INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Speech Emotion Recognition with a Reject Option

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

Speech emotion recognition (SER) for categorical descriptors is a difficult task when the recordings come from everyday spontaneous interactions. The boundaries between emotional classes are less clear, resulting in complex, mixed emotions. Since the performance of a SER system varies across speech recordings, it is important to understand the reliability associated with its prediction. An intriguing formulation in machine learning related to this problem is the reject option, where a classifier only provides predictions over samples with reliability above a given threshold. This paper proposes a classification technique with a reject option using *deep neural networks* (DNNs) that increases its performance by selectively trading its coverage in the testing set. We use two different criteria to develop a SER system with a reject option, where it can accept or reject a sample as needed. Using the MSP-Podcast corpus, we evaluate this idea by comparing different classification performance as a function of coverage. By selectively defining a coverage of 75% of the samples, we obtain relative gains in F1-score of up to 25.71% for a five-class problem and 20.63% for an eight-class problem. The sentences that are rejected are analyzed in the evaluation, confirming that they have lower inter-evaluator agreement.
Index Terms: Emotion Classification, Deep neural Networks, Inter-evaluator agreement, F1-Score

1. Introduction

Speech emotion recognition (SER) is an important research area due to its wide applications in *human computer interactions* (HCIs) and behavioral studies [1]. Therefore, it is important to design SER systems that lead to reliable predictions. An important problem in SER is the classification of categorical emotions, such as happiness, anger and sadness (basic emotion theory [2]). The boundary between categorical emotions becomes less clear with natural, spontaneous recordings, which are characterized by mixed emotions [3–5]. People in everyday interactions exhibit complex emotional behaviors, which cannot be aways described by a single emotional class with absolute certainty. This challenge is one of the key reasons for the low inter-evaluator agreements that are often observed in emotional perceptual evaluations, as different raters have different perceptions after listening to the same recording. SER systems built to recognize categorical emotions also face similar challenges.

Studies have shown that the reliabilities of SER systems vary according to the emotional ambiguity of the speech, where the classification performances increase when they are evaluated with recordings that are consistently annotated [6]. It is important to design architectures that not only have high accuracy, but also provide information about their confidence [7, 8]. For example, it is valuable to have a SER system that can distinguish cases where its confidence is low to make a reliable decision. An appealing formulation in machine learning is to build a classifier with a reject option [9–11], where the system refuses to provide a prediction for uncertain samples. This paper addresses this problem by introducing a SER system with a reject option, balancing the tradeoff between classification accuracy and coverage in the test set (i.e., percentage of accepted samples). While studies have proposed confidence metrics to describe the prediction of the systems, the implementation of a reject option in SER is a novel formulation.

We train a DNN with a reject option using two different criteria to accept or reject samples from the test set. The first criterion learns the rejection function using the *selective guaranteed risk* (SGR) algorithm [12]. This algorithm uses the *softmax response* (SR) mechanism, where a threshold is applied on the softmax probabilities of the output layer of a DNN. The goal is to learn a rejection function that guarantees a desired risk, or error rate, with high probability. The second criterion uses the difference between the two highest softmax outputs to set a threshold. We reject the sample if the difference is below a given threshold. The F1-scores of both methods are evaluated as a function of the coverage of the classifiers. We observe relative gains in F1-score up to 25.71% for a five-class problem and 20.63% for an eight-class problem by rejecting only 25% of the testing set. These results show the benefits of using a reject option in SER problems.

2. Related Work

The concept of a reject option in classifiers has been used in various machine-learning problems including gene-expression classification [13], odor classification [14], and vocal fold paralysis [15]. Studies have included a reject option in binary [16], multi-class [17], and multistage [9] classifications. Studies have often discussed the concept of a reject option in the context of classifiers built with *support vector machines* (SVMs) [10, 16], *nearest neighbor* (KNN) [11], ensemble of classifiers [18], boosting algorithms [19], and DNNs [20, 21]. The use of a classifier with a reject option has not been studied in SER problems.

