Unsupervised Classification of Speaker Profiles as a Point Anomaly Detection Task

Cedric Fayet CEDRIC.FAYET@IRISA.FR
Arnaud Delhay ARNAUD.DELHAY@IRISA.FR
Damien Lolive DAMIEN.LOLIVE@IRISA.FR
Pierre-François Marteau PIERRE-FRANCOIS.MARTEAU@IRISA.FR
IRISA, EXPRESSION Team, France

Editor: Editor's name

Abstract

This paper presents an evaluation of three different anomaly detector methods over different feature sets. The three anomaly detectors are based respectively on Gaussian Mixture Model (GMM), One-Class SVM and isolation Forest. The considered feature sets are built from personality evaluation and audio signal. Personality evaluations are extracted from the BFI-10 Questionnaire, which allows to evaluate five personality traits (Openness, Conscientiousness, Extroversion, Agreeableness, Neuroticism). From the audio signal, we extract a prosodic feature set, which performs well in affective computing. The different combinations of models and feature sets are evaluated on the SSPNET-Personality corpus which has already been used in several experiments, including a previous work on separating two types of personality profiles in a supervised way.

In this work, we propose an evaluation of the three detectors with consideration to the features used. Results show that, regardless of the feature set, GMM based method is the most efficient one (0.96 ROC-AUC score with the best feature set). The prosodic feature set seems to be a good compromise between performance (0.91 ROC-AUC score with GMM based method) and ease of extraction.

Index Terms: Anomaly detection, Gaussian Mixture Model, One Class-Support Vector Machine, Isolation Forest, Threshold Classification, Social Signal, Big Five, Prosody, SSPNET-Personality.

1. Introduction

According to Chandola et al. (2009), "an anomaly is defined as a pattern that does not conform to an expected normal behavior". The main objective of any anomaly detection (AD) system is to identify abnormal states from normal state distributions. The feature sets that describe these states are related to the nature of the input data (continuous, categorical, spatial, or spatio-temporal data), but also to the nature of the anomalies that we aim to track (point-wise, contextual or collective anomalies) Chandola et al. (2009). Finally, to choose an adequate anomaly detection model, we also need to consider the desired output (score or label) and the availability of labeled data (supervised or unsupervised techniques).

In the literature, anomaly detection techniques are applied in intrusion detection Axelsson (2000), fraud detection Abdallah et al. (2016), sensor network Park et al. (2010), monitoring flight safety Li et al. (2011), etc. As far as we know, anomaly detection with

the speech signal is more focused on speech pathology or disorder Alonso et al. (2005) identification, or on the detection of other type of pathology or disease, as cancer for instance Clapham et al. (2012). The speech signal is also used for detecting stress or depression He et al. (2009); Valstar et al. (2016), that could be seen as an anomaly detection in the way that this state of mind can be considered as an abnormal mental state. Handling audio, video or biological signal to infer social information such as personality is part of a field called *Personality Computing*. This field is focused on the recognition of self perceived personality (Automatic Personality Recognition), the prediction of the personality perceived by the others (Automatic Personality Perception), and the generation of artificial personalities (Automatic Personality Synthesis) Vinciarelli and Mohammadi (2014).

The Big Five model has been proposed to describe the speaker personality using the five following personality traits: openness, conscientiousness, extroversion, agreeableness and neuroticism John and Srivastava (1999). Different tests with adaptation to a local context have been conducted on different languages and cultures, and the big five model seems to be generalizable to them Goldberg (1981); Gurven et al. (2013).

SSPNET-Personality corpus was built to experiment the prediction of the big five scale over audio features Mohammadi and Vinciarelli (2012). It is composed of french audio clips extracted from the "Radio Suisse Romane" and recorded by either professional or guest speakers. Schuller et al. (2015) offers a good overview of the different systems using this corpus. A lot of experiments has already been conducted on this corpus and most of them try to predict the value of the big five scales using different audio representations. In addition, some experiments in Mohammadi and Vinciarelli (2012) suggest that the big five representation could be a good predictor of the role of a speaker. The authors propose a supervised SVM method that learns to recognize if a speaker is a professional speaker or a guest. The feature set used in this work is the psychological evaluation of a sample given by annotators using Big five traits. Their predictor reaches an accuracy of about 75%.

The purpose of our work is to compare the use of the big five features to the use of prosodic features to design a personality predictor able to separate professional speakers from unprofessional guests. To this end, we consider a professional speaker as belonging to a normal personality class and a guest to an abnormal one. Contrary to Mohammadi and Vinciarelli (2012), we propose to work in the unsupervised anomaly detection framework and to evaluate three unsupervised strategies. Furthermore, three sets of features are evaluated. Results show that prosodic features perform well and are much less costly compared to manually annotated Big Five traits, even if those features may provide better detection performance.

