IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 5, 2021, accepted March 22, 2021, date of publication April 7, 2021, date of current version April 16, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3071485

Perceptual Borderline for Balancing Multi-Class Spontaneous Emotional Data

LEILA BEN LETAIFA AND M. INÉS TORRES, (Member, IEEE)
Speech Interactive Research Group, Universidad del País Vasco UPV/EHU, 48940 Leioa, Spain

Corresponding author: M. Inés Torres (manes.torres@ehu.eus)

This work was supported by the European Commission Horizon 2020 Research and Innovation Program through the Empathic, Expressive, Advanced Virtual Coach to Improve Independent Healthy-Life-Years of the Elderly (EMPATHIC) Project (Grant number 769872) and through the Mental Health Monitoring Through Interactive Conversations (MENHIR) Marie Skłodowska-Curie Action (Grant number 823907).

ABSTRACT Speech is a behavioural biometric signal that can provide important information to understand the human intends as well as their emotional status. The paper is centered on the speech-based identification of the seniors's emotional status during their interaction with a virtual agent playing the role of a health professional coach. Under real conditions, we can just identify a small set of task-dependent spontaneous emotions. The number of identified samples is largely different for each emotion, which results in an imbalanced dataset problem. This research proposes the dimensional model of emotions as a perceptual representation space alternative to the generally used acoustic one. The main contribution of the paper is the definition of a perceptual borderline for the oversampling of minority emotion classes in this space. This limit, based on arousal and valence criteria, leads to two methods of balancing the data: the Perceptual Borderline oversampling and the Perceptual Borderline SMOTE (Synthetic Minority Oversampling TEchnique). Both methods are implemented and compared to state-of-the-art approaches of Random oversampling and SMOTE. The experimental evaluation was carried out on three imbalanced datasets of spontaneous emotions acquired in human-machine scenarios in three different cultures: Spain, France and Norway. The emotion recognition results obtained by neural networks classifiers show that the proposed perceptual oversampling methods led to significant improvements when compared with the state-of-the art, for all scenarios and languages.

INDEX TERMS Dimensional model of emotions, emotion recognition, multi-class classification, perceptual borderline, speech analysis, speech processing.

I. INTRODUCTION

Speech is a biometric signal that is able to provide information about the identity of the speaker [1], [2], the content of the message [3], [4] or the language used to code it [5]. In addition, speech can be analysed to identify the current emotional status of the speaker, which could be an important cue to improve mood prediction as well as to monitor some mental disorders such as anxiety or depression, among others [6]. However, speech as a behavioural biometric data, can also be affected by many other factors such as the speaker habits, personality, culture or the specific task being performed. As a consequence, any behaviour analysis, prediction or monitoring becomes a challenge.

This work concerns well-being conversational systems for behavioural change, which is a research topic of growing interest [7]. In particular, the EMPATHIC european project[1] develops new interaction paradigms for personalized virtual coaches to promote healthy and independent aging. A natural speech interaction between humans and machines implies, among other things, that the machine understands the emotional state of the user, hence the importance of emotion recognition. The paper is centered on the speech-based identification of the senior's emotional status during their interaction with a virtual agent playing the role of a health professional coach. The knowledge of the senior's emotions allows the system to react accordingly [8]–[10]. In addition, the outcomes of these interactions in terms of moods and emotions could provide useful and real time information to the elderly care support systems as well as to caregivers [11].

This objective needs to face important challenges, of which the more important are derived from the need of processing

The associate editor coordinating the review of this manuscript and approving it for publication was Bijan Najafi.

[1] http://www.empathic-project.eu

spontaneous emotions triggered in real scenarios. Given the difficulty to implement these conditions, the research on machine analysis of human emotions has been carried out over the simulation of the six basic emotions [12] performed by professional actors in the lab [13]. However, the features selected for acted and realistic emotions show significant differences [14]. Furthermore, only a small subset of the emotions defined by Eckman [15] can be distinguished in real scenarios. Moreover, this subset is strongly dependent of the task. For instance, a political debate on TV [16] or a podcast interview [17] cannot be expected to show the same human emotions than a human-machine scenario [10], [18], [19]. Indeed, fear is not expected in none of the mentioned corpus.

Emotions, unlike moods, are triggered by specific events and do not last more than a few seconds after which the basal mood appears again [12], [20]. Importantly, these emotions are usually the ones to be considered for the analysis of human behaviour as for instance in the proposed research scenario where the virtual coach needs to adapt its conversation accordingly [9]. In terms of Artificial Intelligence and Pattern Recognition techniques these facts pose an important problem for the automatic identification of human emotions: the huge difference among the number of samples acquired for each of the spontaneous emotions identified [21].

As a consequence, the number of objects in one class, or emotion, is considerably lower than in other classes. Referred to as an imbalanced dataset problem, classification performance typically degrades in several data mining applications, including pattern recognition, telecommunications and bioinformatics [21], [22]. After two decades of research, learning from imbalanced data is still a focus [23], [24]. In addition, the multi-class imbalanced classification is not well developed as a binary classification [25]. In this paper we are dealing with a more complicated situation. Indeed, when it comes to multi-class imbalanced data, we can easily loose performance on one class while trying to gain on another [22], [24]. Given this problem, a more in-depth understanding of the nature of the class imbalance problem is needed. Recent trends focus on analyzing not only the disproportion between classes, but also other difficulties related to the nature of the data [26].

Speech emotion recognition is naturally a multi-class classification problem, in which the imbalance character of the dataset could affect its performance. Indeed, most of the test segments are assigned to the majority class due to the data skewness. However, all the emotions should have the same importance. Even so, minority classes are a focus of attention in many scenarios [16].

This work examines the intrinsic properties of the speech samples to find a suitable borderline between majority and minority emotion classes. In a previous work [19] we have proposed the dimensional model of emotions (Valence-Arousal-Dominance, or VAD) as an additional space of the parameter representation, which has demonstrated to improve the emotion recognition performance. On the other hand, some studies [27] have shown that the oversampling of borderline samples is an effective way to deal with imbalance of data and remains the state of the art.

The main contribution of this paper is the proposal of the VAD model of emotions as a representation space alternative to the generally used acoustic one to ground the oversampling borderline criteria. To our knowledge, this is the first time that the dimensional parameters are involved in a balancing process of emotional data. The work is based on the following hypothesis: if two emotions are close in the dimensional space, they are acoustically nearby and the emotion classifier can confuse them. It is also assumed that the information extracted from perceptual space is more reliable than that from acoustic space. Indeed, the perceptual space is the result of a manual annotation procedures whereas the acoustic one is automatically estimated by machine.

The perceptual borderline gives rise to oversampling approaches in the context of multi-class oversampling - either by increasing the size of the samples on the frontier (by replication or by artificial synthesis), - or by establishing a variable number of neighbours for the artificial synthesis in order to avoid classes' samples overlapping, resulting in additional contributions [21], [23], [24]. The proposed methodology has been evaluated over three very imbalanced corpus consisting of emotional speech samples acquired in a spontaneous human-machine scenario in three different cultures, namely Spain, France and Norway, resulting in significant improvements of emotion identification rates.

The paper is organized as follows. Section II illustrates related works and Section III describes the dimensional model of emotions. Section IV develops the perceptual oversampling methodology. Then Section V shows the description and analysis of the EMPATHIC data for three languages and Section VI describes the experimental framework. Finally Section VII shows and discusses experimental results and Section VIII presents the main conclusions of the work.

II. RELATED WORK

Studies have shown that for several basic classifiers, a balanced dataset provides improved overall classification performance compared to imbalanced data set [28]. Hence the interest in balancing data. Solutions addressing the problem of imbalanced data tend to focus on the data level and classifier level. Hybrid methods tend to combine their advantages. Undersampling [29] and oversampling [27], [30] are common data sampling methods. Undersampling removes some data from the majority class which can lead to a loss of discriminating samples. Oversampling appends replicated data to the original dataset, so multiple instances of certain examples become ''tied'' leading to overfitting [28]. Furthermore, marginal and noisy examples are also replicated.

In order to introduce some generalization, Synthetic Minority Over-sampling TEchnique (SMOTE) [31] is proposed. This approach aims to overcome imbalance in the original data sets by artificially generating data samples. To this end, it produces synthetic samples between an example and its nearest neighbours. SMOTE method generates the same number of synthetic data samples for each original minority example. This procedure, does not pay attention to neighbour examples, which results in an increase of the occurrence of overlapping between classes [28]. To avoid this effect, various adaptive sampling methods [32] have been put forward. Some representative work include Borderline-SMOTE [33] and Adaptive Synthetic sampling (ADA-SYN) [34]. In the case of Borderline-SMOTE, borderline samples are identified and then over-sampled. On the other hand, ADA-SYN creates different amount of synthetic data according to their distribution: more synthetic data are generated for minority class samples that are harder to learn compared to minority samples that are easier to learn [34].

Algorithm-level methods modify classifiers to alleviate their bias towards majority class. The most popular approaches are grouped as cost-sensitive learning. The classifier is modified to introduce varying penalty for each class. These methods have been used in many classification systems, including boosting, decision trees and Support Vector Machines (SVM) and recently Deep Neural Networks (DNN) [35].

Most of the mentioned work has been carried out in a binary classification context. Multi-class classification of imbalanced data has not been well developed because of the complexity of the task. We can face several difficulties, namely - class overlapping may appear with more than two groups, - class label noise may affect the problem and - borders between classes may be far from being clearly defined. Therefore, data sampling procedures that take into account the variety of characteristics of classes and balanced performance of all of them should be proposed [26].

As mentioned above, corpus of spontaneous emotions acquired in realistic scenarios are scarce because huge effort is needed to their development. They are also more challenging because the emotional state of the speakers is not predetermined. In fact, Speech Emotion Recognition (SER) from spontaneous data is still a challenging task and several further steps must be taken before SER can be considered ready for usage "in the wild" [36]. As a consequence, spontaneous SER systems performs generally worse than acted SER systems [37]–[40]. In addition, they face the problem of imbalance in data.

Few research studies are carried out to balance emotional training data. Generally, small sample environment and acted data, in which ratio between majority and minority classes is not very important, are targeted. For example, a selective SMOTE algorithm based on acoustic parameters is described in [41]. The approach aims to avoid oversampling noisy samples. It is validated on acted small datasets (SAVEE [42], EMO-DB [43] and CASIA [44]) using SVM classifier. In [45], Random and SMOTE oversampling are applied to balance IMPROV [46] and IEMOCAP [47] datasets. In both acted datasets, the majority/minority class ratio is about 10%. SMOTE technique is also employed with emotional Youtube data and evaluated by three learning techniques: multi nominal Naive Bayes, decision trees and SVM [21]. Oversampling and undersampling are combined in [39] to increase movies data. Then SVM algorithm classified samples into fear/not fear.

As sub-mentioned, Random oversampling approach suffers from replicating noisy samples and SMOTE performs blind interpolation with fuzzy class boundaries. The objective of this research is to overcome these drawbacks in the context of multi-class prediction. In this paper we first show the advantage of a borderline based on the intrinsic nature of the data. In this context, both oversampling ways, by replication and synthesis, are investigated. On the other hand, SMOTE varieties (such as Borderline-SMOTE and ADA-SYN) focus on fixing the number of artificial samples to prevent samples overlapping. We then suggest to use a dynamic neighbours number to solve the same issue. This number is based on the perceptual borderline. This work is carried out in the context of a spontaneous emotional dataset in which the minority and majority class ratio is extremely important.