The use of a reject option in neural networks often considers costs or penalties assigned to misclassified samples, or rewards assigned to correctly classified samples [20, 21]. An alternative approach is to formulate the problem in terms of risk and coverage [12, 22]. Geifman et al. [12] proposed a classifier with a reject option for the problem of object recognition on images. They propose a *selective guaranteed risk* (SGR) algorithm to learn a rejection function on the classifier. They introduced the notion of risk and coverage to learn a selection/rejection function that achieves a desired risk with very high probability. The goal of the algorithm is to find the right threshold that minimizes the generalization error (i.e., target risk) of the classifier, while keeping the maximum test coverage. They evaluated the classification performance on the CIFAR and ImageNet databases, achieving a very low error (2%) on the test data using 60% test coverage. Our study follows this formulation.

Our study is closer to the work of Deng et al. [7], where a confidence measure for SER was defined based on human labeler agreement. The approach uses the FAU Aibo emotion corpus, which has emotion labels as well as human agreement scores assigned for each sample. They use an SVM classifier to recognize the emotions. They also train scoring models using the human agreement levels as ground truth. These scoring models estimate the agreement level a sentence. These estimates are combined to define a confidence score, which was shown to be highly correlated with the unweighted average recall of the classifier. Deng et al. [8] explored the reliability of

http://dx.doi.org/10.21437/Interspeech.2019-1842

predictions of a SER system by training classifiers on multiple databases with manually created confidence levels. They used a semi-supervised learning approach to gradually include data from the target domain on the train set during this iterative procedure. Classifiers trained on multiple corpora predict various confidence ratios for each instance from the target domain. They aggregated these confidence ratios to calculate a confidence measure. Both of these studies have used SVMs to design classifiers [7, 8]. However, our approach is implemented with DNNs, since they have achieved better performance than SVMs [23, 24].

3. Resources

3.1. The MSP-Podcast Database

This study uses the MSP-Podcast corpus [25]. The database is a collection of spontaneous speech samples rich in emotional content, obtained from various audio-sharing websites. The speech samples are formed by segmenting the podcasts into speaking turns with duration between 2.75 and 11 seconds, with no background music or overlapped speech [25]. The study uses version 1.4 of the corpus, which consists of 33,262 speaking turns (56h 29m). In this set, we have identified and manually annotated the speaker identity of 30,070 sentences (703 speakers). The database is split into train, test and validation partitions, aiming to obtain speaker independent partitions. The test set has 9,255 samples from 50 speakers, the validation set has 4,300 samples from 30 speakers, and the train set has 19,707 samples from the rest of the speakers, including the segments without speaker information. The database is annotated with emotional labels using *Amazon Mechanical Turk* (AMT). The evaluation uses a variation of the crowdsourcing protocol discussed in Burmania et al. [26] to track the performance of the workers in real-time. Each speech sample is annotated by at least five annotators. While the corpus has categorical and attribute-based descriptors, this study uses primary categorical emotions: anger, sadness, happiness, surprise, fear, disgust, contempt, and neutral states. More details on this corpus is provided in Lotfian and Busso [25].

3.2. Acoustic Features

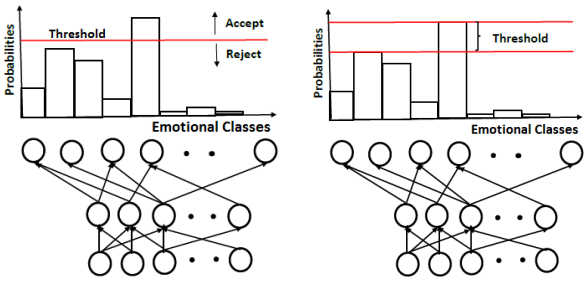
This study uses the acoustic features proposed for the Interspeech 2013 computational paralinguistics challenge [27]. We extract the features using the OpenSmile toolkit [28]. The set consists of *low level descriptors* (LLDs) and *high level descriptors* (HLDs) extracted from each speaking turn. The LLDs consist of frame-based features such as energy, fundamental frequency and *Mel-frequency cepstral coefficients* (MFCCs). The HLDs are sentence-level statistics from the LLDs (e.g., mean of the energy). The approach creates a 6,373 dimension feature vector for each utterance, regardless of its length.

