Over-sampling Emotional Speech Data Based on Subjective Evaluations Provided by Multiple Individuals

Reza Lotfian, *Student Member, IEEE,* and Carlos Busso, *Senior Member, IEEE*

Abstract—A common step in the area of speech emotion recognition is to obtain ground-truth labels describing the emotional content of a sentence. The underlying emotion of a given recording is usually unknown, so perceptual evaluations are conducted to annotate its perceived emotion. Each sentence is often annotated by multiple raters, which are aggregated with methods such as majority vote rules. This paper argues that several labels provided by different individuals convey more information than the consensus labels. We demonstrate that leveraging the information provided by separate evaluations collected by multiple raters can help in building more robust classifiers which maximize the utilization of labeled data. Motivated by the *synthetic minority over-sampling technique* (SMOTE), we present a novel over-sampling approach during training, where the samples with categorical emotion labels are over-sampled according to the labels assigned by multiple individuals. This approach (1) increases the number of sentences from classes with underrepresented consensus labels, and (2) utilizes sentences with ambiguous emotional content even if they do not reach consensus agreement. The experimental evaluation shows the benefits of the approach over a baseline classifier trained with consensus labels, which increases the F1-score by 5.2% (absolute) for the USC-IEMOCAP corpus, and 5.4% (absolute) for the MSP-IMPROV corpus.

Index Terms—Speech emotion recognition, synthetic over-sampling, data augmentation for deep neural network

1 Introduction

Emotion is an important aspect of human social interaction [1]. Machines that can recognize the emotional state of the user will be able to respond more efficiently and naturally to the users. Incorporating emotional capabilities to speech-based *human-machine interfaces* (HMIs) involves studying and understanding the emotional modulation conveyed in expressive speech, as well as designing robust machine-learning frameworks based on the underlying characteristics conveyed in emotional speech.

Emotional databases used for emotion recognition usually rely on subjective evaluations to annotate the emotional content of each stimulus. The annotations are then employed as ground truth to train emotional classifiers. The major challenge in collecting subjective evaluations is to reliably annotate the emotional content. Even experienced and attentive annotators often disagree on the emotion conveyed in a speech sentence due to the subjective nature in perceiving emotions. To address this issue and to obtain more robust labels for training emotion recognition classifiers, each speech sentence is independently evaluated multiple times by different subjects. Then, these labels are aggregated to build a single consensus label with rules such as majority vote or more sophisticated algorithms that aim to correct for the bias and reliability of the evaluators [2], [3]. The drawback of this approach is the increased cost of labeling new data. For a fixed budget, increasing the number of annotations per sentence decreases the number of recordings to be annotated. During the estimation of the aggregated labels, annotations that are different from the consensus label are ignored. Furthermore, in many cases a consensus cannot be reached. For these cases, the common approach is to discard sentences without agreement. This paper proposes that disagreement between annotators provides useful information, and should be leveraged in the training of emotional classifiers.

- *Reza Lotfian and Carlos Busso are in the Erik Jonsson School of Engineering & Computer Science, The University of Texas at Dallas, Richardson TX 75080, USA.*
E-mail: reza.lotfian@utdallas.edu, busso@utdallas.edu

Manuscript received March 12, 2018; Accepted February 14, 2019.

The assumption behind using a consensus label is that the evaluations are noisy and rules such as majority vote give a good approximation of the true emotion of the sentence. However, the perception of emotion is subjective and more than one answer can be right. Furthermore, when the evaluators have many categorical options to select, they might overlook less frequent classes, biasing their assessments to more common classes. For example, *anger* might be often selected as the primary emotion over emotional classes such as *disgust*, since it is a more common emotion during human interaction. For these sentences, it is unlikely that the consensus label will correspond to the less common emotion (in this example *disgust*), even though some annotators may have selected that class. Having classes with few sentences affects the balance of the corpus, affecting the classification task. This problem is more critical for classifiers based on *deep neural networks* (DNN) where a large training set is crucial for robust performance.

This study proposes a novel over-sampling method to train an emotion classifier aiming to recognize categorical emotion classes. The approach mitigates the sparsity of minority classes and avoid discarding sentences due to lack of agreement. The proposed approach relies on over-sampling the training corpus based on the evaluations from multiple raters assigned to speech sentences (i.e., labels before aggregating the annotations). We use the term *sample* to refer to synthetic examples created by oversampling frameworks. Our approach is inspired by the *synthetic minority over-sampling technique* (SMOTE). First, we replicate the sentences by making copies. The labels for these copies are assigned by sampling from the separate evaluations assigned to that sentence. This approach considers emotional labels that differ from the consensus label. Second, we create synthetic examples by linearly moving the replicated samples toward the distribution of their respective classes in the feature space. This step reduces the overlap of the sentences between classes in the feature space caused by adding synthetic samples. It also avoids the problem of having a sample with two different labels, which creates non-separable classes in the recognition task. The proposed approach is a principled framework to augment the data, which has been successfully used to train

deep learning classifiers in emotion recognition [4], and in other domains [5], [6]. Even though over-sampling methods have been previously used on emotion classification [4], to the best knowledge of the authors, this is the first time that an over-sampling approach leverages evaluations provided by multiple raters, which are often ignored after estimating the consensus labels. In contrast, our approach uses all the sentences even when consensus labels cannot be obtained. This is a principled approach to leverage emotional evaluations collected across multiple individuals. This approach is also radically different from the conventional SMOTE, which synthesizes new samples relying on the consensus labels. This key difference distinguishes the proposed method from various implementation of SMOTE in different applications [7], [8], and other techniques for dealing with class imbalance [9], [10].

We extensively evaluate the proposed approach with two emotional databases: the MSP-IMPROV and USC-IEMOCAP corpora. The experimental evaluation clearly shows the benefit of using the proposed over-sampling method over classifiers trained with consensus labels. The proposed approach increases the F1-score by 5.2% (absolute) for the MSP-IMPROV database, and 5.4% (absolute) for the USC-IEMOCAP database over the baseline approach relying on consensus labels. Even after over-sampling only two samples per sentence, the proposed approach obtained significant improvements over a baseline classifier trained with consensus labels. We also show that our method outperforms the conventional SMOTE, where the differences are statistically significant. When compared with SMOTE, the proposed approach leads to absolute improvements in F1-score between 1.2% and 2.7%. The proposed method can also be used to over-sample minority classes to build a balanced training set. The results show the benefit of using our method over alternative feature space based data augmentation methods.

The rest of this paper is organized as follows. Section 2 reviews over-sampling techniques with a focus on emotion recognition. Section 3 briefly introduces the two independent databases and the acoustic features that we use in this study. Section 4 describes the motivation behind using the proposed over-sampling approach for speech emotion classification, describing the framework. Section 5 describes the experimental evaluation and the results. Section 6 concludes this paper, summarizing the main contributions and suggesting future research directions.

2 Related Work

Learning from imbalanced datasets is a well-known problem in machine learning, where the performance of the algorithms is affected by the skewness of the class distribution. Learning from such data requires sampling algorithms that efficiently employ underrepresented classes. Two common approaches are random down-sampling and random over-sampling. An example of random down-sampling is the Tomek links [11], which aims to remove samples from the majority class that are closer to the samples in the minority class. Other examples of down-sampling methods include one sided selection [12] and Hart's condensed nearest neighbor rule [13]. The second approach includes over-sampling the data, which is the main focus of this paper.

2.1 Data Augmentation

The practice of enlarging the size of the training set by artificially making new instances is a common approach in many machine-learning applications [6], [14], [15]. This procedure usually involves employing transformations that generate random variations of the original example while preserving the label. For example in image classification, studies have augmented the data by creating horizontal reflections or changing the intensities of the RGB channels to produce new training samples. These approaches avoid a substantial over-fitting problem due to a small training set and increase robustness by creating variations that are likely to appear in the testing set (e.g., rotation and translation). In speech processing, Bellegarda et al. [16] used data augmentation for speaker adaptation in speaker-dependent speech recognition, where the enrollment data available from a new speaker is rather small to train acoustic models. They proposed to apply piecewise linear transformations that map the reference speakers to a new speaker at the phone level. Cui et al. [5] investigated other approaches for data augmentation such as *vocal tract length perturbation* (VTLP) [17] and *stochastic feature mapping* (SFM) for DNNs and *convolutional neural networks* (CNNs). The results showed that data augmentation effectively decreased the *word error rate* (WER). Ragni et al. [18] also used a data augmentation approach for low resource languages by synthesizing target language speech with methods such as concatenative speech synthesis. In these examples, the data augmentation method was targeted to each application. The transformations were carefully designed to increase data variations without affecting the discriminative characteristic of the primary classification task (e.g., idiosyncratic cues in speaker verification, objects in images, and phonetic content in speech recognition).