III. DIMENSIONAL SPACE OF EMOTIONS

Emotions are traditionally represented by two models: a discrete categorical model and a continuous dimensional one.

Categorical model is based on a set of mutually exclusive discrete "basic" categories (Fear, Surprise,..). Thus, even though a person might undergo two emotions simultaneously, each emotion belongs to one and only one of the basic categories [48]. Several classification proposals are introduced ([12], [15], [49],..) but they don't seem to converge to the same final categories.

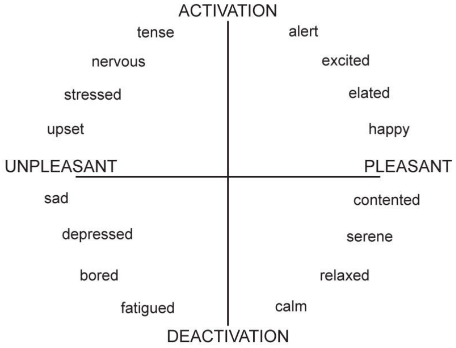


FIGURE 1. The emotional 2-D model from [50] where the horizontal axis represents the valence dimension and the vertical one the arousal dimension.

Dimensional emotion representation is based on some psychological understandings. In the dimensional model, emotions are treated as being dimensional or continuous rather than discrete. The emotion plane is viewed as a continuous space where each point corresponds to a separate emotion state. The VAD is a tridimensional model defined by Valence, Arousal and Dominance axes. However, most models posit the existence of two fundamental dimensions: valence (or pleasantness) and intensity (or arousal) as represented in Figure 1.

Valence varies from −1 (unpleasant) to 1 (pleasant) and therefore it can be characterized as the level of pleasure. Arousal, on the other hand, represents the intensity of the emotional state and it ranges from −1 (passive) to 1 (active) [51].

According to previous studies (see SECTION II), multi-class classification requires a deeper understanding of the intrinsic nature of imbalanced data. In particular, it is interesting to analyse the type of examples present in each class and their relations to the other classes [26]. Dealing with emotions, these analyses can be conducted in an emotional representation space, for instance the 2-D dimensional one. Indeed, the relationships between emotions in the 2D plane are easier to interpret than in the high dimensional acoustic one. In addition, the information extracted from these 2D relationships is more reliable than the one extracted from the acoustic space as it comes from manual annotation of the human perception of the emotional dimensions and not from sets of acoustic parameters automatically calculated.

IV. PERCEPTUAL OVERSAMPLING

We denote by borderline data the minority class samples that are at the frontier with the majority class. Our proposal consists in defining a border based on the intrinsic nature of our data which in this case is based on the human perception of emotions through a manual annotation procedure. More precisely, this frontier is set in the 2-D emotional space where each point represents the valence and arousal values perceived by annotators. For this reason, it is entitled ''Perceptual Borderline''. This border will allow a better over-sampling of the data by replication/synthesis as well as the extension of the SMOTE algorithm to multi-class problems.

A. PERCEPTUAL BORDERLINE OVERSAMPLING (PBO)

The objective of the Perceptual Borderline Oversampling (PBO) approach is to over-sample the borderline samples more than others. This oversampling is performed by replication. The PBO algorithm is described as follows:

```
PBO: Perceptual Borderline Oversampling
  1) All training samples are projected into 2-D space.
     A sample s is represented by its arousal and valence
     co-ordinates s(a,v)
  2) Gravity center of majority class G is computed.
  3) For a sample s of the minority class:
     a) Compute the distance d(s,G) between s and G
     b) Compare d(s,G) to a predefined threshold $\epsilon$.
        • if d(s,G)$\leq \epsilon$, s is considered borderline. It is
          replicated R1 times.
        • if d(s,G)$\geq \epsilon$, s is not borderline. It is replicated
          R2 times with R2 $\leq$ R1
  4) Take an other sample s and go to step 3.
```

Figure 2 represents the projection of minority and majority class data in the 2-D emotional plan. Distances between the center of gravity of the majority class (G) and samples of the minority classes are computed. Then, examples are considered to be borderline samples or not depending on these distances.

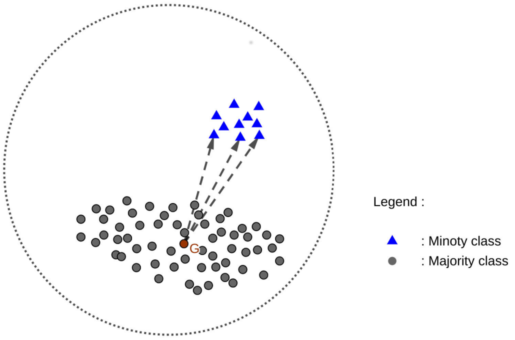


FIGURE 2. **Perceptual borderline oversampling.**

B. PERCEPTUAL BORDERLINE SMOTE (PB-SMOTE)

This method is similar to the previous one except that the oversampling is carried out by the SMOTE algorithm, i.e. the algorithm generates synthetic samples (see Section II). The objective of Perceptual Borderline SMOTE (PB-SMOTE) is to investigate the perceptual borderline impact on synthetic oversampling. The algorithm is described as follows:

```
PB-SMOTE: Perceptual Borderline SMOTE
  1) training samples are projected into 2-D space.
  2) Gravity center of majority class G is computed.
  3) For a sample s of the minority class:
     a) Compute the distance d(s,G) between s and G
     b) Compute the distance d(s,Si) between s and its N
        neighbours Si
     c) Compare d(s,G) to a predefined threshold $\epsilon$.
        • if d(s,G)$\leq \epsilon$, s is considered borderline. It is
          assigned an oversampling rate R=R1.
        • if d(s,G)$\geq \epsilon$, s is not borderline. It's oversampling
          rate R=R2 $\leq$ R1
     d) R synthetic samples s-new are generated between
        s and Si as follows: s-new = s + $\gamma$ d(s,Si).
        $0 \leq \gamma \leq 1$
  4) Take another sample s and go to step 3.
```

Figure 3 shows that more artificial examples are generated for borderline samples than for distant samples.

C. STRETCHY SMOTE (S-SMOTE)

The SMOTE algorithm is based on a previously fixed number of neighbours. Stretchy SMOTE (S-SMOTE), is an extension of the SMOTE algorithm aimed to deal with multi-class oversampling. S-SMOTE method manages a dynamic number of neighbours, which depends on the distance to the majority

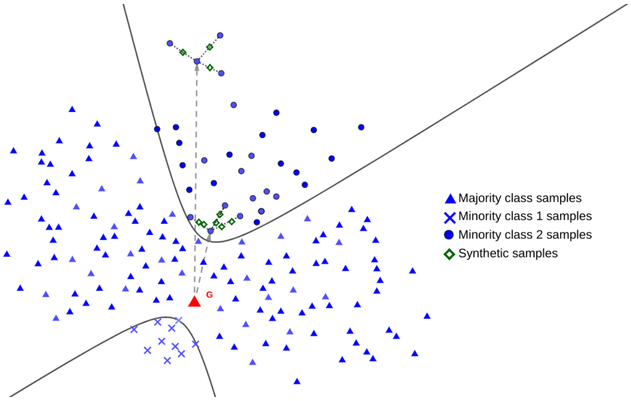


FIGURE 3. Perceptual Borderline SMOTE.

class, instead of considering a fixed number. The algorithm is described as follows:

S-SMOTE: Stretchy SMOTE

```
• Training data are projected into 2-D space.
• Compute R, the ratio between majority and minority classes.
• Gravity center of majority class G is also computed.
• For each sample s of a minority class:
    1) Compute the distance d(s,G) between s and G
    2) Compare d(s,G) to a predefined threshold $\epsilon$.
        – if d(s,G)$\leq \epsilon$, s is considered borderline. It's neighbours number is N=N1.
        – if d(s,G)$\geq \epsilon$, s is not borderline. It's neighbours number is N=N2 $\geq$ N1.
    3) R samples (s-new) are synthesized between s and its N neighbours. s-new = s+$\gamma$ d(s,Si). $d(s, Si)$ is the distance between s and its N nearest neighbours Si. $0 \leq \gamma \leq 1$
```

Figure 4 shows that borderline samples have less neighbours than distant samples, for the same oversampling ratio (for instance 5 in the picture).

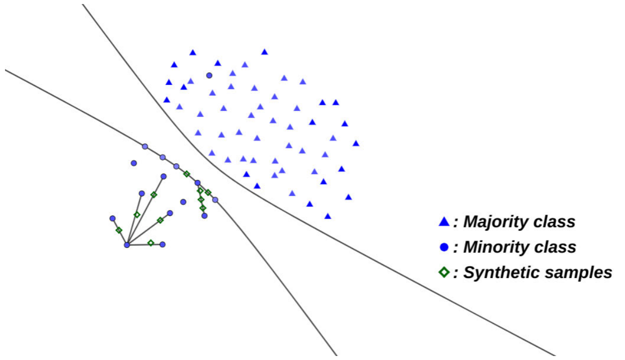


FIGURE 4. Stretchy SMOTE.

V. DATA ANALYSIS

This research is carried out in the context of EMPATHIC project, where seniors interact with a virtual agent playing the role of a health professional coach. In this context, three highly imbalanced corpus of spontaneous emotions were acquired in human-machine scenarios in three different languages and cultures: Spanish, French and Norwegian.

The recordings were carried out in four main regions of Europe: the Basque Country in Spain, Île-de-France and Bourgogne regions in France and Oslo area in Norway. Participants were required to be over 64 years, healthy and living independently.

TABLE 1. Some demographic percentages of the participants in the three countries. All of them were required to be above 64 years, healthy and living independently.

	Spain %	France %	Norway %
Gender: Female	70	85	69
Age: under 75 years	79	78	62
Studies: University	61	65	74
Marital status: Married	60	38	-

Table 1 shows some demographic statistics of the participants. Most of them are women under 75 and have got higher education. Experiments performed in the three countries were approved by the corresponding ethical committees, namely Ethics Committee for Research involving Human Beings[2] of the University of the Basque Country and the Basque Ethical Committee for the Clinical Research[3] in Spain, the National Commission for Computing and Freedoms[4] in France and the Privacy Office of the Oslo University Hospital in Norway.

Through the next subsections we analyse the perception experiments and their results as well as the emotion distribution in the three datasets.

A. PERCEPTUAL ANNOTATION OF EMOTIONS

The audio files of the three datasets were labelled by native collaborators in terms of emotions. Two dialogues are recorded per speaker: the first one implements an introductory session through some general aspects about the senior's lifestyle whereas the second one simulates a coaching session on healthy habits for nutrition. As a result, the corpus consists of 134 Spanish audio files, 76 French and 62 Norwegian files. Each file was annotated by three persons. The time limits that indicate changes in emotional state were also set by the annotators.