4. Proposed Method

The main contribution of this study is the use of a reject option for SER problems, where the classifier can decline to make a prediction when it is not confident. A key challenge in this formulation is to define a meaningful criterion to accept or reject a sample. This study evaluates two criteria, which are explained in this section.

4.1. Criterion 1: Threshold on the Neuronal Activations

The main objective in the first criterion is to minimize the empirical risk of the selective classifier, while keeping the test coverage as high as possible. We use the SGR algorithm introduced by Geifman et al. [12] to construct a selection function (g) that guarantees a desired risk (r^*) with high probability. The learning of a selection function can be formulated as learning an optimal risk bound on the classifier. This goal is achieved by setting an optimal threshold on the softmax probabilities of the output layer that guarantees a desired error rate with high confidence.



(a) Criterion 1 (b) Criterion 2

Figure 1: *Defining thresholds to reject or accept samples using criteria 1 or 2.*

The empirical selective risk ($\hat{r}(f, g|S_m)$) is defined as the ratio of the expected loss ($\frac{1}{m}\sum_{i=1}^{m} l(f(x_i), y_i)g(x_i)$) and the coverage of the selective classifier ($\hat{\phi}(f, g|S_m)$),

$$\hat{r}(f, g|S_m) = \frac{\frac{1}{m}\sum_{i=1}^{m} l(f(x_i), y_i)g(x_i)}{\hat{\phi}(f, g|S_m)} \quad (1)$$

where, f is a classifier, g is a selection function indicating that the classifier accepts a sample when $g(x_i) = 1$ and rejects a sample when $g(x_i) = 0$. S_m is the labeled training samples ($\{(x_i, y_i)\}_{i=1}^{m}$) and $\hat{\phi}(f, g|S_m) \triangleq \frac{1}{m}\sum_{i=1}^{m} g(x_i)$ is the empirical coverage. Since the algorithm uses a 0-1 loss, the empirical risk of the classifier can be interpreted as the generalization error of the classifier.

The algorithm uses the softmax response technique, where the threshold is imposed on the neuronal responses of the layer previous to the final softmax layer. The activation responses are reflected on the final softmax probabilities. Therefore, the threshold on the activations is equivalent to imposing the threshold on the softmax probabilities (Fig. 1(a)). The algorithm chooses an appropriate threshold using the validation set to guarantee that the empirical risk is close to the desired risk with very high confidence (we set the confidence at 99.99%) (Eq.2). The optimal thresholds are used in the test set, defining different levels of coverage.