The second type of over-sampling approach is to synthetically build new instances for certain classes by manipulating their samples in the feature domain. These approaches are more general, since they can be used across domains. A representative approach is the *synthetic minority over-sampling technique* (SMOTE), which is an over-sampling method where the minority class is over-sampled by synthesizing additional examples. The objective of SMOTE is to over-sample a minority class by $N\%$ where N is usually chosen to be a multiple of 100. It starts by finding the K nearest neighbor of the samples in the minority class. Then, depending on the amount of over-sampling N, it selects some of the K nearest neighbors. For example for $N = 300\%$, the approach selects three of the nearest neighbors. For each of these neighbors, a new synthetic sample is generated by randomly finding one point on the line segment in the feature domain connecting the original sample and the selected neighbor. The new point will be added to the minority class increasing the size of the underrepresented class. This process is repeated for all the $N/100$ nearest neighbors. The SMOTE approach has been widely used in many applications where an unbalanced set of data is used for classification. Phua et al. [7] used SMOTE for fraud detection, where the database was skewed towards non-fraud cases. They show that balancing the data with SMOTE improves the detection rate compare to training with the original dataset. The SMOTE framework has also been implemented together with under-sampling methods to avoid over-fitting. For example, Batista et al. [19] proposed SMOTE+Tomek links and SMOTE+ Wilson's *edited nearest neighbor* (ENN) rule [20] to expand minority class clusters.

2.2 Using Data Augmentation in Emotion Recognition

Studies have shown that data augmentation is a useful framework to increase performance in emotion recognition. Schuller and Burkhardt [21] showed that the training set can be increased by using synthetic emotional speech. They showed

that combining synthetic samples and natural training sentences provided better accuracy than using only natural speech. Aldeneh and Provost [4] implemented data augmentation for speech emotion recognition following the approach suggested by Ko et al. [22] for speech recognition, where the speech rate is increased while preserving the spectral properties. Augmenting the data led to statistically significant gains in the classification performance [4]. A similar approach was used by Abdelwahab and Busso [23].

The previous studies on categorical emotion classification consistently face the problem of unbalanced class size with relatively small databases. Therefore, over-sampling methods, especially those using SMOTE, have been widely used in the past to balance the training set by simultaneously over-sampling minority classes and/or down-sampling majority classes [24], [25], [26], [27], [28], [29]. For example, Yildirim et al. [30] applied SMOTE to obtain a more balanced class distribution in a three-way classification problem to avoid biasing the classifiers toward the more prominent class *neutral*. They doubled the number of instances of the minority classes *polite* and *frustrated*. Calix et al. [31] used SMOTE to address the problem of training with unbalanced classes on the SEMAINE database.

2.3 Leveraging Separate Labels in Perceptual Evaluation

In the aforementioned examples of data augmentation, the underlying assumption is that the true class label is known for the original training sentences. Therefore, the newly generated over-sampled instances inherit the class label from the original sample. This assumption is known to be flawed in emotion recognition, since expression of emotions are often perceived differently by different listeners. Therefore, each sample can potentially belong to more than one class (e.g., soft labels). To address this issue, some studies have relied on emotional profiles as labels, which better capture the emotional content of a speech sample [32], [33], [34]. For example, a speech audio that is perceived as *happy* by all the annotators indicates that this sample is a stronger example of the given emotion than a speech sample in which some annotators choose *happy* but others select *neutral*. Studies have explored this problem in the context of emotion recognition. For retrieving categorical emotions, Lotfian and Busso [35] derived a probabilistic relevance score from emotional annotations provided by multiple raters to determine the level in which a target emotion is conveyed in a sentence. The relevance score was higher for sentences that were consistently assigned to the target emotional category. This relevance score was used to create relative labels between two sentences (e.g., sentence one is happier than sentence two), training preference leaning algorithms to rank sentences according to categorical emotions. Steidl et al. [36] suggested that the systematic confusions across evaluators should be accounted for when assessing the performance evaluation of an emotional classifier. They introduced an entropy-based measure that considers common disagreements between evaluations assigned to a sample, showing that with this metric a classifier can perform as well as human labelers.

There are few studies that have leveraged separate evaluations provided by multiple raters (before estimating consensus) to improve the modeling of emotion. Fayek et al. [34] formulated the problem with an ensemble of DNNs, each of them representing an individual annotator. The output of the DNNs were later combined to create a single label. They also suggested training DNNs with soft-labels, reflecting the distribution of the classes assigned to a given speech sample. Their experimental results showed the benefit of considering separate evaluations in building emotion classifiers. Lotfian and Busso [33] derived a new soft-label formulation by modeling the process of perceiving and annotating emotions as a probabilistic model. They showed that their proposed soft-label approach achieved better classification performance compared to classifiers trained with hard labels derived from consensus labels and soft-labels that only reflected the proportions of the classes selected by evaluators.

2.4 Contributions of this Study

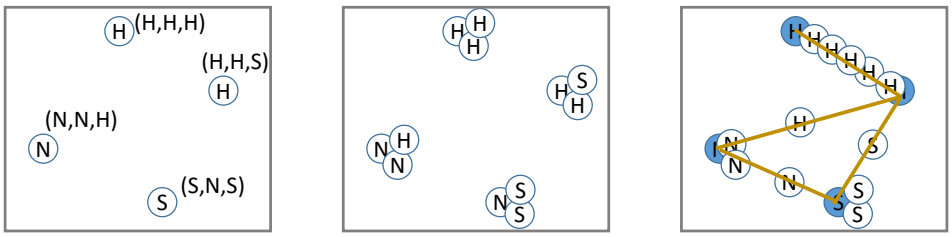
The contribution of this work in comparison to other data augmentation solutions for emotion recognition is that the over-sampling approach considers evaluations assigned to a sample before aggregating the consensus labels. The disagreement between evaluators is not considered as noise, as in most studies, but as useful information. The proposed approach has the following advantages:

- It better utilizes the available information for training the classifiers by exploiting all the annotations assigned to a sample.
- In contrast to SMOTE, a single sentence can be used to synthesize new samples from different classes in proportion to the label distribution assigned to the sentence.
- It creates new samples even when the majority consensus label is not reached due to lack of agreement between evaluators.

The proposed method addresses some well-known problems in the recognition of categorical emotions. Since each sample is labeled multiple times, the number of available labels is larger than the number of sentences, so extending the training set is straightforward. Furthermore, conflicting labels for a sample that do not lead to a consensus can still be used, avoiding discarding sentences due to lack of agreement. This feature of our method also leads to larger training sets. The over-sampling framework is particularly effective for minority classes that are less common in human interaction. Another advantage is that this method addresses the challenge of inter-class variability in categorical emotions. Considering majority vote labels, all the sentences in a class will be treated equally independent of their intensity. In practice, high intensity prototypical behaviors with clear emotional content tend to have higher agreement between evaluators. This over-sampling approach will create multiple samples for the target class, having a stronger impact on the classifier. Ambiguous emotional behaviors, in contrast, will have more disagreement, so the over-sampling method will reflect this ambiguity.

3 Databases and Acoustic Features

This study considers two emotional databases to explore different aspects of the proposed over-sampling method. We consider the USC-IEMOCAP database [37], which includes ten categorical emotions in the perceptual evaluations. The extended number of emotional classes leads to classes with few examples, which challenges the training of emotional classifiers. The second database is the MSP-IMPROV database [38], since it has at least five subjective evaluations per audio sample. The effect of the proposed over-sampling method on these classes will give insight on the benefits of using the proposed method in dealing with such cases. While these corpora are audiovisual databases, this study relies only on speech. This section briefly describes these two databases and the acoustic features used in the evaluation.



(a) Majority vote label (b) IEOS (c) SIEOS

Fig. 1. Example to illustrate the proposed over-sampling method. The figure shows four sentences with evaluations from three raters. (a) The label inside the circle defines the consensus labels. (b) The IEOS step replicates N samples per sentence, assigning the labels provided by separate evaluators (in this case $N = 3$). (c) The SIEOS step moves the samples toward randomly selected neighbor sentences assigned with the target emotion.