The perceived emotion was labelled into both, the categorical and the three-dimensional models. Thus, each emotion is described by four parameters:

- its category : Calm, Sad, Happy, Puzzled and Tense,
- its arousal level : excited (1), slightly excited (0) and neutral (−1),
- its valence: positive (1), neither positive or negative (0) and negative (−1)

[2] Comité de ética de la investigación con seres humanos (CEISH) de la Universidad del País Vasco

[3] Comité ético de investigación clínica (CEIC) de Euskadi

[4] Commission Nationale de l'Informatique et des Libertés (CNIL)

- its dominance level : rather dominant (1), neither dominant nor intimidated (0) and rather intimidated (-1).

For each language and corpus, the intersection of the three annotations is considered as the final label. The segments getting annotator's agreement constitute the samples of the corpora. When only two annotators agree, we evaluate the level of confidence of their agreement to decide if the segment will be included or discarded. To this end, the inter-annotator agreement W is computed as follows: $W = \alpha * V + (1-\alpha) * G$ where V is the *per event* agreement also known as Jaccard Index [52] and G is the Global agreement (intersection over union). For these analyses, α is empirically set to 0.5. The agreement score W varies between 0 (for total disagreement) and 1 (for total agreement). If $W \geq$ *threshold* the intersection is performed, otherwise the segment is removed. In this work, threshold is fixed to 0.6.

B. DATASET IMBALANCE

The duration of the audio files for each annotated emotion and dataset are:

- Spanish dataset: 6 hours and 53 minutes for Calm, about 9 minutes for Happy and 10 minutes for Puzzled.
- French dataset: 3 hours and 10 minutes for Calm, about 4 minutes for Happy and 3 minutes for Puzzled.
- Norwegian dataset: 2 hours and 23 minutes for Calm, about 7 minutes for Happy and 40 seconds for Puzzled.

The number of samples and their percentages are reported in Table 2 for each corpus and emotional category.

TABLE 2. Number and percentage of segments.

		Calm	Sad	Happy	Puzzled	Tense
Spain	segments	5423	17	178	260	14
	%	92.04	0.28	3.02	4.41	0.23
France	segments	3597	0	96	77	0
	%	95.41	0	2.54	2.04	0
Norway	segments	3097	0	166	13	0
	%	94.53	0	5.06	0.4	0

Table 2 shows that the three datasets are highly imbalanced, which is the typical situation when dealing with spontaneous emotions and realistic tasks. In fact, the minority classes percentages are between 5 and 0.2 % of the database, which are very low percentages.

For the sake of comparison, only the common classes among datasets are considered for the experiments, namely Calm, Happy and Puzzled. The Calm category is designated majority class whereas Happy and Puzzled are designated minority classes. Ratios between majority class and minority classes are reported in Table 3.

Table 3 shows that the Norwegian corpus is less balanced than the French one, which is also less balanced than the Spanish one. In addition, Norwegian corpus not only faces the imbalance problem but also the small size of the dataset. The combination of these issues presents a new challenge to the community [28], [53].

TABLE 3. Ratios between each pair of majority and minority classes in the three datasets.

	Calm/Happy	Calm/Puzzled
Spanish	30.46	20.85
French	37.46	46.71
Norwegian	18.65	238.23

C. EMOTIONS DISTRIBUTION

The number of speakers is 67 for Spanish experiments, 38 for French and 31 for Norwegian ones. Let's classify these speakers into four groups: - speakers who are perceived always Calm, and thus labeled Calm -speakers whose audio files include segments labeled as Calm and segments labeled as Happy - speakers who generate both Calm and also Puzzled segments and - speakers whose audio files include segments labeled as Calm, segments labeled as Happy and segments labeled as Puzzled. Each group is represented by a bar in Figure 5. The x-axis indicates the number of speakers and the y-axis represents the percentage of segments per emotion. We notice that most Spanish and French speakers express all three emotions. However, the majority of Norwegians show only two emotions.

Figure 5 also shows the speaker dependency of emotions. In fact, Calm segments are present in all the recordings whereas Happy and Puzzled are concentrated in a subset of speakers. In particular, Puzzled category only appears in six Norwegian speakers'dialogues.

Thus, in this emotion recognition task, taking or skipping some speakers of the training/test set can lead to the presence or absence of emotions in that set. Hence the importance of choosing an adapted experimental protocol.

VI. EXPERIMENTAL PROTOCOL

As explained above, in Subsection V-C, some speakers don't show certain emotions so an experimental protocol such as ''one speaker leave out'' does not guarantee the presence of all emotions in all the folds. As a consequence, we opted for a protocol where all the emotions are present in all the train/test folds. However, this experimental protocol can cause performance losses of the Norwegian system, as we will discuss in Section VII.

Each dataset has been divided into ten partitions. In order to obtain the same distribution of emotions in all partitions, each partition contained 10% of each emotion occurrences. The experimental protocol is based on 10 training/test folds where:

- training/test folds are not overlapping
- folds cover all the dataset.
- each test fold matches a partition
- training set contains the rest

In the following subsections we will analyse emotion distribution per training/test fold in terms of both, categories and continuous dimensions.

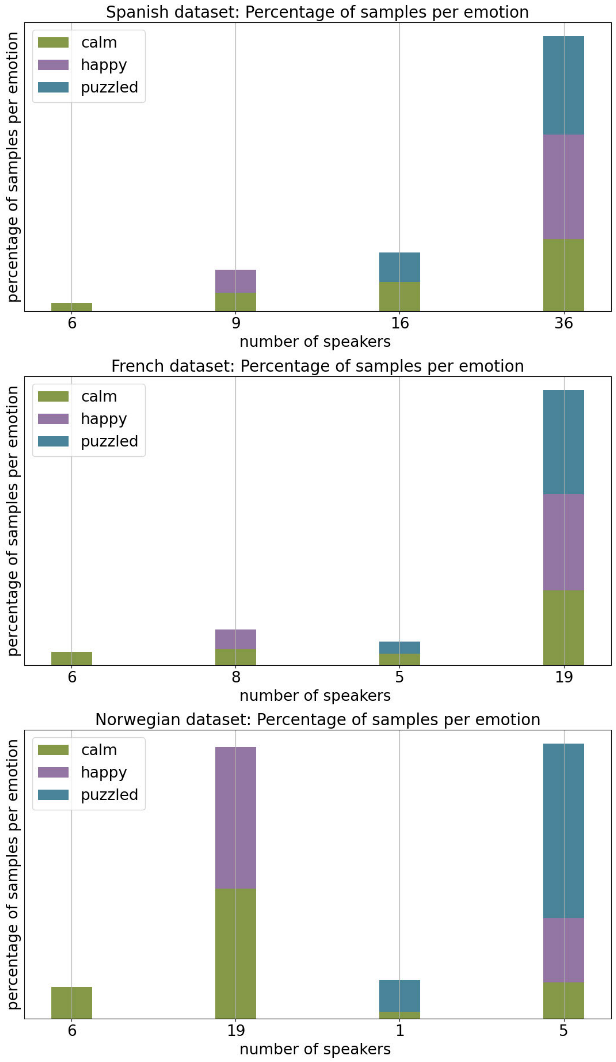
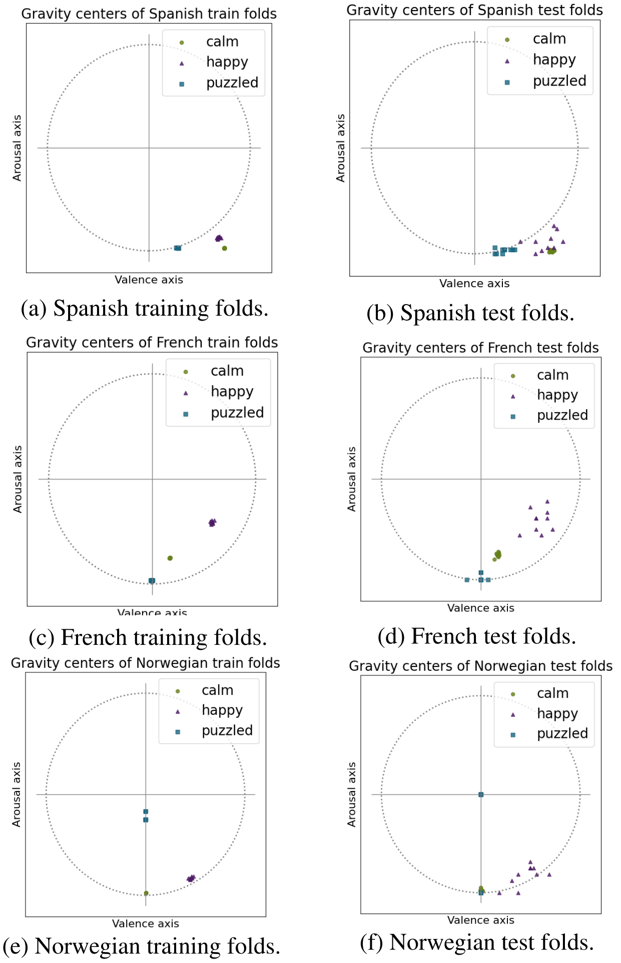


FIGURE 5. Number of samples per speaker and their percentage per emotional category for the three datasets.



(a) Spanish training folds.

(b) Spanish test folds.

(c) French training folds.

(d) French test folds.

(e) Norwegian training folds.

(f) Norwegian test folds.

FIGURE 6. Gravity center of each class and fold represented in the 2-D dimensional space for Spanish, French and Norwegian datasets.

A. CATEGORICAL DISTRIBUTION

The majority class contains enough data for all training/test folds, but minority classes can show a lack of data in some cases.

As defined by the protocol, the distribution of emotion categories is similar for the ten training/test folds. This is convenient for the Spanish and the French datasets but not for the Norwegian one. Indeed, the reduced amount of data of Norwegian dataset results in only one sample of class ‘‘Puzzled’’ per test fold. At the same time, training folds include very few ‘‘Puzzled’’ data, which are not enough to train the emotion recognition models. As a consequence, recognition of ‘‘Puzzled’’ class should be really difficult.

B. DIMENSIONAL DISTRIBUTION

Unlike categorical parameters, which are discrete labels, dimensional parameters correspond to continuous values. Figure 6 shows the gravity center of each class into the 2-D model space, for each of the training and each of the test folds, for the three datasets. Regarding training datasets, gravity centers are quite close to each other. This is due to the abundance of data. For the test data, the situation is the opposite. For test folds, Spanish gravity centers are quite overlapping, i.e. classes are mixed. Nevertheless, French and Norwegian classes do not overlap so much. This bias comes from cultural differences and also from differences on the perception of annotators. We also notice that Spanish gravity centers are outside the circle. This reflects very low values of arousal for all classes along with positive values of valence. This is not the case for the other languages.

Regarding the imbalance of the 2-D model parameters, Figure 6 also shows a clear imbalance of arousal and a slightly more balance in valence samples. Also, high activation values as well as negative values of valence do not appear in these datasets. This is related to the nature of the task since high arousal and negative emotions, such as stress or nervous (see Figure 1), are not expected in the EMPATHIC human-machine interaction task previously described. Moreover, only a small part of the dimensional space is occupied by the annotated data.

For more information on the distribution of classes, the euclidean distances between the center of gravity of the majority class, i.e. the Calm category, and those of each of the minority classes, i.e. Happy and Puzzled, are computed and reported in Tables 4 and 5 for training and test folds respectively. These tables show that Calm is mostly closer to Puzzled for French but closer to Happy for Spanish and Norwegian datasets.