$$Pr_{S_m}\{\hat{r}(f, g|S_m) < r^*\} > 99.99\% \quad (2)$$

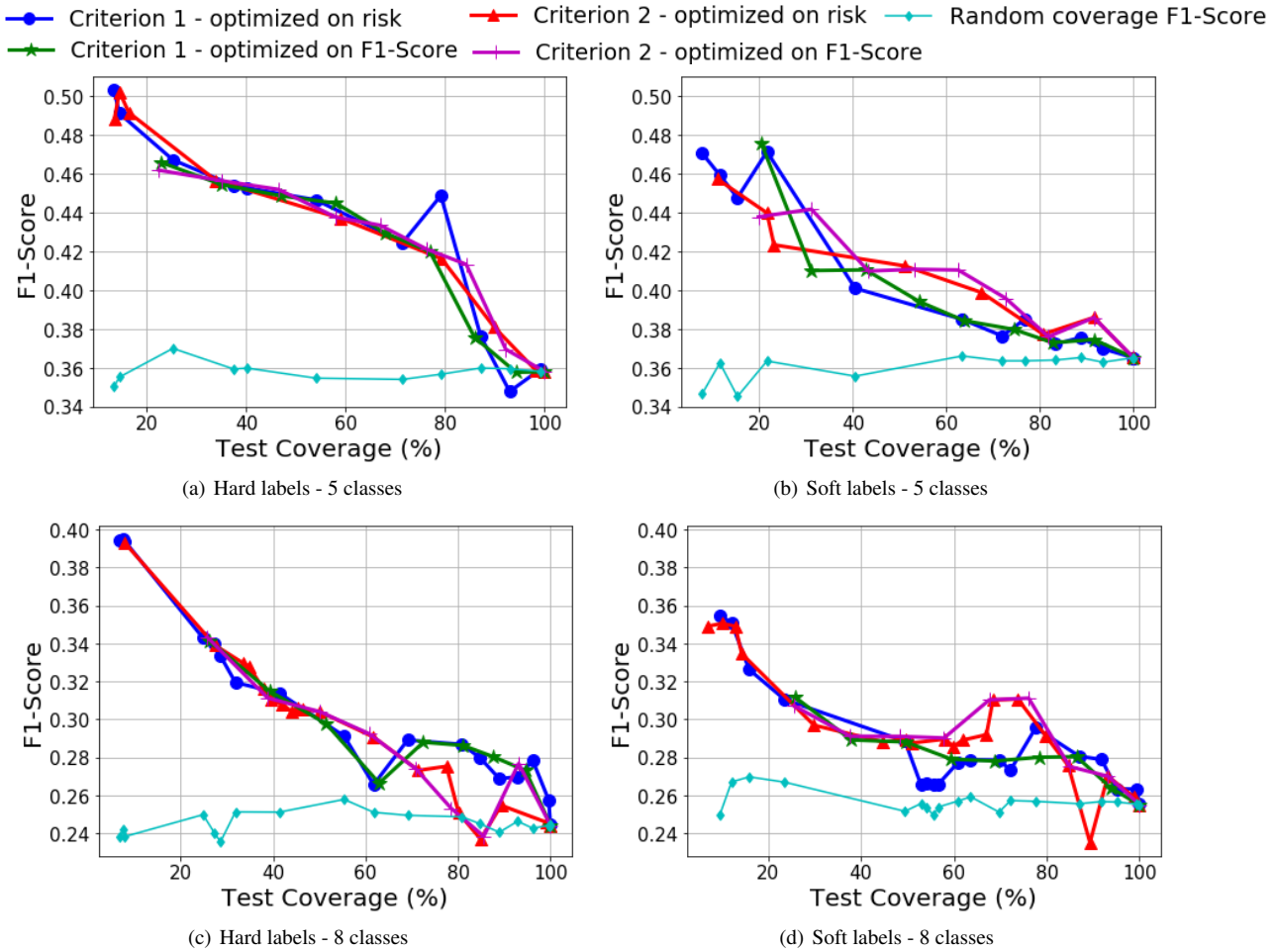
4.2. Criterion 2: Two Highest Prediction Values Difference

The second criterion defines the threshold by estimating the difference between the two predictions with the highest values (Fig. 1(b)). We expect a big difference for confident cases, indicating that the top choice is the clear prediction. This criterion has been used by Mitra and Franco [29] to deal with unseen data in *automatic speech recognition* (ASR).

Since we use a multicalss classification approach to recognize categorical emotions, the softmax output of a trained emotion classifier is a probability vector, indicating the prediction probability for each emotional class. We calculate the difference between the highest and second highest predicted probability values, setting a threshold on this difference (Fig. 1). The thresholds are also optimized on the validation set, defining different levels of coverage.

4.3. Optimization of the Models

We optimize the selective classifier using two different techniques. The first alternative is to optimize the empirical risk of the classifier. We use the SGR algorithm to choose an appropriate threshold, such that the generalization error of the classifier is brought closer to a desired error rate (Eq.2). The two errors are tightly bound with very high probability. The second alternative is to optimize the F1-score achieved on the validation



(a) Hard labels - 5 classes

(b) Soft labels - 5 classes

(c) Hard labels - 8 classes

(d) Soft labels - 8 classes

Figure 2: *F1-score as a function of the test coverage. The figures consider the two criteria to define the reject option, and the two optimization methods.*

set. We use these optimization methods for the two proposed criteria to introduce a reject option in a classifier, creating four different implementations.