3.1 The USC-IEMOCAP database

This study relies on the *interactive emotional dyadic motion capture* (IEMOCAP) database [37] collected to study expressive human interactions under controlled conditions. The corpus contains 12 hours of dyadic conversations recorded in five different sessions between ten trained actors. The emotions were elicited with scripts and spontaneous improvisations to evoke sadness, happiness, anger and frustration. Other emotions were also elicited as dictated by the course of the conversation between the actors. By employing well-established theories and methods of theater, the recording provided emotional displays close to natural interactions, even though the corpus was collected from actors [39]. The spontaneous dialogs were segmented into speaker turns and manually transcribed. The emotional annotations were collected for categorical and attribute-based labels at the turn level. For categorical emotions, three evaluators evaluated each turn by selecting the following 10 options: neutral, happiness, sadness, anger, frustration, excitement, surprise, fear, disgust or other. The corpus is not emotionally balanced, where some of these classes have few sentences. Therefore, we investigate the effect of the proposed over-sampling method when the classes are unbalanced. The speaking turns with overlapped speech were excluded from the evaluation. In total, 5,496 speaking turns are used in this study. Further information about the database is provided in Busso et al. [37].

3.2 The MSP-IMPROV database

The MSP-IMPROV database [38] is an audiovisual corpus collected to explore emotional behaviors during dyadic conversational interactions. The data collection scenarios were designed to promote natural recordings, while keeping control over the lexical content. This goal was achieved by creating specific scenarios per emotion, which led one actor to utter a target sentence. This elicitation approach facilitated the recording of spontaneous sentences conveying the same lexical content under different emotions. In addition to the target sentences, the corpus includes all the turns during the improvisation that led one of the actors to utter the target sentence. The corpus also includes the conversations during the break, since the microphones and cameras were not stopped. In total, this corpus consists of 8,438 speaking turns (over 9 hours). The emotional labels of the corpus were collected through a crowdsourcing perceptual evaluation described in Burmania et al. [40]. The primary emotional classes used in the evaluation are anger, happiness, sadness, neutral and other. The key feature of this corpus is that each turn was annotated by at least five raters, providing the ideal resource to study the effect of the over-sampling rate on the performance of classifiers. For further information about this corpus the readers are referred to Busso et al. [38].

3.3 Feature set

This study uses the *Geneva minimalistic acoustic parameter set* (GeMAPS) [41], which includes standard acoustic parameters carefully selected for paralinguistic tasks. This study uses the extended version (eGeMAPS), which contains 88 parameters extracted using the OpenSMILE toolkit [42]. The set includes features related to frequency, energy/amplitude and spectral parameters. The acoustic features are extracted for each speaking turn independent of its length. The readers are referred to Eyben et al. [41] for a detailed description of this feature set.

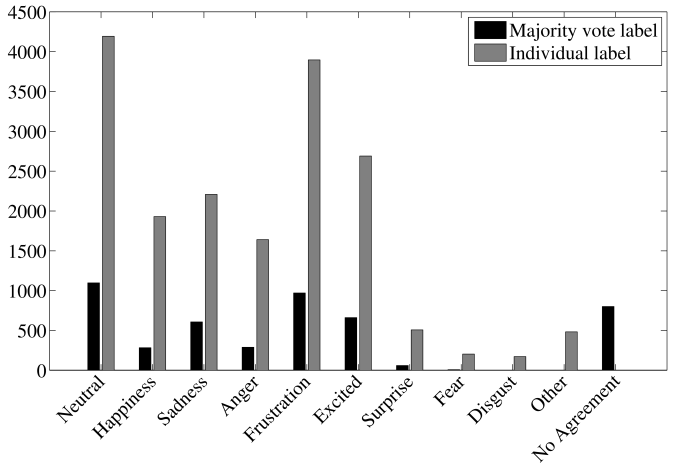
4 Proposed Over-Sampling Method

This section describes the proposed over-sampling approach, which considers the conventional perceptual evaluation process used to annotate emotional databases. The intuition behind our method is that a speech sample in spontaneous conversations often conveys more than one emotional trait [43]. We have observed in our previous evaluations that listeners often identify more than one emotional classper sample [33], [38], [44]. Furthermore, they often disagree between themselves while evaluating the same speech file. We propose to take this phenomenon into consideration by sampling speech samples from the class labels assigned by each annotator during the perceptual evaluation. This section describes our proposed over-sampling approach to leverage this information and improve the performance of categorical emotional classifiers.

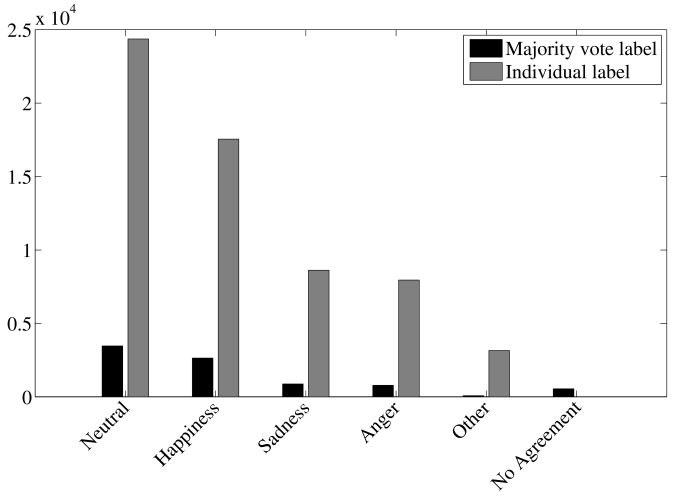
4.1 Synthetic Individual Evaluation Over-Sampling (SIEOS) Method

We motivate the proposed approach with the illustration in Figure 1. Figure 1(a) shows that there are four speech sentences in this example, represented with circles, each receiving three independent evaluations (*H*: happiness, *S*: sadness, *N*: neutral). The speech sentences are plotted in two dimensions which represents the feature space (in our implementation this process is applied to a 88-dimensional feature space). The circles are labeled with the consensus labels obtained with the majority vote rule. Figure 1(a) shows that we initially have two *happy* sentences (H), one *sad* sentence (S) and one *neutral* sentence (N). The consensus labels ignore the fact that some evaluators did not agree with the aggregated label, even when these sentences may convey emotional traits of the secondary emotions. Our over-sampling method aims to leverage this information.

The first step of the proposed over-sampling framework replicates the speech sentences N times, assigning labels by

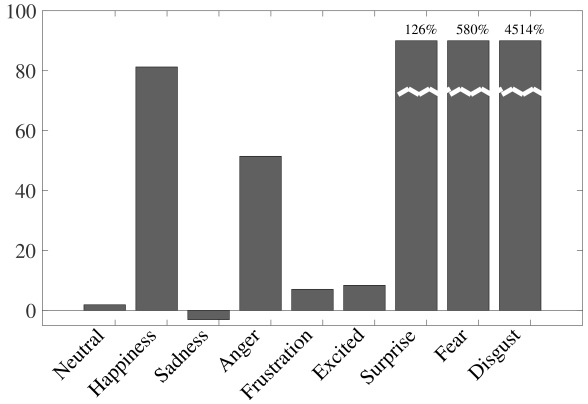


(a) USC-IEMOCAP

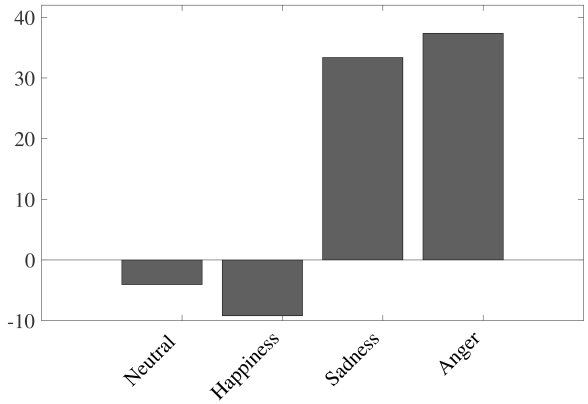


(b) MSP-IMPROV

Fig. 2. Distribution of consensus and individual labels for the USC-IEMOCAP and MSP-IMPROV databases. Some of the minority labels are selected by the individual evaluators, but they are not reflected in the consensus labels.



(a) USC-IEMOCAP



(b) MSP-IMPROV

Fig. 3. The relative changes in the proportions of the emotional categories over the total number of samples after using the proposed IEOS method. The relative changes are estimated with respect to the proportion observed when using the majority vote rule. The relative changes in the proportions of minority classes are higher after applying the proposed IEOS method.

randomly sampling without replacement from the labels provided by the annotators. Figure 1(b) shows the outcome of this procedure when $N = 3$, creating 12 samples. We refer to this step as the *individual evaluation over-sampling* (IEOS) process. In this example, we choose N to be equal to the number of individual annotations. However, N is an arbitrary number since the proposed method relies on N random inquiries from the finite sample space of individual evaluations. The probability of each outcome is proportional to the frequency of that outcome among the individual annotations for that sample. If a sample is consistently labeled with a given class, then the N copies of the sample will have the same label. If the emotional content of the sample is ambiguous and the evaluations do not agree, the new copies will be labeled with different, but relevant classes. This step preserves the distribution of the emotional classes assigned to the sentences. Notice that we do not have to replicate all the individual evaluations, as shown on Figure 1. An alternative approach is to sample from the individual evaluations until reaching the target number of synthetic samples (i.e., N).

