SMM19, Workshop on Speech, Music and Mind 2019
14 September 2019, Vienna, Austria

No Sample Left Behind: Towards a Comprehensive Evaluation of Speech Emotion Recognition Systems

Pablo Riera[1], *Luciana Ferrer*[1], *Agustín Gravano*[1,2], *Lara Gauder*[1,2]

[1]Instituto de Investigación en Ciencias de la Computación (ICC), CONICET-UBA, Argentina
[2]Departamento de Computación, FCEyN, Universidad de Buenos Aires (UBA), Argentina
priera@dc.uba.ar, lferrer@dc.uba.ar, gravano@dc.uba.ar, mgauder@dc.uba.ar

Abstract

Human agreement for the task of labeling speech utterances with emotion information is usually low, especially for natural speech, where emotions could be ambiguous or subtle. For this reason, datasets of emotional speech are generally labeled by several human annotators. The common practice in speech emotion recognition (SER) literature is to summarize the multiple labels provided by the annotators for a sample into a single one by choosing the majority label. The problem with this approach is that a significant proportion of samples may not be assigned a majority label. These samples are usually ignored for system evaluation, along with any samples initially labeled by the annotators as being from emotions other than the emotions of interest for the specific dataset. This implies that the estimation of emotion recognition performance is incomplete. We do not know how the system will behave when presented with those ambiguous samples, which will certainly appear in practice. In this paper, we analyze the effects that these samples have in system performance and propose different ways to use the multiple labels available from the annotators during evaluation and to assess system performance without discarding any samples.

Index Terms: emotion recognition, deep neural networks, performance metrics.

1. Introduction

The task of recognizing emotions from speech is generally framed as a classification problem. The aim is to determine which emotion, within a predefined set (e.g., angry, sad, happy and neutral), is present in a certain speech utterance [1, 2, 3]. In doing this, an assumption is made that an utterance can only correspond to one and only one of the emotions the system was trained to detect. Traditionally, datasets used for development and evaluation of speech emotion recognition (SER) systems are composed of acted speech [2]. Subjects, generally actors, are asked to read phrases using clearly-defined emotions. This scenario fully complies with the assumptions made by classification systems: each utterance corresponds to a single emotion selected from a predefined set. While these datasets are convenient for research and development of SER systems, the speech they contain is not representative of most of the speech the systems would encounter when deployed.

Emotion datasets composed of relatively more natural speech compared to the ones described above are also available [4, 5, 6]. In some datasets, subjects are presented with a task that is meant to elicit certain emotions [7, 6]. In others, they are asked to act out emotional scenarios (scripted or improvised) without indicating the emotion they should use for each phrase [4]. Finally, other datasets are composed of speech harvested from previously existing datasets [5]. In all these cases, the labels are obtained from human listeners. A few annotators are asked to label the utterances with the perceived emotion, selected from a predefined set of emotions of interest. In many cases, annotators are allowed to assign more than one label per utterance [4, 5, 8] or a label of "other" for utterances they cannot assign to any of the emotions of interest. In other cases, labels are continuous (valence and activation levels) [5] or assigned at very short intervals of time, or over words [6]. In general, in all these datasets, the agreement across annotators is low [4].

In order to use these datasets for development and evaluation of emotion classification systems, the labels given by the annotators have to be manipulated to obtain a single emotion per utterance. This is usually done by selecting the emotion that was chosen the most by annotators, the majority label. Yet, since the agreement across annotators is low, a majority label may not exist for all utterances (e.g., if an utterance is labelled with three different emotions by three different annotators). The standard procedure to deal with this issue is to simply discard this samples both for training and evaluation of the systems [9, 10]. Furthermore, samples labelled by the annotators as "other" are also generally discarded. As a consequence, the performance metrics presented for the SER systems in the literature ignore a significant percent of samples that are ambiguously labelled or from emotions other than the ones defined by the authors of the dataset as emotions of interest. This means that we do not actually know what the performance of these systems would be in practice, when systems are deployed and used to classify natural speech utterances.