5. Experimental Evaluation

The experimental evaluation considers the two criteria to introduce a reject option with the optimization methods. We evaluate the F1-score of the models versus the test coverage, which decreases as we reject more samples. To compute the F1-score, we estimate the precision and recall rates for each class, calculating their average across classes. Then, we use the average precision and average recall rates to estimate the F1-score. This approach gives the same weights to each emotional class, which is important as the corpus is not balanced across emotions.

The recognition of categorical emotions is formulated as a classification problem implemented with DNNs. We consider two classification problems. The first problem is an eight-class problem (happiness, sadness, anger, surprise, fear, disgust, contempt and neutral state). The second problem is a five-class problem (happiness, sadness, anger, disgust and neutral state), removing some of the classes with fewer samples.

We implement a DNN with three dense layers with $1,024$ nodes per layer. The DNNs are implemented with *rectified linear unit* (ReLU) for intermediate layers and a softmax activation for the output layer. We use a dropout of $p = 0.5$ and a batch normalization for the hidden layers. Batch normalization has been proven to improve the performance of DNNs [24, 30, 31] by reducing the covariance shift and normalizing the output of each layer, leading to better gradient flow and faster training. We use *adaptive moment estimation* (ADAM) with a learning rate of 0.0002 to optimize the parameters of the network. We train the network to minimize the cross-entropy loss function. The input to the network is the 6,373 dimension feature vector described in Section 3.2. The features are standardized using the mean and standard deviation values estimated over the training samples. The softmax layer outputs a vector with the probabilities associated with the emotional classes. We train the models with hard and soft labels. The hard labels are obtained by estimating the consensus across the annotations of each speaking turn, using the majority vote rule. An alternative method is to use soft labels to estimate the proportion of the labels assigned to the speaking turns (e.g., happiness=0.7; neutral=0.3; other emotions=0). Soft-labels have been successfully used in SER problems [32, 33].

6. Results and Analysis

This section evaluates the use of a reject option in SER problems. We create a baseline system for the five-class and eight-class problems using a similar DNN structure, without the reject option (i.e., 100% coverage). We train the classifier for 200 epochs, optimizing its performance on the validation set. We evaluate the performance of the classifier with the reject option by calculating the F1-score on accepted samples from test set.

6.1. Five-Class Problem

We select the speaking turns with consensus labels belonging to the five emotions (happiness, anger, sadness, disgust, and neu-

Table 1: *Relative gains in F1-scores achieved for a 75% coverage on the test data. The table lists the results for the two criteria to reject subjects, using the two optimization methods. It also lists the results when we select random samples until achieving the desired coverage.*

Proposed methods	Gain in F1-Score (%) for a 75% test coverage			
	Hard 5	Soft 5	Hard 8	Soft 8
Criterion 1 optimized on risk	25.71	6.30	16.39	17.64
Criterion 1 optimized on F1-Score	17.00	4.10	17.62	10.00
Criterion 2 optimized on risk	19.90	6.20	12.70	18.82
Criterion 2 optimized on F1-Score	17.27	6.84	6.55	20.63
Random Sampling	-0.84	-0.68	1.80	0.70

tral states). For consistency, we remove the samples in the corpus without consensus, even when using soft labels to train and evaluate our models.

Figures 2(a) and 2(b) show the performance of the classifier with a reject option learned using the two different criteria. The performance of the baseline corresponds to the point in the figures when the coverage is 100%. As expected, we observe that the gain in performance increases as the test coverage decreases. As the classifier rejects ambiguous samples, its confidence in predicting the accepted samples increases. The performances across criteria and optimization methods are very similar. The two methods that tend to have better F1-score are criterion 1 with risk optimization, and criterion 2 with F1-score optimization. In general, training with hard labels leads to better performance. The line labelled as "Random coverage F1-Score" corresponds to the performance of a classifier without a reject option, evaluated on test samples selected at random until matching the test coverage of the classifier with the reject option. The results show that the classifier with reject option consistently performs better than the one without a reject option for different coverage of the test data, indicating that the classifier is selective in rejecting the samples.