Figure 2 shows the distribution of the emotional classes for the USC-IEMOCAP and MSP-IMPROV databases. The black bars represent the consensus labels obtained with the majority vote rule. The gray bars represent the individual labels assigned to the sentences. If the IEOS approach is applied to consider all the individual evaluations, the figure shows that emotional classes with few sentences would be better represented. Figure 2(a) shows that the increase in samples is more prevalent in minority classes such as surprise, fear, and disgust. These classes are less common in daily interactions, so evaluators converged to more common emotions, discarding these traits after estimating consensus labels. To explore further the benefits of the IEOS method on minority classes, we measure the relative changes in the proportion of the classes when using either consensus labels (majority vote) or the proposed IEOS method. For example, if a class account for 3% of the data when using the consensus labels, and 6% when using the IEOS method, its relative change is 100%. Figure 3 shows the results of this analysis. The emotional classes in the x-axis are ordered based on the popularity of the samples according to the consensus labels. This figure shows that the emotional classes that benefit the most from the IEOS framework are the underrepresented classes. For both databases, the representation of less frequent categories has increased by considering individual evaluations. This result validates the hypothesis that annotators tend to select more common emotions, which make it difficult to reach agreement in minority emotional classes.

One concern in using the proposed IEOS framework is the increased overlap between emotional classes as repeated sam-

ples will have different labels. The overlap between classes will negatively affect the training of emotional classes. For example, it is not possible to find a classifier to correctly separate all these 12 examples in Figure 1(b). The second step of the proposed over-sampling approach addresses this issue by moving the repeated samples in the feature space toward other sentences labeled with the target emotion. This step is motivated by the *synthetic minority over-sampling technique* (SMOTE) [24], which generates new synthetic samples by considering real sentences with similar labels. For this purpose, we rely on the majority vote labels to be the confident point of the emotional classes. For a given repeated sample, we find the K nearest neighbor sentences belonging to the target emotional class. From these K sentences, we randomly select one of the them. We find the line between this sentence and the repeated sample in the feature space. The new location of the repeated sample is determined by selecting a random point along this segment. The feature vector of the new sample s_{sieos} is obtained by interpolating the feature vector of the repeated sample, $s_{repeated}$, and the feature vector of the selected neighbor sentence, s_{nn}, using Equation 1:

$$s_{sieos} = \alpha \times s_{repeated} + (1 - \alpha) \times s_{nn} \quad (1)$$

where α is selected at random between 0 and 1. The variable α determines how to interpolate the synthetic sample. As our method can generate two or more synthetic samples with the same emotion, it is important than this parameter is selected at random for each sample. As opposed to SMOTE, the replacement step is less vulnerable to over-fitting issues [45] since the duplicated samples are not necessarily from the same class. The interpolation step in the SIEOS algorithm will have similar problems as the interpolation in SMOTE when the density of the samples is low. Both of them will have higher confidence when more samples are available since the shift of the samples is limited by the distance to the selected sentence within the K nearest neighbor. Figure 1(c) illustrates the effect of this procedure on the 12 samples created with IEOS. We refer to this method as the *synthetic individual evaluation over-sampling* (SIEOS) approach. Algorithm 1 formally describes the implementation of the SIEOS method.

Algorithm 1 *Synthetic individual evaluation over sampling* (SIEOS) for sentence i

Input: Number of Synthetic samples per sentence N
  Number of nearest neighbors K
Output: $OSfeatures_i[\ ][\ ]$: array for over-sampled features for sentence i
  $OSlabel_i[]$: array for over-sampled labels for sentence i
define: $features[\ ][\ ]$ array of features of original sentences
define: $S_i(m) : m^{th}$ individual evaluation for sentence i
for $n <= N$ **do**
  Randomly select m^{th} individual evaluation for sentence i
  Find K nearest neighbors of sentence i with consensus label $S_i(m)$.
  Save indices in $nnarray$
  $Populate(i, nnarray)$

$Populate(i, nnarr)$ ($*$ Function to generate the synthetic samples. $*$)
$OSlabel_i[n] = S_i(m)$
Choose nn= a random sentence from the K nearest neighbors of sentence i
Compute: α = a random number in [0 1]
for j <# of features **do**
  $OSfeatures_i[n][j] = \alpha \times features[i][j] + (1 - \alpha) \times features[nn][j]$
return ($*$ End of Populate. $*$)

4.2 Differences with SMOTE

Although our method is inspired by the SMOTE algorithm, there are fundamental differences between them. SMOTE is a general algorithm that is applicable to any classification problem with labeled data. The proposed SIEOS approach is suitable for classification problems where a set of (possibly distinct) labels are available for each sentence and the final label is obtained with consensus labels. These kinds of problems arise when the class labels correspond to ambiguous or subjective concepts such as categorical emotion recognition. Therefore, in SIEOS a single sentence can be used to synthesize new samples with conflicting labels if the annotations suggest such a distribution. In contrast, the over-sampling in SMOTE is separately done for each class by considering only sentences from the same class. As we will show in the experimental section, the SIEOS approach is more effective for categorical emotion recognition.

Another difference is that over-sampling with SMOTE does not change the center of the clusters of classes in the feature space, since it only involves linear transformations on the features of sentences randomly selected from the same class. In contrast, the SIEOS approach changes the centers of the clusters to reflect the disagreements of annotators over related classes that are interchangeably chosen by evaluators. We estimate the mean value of the 88-dimensional feature vector for each emotion in the USC-IEMOCAP database. We measure the average distance between the center of each emotion before and after over-sampling using SMOTE and our proposed over-sampling methods. The average change in the centers of the clusters for the SMOTE method is relatively small (0.46) compared to the changes in the centers of the clusters for the IEOS method (3.54) and the SIEOS method (3.08). Both methods move the center of the emotional classes by considering the individual evaluations, including dissident labels that do not agree with the consensus labels. We believe this disagreement is not merely caused by human error. Instead, it conveys relevant information to better describe the emotion content of a recording. While information from dissident labels is often discarded in conventional approaches, our methods use this valuable information to shift the hyperplanes of the classifier improving the performance of the systems. The experimental evaluation in Section 5 demonstrates that moving the clusters for each emotion using the SIEOS approach improves the classification performance.

The next section assesses the performance of multi-class emotion classifiers trained with the proposed method.

5 Experimental Results

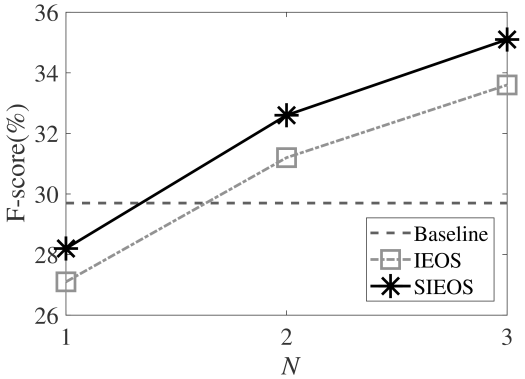
This section evaluates the effect of the proposed SIEOS method in the classification performance of speech emotion recognizers. We follow a *leave-one-speaker-out* (LOSO) cross-validation procedure to select the train, test, and validation sets. For each fold, we assign sentences from one speaker as the test set. Then, we randomly select one of the remaining speakers as the validation set. The sentences from the rest of the speakers form the train set. We aggregate the results of different folds by adding up the confusion matrices of the results.