In this paper, we propose different ways to evaluate system performance without having to discard samples that cannot be assigned to one of the emotions of interest with the majority rule. First, we show the effect that the samples without majority label would have in performance metrics usually used to evaluate SER systems. We then propose that the emotion recognition problem might be better solved as a set of detection tasks and show system performance using this approach.

Note that we do not explore the issue of how to use samples without agreement during training since this is an area that has been more thoroughly investigated in the literature compared to the issue of evaluation. In the future we plan to compare different training methods that include samples without agreement using our proposed metrics.

2. Related Work

Some papers explored ways of using multiple annotator labels during training rather than using the majority label [11, 12, 13, 14]. The approaches generally involve turning the set of labels given by the annotators into a vector of frequencies for each emotion (soft labels). A metric like cross-entropy or a standard or weighted distance between the posterior probabilities predicted by the system and the soft labels can then be used as objective function to train the models. These same metrics can

10.21437/SMM.2019-3

also then be used to evaluate the system performance, even on samples that do not have a majority label. Yet, these metrics are not easy to interpret. They can be used to compare performance across systems, but they cannot be directly used to understand how frequently our system will make errors of one kind or another. Mower et al. [15] proposed an SVM-based method and evaluated it on samples without a majority label using standard SER metrics by considering any of the labels assigned by the annotators as correct. Yet, they still discard samples that do not have any of the target emotions considered in the work. Kim et al. [13] showed accuracy results on samples without a majority label as well as cross-entropy between their predicted posteriors and the frequency of labels assigned by the annotators. Finally, Li et al proposed to use detection metrics to evaluate performance on a sentiment analysis task [16]. Our work builds upon these previous works doing further analysis and proposing new metrics to evaluate performance on complete rather than partial datasets.

3. Methods

In this section we describe the data, systems and metrics used in our experiments. As stated above, we use standard SER systems since the goal of this work is to investigate the evaluation of SER systems rather than to optimize them.

3.1. Dataset

We use the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset [4]. IEMOCAP has a length of approximately 12 hours and is comprised of scripted and improvised dialogues. It is divided into 5 sessions with 2 actors each. Annotators were asked to label each sample choosing one or more labels from a pool of emotions including angry, happy, sad, neutral, frustrated, excited, fearful, surprised, disgusted, and other. In this work, following [9], we relabel excited as happy, and fearful, surprised and disgusted as "other".

In most of the SER literature, the labels provided by the annotators are summarized into a single label using the majority rule: the emotion that was chosen by more than half of the annotators is selected as the label for the sample. If no emotion fulfills this requirement, the label is considered unassigned and discarded from evaluation (and, generally, also from training). With the way of grouping emotions described above the percent of samples without a majority label for IEMOCAP is 20%.

One of the main goals of this paper is to study the problem of evaluating the performance of a SER system on the unassigned samples. To this end, we use the labels provided by all annotators by converting the set of labels for each sample into a vector of frequencies as done, for example, in [13]. If an annotator selected N emotions, each of them counts as $1/N$ when computing the vector of frequencies. These frequencies can be used directly for some metrics, or thresholded, labelling each sample with all the emotions for which the label frequency was larger than a certain *annotation* threshold. This means that depending on the threshold, samples can have more than one "correct" emotion. Table 1 shows examples of different sets of labels. The majority label for a sample coincides with the one selected with an annotation threshold of 0.5. For our exploration, we compare results obtained with the 0.5 threshold and a more permissive one of 0.2. This value is chosen since we wanted to allow for samples to be labelled with three different labels, one from each of the annotators, and a few other less frequent combinations.

The proposed way of using the multiple labels scales up when the number of annotator increases since it is based on the frequency of appearance of each emotion. If the rule had been that all emotions chosen by any annotator are considered correct (as done in [15]), then the more annotators are used, the more likely it is that all possible emotions will eventually be considered correct for a certain sample.