Table 1 gives a comparison of the relative gains in F1-scores achieved over the baselines by using different criteria to introduce a reject option for a coverage of 75% (e.g., we reject 25% of the test set). The first two columns of the table lists the performance for the five-class problem. We observe relative gains up to 25.71% with hard labels (criterion 1, risk optimization) and 6.84% with soft labels (criterion 2, F1-score optimization).

6.2. Eight-Class Problem

For the eight-class problem using hard and soft labels, we use samples in the database with consensus labels associated to happiness, sadness, anger, surprise, fear, disgust, contempt and neutral state.

Figures 2(c) and 2(d) lists the results showing consistent results as the one observed for the five-class problem. The classifiers with a reject option achieve better classification as the coverage decreases. When the test coverage is between 60% and 80%, the classifier trained with soft labels lead to better performance. The last two columns of Table 1 show that for a 75% test coverage, we achieve relative gains over the baseline up to 17.62% for hard labels (criterion 1, F1-score optimization), and 20.63% for soft labels (criterion 2, F1-score optimization). Notice that the performance gains are always higher than the ones observed for the random sampling condition, indicating that the classifiers are able to effectively identify samples that lead to higher performance.

Table 2: *Inter-evaluator agreement of the accepted and rejected samples for different test coverages.*

		Inter-evaluator agreements (Fleiss Kappa)	
	Test Coverage (%)	Accepted Samples	Rejected Samples
Hard labels 5 classes	100	0.2642	-
	75	0.2773	0.2590
	50	0.2897	0.2651
	25	0.3080	0.2633
Soft labels 8 classes	100	0.2680	-
	75	0.2723	0.2450
	50	0.2842	0.2496
	25	0.2983	0.2563

6.3. Analysis of Accepted and Rejected Samples

This section studies the inter-evaluator agreement of the sentences that are either accepted or rejected by looking at their annotations. We use the Fleiss' Kappa statistic to quantify the agreement between annotations. For the analysis, we only consider the best system for the five-class problem (hard label, criterion 1, risk optimization) and the eight-class problem (soft label, criterion 2, F1-score optimization).

Table 2 shows the inter-evaluator agreement scores obtained on the accepted and rejected samples for different levels of coverage on the test data. As the test coverage decreases, we consistently observe that the agreement for accepted samples increases. The inter-evaluator agreement for accepted samples is always higher than the inter-evaluator agreement for rejected samples. This result suggests that the rejected samples tend to be more emotionally ambiguous, where even human evaluators disagree. The two criteria used to reject samples (Sec. 4) are effective in rejecting these samples, where many of the accepted samples with high inter-evaluator agreement satisfy both of these criteria.

7. Conclusions

This study showed that a classifier with an option to abstain from giving a prediction when its confidence is low can be successfully designed without compromising much on its test coverage. The reject option is a valuable feature, increasing the confidence in the SER system when it gives a prediction. The study evaluated two criteria, which dictate the acceptance or rejection of a sample. The first criterion set a threshold on the neuronal activations of the output layer. The second criterion set a threshold on the difference between the top two classes with the highest probabilities predicted by a classifier. The results validated the proposed criteria to reject samples, observing better performance as we allow the network to reject more samples. When the coverage is 75% of the test set, the relative gains in F1-score were up to 25.71% for a five-class problem, and 20.63% for an eight-class problem. The analysis of the annotations revealed a lower inter-evaluator agreement for the rejected samples, suggesting that the proposed criteria were effective in rejecting samples with ambiguous emotions.

Introducing a classifier with a reject option in SER problems can have important implications, such as improving the precision and efficiency of SER models implemented to solve real world problems in behavioral or psychological studies. As future work, we intend to extend this framework to regression problems that predict the emotional attributes arousal, valence and dominance. We are also exploring alternative rejection option criteria that increase the F1-score while maintaining the coverage as high as possible.