The experimental evaluation relies on fully connected feed forward *deep neural networks* (DNNs) with two hidden layers with utterance-level features as input (Section 3.3). Each hidden

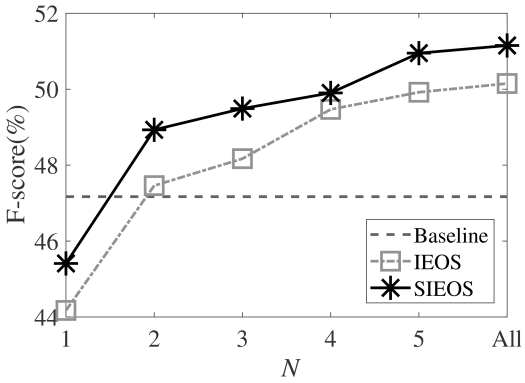
layer consists of 256 nodes with *rectified linear unit* (ReLU) as activation function. This activation function and number of hidden layers has been shown to be competitive in learning emotion recognition networks [23]. Similar DNN architectures have been used in other studies on classification of emotions with utterance-level features [46], [47], [48], [49]. The last layer is a Softmax layer, which gives a vector where the elements correspond to the score of each emotional class. We use Keras with TensorFlow as backend to implement and train the models. We rely on *adaptive moment estimation* (ADAM) [50] for the optimization of the network parameters. We use different learning rates (0.01, 0.001, 0.0001), and evaluate the model on the validation set. The results demonstrated that a learning rate of 0.001 works better. We train all the models with 50 epochs.

We investigate three conditions: 1. Majority vote (baseline) 2. Individual evaluation over-sampling (IEOS) 3. Synthetic individual evaluation over-sampling (SIEOS). For the baseline method using the consensus labels, we remove the sentences in the training set without agreement. Notice that we have to discard 801 sentences in the USC-IEMOCAP corpus and 554 sentences in the MSP-IMPROV corpus. For the IEOS and SIEOS methods, we do not have to remove these sentences, since the labels of the samples are obtained from the individual evaluations. This is an important advantage of the IEOS and SIEOS methods, which can use sentences with ambiguous emotional content without consensus agreement. For the test set, we remove speaker turns in which the evaluators do not reach majority vote agreement for all the conditions, creating fair comparisons.

We report the results with the F1-score calculated as follows. First, we estimate the precision and recall rates for each emotional class. Then, we estimate the average precision ($\bar{P}$) and average recall ($\bar{R}$) rates across emotional classes. Finally, the F1-score is computed using $\bar{P}$ and $\bar{R}$.



(a) The USC-IEMOCAP corpus



(b) The MSP-IMPROV corpus

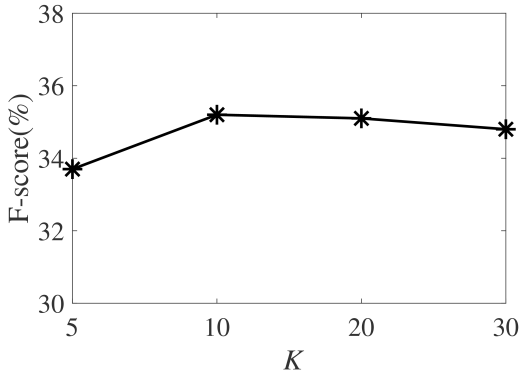
Fig. 4. F1-score of different methods as a function of the number of synthetic samples created for each sentence. Even with $N = 2$, the over-sampling methods obtain better classification performance than classifiers trained with consensus labels.

5.1 Performance as a Function of the Number of Samples Created per Sentence

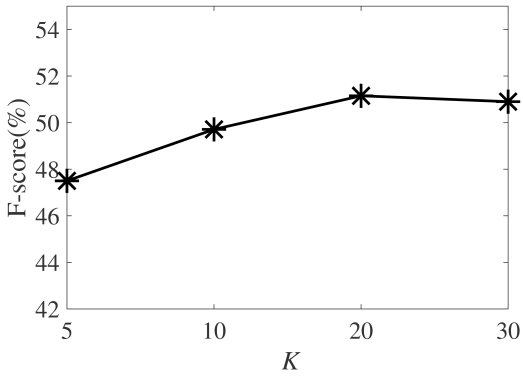
This section evaluates the classification performance as a function of N. The SIEOS approach is implemented with $K = 20$ (i.e., number of nearest neighbor sentences used in the interpolation). For the USC-IEMOCAP database, the classification problem has eight classes where we remove the classes disgust and other (the corpus has only one sentence labeled as disgust). In the USC-IEMOCAP database each sentence has at least three evaluations. Since we are sampling without replacement, we can only generate three synthetic samples per sentence. We randomly choose one, two, or three annotations for over-sampling. Figure 4(a) shows that increasing N consistently increases the classification performance. To compare the results of our over-sampling methods with the results of the baseline approach trained with consensus labels, we conducted statistical significance tests using one-tail population mean z-tests, asserting significance at p-value = 0.05. An interesting observation is that the results are better than the baseline, even when we create only two replicas per sentence. For the IEOS method, the improvement over the baseline is statistically significant for $N = 3$ (p-value=0.038). For the SIEOS method, the improvements over the baseline are statistically significant for $N \geq 2$ (p-value=0.042 for $N = 2$ and p-value=0.029 for $N = 3$). We observe that the SIEOS method has a higher F1-score than the IEOS approach.

We repeat the same experiment using the MSP-IMPROV corpus. In this corpus, each speaking turn has at least five evaluations, so we can increase the value of N even more. Some sentences have more than ten evaluations. For these sentences, we randomly choose ten of them and discard the rest. Therefore, each sentence has between five and ten evaluations. Figure 4(b) shows the unweighted F1-score for a four-class classification problem for the MSP-IMPROV database (anger, happiness, sadness and neutral). The baseline approach considers the consensus labels, and is represented by the horizontal dashed line. We implement the IEOS and SIEOS methods by randomly choosing one, two, three, four, five, or all the individual annotations (up to 10). Figure 4(b) shows similar patterns as the results obtained on the USC-IEMOCAP corpus. The IEOS and SIEOS methods outperform the baseline even with $N = 2$. Burmania and Busso [51] studied the effect of adding extra evaluators in the consensus labels on a portion of the MSP-IMPROV corpus. An interesting finding was that around 45% of the sentences preserve the labels assigned by the first annotator. This result suggests an important overlap between the consensus labels and the labels sampled from the individual annotations, explaining the competitive performance of the system trained even with $N = 2$. For the SIEOS approach, the improvements are statistically significant when $N \geq 2$. For the IEOS approach, the differences are statistically significant for $N \geq 4$. When we use all the individual evaluations, the differences are very clear. Figure 4(b) consistently shows better performance for the SIEOS method over the IEOS method, indicating that moving the created samples toward their target classes increases the separation of the classes, and, therefore, the classification performance.

Each sentence has been evaluated by multiple annotators, where the consensus labels are estimated with majority vote



(a) The USC-IEMOCAP corpus



(b) The MSP-IMPROV corpus

Fig. 5. Performance of the classifiers as a function of the number of nearest neighbor sentences used for the interpolation of the synthetic samples in the SIEOS approach (K in Algorithm 1).

to obtain prominent emotions, discarding dissident labels. The experimental evaluation shows that our over-sampling methods can effectively use these individual evaluations to improve the performance of the classifiers. Unless stated differently, we consider five oversampled instances per entry for the MSP-IMPROV corpus (i.e., $N = 5$), and three oversampled instances per entry for the USC-IEMOCAP corpus (i.e., $N = 3$) for the rest of the evaluations.

5.2 Analysis of Number of Nearest Neighbor Sentences

This section evaluates the selection of the number of nearest neighbor sentences used for the interpolation of the synthetic samples in the SIEOS approach (i.e., parameter K). The selection of K sentences that are near to the target sentence is implemented to increase the randomness in the selection of samples to be interpolated. We analyze the sensibility of the SIEOS approach by implementing the algorithm with different values of K. We consider the following options: 5, 10, 20, and 30. Figure 5 shows the results for the USC-IEMOCAP and MSP-IMPROV corpora. The F1-scores are lower when K is set to 5, which reduces the randomness in the selection of the sentences used for interpolation. The F1-score improve as we increase the value for K. The figure shows that for both corpora the performance of the classifiers are very stable when K is over 10. We set $K = 20$ in the evaluations discussed in the rest of the paper.