TABLE 4. Distances between the gravity centers for pairs (Calm/Happy) and (Calm/Puzzled) for training folds.

Fold	Spain		France		Norway	
	Happy	Puzzled	Happy	Puzzled	Happy	Puzzled
1	0.10	0.45	0.53	0.27	0.46	0.72
2	0.12	0.46	0.52	0.27	0.48	0.81
3	0.12	0.46	0.52	0.28	0.46	0.72
4	0.11	0.45	0.54	0.27	0.48	0.80
5	0.12	0.46	0.53	0.27	0.48	0.72
6	0.10	0.46	0.53	0.28	0.50	0.72
7	0.10	0.47	0.53	0.27	0.46	0.80
8	0.12	0.45	0.55	0.27	0.44	0.72
9	0.12	0.47	0.51	0.28	0.46	0.72
10	0.11	0.45	0.51	0.26	0.46	0.72
Nearest	Happy		Puzzled		Happy	

TABLE 5. Distances between the gravity centers of pairs (Calm/Happy) and (Calm/Puzzled) for test folds.

Fold	Spain		France		Norway	
	Happy	Puzzled	Happy	Puzzled	Happy	Puzzled
1	0.20	0.50	0.55	0.28	0.55	0.97
2	0.16	0.48	0.60	0.30	0.32	0.05
3	0.06	0.41	0.61	0.23	0.58	0.98
4	0.20	0.54	0.46	0.31	0.36	0.02
5	0.12	0.47	0.50	0.33	0.40	0.96
6	0.29	0.43	0.51	0.24	0.19	0.97
7	0.24	0.36	0.53	0.30	0.54	0.01
8	0.10	0.54	0.29	0.29	0.69	0.98
9	0.05	0.36	0.65	0.23	0.57	0.96
10	0.13	0.52	0.72	0.40	0.56	0.97

C. EVALUATION METRICS

Traditionally, the most frequent metrics for classification tasks are Accuracy and Error Rate, also noted as $1-Accuracy$. Additional information can also be obtained from the analysis of the confusion matrix. Furthermore, Precision, Recall and F value provide specific information about quality and sensitivity of the model. By convention the class label of minority class is positive and the class label of the majority is negative. *TP* and *TN* denote the number of positive and negative examples that are classified correctly whereas *FN* and *FP* denote the number of misclassified positive and negative examples respectively [33].

$$Accuracy = \frac{TP+TN}{TP+FN+FP+FN}$$

$$Recall = \frac{TP}{TP+FN} \quad Precision = \frac{TP}{TP+FP}$$

$$F_{value} = (1+\beta^2)\frac{Recall \times Precision}{(\beta^2.Precision)+Recall}$$

β is a coefficient to adjust the relative significance of Precision versus Recall. If $\beta < 1$ then more significance is given to Precision whereas if $\beta > 1$ then more significance is given to Recall. For $\beta = 1$, both are equally significant.

Many representative works of the ineffectiveness of Accuracy in the imbalanced learning scenario can be found in the literature [28], [54]. Indeed, if the dataset is imbalanced, even when the model classifies all the majority examples correctly and misclassifies all the minority examples, the Accuracy of the model is still high because there are much more majority examples than minority examples.

F value is a popular evaluation metric that lessens this effect. It combines Precision and Recall, which are effective metrics for information retrieval community where the imbalance problem exists [33].

Nevertheless, in the context of emotion recognition of imbalanced datasets, average Recall denoted as Unweighted Accuracy (*UA*) is commonly used [38], [40], [45], [55]. It is defined as the average of Recall obtained on each class:

$$UA = \sum_{i=1}^{N} Recall_i/N$$

where $Recall_i$ is the *Recall* value for class i and N is the number of classes. Unlike Accuracy, *UA* gives the same importance to all classes. For this research Unweighted Accuracy as well as F value are used as evaluation metrics. We consider $F2$ score which is F value when $\beta = 2$.

Confidence interval measurement is introduced and commonly used in speech recognition [56], [57] to measure the reliability of the error of recognition rate. This measure can be applied in many classification problems including speech emotion recognition [54]. According to [56], the 'true' classification rate has a probability x of falling in the confidence interval [P+,P−] where P is the measured classification rate. The limits of the confidence interval [P+,P−] are obtained as follows:

$$\frac{P + \frac{z_x^2}{N} \pm z_x\sqrt{\frac{P(1-P)}{N} + \frac{z_x^2}{4N^2}}}{1 + \frac{z_x^2}{N}}$$

where N is the number of tests, $z_{90\%} = 1.64$, $z_{98\%} = 2.33$, etc and ± being + for P+ and − for P−.

In this work, *UA* represents the classification rate P. Confidence levels are calculated for the best results reported in Subsection VII-F through Tables 16, 17, 18, 19 and 20.

VII. EXPERIMENTS AND RESULTS

Inspired by the source-filter model of speech production, researchers have extracted a large number of acoustic parameters to analyse speech content. Some of these acoustic features have subsequently been used to recognize emotions [58]. These are parameters derived from the time domain (e.g. speech rate), frequency domain (e.g. pitch), amplitude domain (e.g. energy) and spectral distribution domain (e.g. relative energy in different frequency bands) [59], [60]. Recently, the raw audio as well as the spectrograms [61], [62] have been supplied to the input of a Convolutional Neural Netwok to extract the acoustic characteristics for emotion recognition purposes [63]. In this research, we use the following acoustic parameters: "zero crossing rate", "energy", "energy entropy", "spectral centroid", "spectral spread", "spectral entropy", "spectral flux", "spectral rolloff", "13 mfcc coefficients", "13 chroma features", "log(F0)" and "Harmonic Noise Ratio (HNR)". These parameters are extracted using the tools parselmouth [64] for F0 and HNR and pyaudioanalysis [65] for the rest. Based on frame-by-frame extraction [36], acoustic features are computed for each 50 milliseconds of signal. Then the average and the standard deviation are computed for all segment features leading to only one vector per speech segment. Hence each sample is represented by a 72 dimensional vector of acoustic information.

The experiments are carried out using a deep neural network predictor implemented on the Theano library [66]. Different hyper-parameters have been tested empirically on a development set (10% of fold0 of training set). As a result of this procedure, the first two layers of the multi-layer perceptron have 50 neurons each and ReLu activation function and the output layer is Softmax. RmsProp algorithm [67] is used for optimization. The number of epochs is set to 100 after some training/dev evaluations to avoid overfeed. Batch normalization is performed with a small set of 16 samples. All experiments were run on a GPU machine (with Nvidia TITAN Xp graphics card). The average time is one hour per experiment. A "seed" function for random values is set to 1 in order to make experiments reproducible.

It is important to mention that the oversampling process concerns only the training data. So, even if the test set is also highly imbalanced, no test data oversampling is performed. For comparison purposes, the same oversampling rate R is adopted in all the experiments, which is set to the ratio of the number of samples for majority/minority classes.

- For Happy class: $R = \#Calm \,/\, \#Happy$
- For Puzzled class: $R = \#Calm \,/\, \#Puzzled$

A. BASELINE EXPERIMENTS

A preliminary set of experiments was carried out without oversampling for the three datasets. The evaluation performance is reported in Table 6 in terms of Precision and Recall per class. In addition, the overall results per dataset are also reported in terms of average of Precision, UA, F2 value and Accuracy scores.

In the absence of oversampling, the model is overfed to the majority class (Calm) and almost all test samples are classified Calm. Hence, Calm scores are relevant. Overall Accuracy, which is more related to majority class identification, is also high. Regarding minority classes, they are rarely recognized and the Recalls are very low. In fact, Precision values for the minority classes are not informative because almost no test samples are assigned to Happy or Puzzled categories. Thus, Precision values are computed as a ratio of two very little values. Nevertheless, the tasks involving recognition of spontaneous emotions frequently need to focus on minority classes.

TABLE 6. Emotion recognition performance without oversampling is showed in terms of Precision and Recall per class and country. The overall results for each country dataset are also reported in terms of the average of Precision, UA, F2 value and Accuracy scores.

	Prec.	Recall		
	Spain			
Calm	92.80	99.90		
Happy	33.33	5e-11		
Puzzled	49.33	3.84		
	avg. Pre.	**UA**	F2	Accuracy
Overall	58.48	**34.54**	37.61	**92.73**
	France			
Calm	95.44	99.99		
Happy	46.66	4.16		
Puzzled	33.33	2e-10		
	avg. Pre.	**UA**	F2	Accuracy
Overall	58.48	**34.72**	37.79	**95.44**
	Norway			
Calm	94.81	99.99		
Happy	39.99	0.62		
Puzzled	33.33	9e-10		
	avg. Pre.	**UA**	F2	Accuracy
Overall	56.04	**33.54**	36.45	**94.81**

B. BORDERLINE SAMPLES DISTRIBUTION

The distribution of borderline samples is computed a priory in order to understand the impact of the proposed perceptual borderline sampling. To this end, we noted the total number of borderline samples. Then, we deduced the proportion of each class (see Figure 7).

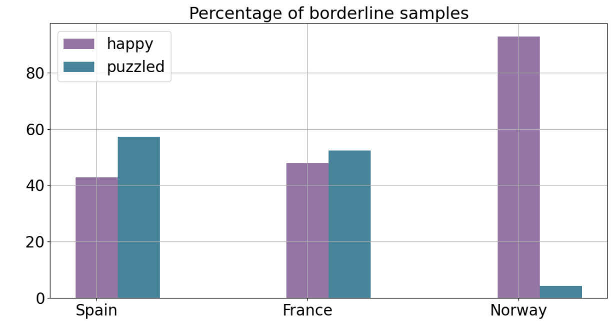


FIGURE 7. Borderline samples percentages for the training folds.

For Spanish and French, there are more Puzzled borderline samples than Happy ones. Nevertheless, Table 4 shows that the centers of gravity of the Spanish folds for Calm are closer to the centers of gravity of Happy folds than to the centers of gravity of Puzzled folds. This can be due to the higher number of Puzzled samples for Spanish. In addition, their distribution could be somewhat flat because of the coarse annotation.

C. PERCEPTUAL BORDERLINE (PB)

Random oversampling (RO) and borderline perceptual oversampling (PBO) experiments are performed in order to increase the size of the minority class training data. Both are carried out under the same conditions, as follows:

- RO algorithm: all the samples of minority classes are replicated R times.
- PBO algorithm described in Subsection IV-A deals with a pair of (minority/majority) classes. It is adapted to the three classes context as follows:
  - Happy class: borderline samples are replicated $R1 = \alpha * R$ and other samples $R2 = (2\text{-}\alpha) * R$
  - Puzzled class: borderline samples are replicated $R1 = \beta * R$ and other samples $R2 = (2\text{-}\beta) * R$

Note that RO is equivalent to PBO for $\alpha = \beta = 1$

1) SPANISH RESULTS

Unweighted Accuracy (UA) is reported in Table 7 for α and β values ranking from 0.25 up to 1.75.

TABLE 7. Unweighted Accuracy (UA) of PBO Spanish emotion recognition experiments for different relations between α and β. $\alpha = \beta = 1$ corresponds to the RO method.