Table 1: *Examples of the different labelling criteria. First column has raw labels from the three annotators, second column has frequencies of those labels for anger, frustration, happy, neutral, and sadness (in that order), third and fourth columns shows thresholded labels with levels of 0.5 and 0.2, respectively. Labels selected with a threshold of 0.5 are the majority labels.*

Labels	**Label freq.**	**Thr=0.5**	**Thr=0.2**
neu, neu, neu	0, 0, 0, 1, 0	neu	neu
fru, ang, fru	1/3, 2/3, 0, 0, 0	fru	fru,ang
sad, fru/sad, neu	0, 1/6, 0, 1/3, 1/2	sad	sad,neu
hap, neu, fru	0, 1/3, 1/3, 1/3, 0	unassigned	hap,neu,fru

3.2. Emotion Recognition Systems

We focus on the problem of evaluation of SER systems, assuming a system that assigns a set of scores for each sample, one for each emotion of interest. For our experiments we use two systems, one based on a Time Delay Neural Network (TDNN) and another one based on Support Vector Machines (SVM).

In the SVM system, the classification problem is treated as one-vs-rest, a common practice in multi-label problems [17]. Each of the SVMs is then a detector of one of the emotions of interest. We use a radial basis function (RBF) kernel and a C parameter equal to 1.0. The features used for the SVM are the 2016 updated version of those proposed for the INTERSPEECH 2013 Computational Paralinguistics Challenge (ComParE) [18]. They were extracted using OpenSMILE [19].

The architecture for the TDNN is a simpler version of [20]. The network comprises two time delay layers with contexts (512 nodes and context of $[-2, -1, 0, 1, 2]$ for the first layer, and 256 nodes and context $[-1, 0, 1]$ for the second layer), a pooling layer with mean and standard deviation, and a softmax output layer. Batch normalization and dropout with probability 0.5 was used at the output of all layers except the last one. The features used for the TDNN are based on [10] and are the mel frequency cepstral coefficients (MFCC), pitch, energy and voice quality features with first order deltas extracted every 10ms.

Both systems are trained using only the subset of the training data for which a majority label is available. Tuning of all parameters for both systems was done using the fifth session. The results in this paper are presented over the other four sessions using cross-validation leaving one session out at a time. The imbalance across classes is compensated for during training of both systems by weighting the objective function with the inverse of the frequency of the classes for each sample.

3.3. Metrics for Classification

The most commonly used metrics in SER publications are *accuracy* and *average recall* (sometimes referred to as weighted and unweighted average recall, respectively). These two metrics are computed assuming that a single emotion is predicted for each sample. The prediction is generally done by choosing the emotion with the highest score (e.g., posterior or distance from the hyperplane) from the system. Given these predictions, the accuracy is computed as the percent of total samples that are correctly labelled. When using an annotation threshold to define multiple correct emotions, any of those labelled emotions is considered correct. The recall is defined as the percent of positive samples correctly detected as positive. For SER, we compute a separate recall for each emotion, considering only samples labelled with that emotion as positive and the rest as negative. The average recall is simply the average of all these emotion-dependent recall values (the “other” category is excluded in this average). Note that the accuracy is sensitive to the prior of the classes, while average recall is not. Also note that samples labelled as “other” are never correctly labelled since our systems were not trained to predict this label.

The naïve baseline for the accuracy is given by a system that selects always the most frequent emotion. This will depend on the way the samples are labelled (whether by selecting the majority label or by a threshold). In our results, we use the best baseline for each experiment: the one that selects the most frequent emotion for the specific labelling method being used on the subset of data being evaluated. The baseline for average recall is simply the inverse of the number of classes and it corresponds to a system that selects always the same emotion (any of them), since the recall for that emotion would be 1 and for the rest would be 0, averaging $1/N$, where N is the number of target emotions.

3.4. Metrics for Detection

We propose an alternative way of evaluating SER systems by considering the problem as a set of detection tasks, one for each target emotion (excluding “other”). That is, the scores from the system for each emotion can be used to make a binary decision about whether that emotion is present in the sample or not. The decision is made by comparing the score with a threshold. If the score is above the threshold, the emotion is considered to be present in the sample. Otherwise, it is considered missing from the sample. This may result in more than one emotion detected for each sample. We believe this is a reasonable output for a system, since humans may also label a sample with more than one emotion. Interestingly, the same idea of posing the task as a set of detection problems was independently proposed for sentiment analysis for a call center in a very recent paper [16].