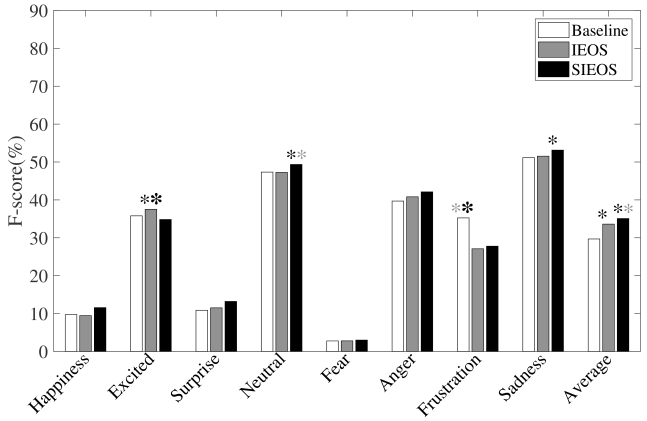
5.3 Classification Performance per Emotional Category

This section investigates the effect of using the SIEOS framework on the performance of individual emotional categories (Fig. 6). We are particularly interested on minority classes. First, we analyze the F1-score for each emotion in the USC-IEMOCAP database. This corpus has multiples emotional categories, as explained in Section 3.1, some of them with few sentences (284 for happiness, 60 for surprise and 8 for fear). Figure 6(a) shows the average F1-score per emotion for the USC-IEMOCAP database, using the baseline trained with consensus labels, and the proposed IEOS and SIEOS methods. The color-coded asterisks denote statistically significant differences between conditions (one-tailed population mean z-test, asserting significance at p=0.05). An asterisk over a bar indicates that the result for a method is significantly higher than the one for the bar with that color. With the exemption of frustration, one of the proposed approaches always has a higher F1-score than the baseline method. For the three minority classes, the SIEOS provides the highest F1-score rate, although the differences are not statistically significant. The analysis of the average F1-scores indicates that the SIEOS framework obtains significantly better results than the IEOS approach and the baseline approach (last set of bars in Figure 6(a)). The IEOS framework is also significantly better than the baseline method. On average, the IEOS and SIEOS methods provide 8.3% and 9.2% relative improvement over the baseline framework. Among the minority classes, fear does not benefit from the proposed over-sampling method, suggesting that the method is not beneficial when there are extremely few examples available from a class (eight sentences in this case).

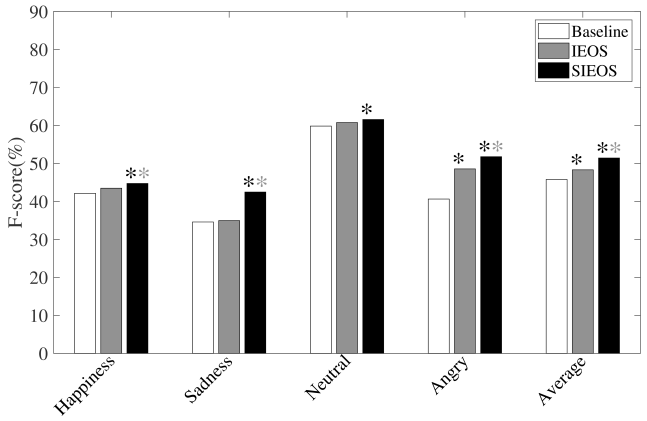
We also find the F1-score of each emotional class using the MSP-IMPROV database. This database is more balanced without a minority class. Figure 6(b) shows the results. When compared to the baseline, we observe statistically significant improvements in F1-score for all the emotional classes when using the proposed SIEOS method. When using the SIEOS framework, we observe the best F1-score improvements over the baseline for the emotions *sadness* and *anger*, which are the classes with fewer samples on the MSP-IMPROV.

5.4 Analysis of Overlap Between Classes

We investigate the overlap between emotional classes generated by the proposed approach. We use the Bhattacharyya coefficient [52], which is a metric that has been widely used to compare two distributions. It measures the amount of overlap between two statistical populations. For our analysis, lower Bhattacharyya coefficient is desired since we employ this coefficient to compare the distribution across distinct emotional categories in the feature space. We find the Bhattacharyya coefficient between every pair of categorical emotions. To account for the randomness in the algorithm, we implement the IEOS and SIEOS approaches 20 times. For comparison, we also implement the SMOTE approach 20 times. Table 1 reports the mean and standard deviation of the results computed across emotions and implementations. The results are separately tabulated for the USC-IEMOCAP and MSP-IMPROV databases. For the USC-IEMOCAP, the overlap according to this measure is 0.524 for the original sentences. Using the IEOS approach, the overlap increases to 0.628. With the IEOS approach, a given sentence is repeated multiple times with different labels (as assigned by multiple evaluators). Therefore, N samples will have the same features, but with different labels. This step clearly increases the overlap between the classes. The SIEOS method decreases the Bhattacharyya coefficient to 0.553, since it uses interpolation to move these samples toward their classes, reducing the overlap between the classes. By moving repeated samples towards sentences with consensus labels, matching



(a) The USC-IEMOCAP corpus



(b) The MSP-IMPROV corpus

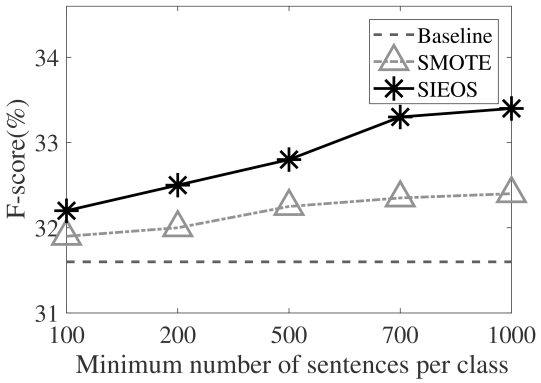
Fig. 6. F1-score obtained for each emotional class (baseline, IEOS method, SIEOS method). The figure also shows the average results across emotion. An asterisk denotes that the F1-score value for that condition is significantly higher (p = 0.05) than the F1-score of the approach identified by the asterisk's color.

TABLE 1
Average Bhattacharyya coefficient across pairs of categorical emotions. The table reports the mean and *standard deviation* (STD) across emotions for 20 implementations of the sampling methods.

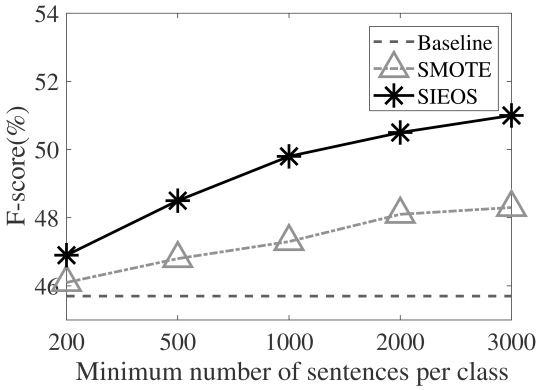
	USC-IEMOCAP (N=3)		MSP-IMPROV (N=5)	
	Mean	STD	Mean	STD
Baseline	0.524	0.000	0.367	0.000
SMOTE	0.495	0.003	0.359	0.003
IEOS	0.628	0.002	0.431	0.002
SIEOS	0.553	0.004	0.401	0.002

the target class (function *populate* in algorithm 1), the overlap between emotions is reduced. This analysis confirms the role of SIEOS in decreasing the overlap between distinct classes. For comparison, over-sampling with SMOTE reduces the overlap between classes to 0.495. SMOTE generates samples by interpolating sentences with consistent labels, which reduces the overlap. Similar results are observed in the MSP-IMPROV corpus.

The SIEOS method introduces more overlap between the classes, which is not something desirable from a machine-learning perspective. The key observation to understand the benefits of the proposed approach is that speech emotion recognition, or any other task where the ground-truth labels are coming from perceptual evaluations describing subjective judgments, imposes specific challenges in classification tasks.



(a) USC-IEMOCAP



(b) MSP-IMPROV

Fig. 7. Effect of balancing emotional classes with synthetic samples. The figure shows the F1-score by increasing the number of samples per emotional class. The additional samples come from over-sampling the data using either SMOTE or the SIEOS approach.

Generally, a classifier has well defined classes. Samples from different classes may overlap in the feature space, so it may not be possible to obtain a perfect hyperplane to separate the classes. However, we know where each sample belongs. In speech emotion recognition, the emotional content in conversational speech is often ambiguous with mixed emotion [53], [54], [55], [56], [57]. A sentence may convey more than one emotion, where one person may perceive "anger", while another may perceive "frustration". We argue that both judgments may be right as speech may convey traits of both emotions. Notice that the overlap is intrinsically related to the labels. This problem leads to the question on how to train a classifier for speech emotion recognition. In this paper, we propose an oversampling framework that considers the evaluations provided by multiple raters, before estimating the consensus label. As a result, the formulation captures the emotional traits that are not represented by the consensus labels, which leads to clear improvement in classification, even with the increased overlap between classes.

5.5 Comparison to SMOTE over-sampling

This section compares the proposed approaches with SMOTE, which is the most popular over-sampling method. As discussed earlier, many of the emotion databases have unbalanced emotional content. Therefore, SMOTE has been successfully used to balance the size of the classes in the training set, improving the overall classification performance.