β	α						
	0.25	0.50	0.75	1.00	1.25	1.50	1.75
0.25	54.90	57.13	55.99	55.57	54.99	55.35	54.23
0.50	57.66	57.45	56.08	56.89	56.48	57.49	55.78
0.75	56.91	57.59	**58.69**	56.63	55.65	56.69	56.33
1.00	58.65	56.27	57.03	**57.09**	57.12	56.08	55.65
1.25	57.26	56.68	55.72	57.62	56.61	56.87	55.84
1.50	55.96	56.98	57.34	58.62	55.45	56.36	55.73
1.75	56.25	53.98	56.80	55.85	55.78	56.15	55.80

We notice that for each pair of (α, β), UA is higher when α and β are close to each other and $\beta \geq \alpha$. Indeed, according to Figure 7, there are more Puzzled samples at the borderline than Happy samples. The best PBO performance (58.69%) is reached for $\alpha = \beta = 0.75$. This corresponds to an absolute gain of about 1.6% compared to RO for which UA=57.09% ($\alpha = \beta = 1$).

Table also shows a couple of exceptions when $\alpha \geq$1.5 and $\beta \leq 0.75$ and the opposite. To explain this phenomena lets examine an example.

$\alpha = 1.75$, $\beta = 0.5$

- UA Calm=50.62%
- UA Happy=60.58%
- UA Puzzled=56.15%
- UA all=55.78%

$\alpha = 0.5$, $\beta = 1.75$

- UA Calm=55.20%
- UA Happy=50.58%
- UA Puzzled=56.05%
- UA all=53.58%

Increasing too much α and decreasing β improves Happy performance and degrades the Calm one. As Calm is closer to Happy than to Puzzled (Table 4), it seems that the model is more overfed to Happy than to Calm.

TABLE 8. Unweighted Accuracy (UA) of PBO French emotion recognition experiments for different relations between α and β. $\alpha = \beta = 1$ corresponds to the RO method.

β	α						
	0.25	0.50	0.75	1.00	1.25	1.50	1.75
0.25	61.05	59.81	60.46	59.04	57.83	57.67	58.98
0.50	60.64	60.56	60.51	58.24	58.33	58.24	59.93
0.75	60.67	60.06	59.34	60.02	58.51	60.07	59.43
1.00	**62.41**	62.23	60.72	**59.33**	59.79	60.19	58.39
1.25	60.94	59.64	59.54	60.79	58.36	55.71	58.97
1.50	58.79	59.14	59.66	60.33	60.12	59.12	57.58
1.75	57.50	58.75	59.83	59.62	60.69	59.92	58.63

2) FRENCH RESULTS

Unweighted Accuracy of French emotions experiments are reported in Table 8.

Table 4 shows that the centers of gravity of the French folds for Calm are closer to the centers of gravity of Puzzled folds than to the centers of gravity of Happy folds. In addition, Figure 7 shows a higher number of borderline Puzzled samples than Happy samples. Hence $\alpha \leq \beta$. Indeed, UA is mostly higher under the diagonal. The best PBO performance is obtained for $\alpha = 0.25$ and $\beta = 1.0$. This corresponds to an absolute gain of about 3% compared to RO.

Only two exceptions are noticed when $\alpha \gg \beta$ or $\beta \ll \alpha$.

3) NORWEGIAN RESULTS

Unweighted Accuracy of Norwegian emotions experiments are reported in Table 9.

TABLE 9. Unweighted Accuracy (UA) of PBO Norwegian emotion recognition experiments for different relations between α and β. $\alpha = \beta = 1$ corresponds to the RO method.

β	α						
	0.25	0.50	0.75	1.00	1.25	1.50	1.75
0.25	51.20	52.13	51.86	52.84	52.12	51.98	**55.07**
0.50	50.43	48.75	49.11	52.47	51.08	48.75	47.90
0.75	48.70	49.80	51.38	50.01	52.03	49.38	48.74
1.00	48.53	49.74	47.69	**49.01**	48.55	51.36	52.20
1.25	45.48	49.99	48.44	46.26	47.85	48.76	48.82
1.50	48.65	49.02	50.29	45.18	51.74	48.18	51.76
1.75	45.36	47.88	50.69	49.02	50.59	50.95	50.68

According to Table 4, the centers of gravity of the Norwegian folds for Calm are closer to the centers of gravity of Happy folds than to the centers of gravity of Puzzled folds, as for Spanish experiments. In addition, Figure 7 shows a considerably higher number of borderline Happy samples than Puzzled samples. Hence $\alpha \geq \beta$. Indeed, UA is mostly higher above the diagonal.

The test folds contain only one Puzzled sample. Moreover, training set contains a very limited number of samples (12) which justifies low scores and more exceptions than in the other datasets (mainly when $\beta \geq 1.25$). The best PBO performance is obtained for $\alpha = 1.75$ and $\beta = 0.25$. This corresponds to an absolute gain of about 6% compared to RO.

D. PERCEPTUAL BORDERLINE SMOTE (PB-SMOTE)

We now apply the perceptual borderline algorithm in Subsection IV-B. This algorithm is similar to PBO algorithm in Subsection IV-A, which was applied in experiments of previous section, except that now new samples are artificially generated instead of being generated by replication. SMOTE algorithm is the state of art oversampling method by artificial synthesis. The number of neighbours is set to 5 for these experiments. SMOTE is equivalent to Borderline Perceptual SMOTE (PB-SMOTE) when $\alpha = \beta = 1$

1) SPANISH RESULTS

Unweighted Accuracy (UA) is reported in Table 10 for α and β values ranking from 0.25 up to 1.75.

TABLE 10. Unweighted Accuracy (UA) of PB-SMOTE Spanish emotion recognition experiments for different relations between α and β. $\alpha = \beta = 1$ corresponds to the SMOTE algorithm.

β	α						
	0.25	0.50	0.75	1.00	1.25	1.50	1.75
0.25	55.89	55.98	56.33	55.40	55.88	56.48	55.32
0.50	56.42	56.37	57.27	56.25	56.25	56.75	55.87
0.75	56.00	56.18	55.45	55.78	56.39	55.58	53.65
1.00	55.85	56.47	57.08	**56.33**	55.75	55.39	56.81
1.25	55.45	55.41	56.89	55.64	56.75	56.04	56.21
1.50	55.16	55.71	57.50	55.67	55.38	57.55	54.83
1.75	55.67	55.83	56.02	**57.76**	55.52	55.17	55.42

Table 10 shows that UA values are generally higher when $\beta \geq \alpha$, as in the case of the PBO algorithm. However, there are more exceptions, when α, $\beta > 1$. The best PB-SMOTE performance is obtained for $\alpha = 1.0$ and $\beta = 1.75$. This corresponds to an absolute gain of about 1.4% compared to SMOTE.

2) FRENCH RESULTS

Unweighted Accuracies (UA) obtained through French emotion recognition experiments are reported in Table 11 for α and β values ranking from 0.25 up to 1.75.

TABLE 11. Unweighted Accuracy (UA) of PB-SMOTE French emotion recognition experiments for different relations between α and β. $\alpha = \beta = 1$ corresponds to the SMOTE algorithm.

β	α						
	0.25	0.50	0.75	1.00	1.25	1.50	1.75
0.25	60.06	58.69	59.12	57.98	59.84	58.65	57.87
0.50	59.56	57.07	60.09	57.81	60.81	59.84	60.63
0.75	**62.38**	60.71	60.37	58.84	62.37	58.79	59.63
1.0	59.70	58.28	61.47	**59.29**	58.95	58.90	58.34
1.25	60.72	60.90	58.52	59.17	58.40	57.72	58.25
1.50	57.62	59.87	56.68	60.10	60.36	60.39	57.84
1.75	58.02	58.00	60.36	58.17	59.18	57.82	56.41

Table 11 shows that UA is generally higher when $\beta \geq \alpha$, as in the case of the PBO algorithm. There are some exceptions when $\alpha \geq$1.5 and/or $\beta \geq$1.5. This Table also shows that high oversampling ratios for both classes, Happy and Puzzled, result also in a high number of synthetic samples generated in the same area. This could lead to the overlapping of the synthetic samples.

The best PB-SMOTE performance is obtained for $\alpha = 0.25$ and $\beta = 1.0$. This corresponds to an absolute gain of more than 3% compared to SMOTE.

3) NORWEGIAN RESULTS

For highly imbalanced datasets, the minority class is often poorly represented and lacks a clear structure. Hence, methods that rely on relations between minority objects (like SMOTE) tend to fail [26]. In this case, the imbalance rate for Puzzled class is greater than 220, which is very high. Lets now see the impact of perceptual borderline on SMOTE performance in such a case. To this end, Unweighted Accuracy is reported in Table 12.

TABLE 12. Unweighted Accuracy (UA) of PB-SMOTE Norwegian emotion recognition experiments for different relations between α and β. $\alpha = \beta = 1$ corresponds to the SMOTE method.

β	α						
	0.25	0.50	0.75	1.00	1.25	1.50	1.75
0.25	52.93	52.80	**53.67**	52.02	52.45	51.91	50.54
0.50	52.38	52.48	51.46	48.22	50.57	51.48	48.70
0.75	50.42	48.95	48.68	51.78	47.62	44.75	48.07
1.0	49.51	49.68	48.73	**47.53**	49.53	48.33	47.76
1.25	49.74	49.08	50.76	47.76	46.95	47.46	44.80
1.50	50.08	50.53	52.32	48.33	45.44	47.44	48.48
1.75	48.14	47.38	51.89	51.23	46.00	44.44	45.88

UA is generally higher above the diagonal. When $\alpha \geq 1.25$ and $\beta \geq 1.25$, UA is often lower. In this case, we are generating more samples in the borderline area than in the rest of the acoustic space. The UA degradation could be due to a number of new artificial samples of Puzzled class that overlapping the new Happy ones.

E. STRETCHY SMOTE (S-SMOTE)

S-SMOTE algorithm (See Subsection IV-C) fixes the oversampling ratio R as the ratio between majority and minority classes, which is the same for all samples. As a consequence, the borderline is not used to select samples for which a different ratio has to be applied as in PBO (Subsection IV-A) and PB-SMOTE (Subsection IV-B) algorithms.

When new samples are generated by the SMOTE technique, a fixed number of neighbours N has to be considered. We want now to verify the hypothesis that borderline samples should have less neighbours than the others, as the S-SMOTE algorithm proposes. To this end, we denote by $N1$ the number of neighbours of the borderline samples and by $N2$ the number of neighbours to be considered for other samples.

When $N1 = N2$, Stretchy SMOTE is equivalent to SMOTE.

In order to compare S-SMOTE to SMOTE, the same total number of neighbours is considered. To this end,

TABLE 13. Unweighted Accuracy (UA) for Spanish emotion recognition using SMOTE and S-SMOTE algorithms. Each pair of columns represents the N1 and N2 values and the UA obtained for the particular experiment.

