeled segments, matching the number of sentences in the test set. The unlabeled data is independently selected for each of the 10 trials. This setting is referred to as *T-S framework (unlabeled)*. Table 1 shows that the performances between *T-S framework (unlabeled)* and *T-S framework (test)* are not statistically significant with a relative difference less than 0.4%. These differences are not statistically significant. Notice that the *T-S framework (unlabeled)* is also significantly better than other settings in the table, with the exception of the *T-S framework (test)*, validating the generalization claim of our formulation. We also explore two extra settings. First, we evaluate the use of pseudo labels, where predictions by the student ensemble on the test set are used as labels, augmenting the training set to retrain the models. The second setting restricts the unlabeled data used to train the student ensemble by considering only the samples from the test set with the lowest standard deviation estimated from the N teachers and 100 MC dropout samples. We only consider the top 75% sentences with the lowest uncertainty. The table shows that these two strategies do not lead to improvement over the *T-S framework (test)* model.

5.2. Analysis of Uncertainty in Predictions

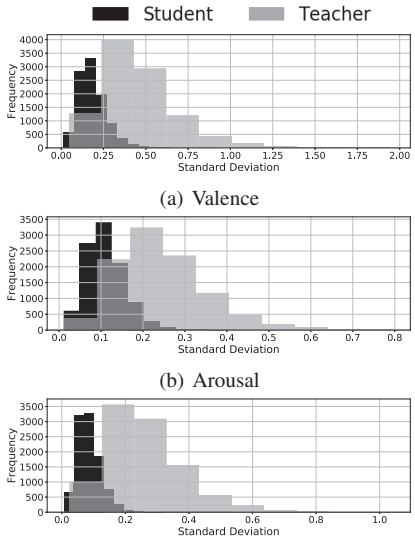
This section analyzes the consistency of the predictions of the teacher and student ensembles. We quantify the consistency of the predictions by estimating the standard deviations in the predictions. For the teacher models, we select one of the MC dropout samples, creating one prediction per teacher. Then, we estimate the standard deviation across the ensemble of five teachers. For the student models, we just estimate the standard deviation across the ensemble of five students.

Figure 2 shows the distribution of the standard deviations in predicting emotional attributes using the teacher and student ensembles. The standard deviations for the teacher are higher and more dispersed, showing that the consistency of the student ensemble is higher than the consistency in the teacher ensemble. These results show that using the average teacher embedding derived from the MC dropout is effective to guide the learning of the student ensemble. This analysis demonstrates that our proposed T-S framework not only improves the performance of the system, but also increases the consistency in the results.

5.3. Ablation Study

We perform an ablation study to understand the contributions of each feature of our method. We systematically remove components of our model, reporting the changes in performance. We compare whether the differences in performance between each case and the full model are significant using a paired one-tailed t-test over 10 trials, asserting significance at p-value$<$0.01.

Table 2 shows that the best performance is achieved by using all the components of our proposed method with both labeled and unlabeled data. The unlabeled data in this analysis



(a) Valence

(b) Arousal

(c) Dominance

Figure 2: *Consistency in the predictions of the teacher and student ensembles. The figures show the standard deviation histograms of the predictions from the N teachers and N students.*

corresponds to the test data, which is used during training without emotional labels. The second row of the table shows that the CCC values slightly decrease for all the emotional attributes without unlabeled data. The third row demonstrates the importance of using MC dropout to obtain an approximation of the probabilistic distributions of the teacher embeddings. The last three rows of the table show the drop in performance observed by training the system with a single T-S model (i.e., no ensembles). The last row is particularly interesting, showing that without MC dropout and the ensemble, the CCC values have a relative loss between 6.4% and 17.2% compared to the full model.