The first evaluation increases the number of samples of the minority classes until reaching a target number of samples per class (200, 500, 1,000, 2,000, and 3,000). We do not create

synthetic samples for a class that has more sentences than the target number of sentences. As a result, all the categorical emotions include at least the minimum target size in the over-sampled training set. This approach is different from the approach presented in previous sections where a fixed number of samples is synthesized per sentence (i.e., N). For this evaluation, we collect all the individual annotations selecting a given emotional class. A sentence can be represented multiple times if several raters perceived and annotated this class. From this pool of evaluations, we randomly select the required number of sentences with replacement. We conduct this analysis on the USC-IEMOCAP and MSP-IMPROV databases. We follow the standard procedure for SMOTE [24]. Figure 7 shows the F1-score of emotion classifiers trained with the over-sampled training set created with SMOTE and the SIEOS method. For comparison, the figure also includes the baseline results using the original unbalanced training set considering the consensus labels (horizontal dashed line). The x-axis in the figures shows the minimum number of samples for each categorical class after over-sampling. The results show that the approaches with over-sampling improve the F1-score performance over the baseline, highlighting the importance of data augmentation in speech emotion recognition. When we compare SMOTE and the SIEOS framework, we can clearly observe improvements with our proposed approach on both corpora. The key benefit of the SIEOS framework is the use of individual evaluations in the over-sampling process. The results indicate the clear benefits of this simple, but powerful approach.

The second evaluation explores the benefit of combining SMOTE with the proposed over-sampling frameworks. Table 2 presents the results for the USC-IEMOCAP corpus, and Table 3 presents the results for the MSP-IMPROV corpus. Tables 2 and 3 list the results in terms of average precision, average recall and F1-score. We provide the 95% confidence interval for a population proportion. For both tables, the first row presents the results for the baseline method trained with consensus labels. The second and third rows present the results for the IEOS and SIEOS methods, creating either three (USC-IEMOCAP corpus) or five (MSP-IMPROV corpus) synthetic samples per sentence. The fourth row presents the results for SMOTE. Tables 2 and 3 show that the baseline using the consensus labels give the worst F1-scores. Using over-sampling consistently increases the classification performance over the baseline. The absolute improvements obtained with the SIEOS method over the baseline are 5.4% for the USC-IEMOCAP corpus and 5.2% for the MSP-IMPROV corpus. Both tables demonstrate the superior performance of the IEOS and SIEOS methods over SMOTE, showing improvements between 1.2% to 2.7%.

The last two rows of Tables 2 and 3 show the results when we combine the proposed approaches with SMOTE. After generating N samples per sentence, we balance the corpus with SMOTE by making sure that the minimum number of sentences per emotional class is either 3,000 for the MSP-IMPROV corpus, or 1,000 for the USC-IEMOCAP corpus. The last two rows of Tables 2 and 3 show that combining SMOTE after applying the proposed over-sampling approach does not produce any additional gain in classification performance, suggesting that the SIEOS framework is enough to augment the training set.

TABLE 2
Classification performance of the proposed over-sampling methods on the USC-IEMOCAP database. The table also lists the results for SMOTE, and the combination of the IEOS and SIEOS methods with SMOTE.

	Recall [%]	Precision [%]	F1-score [%]
Baseline	29.1±1.2	30.5±1.2	29.7±1.2
IEOS (N=3)	32.4±1.2	34.1±1.2	33.6±1.2
SIEOS (N=3)	34.6±1.2	36.9±1.3	35.1±1.2
SMOTE	31.1±1.2	34.2±1.2	32.4±1.2
IEOS (N=3) + SMOTE	31.1±1.2	36.9±1.3	33.2±1.2
SIEOS (N=3) + SMOTE	31.7±1.2	39.1±1.3	34.6±1.2

TABLE 3
Classification performance of the proposed over-sampling methods on the MSP-IMPROV database. The table also lists the results for SMOTE, and the combination of the IEOS and SIEOS methods with SMOTE.

	Recall [%]	Precision [%]	F1-score [%]
Baseline	43.2±1.1	48.5±1.1	45.7±1.1
IEOS (N=5)	47.0±1.1	53.5±1.1	50.1±1.1
SIEOS (N=5)	47.7±1.1	54.6±1.1	50.9±1.1
SMOTE	45.1±1.1	52.4±1.1	48.3±1.1
IEOS (N=5) + SMOTE	46.2±1.1	58.2±1.0	51.6±1.1
SIEOS (N=5) + SMOTE	45.9±1.1	56.8±1.1	50.8±1.1

6 Conclusions and Future Work

This study introduced an over-sampling method for machine-learning problems where the ground truth is obtained by estimating consensus labels from individual annotations. The formulation creates repeated samples for each sentence by sampling the labels assigned by individual evaluators. The new samples are then transformed in the feature space by randomly moving their values close to real sentences with the target emotion. This is a simple, but powerful method to leverage individual evaluations, which are commonly ignored after obtaining consensus labels. The study evaluated the proposed approach for speech emotion recognition of categorical classes using the USC-IEMOCAP and MSP-IMPROV corpora. The experimental results showed that using this over-sampling method significantly improves the classification performance. The evaluation shows that the proposed framework is better than SMOTE, which is the most common over-sampling method. Applying SMOTE after creating N samples per sentence using the proposed approach does not lead to significant improvements in classification performance, indicating that the proposed over-sampling approach is enough to augment the training set.

This paper shows the benefits of considering individual annotations in speech emotion recognition, which opens multiple research directions. While the formulation assumes that the reliability of all the annotators is similar, some evaluators are more reliable than others. Therefore, it is reasonable to investigate the trade-off between the quality and quantity of the annotations [58]. For example, we can prioritize annotations from evaluators who are considered reliable. Another open question is to explore better machine-learning methods to handle overlap between classes. A drawback of our approach is the increased overlap between the classes, so it is important to find robust training methods for this case. Finally, we will explore the performance of the proposed approach using more complex DNNs and with other feature sets.

Acknowledgments

This study was funded by the National Science Foundation (NSF) CAREER grant IIS-1453781.

References

[1] R. W. Picard, "Affective computing," Technical Report 321, MIT Media Laboratory Perceptual Computing Section, Cambridge, MA,USA, November 1995.

[2] A. Dawid and A. Skene, "Maximum likelihood estimation of observer error-rates using the EM algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 20–28, 1979.

[3] C. Gao and D. Zhou, "Minimax optimal convergence rates for estimating ground truth from crowdsourced labels," *ArXiv e-prints (arXiv:1310.5764)*, pp. 1–38, October 2013.

[4] Z. Aldeneh and E. Mower Provost, "Using regional saliency for speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, New Orleans, LA, USA, March 2017, pp. 2741–2745.

[5] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1469–1477, September 2015.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NIPS 2012)*, Lake Tahoe, CA, USA, December 2012, vol. 25, pp. 1097–1105.

[7] C. Phua, D. Alahakoon, and V. Lee, "Minority report in fraud detection: classification of skewed data," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 50–59, June 2004.

[8] J. Burez and D.Van den Poel, "Handling class imbalance in customer churn prediction," *Expert Systems with Applications*, vol. 36, no. 3, pp. 4626–4636, April 2009.

[9] G. Wu and E. Y. Chang, "KBA: kernel boundary alignment considering imbalanced data distribution," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 6, pp. 786–795, June 2005.

[10] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J.M. Buhmann, "The balanced accuracy and its posterior distribution," in *International Conference on Pattern Recognition (ICPR 2010)*, Istanbul, Turkey, August 2010, pp. 3121–3124.

[11] I. Tomek, "Two modifications of CNN," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-6, no. 11, pp. 769–772, November 1976.

[12] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: one-sided selection," in *International Conference on Machine Learning (ICML 1997)*, Nashville, TN, USA, July 1997, pp. 179–186.

[13] P. Hart, "The condensed nearest neighbor rule," *IEEE Transactions on Information Theory*, vol. 14, no. 3, pp. 515–516, May 1968.

[14] L. S. Yaeger, R. F. Lyon, and B. J. Webb, "Effective training of a neural network character classifier for word recognition," in *Advances in Neural Information Processing Systems (NIPS 1997)*, Denver, CO, USA, December 1996, vol. 9, pp. 807–816.

[15] P.Y. Simard, D. Steinkraus, and J.C. Platt, "Best practices for convolutional neural networks applied to visual document analysis," in *International Conference on Document Analysis and Recognition (ICDAR 2003)*, Edinburgh, UK, August 2003, pp. 958–963.

[16] J. R. Bellegarda, P. V. de Souza, A. Nadas, D. Nahamoo, M. A. Picheny, and L. R. Bahl, "The metamorphic algorithm: a speaker mapping approach to data augmentation," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 3, pp. 413–420, July 1994.