SMOTE		S-SMOTE											
N1 N2	UA	N1 N2	UA										
3 3	55.34	1 5	56.00	2 4	56.16								
		5 1	55.52	4 2	56.02								
4 4	56.50	1 7	56.08	2 6	55.69	3 5	55.13						
		7 1	55.65	6 2	55.34	5 3	56.98						
5 5	56.33	1 9	54.75	2 8	54.93	3 7	55.16	4 6	56.44				
		9 1	57.09	8 2	54.67	7 3	55.02	6 4	54.59				
6 6	**56.71**	1 11	55.13	2 10	56.22	3 9	56.65	4 8	**57.38**	5 7	55.64		
		11 1	56.09	10 2	57.18	9 3	54.73	8 4	56.75	7 5	56.45		
7 7	54.69	1 13	56.32	2 12	56.68	3 11	56.51	4 10	55.59	5 9	54.93	6 8	56.61
		13 1	56.14	12 2	55.81	11 3	56.55	10 4	57.34	9 5	54.50	8 6	54.82

TABLE 14. Unweighted Accuracy (UA) for French emotion recognition using SMOTE and S-SMOTE algorithms. Each pair of columns represents the N1 and N2 values and the UA obtained for the particular experiment.

SMOTE		S-SMOTE											
N1 N2	UA	N1 N2	UA										
3 3	59.01	1 5	59.86	2 4	60.59								
		5 1	57.08	4 2	58.52								
4 4	59.01	1 7	60.00	2 6	60.94	3 5	60.28						
		7 1	60.18	6 2	57.36	5 3	57.24						
5 5	59.33	1 9	57.97	2 8	59.44	3 7	**62.15**	4 6	59.53				
		9 1	55.05	8 2	57.58	7 3	59.47	6 4	58.11				
6 6	58.40	1 11	58.42	2 10	58.28	3 9	61.86	4 8	60.80	5 7	59.26		
		11 1	56.55	10 2	60.74	9 3	59.53	8 4	56.91	7 5	58.94		
7 7	**60.43**	1 13	60.11	2 12	57.59	3 11	59.74	4 10	58.94	5 9	59.49	6 8	59.15
		13 1	57.73	12 2	57.49	11 3	59.51	10 4	59.38	9 5	56.42	8 6	57.46

TABLE 15. Unweighted Accuracy (UA) for Norwegian emotion recognition using SMOTE and S-SMOTE algorithms. Each pair of columns represents the N1 and N2 values and the UA obtained for the particular experiment.

SMOTE		S-SMOTE											
N1 N2	UA	N1 N2	UA										
3 3	**49.19**	1 5	48.45	2 4	48.10								
		5 1	46.20	4 2	44.78								
4 4	48.46	1 7	48.05	2 6	47.49	3 5	48.98						
		7 1	48.49	6 2	47.87	5 3	48.03						
5 5	48.74	1 9	44.01	2 8	44.84	3 7	48.90	4 6	48.67				
		9 1	44.61	8 2	47.16	7 3	48.43	6 4	48.64				
6 6	48.52	1 11	48.79	2 10	48.57	3 9	**49.92**	4 8	48.45	5 7	48.79		
		11 1	49.51	10 2	48.42	9 3	45.25	8 4	47.54	7 5	45.22		
7 7	48.13	1 13	47.76	2 12	43.96	3 11	48.47	4 10	46.75	5 9	49.14	6 8	48.45
		13 1	49.03	12 2	48.37	11 3	49.04	10 4	45.12	9 5	48.11	8 6	47.70

$N1$ and $N2$ are set to N and then the SMOTE algorithm is applied. For the application of the S-SMOTE algorithm we sweep the values of $N1$ and $N2$ but always keeping the relationship $N1+N2=2N$. For each pair $(N1, N2)$, the Unweighted Accuracy is shown for the three languages in Tables 13, 14 and 15 respectively.

1) SPANISH RESULTS

For Spanish data we obtained:

- if $N1 \leq N2$ and the difference $N1 - N2$ is small then generally $UA(N1/N2) \geq UA(N2/N1)$.

- if $N1 \geq N2$ then generally $UA(N1/N2) \leq UA(N2/N1)$

As mentioned in Figure 6, classes are close to each others. Therefore, when the number of neighbours is high, a new synthetic sample of one class may overlap another sample of a different class. Best performance of S-SMOTE corresponds to UA $= 57.36\%$ and is reached for $N1 = 4$ and $N2 = 8$. It corresponds to an absolute gain of 0.7% compared to SMOTE best result which is UA=56.71%.

2) FRENCH RESULTS

For the French dataset we got that if $N1 < N2$ then $UA(N1/N2) \geq UA(N2/N1)$. According to Figure 6, French classes are more distant from each other than the Spanish classes. So they are less confused when the number of neighbours is high.

S-SMOTE best performance (UA=62.16%) is obtained for $N1 = 3$ and $N2 = 7$. This corresponds to an absolute gain of about 2% compared to SMOTE best performance (UA=60.43%), which is achieved for $N1 = N2 = 7$.

3) NORWEGIAN RESULTS

We can draw roughly the same conclusions for the Norwegian data that we got from the Spanish ones. However, the scores are quite lower.

S-SMOTE best performance corresponds to UA=49.92% for $N1 = 4$ and $N2 = 8$, while SMOTE higher score is UA=49.19% for $N1 = N2 = 3$. The absolute improvement is about 0.7%.

F. DISCUSSION OF RESULTS

Figure 5 shows that some emotions can be only found in a reduced set of speakers. As a consequence, a classical experimental protocol such as *one speaker left out* would lead to the lack of certain emotions in the training or in the test set. In addition, the distribution of data between training and test folds could be different. Hence, we opted for a protocol where all emotions are present in all the training/test folds as described in Section VI, which it is now named Protocol 1. However, Protocol 1 results in very low scores of the Norwegian system (Subsections VII-C, VII-D and VII-E). Indeed, the cardinal of the test set is always equal to 1 for Puzzled class. In this Subsection we analyze the best results obtained

TABLE 16. Comparison between RO and PBO emotion recognition results for **Protocol 1**: Precision and Recall values are reported per class and country. Then, the overall results per country in terms of average of Precision, Unweighted Accuracy (UA) and F2 values are also showed.

	RO			PBO		
	Pre.	Rec.		Pre.	Rec.	
	Spain					
Calm	96.32	52.84		96.55	57.28	
Happy	7.52	58.82		7.97	57.64	
Puzzled	10.29	59.611		11.61	61.15	
	avg. Pre.	**UA**	F2	avg. Pre.	**UA**	F2
Overall	38.04	**57.09**	51.82	38.04	**58.69**	53.16
	France					
Calm	98.10	60.05		98.39	62.64	
Happy	6.19	52.22		7.60	53.33	
Puzzled	6.21	65.71		6.12	71.42	
	avg. Pre.	**UA**	F2	avg. Pre.	**UA**	F2
Overall	36.84	**59.33**	52.79	37.37	**62.41**	54.97
	Norway					
Calm	97.23	73.23		97.65	62.10	
Happy	12.22	63.74		10.63	73.12	
Puzzled	0.07	10.00		1.50	30.00	
	avg. Pre.	**UA**	F2	avg. Pre.	**UA**	F2
Overall	36.72	**49.00**	45.60	36.59	**55.07**	49.46

TABLE 17. Comparison between RO and PBO emotion recognition results for **Protocol 2**: Precision and Recall values are reported per class and country. Then, the overall results per country in terms of average of Precision, Unweighted Accuracy (UA) and F2 values are also showed.

	RO			PBO		
	Pre.	Rec.	F2	Pre.	Rec.	F2
	Spain					
Calm	96.38	51.36		95.98	58.00	
Happy	7.0	50.89		8.17	47.16	
Puzzled	10.0	53.83		10.69	54.54	
	avg. Pre.	**UA**	F2	avg. Pre.	**UA**	F2
Overall	37.79	**52.03**	48.29	38.28	**53.90**	49.79
	France					
Calm	97.84	57.90		97.56	58.14	
Happy	5.49	43.44		5.94	57.44	
Puzzled	6.66	75.60		6.70	71.18	
	avg. Pre.	**UA**	F2	avg. Pre.	**UA**	F2
Overall	36.66	**58.98**	52.10	36.73	**62.25**	54.23
	Norway					
Calm	96.72	64.83		96.60	59.38	
Happy	10.41	66.84		9.49	67.74	
Puzzled	65.0	61.42		61.29	71.42	
	avg. Pre.	**UA**	F2	avg. Pre.	**UA**	F2
Overall	57.37	**64.36**	62.69	55.79	**66.18**	63.52

TABLE 18. Comparison between SMOTE and PB-SMOTE emotion recognition results for **Protocol 1**: Precision and Recall values are reported per class and country. Then, the overall results per country in terms of average of Precision, Unweighted Accuracy (UA) and F2 values are also showed.

	SMOTE			PB-SMOTE		
	Pre.	Rec.	F2	Pre.	Rec.	F2
	Spain					
Calm	96.40	51.10		96.27	49.72	
Happy	6.89	50.58		7.65	55.88	
Puzzled	10.07	67.30		9.85	67.69	
	avg. Pre.	**UA**	F2	avg. Pre.	**UA**	F2
Overall	37.79	**56.33**	51.29	37.92	**57.76**	52.28
	France					
Calm	98.19	66.65		98.04	69.30	
Happy	7.42	51.11		8.16	57.77	
Puzzled	6.39	59.99		8.29	60.00	
	avg. Pre.	**UA**	F2	avg. Pre.	**UA**	F2
Overall	37.27	**59.25**	52.89	38.16	**62.36**	55.30
	Norway					
Calm	97.14	68.22		97.29	83.52	
Happy	10.25	64.37		19.65	57.49	
Puzzled	1.11	10.00		1.29	20.00	
	avg. Pre.	**UA**	F2	avg. Pre.	**UA**	F2
Overall	36.16	**47.53**	44.11	39.41	**53.67**	49.69

per class in the framework of Protocol 1, in terms of Precision, Recall and F2. For comparison purposes, we repeat the experiments trough a classical *one speaker left out* protocol that we call Protocol 2. The number of tests carried out through these protocols is: 5,850 Spanish tests, 3,750 French tests and 3,260 Norwegian tests for Protocol 1, and 5,931 Spanish tests, 3,553 French tests and 3,216 Norwegian tests for Protocol 2.

1) PERCEPTUAL BORDERLINE OVERSAMPLING

Tables 16 and 17 report emotion recognition Precision (Pre.) and Recall (Rec.) per class for Random (RO) and Perceptual Borderline Oversampling (PBO) when Protocol 1 and Protocol 2 are used, respectively. In addition, the overall results per dataset are also reported in terms of average of Precision scores, UA and F2 value. Confidence intervals are calculated for UA as described in Subsection VI-C, resulting in a different confidence level for each protocol and dataset.

The best performances are obtained for the French dataset, which are similar for both protocols. In fact, in this case, the three classes are more distant each other (see Figure 6) and the confidence level is 92% for both protocols. The Spanish dataset is less imbalanced (Table 3) than the French and Norwegian datasets but classes overlap more. Protocol 1 provides better performance than Protocol 2, but the confidence level is lower than for French, namely 80% for Protocol 1 and 85% for Protocol 2. Norwegian results vary depending on the protocol. This dataset is extremely imbalanced and contains very few samples of Puzzled class (see Table 3), which results in a very difficult task and, consequently, lower results are expected. The confidence level for Protocol 1 is 98%. Even if the UA is low the system classifies well two classes, but not three because of the lack of samples of one class. On the contrary, the UA for Protocol 2 is higher because the absence of Puzzled samples in many test folds, which explains the good results. The confidence level of 80%.

Let us now compare the results in Table 6 obtained without oversampling with the results in Tables 16 and 17, in terms of Recall per class and UA. We notice that Recall decreases for the majority class and increases for minority ones resulting in a significant improvement of the UA scores, when oversampling is applied.