Detection systems are generally evaluated using metrics extracted from the receiver operating characteristic (ROC) defined as the set of true positive rate (TPR) versus false positive rates (FPR) obtained by sweeping the decision threshold applied to the scores. In this paper we show a modified version of the ROC called DET curve, commonly used for speaker recognition [21]. Several metrics can be extracted from these curves like the area under the curve, the equal error rate or a linear combinations between the two types of error called cost of detection or Cdet.

By posing the SER problem as one of multiple detections, we have the advantage of being able to choose the system’s operating point. That is, we can tune the detection threshold to achieve a certain trade off between one type of error (FNR) and the other (FPR). This could be useful in practical scenarios. For example, if we are designing a system for a call center to detect anger or frustration, then we probably want to have a low FNR so that no angry or frustrated call goes unnoticed. Yet, we can probably handle having some percent of false positives, which would be scanned by a human and discarded as uninteresting. On the other hand, if we are designing a system to search for examples of a certain emotion within a dataset (say, happy scenes in YouTube videos), we probably want to aim for a low FPR, since the total number of negative samples is so large that even a small FPR would result in a very large total number of false positive detections. In these two very different scenarios one would want to choose very different operating points for the system, optimizing the threshold to reduce one type of error at the expense of the other.

4. Results and Analysis

In this section we show results on the classic SER metrics and on detection metrics on different subsets of data and using different annotation criteria.

4.1. Classification Results

All results in this section assume that a single decision is made for each sample, corresponding to the emotion with maximum score for that sample. The left plot in Figure 1 shows the accuracy for both systems for different subsets of data for two annotation thresholds (0.2 and 0.5). Gray regions indicate the baseline performance and numbers above the bars are the relative improvement of the system with respect to the baseline.

The subsets are: the *with-majority* subset, formed by the samples for which a majority label can be selected; the *without-majority* subset, composed of all other evaluation samples; and the union of both sets. The latter set contains all evaluation samples, regardless of their annotation. No sample is discarded. The *with-majority* subset is the one traditionally used in SER publications to show accuracy and average recall results. Results on the subset without majority labels can only be computed when the annotation threshold is low. For the 0.5 annotation threshold, those samples all belong to the class “other“ and both metrics would be 0.

The last two group of bars corresponds to the 0.5 annotation threshold, which is equivalent to the majority label criteria. In this case including the samples without a majority label reduces the accuracy because we are just adding samples that can not be classified correctly by the system, since the system was only trained to predict the five target emotions and is unable to decide that a sample does not belong to any of these emotions.

For the first two groups of bars the annotation threshold is 0.2, for which samples can have correct predictions for multiple emotions. This raises the accuracy for all subsets, but the performance on the subset without majority labels is worse than that of a baseline system that always chooses the majority label within that subset (considering the labels obtained with a threshold of 0.2), appearing to indicate that the system does not provide any useful information about those samples. Nevertheless, this is only because accuracy is affected by the priors of the classes, which are extremely imbalanced in this subset.

The right plot in Figure 1 shows the average recall for the same conditions as discussed for the accuracy. In this case the baseline is the inverse of the number of classes and does not depend on the conditions. In contrast to what happens for the accuracy, the average recall on the subset without majority labels is significantly higher than the baseline. This means that the system is providing useful information about these samples and is able to predict some of the labels provided by the annotators. Furthermore, by adding new labels to be predicted, we degrade the average recall on the samples with a majority label. This is expected since, when doing classification, we are only allowing the system to detect a single emotion per sample. Note that for the 0.5 threshold, the blue and orange bars are identical because the samples added in the “All samples” case are not labelled as any target emotion and, hence, do not affect the recall.