[17] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (VTLP) improves speech recognition," in *Workshop on Deep Learning for Audio, Speech, and Language Processing at International Conference on Machine Learning (ICML 2013)*, Atlanta, GA, USA, June 2013, pp. 1–5.

[18] A. Ragni, K. M. Knill, S. P. Rath, and M. Gales, "Data augmentation for low resource languages," in *Interspeech 2014*, Singapore, September 2014, pp. 810–814.

[19] G. Batista, R. Prati, and M. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 20–29, June 2004.

[20] D.L. Wilson, "Asymptotic properties of nearest neighbor rules using edited data," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-2, no. 3, pp. 408–421, July 1972.

[21] B. Schuller and F.Burkhardt, "Learning with synthesized speech for automatic emotion recognition," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2010)*, Dallas, TX, USA, March 2010, pp. 5150–5153.

[22] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Interspeech 2015*, Dresden, Germany, September 2015, pp. 3586–3589.

[23] 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.

[24] N.V. Chawla, K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal Of Artificial Intelligence Research*, vol. 16, pp. 321–357, June 2002.

[25] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Interspeech 2009 - Eurospeech*, Brighton, UK, September 2009, pp. 312–315.

[26] T. Vogt and E. André, "Exploring the benefits of discretization of acoustic features for speech emotion recognition," in *Interspeech 2009*, Brighton, UK, September 2009, pp. 328–331.

[27] A. Rosenberg, "Classifying skewed data: Importance weighting to optimize average recall," in *Interspeech 2012*, Portland, OR, USA, September 2012, pp. 2242–2245.

[28] Y. Attabi and P. Dumouchel, "Anchor models for emotion recognition from speech," *IEEE Transactions on Affective Computing*, vol. 4, no. 3, pp. 280–290, July-September 2013.

[29] L. Martin and P. Pu, "Prediction of helpful reviews using emotions extraction," in *AAAI Conference on Artificial Intelligence (AAAI 2014)*, Qubec City, QC, Canada, July 2014, pp. 1551–1557.

[30] S. Yildirim, S.S. Narayanan, and A. Potamianos, "Detecting emotional state of a child in a conversational computer game," *Computer Speech and Language*, vol. 25, no. 1, pp. 29–44, January 2011.

[31] R.A. Calix, M.A. Khazaeli, L. Javadpour, and G.M. Knapp, "Dimensionality reduction and classification analysis on the audio section of the SEMAINE database," in *Affective Computing and Intelligent Interaction (ACII 2011)*, S. D'Mello, A. Graesser, B. Schuller, and J.-C. Martin, Eds., vol. 6975/2011 of *Lecture Notes in Computer Science*, pp. 323–331. Springer Berlin / Heidelberg, Memphis, TN, USA, October 2011.

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

[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.

[34] 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.

[35] R. Lotfian and C. Busso, "Retrieving categorical emotions using a probabilistic framework to define preference learning samples," in *Interspeech 2016*, San Francisco, CA, USA, September 2016, pp. 490–494.

[36] S. Steidl, M. Levit, A. Batliner, E. Nöth, and H. Niemann, ""Of all things the measure is man" automatic classification of emotions and inter-labeler consistency," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)*, Philadelphia, PA, USA, March 2005, vol. 1, pp. 317–320.

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

[38] 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.

[39] C. Busso and S.S. Narayanan, "Recording audio-visual emotional databases from actors: a closer look," in *Second International Workshop on Emotion: Corpora for Research on Emotion and Affect, International conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May 2008, pp. 17–22.

[40] 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.

[41] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. André, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. Truong, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice

research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, April-June 2016.
[42] 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.
[43] K.R. Scherer and G. Ceschi, "Lost luggage: A field study of emotion antecedent appraisal," *Motivation and Emotion*, vol. 21, no. 3, pp. 211–235, September 1997.
[44] 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.
[45] N. Japkowicz, "The class imbalance problem: Significance and strategies," in *International Conference on Artificial Intelligence (ICAI 2000)*, Las Vegas, Nevada, USA, June 2000, pp. 111–117.
[46] 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.
[47] M.W. Bhatti, Y. Wang, and L. Guan, "A neural network approach for human emotion recognition in speech," in *IEEE International Symposium on Circuits and Systems (ISCAS 2004)*, Vancouver, BC, Canada, May 2004, vol. 2, pp. 181–184.
[48] K. Soltani and R.N. Ainon, "Speech emotion detection based on neural networks," in *International Symposium on Signal Processing and Its Applications (ISSPA 2007)*, Sharjah, United Arab Emirates, June 2007, pp. 1–3.
[49] Y. Kim, H. Lee, and E. Mower Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)*, Vancouver, BC, Canada, May 2013, pp. 3687–3691.
[50] D.P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, San Diego, CA, USA, May 2015, pp. 1–13.
[51] A. Burmania and C. Busso, "A stepwise analysis of aggregated crowdsourced labels describing multimodal emotional behaviors," in *Interspeech 2017*, Stockholm, Sweden, August 2017, pp. 152–157.
[52] T. Kailath, "The divergence and Bhattacharyya distance measures in signal selection," *IEEE Transactions on Communication Technology*, vol. 15, no. 1, pp. 52–60, February 1967.
[53] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth, "Desperately seeking emotions or: actors, wizards and human beings," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, Newcastle, Northern Ireland, UK, September 2000, pp. 195–200.
[54] R. Cowie, E. Douglas-Cowie, and C. Cox, "Beyond emotion archetypes: Databases for emotion modelling using neural networks," *Neural Networks*, vol. 18, no. 4, pp. 371–388, May 2005.
[55] 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.
[56] E. Douglas-Cowie, L. Devillers, J.C. Martin, R. Cowie, S. Savvidou, S. Abrilian, and C. Cox, "Multimodal databases of everyday emotion: Facing up to complexity," in *9th European Conference on Speech Communication and Technology (Interspeech'2005)*, Lisbon, Portugal, September 2005, pp. 813–816.
[57] E. Mower, A. Metallinou, C.-C. Lee, A. Kazemzadeh, C. Busso, S. Lee, and S.S. Narayanan, "Interpreting ambiguous emotional expressions," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2009)*, Amsterdam, The Netherlands, September 2009, pp. 1–8.
[58] A. Burmania, M. Abdelwahab, and C. Busso, "Tradeoff between quality and quantity of emotional annotations to characterize expressive behaviors," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, March 2016, pp. 5190–5194.

Reza Lotfian (SM'17) received his BS degree (2006) with high honors in Electrical Engineering from the Department of Electrical Engineering, Amirkabir University, Tehran, Iran, the MS degree (2010) in Electrical Engineering from the Sharif University (SUT), Tehran, Iran, and the PhD degree (2018) in electrical engineering from the University of Texas at Dallas (UTD). He is currently a research scientist at Cogito Corp at Boston, Massachusetts, USA. His research interest includes the area of speech signal processing, affective computing, human machine interaction, and machine learning.

Carlos Busso (S'02-M'09-SM'13) received the B.S. and M.S. degrees (Hons.) in electrical engineering from the University of Chile, Santiago, Chile, in 2000 and 2003, respectively, and the Ph.D. degree in electrical engineering from the University of Southern California (USC), Los Angeles, in 2008. He is an associate professor at the Electrical Engineering Department of The University of Texas at Dallas (UTD). He was selected by the School of Engineering of Chile as the best electrical engineer graduated in 2003 across Chilean universities. At USC, he received a provost doctoral fellowship from 2003 to 2005 and a fellowship in Digital Scholarship from 2007 to 2008. At UTD, he leads the Multimodal Signal Processing (MSP) laboratory [http://msp.utdallas.edu]. He is a recipient of an NSF CAREER Award. In 2014, he received the ICMI Ten-Year Technical Impact Award. In 2015, his student received the third prize IEEE ITSS Best Dissertation Award (N. Li). He also received the Hewlett Packard Best Paper Award at the IEEE ICME 2011 (with J. Jain), and the Best Paper Award at the AAAC ACII 2017 (with Yannakakis and Cowie). He is the co-author of the winner paper of the Classifier Sub-Challenge event at the Interspeech 2009 emotion challenge. His research interest is in human-centered multimodal machine intelligence and applications. His current research includes the broad areas of affective computing, multimodal human-machine interfaces, nonverbal behaviors for conversational agents, in-vehicle active safety system, and machine learning methods for multimodal processing. His work has direct implication in many practical domains, including national security, health care, entertainment, transportation systems, and education. He was the general chair of ACII 2017. He is a member of ISCA, AAAC, and ACM, and a senior member of the IEEE.