Finally, we highlight that the proposed approach (PBO) improves the classification performance in both protocols.

2) PERCEPTUAL BORDERLINE SMOTE

Tables 18 and 19 report emotion recognition Precision (Pre.) and Recall (Rec.) per class for SMOTE and Perceptual

TABLE 19. Comparison between SMOTE and PB-SMOTE emotion recognition results for **Protocol 2**: Precision and Recall values are reported per class and country. Then the overall results per country in terms of average of Precision, Unweighted Accuracy (UA) and F2 values are also showed.

	SMOTE			**PB-SMOTE**		
	Pre.	Rec.	F2	Pre.	Rec.	F2
	Spain					
Calm	96.05	48.78		96.02	52.13	
Happy	6.41	43.68		7.25	49.72	
Puzzled	8.38	60.17		9.05	58.22	
	avg. Pre.	**UA**	F2	avg. Pre.	**UA**	F2
Overall	36.95	**50.88**	47.17	37.44	**53.35**	49.08
	France					
Calm	96.99	59.82		97.51	60.16	
Happy	4.60	39.16		6.92	65.16	
Puzzled	6.98	57.45		6.02	66.25	
	avg. Pre.	**UA**	F2	avg. Pre.	**UA**	F2
Overall	36.19	**52.14**	47.54	36.82	**63.86**	55.15
	Norway					
Calm	96.75	62.72		96.67	61.91	
Happy	9.71	64.47		9.35	63.67	
Puzzled	62.00	61.42		65.58	71.42	
	avg. Pre.	**UA**	F2	avg. Pre.	**UA**	F2
Overall	56.15	**62.87**	61.34	57.20	**65.67**	63.39

TABLE 20. Comparison between SMOTE and S-SMOTE emotion recognition results for **Protocol 1**: Precision and Recall values are reported per class and country. Then, the overall results per country in terms of average of Precision, Unweighted Accuracy (UA) and F2 values are also showed.

	SMOTE			**S-SMOTE**		
	Pre.	Rec.	F2	Pre.	Rec.	F2
	Spain					
Calm	96.36	49.61		96.48	49.22	
Happy	7.01	58.23		7.33	52.94	
Puzzled	10.0	62.30		9.9	69.99	
	avg. Pre.	**UA**	F2	avg. Pre.	**UA**	F2
Overall	37.78	**56.71**	51.51	37.90	**57.38**	52.01
	France					
Calm	98.13	62.72		98.28	60.58	
Happy	6.91	59. 9		6.75	64.44	
Puzzled	6.24	58.57		6.76	61.42	
	avg. Pre.	**UA**	F2	avg. Pre.	**UA**	F2
Overall	37.09	**60.43**	53.63	37.26	**62.15**	54.63
	Norway					
Calm	97.31	70.71		97.40	69.80	
Happy	11.44	66.87		1.0	10.0	
Puzzled	0.83	10.00		36.71	49.51	
	avg. Pre.	**UA**	F2	avg. Pre.	**UA**	F2
Overall	36.53	**49.19**	45.74	36.71	**49.51**	46.01

Borderline SMOTE (PB-SMOTE) when Protocol 1 and Protocol 2 are used, respectively. In addition, the overall results per dataset are also reported in terms of average of Precision scores, UA and F2 value.

In general terms, we can draw similar conclusions to those from the previous tables. Moreover, we can conclude that PB-SMOTE performs better than SMOTE in terms of UA with confidence levels 80% and 92% for Spanish data, 92% and 98% for French data, and 98% and 90% for Norwegian data. In addition, we notice that SMOTE performs slightly worse than RO for all languages and protocols. Furthermore, PBO is generally the most efficient except for French and Protocol 2.

3) STRETCHY SMOTE

Table 20 shows the best emotion recognition scores for SMOTE and Stretchy SMOTE (S-SMOTE) and Protocol1.

This Table also shows that overall performances obtained by S-SMOTE are close to those obtained by PB-SMOTE, resulting in confidence levels arround 80%. Notice that in S-SMOTE, α and β are fixed to 1 and thus only the number of neighbours number varies. As a consequence, borderline samples are over-sampled at the same ratio than the other samples. As a result, SMOTE blind interpolation issue rises and the new synthetic samples of one class could overlap some examples of other classes.

VIII. CONCLUDING REMARKS

This research addresses the understanding of human behaviour through the analysis of the speech signal in the context of human-machine interaction. More specifically, we focus on the speech-based identification of the seniors's emotional status during their interaction with a virtual agent playing the role of a health professional coach. In this convincing scenario, only a very reduced set of spontaneous emotions was identified in human perception experiments, which shows a huge difference among the number of samples resulting in an imbalanced dataset problem.

To deal with this issue, several balancing data algorithms are proposed in the context of binary classification. Nevertheless, most of these methods are inefficient to deal with a multi-class classification problem. Indeed, we may loose performance on one class while trying to gain it on another. Consequently, there is a need for new methods that consider the specific characteristics of the data and its nature. Some examples of new information to be additionally considered are the distribution of classes or their boundaries.

Emotions are often represented in a continuous dimensional space. So our contribution is to take advantage of this perceptual, visual and easy to interpret dimensional space to examine classes borderlines in this space. This limit, based on the perceived arousal and valence, leads to two methods of balancing the data: the Perceptual Borderline Oversampling where the over-sampling is carried out by replication and the Perceptual Borderline SMOTE that generates artificial samples. In this case, the proposed methods avoid the overlap of the samples of the different classes.

The additional contribution of this work, denoted by Stretchy SMOTE, is an alternative to Perceptual Borderline SMOTE, where the same oversampling rate is applied to the data of the same class but the number of neighbors is variable. This number depends on the distance between the minority class samples and the majority class, which is also calculated in the perceptual emotion plane.

These proposals are implemented and compared to state-of-the-art approaches, namely Random Oversampling and Synthetic Minority Oversampling. The experimental evaluation was carried out on three extremely imbalanced datasets of spontaneous emotions acquired in human-machine scenarios in three different cultures: Spain, France and Norway. The emotion recognition results obtained by neural networks classifiers show that the proposed perceptual oversampling methods lead to significant improvements when compared to the state-of-the art, for all scenarios and languages.

REFERENCES

[1] R. Chakroun, M. Frikha, and L. B. Zouari, "New approach for short utterance speaker identification," *IET Signal Process.*, vol. 12, no. 7, pp. 873–880, Sep. 2018.

[2] R. Chakroun, L. Beltaïfa, and M. Frikha, "An improved approach for text-independent speaker recognition," *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 8, pp. 343–348, 2016.

[3] A. López-Zorrilla, N. Dugan, M. Torres, C. Glackin, G. Chollet, and N. Cannings, "Some ASR experiments using deep neural networks on Spanish databases," in *Proc. IberSpeech*, Lisbon, Portugal, 2016, pp. 149–158.

[4] L. Zouari and G. Chollet, "Speech transcription for embodied conversational agent animation," *J. Acoust. Soc. Amer.*, vol. 123, no. 5, p. 3886, May 2008.

[5] V. G. Guijarrubia and M. I. Torres, "Text- and speech-based phonotactic models for spoken language identification of basque and Spanish," *Pattern Recognit. Lett.*, vol. 31, no. 6, pp. 523–532, Apr. 2010.

[6] S. Khorram, M. Jaiswal, J. Gideon, M. McInnis, and E. Provost, "The priori emotion dataset: Linking mood to emotion detected in-the-wild," in *Proc. Interspeech*, 2018, pp. 1903–1907.

[7] J. Pereira and Ó. Díaz, "Using health chatbots for behavior change: A mapping study," *J. Med. Syst.*, vol. 43, no. 5, pp. 109–127, May 2019.

[8] A. Esposito, T. Amorese, M. Cuciniello, A. Troncone, M. Torres, S. Schlögl, and G. Cordasco, "Seniors' acceptance of virtual humanoid agents," in *Ambient Assisted Living* (Lecture Notes in Electrical Engineering), vol. 544, A. Leone, A. Caroppo, G. Rescio, G. Diraco, and P. Siciliano, Eds. Cham, Switzerland: Springer, 2019, pp. 429–443.

[9] M. I. Torres *et al.*, "The empathic project: Mid-term achievements," in *Proc. 12th ACM Int. Conf. Pervasive Technol. Related Assistive Environ.* New York, NY, USA: Association for Computing Machinery, 2019, pp. 629–638.

[10] R. Justo, L. B. Letaifa, C. Palmero, E. Gonzalez-Fraile, A. T. Johansen, A. Vázquez, G. Cordasco, S. Schlögl, B. Fernández-Ruanova, M. Silva, S. Escalera, M. deVelasco, J. Tenorio-Laranga, A. Esposito, M. Korsnes, and M. I. Torres, "Analysis of the interaction between elderly people and a simulated virtual coach," *J. Ambient Intell. Humanized Comput.*, vol. 11, no. 12, pp. 6125–6140, Dec. 2020.

[11] S. Khanal, A. Reis, J. Barroso, and V. Filipe, "Using emotion recognition in intelligent interface design for elderly care," in *Trends and Advances in Information Systems and Technologies*, Á. Rocha, H. Adeli, L. P. Reis, and S. Costanzo, Eds. Cham, Switzerland: Springer, 2018, pp. 240–247.

[12] R. J. Davidson and P. A. Ekman, *Nature of Emotion: Fundamental Questions*. (Oxford University Press). New York, NY, USA: Springer, 1994.

[13] H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition," *Int. J. Synth. Emotions*, vol. 1, no. 1, pp. 68–99, Jan. 2010.

[14] T. Vogt and E. Andre, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition," in *Proc. IEEE Int. Conf. Multimedia Expo*, Amsterdam, The Netherlands, Jul. 2005, pp. 474–477.

[15] K. R. Scherer, "On the nature and function of emotion: A component process approach," in *Approaches to Emotion*, K. Scherer and P. Ekman, Ed. New York, NY, USA: Taylor & Francis, 1984, pp. 293–318.

[16] M. de Velasco Vazquez, R. Justo, A. L. Zorrilla, and M. I. Torres, "Can spontaneous emotions be detected from speech on TV political debates?" in *Proc. 10th IEEE Int. Conf. Cognit. Infocommunications (CogInfoCom)*, Naples, Italy, Oct. 2019, pp. 289–294.

[17] R. Lotfian and C. Busso, "Predicting categorical emotions by jointly learning primary and secondary emotions through multitask learning," in *Proc. Interspeech*, 2018, pp. 951–955.

[18] R. L. B. Letaifa, M. de Velasco, and M. Torres, "First steps to develop a corpus of interactions between elderly and virtual agents in Spanish with emotion labels," in *Proc. 7th Int. Conf. Stat. Lang. Speech Process.*, Slovenia, Balkans, 2019.

[19] L. B. Letaifa, M. I. Torres, and R. Justo, "Adding dimensional features for emotion recognition on speech," in *Proc. 5th Int. Conf. Adv. Technol. Signal Image Process. (ATSIP)*, Sousse, Tunisia, Sep. 2020, pp. 1–6.

[20] C. Beedie, P. Terry, and A. Lane, "Distinctions between emotion and mood," *Cognition Emotion*, vol. 19, no. 6, pp. 847–878, Sep. 2005, doi: 10.1080/02699930541000057.