To further dissect the performance of the system for the different subsets of data, we computed the average recall as a function of the entropy of the system’s posteriors. The goal is to determine whether the entropy of the vector of posteriors generated for a certain sample indicates something about the quality of the system’s prediction. To this end, we divide the data based on the entropy of the posteriors generated by the TDNN system normalized by the log of the number of classes. Samples are

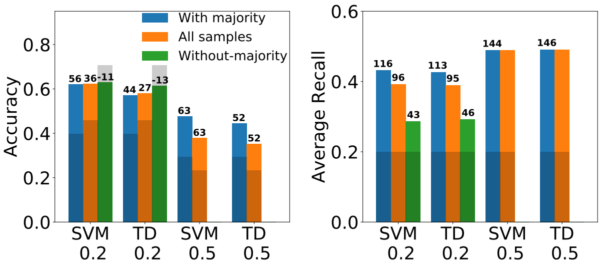

Figure 1: *Accuracy and average recall for different subset of data, different systems and different annotation thresholds (bottom label in the x-axis). The numbers above the bars indicate the relative gain with respect to the baseline (gray bars).*

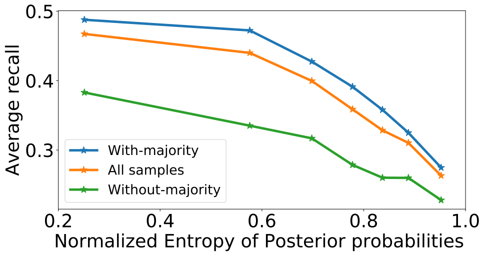

Figure 2: *Recall as a function of the entropy of the posteriors output by the TDNN SER system using an annotation threshold of 0.2. Samples are binned by their entropy, with bins defined to contain the same number of samples for the All set. The average recall is computed for each bin and marked with $\star$ at its center.*

binned by their entropy value and the recall for samples within each bin is computed using an annotation threshold of 0.2. Figure 2 shows that, for all three original subsets the recall goes down in absolute and relative value with respect to the baseline, which is constant and equal to 0.2, as the entropy increases. This means that as the entropy of the system's posteriors increases the model is less capable of making good decisions. That is, the maximum posterior decision when the posteriors are somewhat evenly distributed is unlikely to be a good decision. This is true both for the samples with and without a majority label. This suggests that perhaps a decision should not be made when the system is uncertain, as measured by the entropy of the system's posteriors. This may or may not be feasible or useful depending on the application.

4.2. Detection Results

In this section we show results obtained when doing emotion detection rather than classification. In this case, the system consists of one detector for each target emotion. Each detector makes its decision by thresholding the system's posterior for that emotion. If the posterior is larger than the threshold, the emotion is detected. For each detector, the FPR and FNR can be computed for each possible threshold. By averaging these rates over all emotions for each threshold between 0.0 and 1.0, we obtain a curve of average FPR versus average FNR.

Figure 3 shows DET curves corresponding to the TDNN system for different subsets and annotation thresholds (the curves for the SVM model are very similar and are not shown due to lack of space). For ease of comparison, the figure includes markers indicating the average FNR and FPR obtained with the maximum posterior decision as in Figure 1 (note that average recall is equal to 1 minus average FNR).

As for the average recall when doing classification, the performance on the samples that do not have a majority label is worse than on those that have a majority label, for the same

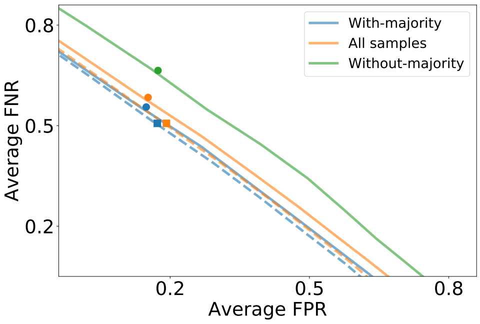

Figure 3: *Average FPR vs FNR curves for the TDNN system. Solid lines correspond to an annotation threshold of 0.2, while dashed lines correspond to 0.5. The markers correspond to the FPR and FNR obtained when doing classification with the maximum posterior decision, on the subset of the same colored lines for an annotation threshold of 0.2 (round) and 0.5 (square).*

annotation threshold of 0.2 (green solid line vs the blue solid lines). These samples are, undoubtedly, harder to classify for our system. This might be due to the fact that our system was trained using only training samples with a majority label. In the future, we will explore whether some of the methods proposed in the literature to take advantage of soft annotation labels are able to improve performance on these samples.