8. Acknowledgements

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

9. References

[1] P. Georgiou, M. Black, and S. Narayanan, "Behavioral signal processing for understanding (distressed) dyadic interactions: some recent developments," in *joint ACM workshop on Human gesture and behavior understanding (J-HGBU 2011)*, Scottsdale, Arizona, USA, December 2011, pp. 7–12.

[2] P. Ekman, "An argument for basic emotions," *Cognition and Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

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

[7] 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.

[8] 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.

[9] P. Pudil, J. Novovicova, S. Blaha, and J. Kittler, "Multistage pattern recognition with reject option," in *IAPR International Conference on Pattern Recognition (ICPR 1992)*, vol. 2, The Hague, Netherlands, August-September 1992, pp. 92–95.

[10] G. Fumera and F. Roli, "Support vector machines with embedded reject option," in *Pattern Recognition with Support Vector Machines: International Workshop on Support Vector Machines*, ser. Lecture Notes in Computer Science, S. Lee and A. Verri, Eds. Niagara Falls, ON, Canada: Springer Berlin Heidelberg, 2002, vol. 2388, pp. 68–82.

[11] M. Hellman, "The nearest neighbor classification rule with a reject option," *IEEE Transactions on Systems Science and Cybernetics*, vol. 6, no. 3, pp. 179–185, July 1970.

[12] 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.

[13] B. Hanczar and E. Dougherty, "Classification with reject option in gene expression data," *Bioinformatics*, vol. 24, no. 17, pp. 1889–1895, September 2008.

[14] 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.

[15] C. Kotropoulos and G. Arce, "Linear classifier with reject option for the detection of vocal fold paralysis and vocal fold edema," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, pp. 1–13, July 2009.

[16] Y. Grandvalet, A. Rakotomamonjy, J. Keshet, and S. Canu, "Support vector machines with a reject option," in *Advances in neural information processing systems (NIPS 2009)*, Vancouver, BC, Canada, December 2009, pp. 537–544.

[17] D. Tax and R. Duin, "Growing a multi-class classifier with a reject option," *Pattern Recognition Letters*, vol. 29, no. 10, pp. 1565–1570, July 2008.

[18] K. R. Varshney, "A risk bound for ensemble classification with a reject option," in *IEEE Statistical Signal Processing Workshop (SSP 2011)*, Nice, France, 2011, pp. 769–772.

[19] C. Cortes, G. DeSalvo, and M. Mohri, "Boosting with abstention," in *Advances in Neural Information Processing Systems (NIPS 2016)*, Barcelona, Spain, December 2016, pp. 1660–1668.

[20] C. De Stefano, C. Sansone, and M. Vento, "To reject or not to reject: that is the question-an answer in case of neural classifiers," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 30, no. 1, pp. 84–94, February 2000.

[21] L. P. Cordella, C. De Stefano, F. Tortorella, and M. Vento, "A method for improving classification reliability of multilayer perceptrons," *IEEE Transactions on Neural Networks*, vol. 6, no. 5, pp. 1140–1147, September 1995.

[22] 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.

[23] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: raising the benchmarks," in *IEEE international conference on acoustics, speech and signal processing (ICASSP 2011)*, Prague, Czech Republic, May 2011, pp. 5688–5691.

[24] 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.

[25] 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. To appear, 2019.

[26] 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.

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

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

[29] V. Mitra and H. Franco, "Interpreting DNN output layer activations: A strategy to cope with unseen data in speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, Calgary, AB, Canada, April 2018, pp. 5724–5728.

[30] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning (PMLR 2015)*, vol. 37, Lille, France, July 2015, pp. 448–456.

[31] H. Fayek, M. Lech, and L. Cavedon, "On the correlation and transferability of features between automatic speech recognition and speech emotion recognition," in *Interspeech 2016*, San Francisco, CA, USA, September 2016, pp. 3618–3622.

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

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