[21] P. Sarakit, T. Theeramunkong, and C. Haruechaiyasak, "Improving emotion classification in imbalanced YouTube dataset using SMOTE algorithm," in *Proc. 2nd Int. Conf. Adv. Inform., Concepts, Theory Appl. (ICAICTA)*, Aug. 2015, pp. 1–5.

[22] S. Vluymans, A. Fernández, Y. Saeys, C. Cornelis, and F. Herrera, "Dynamic affinity-based classification of multi-class imbalanced data with one-versus-one decomposition: A fuzzy rough set approach," *Knowl. Inf. Syst.*, vol. 56, pp. 55–84, Oct. 2018.

[23] P. Kaur and A. Gosain, "Comparing the behavior of oversampling and undersampling approach of class imbalance learning by combining class imbalance problem with noise," in *ICT Based Innovations*, A. K. Saini, A. K. Nayak, and R. K. Vyas, Eds. Singapore: Springer, 2018, pp. 23–30.

[24] J. Bi and C. Zhang, "An empirical comparison on state-of-the-art multi-class imbalance learning algorithms and a new diversified ensemble learning scheme," *Knowl.-Based Syst.*, vol. 158, pp. 81–93, Oct. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S095070511830282X

[25] A. Fernández, C. J. Carmona, M. J. Del Jesus, and F. Herrera, "A Pareto-based ensemble with feature and instance selection for learning from multi-class imbalanced datasets," *Int. J. Neural Syst.*, vol. 27, no. 6, Sep. 2017, Art. no. 1750028.

[26] B. Krawczyk, "Learning from imbalanced data: Open challenges and future directions," *Prog. Artif. Intell.*, vol. 5, no. 4, pp. 221–232, Nov. 2016.

[27] H. M. Nguyen, E. W. Cooper, and K. Kamei, "Borderline over-sampling for imbalanced data classification," in *Proc. Int. Workshop Comput. Intell. Appl.*, 2009, pp. 24–29.

[28] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[29] M. V. Joshi and V. Kumar, "Credos: Classification using ripple down structure a case for rare classes," in *Proc. 4th SIAM Int. Conf. Data Mining*, 2004, pp. 321–332.

[30] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *J. Big Data*, vol. 6, no. 1, pp. 1–54, Dec. 2019.

[31] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[32] W. Xie, G. Liang, Z. Dong, B. Tan, and B. Zhang, "An improved oversampling algorithm based on the samples' selection strategy," *Hindawi Math. Problems Eng. Classifying Imbalanced Data*, vol. 2019, p. 13, 2019, Art. no. 3526539, doi: 10.1155/2019/3526539.

[33] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-smote: A new oversampling method in imbalanced data sets learning," in *Proc. Int. Conf. Intell. Comput. (ICIC)*, 2005, pp. 878–887.

[34] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Proc. IEEE Int. Joint Conf. Neural Netw., IEEE World Congr. Comput. Intell.*, Jun. 2008, pp. 1322–1328.

[35] C. Huang, Y. Li, C. L. Chen, and X. Tang, "Deep imbalanced learning for face recognition and attribute prediction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 11, pp. 2781–2794, Nov. 2020.

[36] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Commun. ACM*, vol. 61, no. 5, pp. 90–99, Apr. 2018.

[37] M. Neumann and N. T. Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," in *Proc. Interspeech*, 2017, pp. 1263–1267.

[38] J. Parry, D. Palaz, G. Clarke, P. Lecomte, R. Mead, M. Berger, and G. Hofer, "Analysis of deep learning architectures for cross-corpus speech emotion recognition," in *Proc. Interspeech*, 2019, pp. 1656–1660.

[39] X. Zhang, X. Cheng, M. Xu, and T. F. Zheng, "Imbalance learning-based framework for fear recognition in the mediaeval emotional impact of movies task," in *Proc. Interspeech*, 2018, pp. 3678–3682.

[40] V. Dissanayake, H. Zhang, M. Billinghurst, and S. Nanayakkara, "Speech emotion recognition 'in the wild' using an autoencoder," in *Proc. Interspeech*, 2020, pp. 526–530.

[41] Z.-T. Liu, B.-H. Wu, D.-Y. Li, P. Xiao, and J.-W. Mao, "Speech emotion recognition based on selective interpolation synthetic minority oversampling technique in small sample environment," *Sensors*, vol. 20, no. 8, p. 2297, Apr. 2020.

[42] S. Haq and P. Jackson, "Speaker-dependent audio-visual emotion recognition," in *Proc. Int. Conf. Auditory-Visual Speech Process.*, 2009, pp. 53–58.

[43] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. Interspeech*, 2005, pp. 1–4.

[44] J. Tao, F. Liu, and M. Zhang, "Design of speech corpus for mandarin text to speech," in *Proc. 4th Workshop Blizzard Challenge*, 2008, pp. 1–4.

[45] C. Etienne, "Apprentissage profond appliqué à la reconnaissance des émotions dans la voix," Ph.D. dissertation, Univ. Paris-Saclay, Gif-sur-Yvette, France, 2019.

[46] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," *IEEE Trans. Affect. Comput.*, vol. 8, no. 1, pp. 67–80, Jan. 2017.

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

[48] J. A. Russell and L. F. Barrett, "Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant," *J. Personality Social Psychol.*, vol. 76, no. 5, pp. 805–819, 1999.

[49] R. Plutchik, *Emotion: A Psycho Evolutionary Synthesis*. New York, NY, USA: Harper Collins, 1980.

[50] A. J. Gerber, J. Posner, D. Gorman, T. Colibazzi, S. Yu, Z. Wang, A. Kangarlu, H. Zhu, J. Russell, and B. S. Peterson, "An affective circumplex model of neural systems subserving valence, arousal, and cognitive overlay during the appraisal of emotional faces," *Neuropsychologia*, vol. 46, no. 8, pp. 2129–2139, Jul. 2008.

[51] T. Giannakopoulos, A. Pikrakis, and S. Theodoridis, "A dimensional approach to emotion recognition of speech from movies," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Taipei, Taiwan, Apr. 2009, pp. 65–68.

[52] P. Jaccard, "Nouvelles recherches sur la distribution florale," *Bull. de la Société Vaudoise des Sci. Naturelles*, vol. 44, no. 163, pp. 223–270, 1908.

[53] R. Caruana, "Learning from imbalanced data: Rank metrics and extra tasks," in *Proc. Amer. Assoc. Artif. Intell. (AAAI) Conf*, 2000, pp. 51–57.

[54] W. Lu, Z. Li, and J. Chu, "Adaptive ensemble undersampling-boost: A novel learning framework for imbalanced data," *J. Syst. Softw.*, vol. 132, pp. 272–282, Oct. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0164121217301395

[55] M. Glodek, S. Tschechne, G. Layher, M. Schels, S. Scherer, M. Kächele, M. Schmidt, H. Neumann, G. Palm, and F. Schwenker, "Multiple classifier systems for the classification of audio-visual emotional states," in *Proc. Int. Conf. Affect. Comput. Intell. Interact.*, 2011, pp. 359–368.

[56] G. Chollet, "Evaluation of ASR systems algorithms and databases," in *Speech Recognition and Coding: New Advances and Trends*. Berlin, Germany: Springer, 1995.

[57] L. Zouari, "Vers le temps réel en transcription automatique de la parole grand vocabulaire," Ph.D. dissertation, Institut Polytechnique de Paris, Télécom ParisTech, Paris, France, 2007.

[58] C. Clavel and G. Richard, "Reconnaissance acoustique des émotions," in *Système d'intéraction émotionnelle*, C. Pelachaud, Ed. Hermês Science Publications, 2010.

[59] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andre, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, Apr. 2016.

[60] A. Tursunov, S. Kwon, and H.-S. Pang, "Discriminating emotions in the valence dimension from speech using timbre features," *Appl. Sci.*, vol. 9, no. 12, p. 2470, Jun. 2019.

[61] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 5200–5204.

[62] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning," in *Proc. Interspeech*, 2017, pp. 1089–1093.

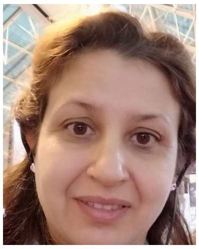
[63] M. deVelasco, R. Justo, L. B. Letaifa, and M. Torres, "Contrasting the emotions identified in spanish tv debates and in human-machine interactions," in *Proc. IberSPEECH*, 2020, pp. 1–5.

[64] Y. Jadoul, B. Thompson, and B. de Boer, "Introducing parselmouth: A Python interface to Praat," *J. Phonetics. Special Issue, Emerg. Data Anal. Phonetic Sci.*, vol. 42, pp. 1–15, Nov. 2018.

[65] T. Giannakopoulos, "PyAudioAnalysis: An open-source Python library for audio signal analysis," *PLoS ONE*, vol. 10, no. 12, Dec. 2015, Art. no. e0144610.

[66] F. Bastien and P. Lamblin, "Theano: New features and speed improvements," in *Proc. NIPS Workshop Deep Learn. Unsupervised Feature Learn.*, 2012, pp. 1–10.

[67] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA, Neural Netw. Mach. Learn.*, vol. 4, no. 2, pp. 26–31, 2012.



LEILA BEN LETAIFA received the M.Eng. degree in electrical engineering and the M.S. degree in signal processing from the National Engineering School of Tunis, Tunisia, in 1997 and 1999, respectively, and the Ph.D. degree in signal and image processing from Télécom Paris-Tech, France, in 2007. From 2007 to 2008, she was contracted as a Research Engineer with the CNRS-LTCI Laboratory, Télécom Paris-Tech. Since 2008, she has been an Assistant Professor of Computer Science in Tunisia. She practiced with the Higher School of Computer Science and Management, the National Engineering School of Sousse, and the National Engineering School of Carthage. She taught different courses and supervised research works related to signal processing, artificial intelligence, and machine learning. Since 2019, she has been contracted as a Researcher with the Speech Interactive Research Group, University of the Basque Country UPV/EHU, Spain. Her research interests include speech, speaker and emotion processing, machine/deep learning techniques, pattern recognition, audio-visual speech processing, spoken dialog systems, and speech pathology.

M. INÉS TORRES (Member, IEEE) received the Ph.D. degree in physics from the Universidad del País Vasco-UPV/EHU, Spain, in 1990, including an internship with the Centre National d'Études des Télécommunications, France, in 1988. In 1990, she founded the Speech-Interactive (SPIN) Research Group, which she has been leading since then. She was a Visiting Researcher with the Polytechnic University of Valencia, Spain, from 1991 to 1992. She was also a Visiting Faculty with Language Technologies Institute, Carnegie Mellon University, in 2012, and then a Visitor in 2013. She was a Visiting Professor with the University of California granted by the Spanish Minister and the Fulbright program in 2018. She is currently a Full Professor, ranked in the highest excellent level, of Computer Science with UPV/EHU where she has also held several academic management positions. She has a multidisciplinary academic and industrial experience in the fields of pattern recognition and machine learning, speech processing, identification of emotional cues in speech and language, speech recognition and understanding, sentiment analysis from social media and human–machine interaction. She was a member of the Board of the Spanish Association of Pattern Recognition (IAPR), from 1995 to 2008.