A very interesting observation from these curves is that they do not show the trend observed on the average recall where the performance on the samples with majority label degrades when lowering the annotation threshold (that is, the two blue lines almost overlap). This indicates that for the 0.2 annotation threshold it is not a good idea to restrict the system to make a single decision per sample. This makes sense, since we need to let the system try to predict the multiple labels that the samples may have. Notably, the fact that the two curves almost overlap indicates that, when allowed to predict more than one label per sample, the system predicts equally well all the labels provided by the annotators. This is a conclusion that could not have been extracted when forcing the system to make a single decision per sample and using classification metrics.

5. Conclusions

We investigated different ways of evaluating a SER model that take all samples into account, even those that cannot be assigned a label when using the majority rule commonly used in the SER literature. We show that, for several different metrics, the performance of two different systems on the samples without majority label is significantly worse than on the samples with majority label. Further, we propose that allowing the system to make multiple decisions per sample by doing separate detection for each emotion (perhaps with scores generated by a system trained to do classification) might be a good idea for SER. This is supported by our results which show that the system is capable of predicting the alternative labels from the annotators with the same performance as the majority labels on the samples that have such majority label.

6. Acknowledgements

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-18-1-0026. We gratefully acknowledge the support of NVIDIA Corporation for the donation of the Titan Xp GPU used for this research.

7. References

[1] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011. [Online]. Available: http://dx.doi.org/10.1016/j.patcog.2010.09.020

[2] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," *International journal of speech technology*, vol. 15, no. 2, pp. 99–117, 2012.

[3] C. N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011," *Artificial Intelligence Review*, vol. 43, no. 2, pp. 155–177, 2012.

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

[5] R. Lotfian and C. Busso, "Formulating emotion perception as a probabilistic model with application to categorical emotion classification," *2017 7th International Conference on Affective Computing and Intelligent Interaction, ACII 2017*, vol. 2018-January, pp. 415–420, 2018.

[6] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In: Automatic Face and Gesture Recognition (FG)," *the 10th IEEE International Conference and Workshops*, no. i, 2013.

[7] S. Steidl, *Automatic classification of emotion related user states in spontaneous children's speech*. University of Erlangen-Nuremberg Erlangen, Germany, 2009.

[8] K. Audhkhasi and S. S. Narayanan, "Emotion classification from speech using evaluator reliability-weighted combination of ranked lists," in *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2011, pp. 4956–4959.

[9] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for Speech Emotion Recognition," *Neural Networks*, vol. 92, pp. 60–68, 2017. [Online]. Available: http://dx.doi.org/10.1016/j.neunet.2017.02.013

[10] C. Z. Seyedmahdad Mirsamadi, Emad Barsoum, "Automatic Speech Emotion Recognition Using Recurrent Neural Networks With Local Attention Center for Robust Speech Systems , The University of Texas at Dallas , Richardson , TX 75080 , USA Microsoft Research , One Microsoft Way , Redmond , WA 98052 , USA," *ICASSP 2017, Proceedings 42nd IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 2227–2231, 2017.

[11] H. M. Fayek, M. Lech, and L. Cavedon, "Modeling subjectiveness in emotion recognition with deep neural networks: Ensembles vs soft labels," *Proceedings of the International Joint Conference on Neural Networks*, vol. 2016-October, pp. 566–570, 2016.

[12] J. Han, Z. Zhang, M. Schmitt, M. Pantic, and B. Schuller, "From hard to soft: Towards more human-like emotion recognition by modelling the perception uncertainty," in *Proceedings of the 25th ACM international conference on Multimedia*. ACM, 2017, pp. 890–897.

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

[14] C. Thiel, "Classification on soft labels is robust against label noise," in *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*. Springer, 2008, pp. 65–73.

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

[16] B. Li, D. Dimitriadis, and A. Stolcke, "Acoustic and lexical sentiment analysis for customer service calls," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5876–5880.

[17] A. C. de Carvalho and A. A. Freitas, "A tutorial on multi-label classification techniques," in *Foundations of Computational Intelligence Volume 5*. Springer, 2009, pp. 177–195.

[18] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*, 2013.

[19] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[20] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[21] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The det curve in assessment of detection task performance," National Inst of Standards and Technology Gaithersburg MD, Tech. Rep., 1997.