6. Conclusions

This study introduced a novel T-S framework that not only improves prediction of emotional attributes, but also leads to more consistent results. The approach includes an ensemble of teachers and ensemble of students. Under the T-S framework, the students are guided with feature embeddings from the teachers trained with MC dropout. The approach models uncertainty in the predictions, leading the student to obtain more consistent predictions. The experimental evaluation demonstrated that using MC dropout to obtain probabilistic teacher predictions increases the generalization capacity of the entire system by producing significant improvement in emotional prediction scores. As a T-S framework, we can also leverage unlabeled data, which leads to further improvements as demonstrated by the ablation study. We achieved relative gains in CCC up to 4.25% for arousal, 2.67% for valence and 4.98% for dominance over a baseline model. The experimental evaluation also showed that the student ensemble creates predictions with lower standard deviations than the teacher ensemble, improving its consistency.

For our future work, we would like to extend these ideas to other SER tasks (e.g., categorical classification, ranking). We will also explore alternative feature representations extracted directly from the raw input features that can describe dynamic variation along a speech segment, and can facilitate end-to-end training of the SER system.

7. Acknowledgements

Study supported by NSF (CNS-1823166; IIS-1453781).

8. References

[1] C. Busso, M. Bulut, and 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. New York, NY, USA: Oxford University Press, November 2013, pp. 110–127.

[2] 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.

[3] ——, "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.

[4] Z. Zhang, J. Deng, E. Marchi, and B. Schuller, "Active learning by label uncertainty for acoustic emotion recognition," in *Interspeech 2013*, Lyon, France, August 2013, pp. 2856–2860.

[5] F. Tao, G. Liu, and Q. Zhao, "An ensemble framework of voice-based emotion recognition system for films and TV programs," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, Calgary, AB, Canada, April 2018, pp. 6209–6213.

[6] 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.

[7] 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.

[8] 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.

[9] R. Lotfian and C. Busso, "Predicting categorical emotions by jointly learning primary and secondary emotions through multi-task learning," in *Interspeech 2018*, Hyderabad, India, September 2018, pp. 951–955.

[10] 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.

[11] R. Xia and Y. Liu, "A multi-task learning framework for emotion recognition using 2D continuous space," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 3–14, January-March 2017.

[12] 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.

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

[14] 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.

[15] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *In Advances in Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, December 2017, pp. 6402–6413.

[16] G. Pereyra, G. Tucker, J. Chorowski, Ł. Kaiser, and G. Hinton, "Regularizing neural networks by penalizing confident output distributions," *ArXiv e-prints (arXiv:1701.06548)*, pp. 1–11, January 2017.

[17] A. Malinin and M. Gales, "Predictive uncertainty estimation via prior networks," in *Advances in Neural Information Processing Systems (NIPS 2018)*, Montreal, QC, Canada, December 2018, pp. 7047–7058.

[18] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *Deep Learning and Representation Learning Workshop: NIPS 2014*, Montreal, QC, Canada, December 2014, pp. 1–9.

[19] C. Gurau, A. Bewley, and I. Posner, "Dropout distillation for efficiently estimating model confidence," *ArXiv e-prints (arXiv:1809.10562)*, pp. 1–11, September 2018.

[20] J. Wong and M. Gales, "Multi-task ensembles with teacher-student training," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2017)*, Okinawa, Japan, December 2017, pp. 84–90.

[21] S. Sun, Y. Cheng, Z. Gan, and J. Liu, "Patient knowledge distillation for BERT model compression," in *Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019)*, Hong Kong, China, November 2019, pp. 4323–4332.

[22] S. Park and N. Kwak, "FEED: Feature-level ensemble for knowledge distillation," *ArXiv e-prints (arXiv:1909.10754)*, pp. 1–8, September 2019.

[23] C. Zhang and Y. Peng, "Better and faster: knowledge transfer from multiple self-supervised learning tasks via graph distillation for video classification," in *International Joint Conference on Artificial Intelligence (IJCAI 2018)*, Stockholm, Sweden, July 2018, pp. 1135–1141.

[24] S. Albanie, A. Nagrani, A. Vedaldi, and A. Zisserman, "Emotion recognition in speech using cross-modal transfer in the wild," in *ACM international conference on Multimedia (MM 2018)*, Seoul, South Korea, October 2018, pp. 292–301.

[25] E. Provost and S. Narayanan, "Simplifying emotion classification through emotion distillation," in *Asia-Pacific Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Hollywood, CA, USA, December 2012, pp. 1–4.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

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

[33] ——, "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.