The remainder of the paper is structured as follows. The anomaly detection methodology we used is described in section 2. In section 3, the feature sets we evaluate are detailed before describing the experimental setup in section 4. Finally, the results are presented and discussed in section 5.

2. Methodology

This work aims at determining if a sample is normal (professional speaker) or abnormal (guest speaker) based on a train set containing only normal audio clips. Each audio clip is

described by a feature vector summarizing the time evolution of the features as described in section 3.

In this section, we describe the three methods that we have used to perform the anomaly detection: a Gaussian Mixture Model (GMM) approach, an Isolation Forest (iForest) approach and a One-Class Support Vector Machine approach (OC-SVM). The OC-SVM has been used to ensure a comparison baseline with the SVM method used in Mohammadi and Vinciarelli (2012).

2.1. Gaussian-Mixture Model approach (GMM)

The Gaussian-mixture model approach for detecting anomaly can be decomposed in two steps:

1. learning a Gaussian Mixture Model (GMM) to model the feature space distribution for the normal samples,

2. choosing a threshold based on the likelihood to decide whether or not a sample has been generated by the learned distribution, and thus if the sample could be considered as normal.

Considering the set $\mathbf{X}$ of feature vectors $\mathbf{x}$, each representing an audio clip, a GMM $\mathcal{M}_{\mathbf{X}}$ with M Gaussians is chosento model the dataset $\mathbf{X}$. Its probability distribution is given by

$$P(\mathbf{x}|\mathbf{\Theta}) \quad = \quad \sum_{m=1}^{M} \alpha_m P(\mathbf{x}|\theta_m)$$

where $\mathbf{\Theta} = (\alpha_1, \ldots, \alpha_m, \theta_1, \ldots, \theta_M)$ is the parameter vector. α_m is the mixing coefficient associated to the m^{th} Gaussian with parameters $\theta_m = (\mu_m, \Sigma_m)$ that represent respectively the mean vector and the covariance matrice of distribution $P(\mathbf{x}|\theta_m)$. The *Expectation-Maximization (EM)* algorithm is used to learn the GMM parameters from unlabeled data and maximizes the loglikelihood of the data and the model Dempster et al. (1977).

By using the distribution learned at the first step, and considering a new sample that we want to label either as normal or abnormal, its likelihood is evaluated and compared to a threshold value. If the likelihood of the sample is above the threshold, the sample is considered as normal. Otherwise, it is considered to be abnormal.

The structure of the covariance matrix used for each Gaussian component is considered to be an hyper-parameter of the model. Consequently, we have three hyper-parameters : the covariance matrix structure, the number of components (both are related to the Gaussian distribution) and the threshold value.

2.2. Isolation Forest (iForest)

Isolation Forest is an ensemble learning method dedicated to detect anomaly. The particularity of this method is that it explicitly isolates anomalies rather than learns a model for normal instances Liu et al. (2008). The main assumption behind iForest is that a normal sample is hard to isolate from other samples, and on the contrary, an anomaly is more easily isolated from other samples. iForest is composed of T iTrees, each one built on a

random selection of ψ samples from the training set. From this subset of samples, an iTree is constructed by a random recursive partitioning, until all the samples are isolated or until a stop criterion is reached (a depth limit for example). The partitioning is realized by the random selection of an attribute (feature) and the random choice of a pivot value in the range of the selected attribute. For an iTree, the sample score is computed as the distance (path length) between the leaf node containing the sample and the root node of the tree.

Let x be a sample, n the number of samples on which the iTrees are built. Let f be the iForest, with $f = \{t_1, t_2, ...t_T\}$. Let $h(t, x)$ be the number of edges of the t iTree between the root and the leaf which contains (or isolates) x. Let $c(n)$ be the average path length of unsuccessful search in a Binary Search Tree. $c(n)$ estimates the average path length of an iTree.

The anomaly score s of an instance x estimated with the iForest f is given by:

$$s(\mathbf{x}, \mathbf{f}, \mathbf{n}) = 2^{\frac{-\sum_{k=1}^{T} h(\mathbf{f_k}, \mathbf{x})}{T * c(\mathbf{n})}}$$

2.3. One-Class SVM Approach (OC-SVM)

In short, a SVM classifier learns a boundary which maximizes the margin between classes. This well-known approach has been shown to be very effective on many classification problems. OC-SVM is an adaptation of SVM to the one-class problem. After transforming the feature via a kernel, OC-SVM considers as a starting point, all the available data as member of a single class $C_{inliers}$ and the origin as the only member of a class $C_{outliers}$ Schlkopf et al. (2001).

During the training, the hyper-parameter ν corresponds to a penalizing term which represents a trade-off between inliers and outliers. With the SVM approaches, the choice of the kernel is important to improve the results. The most widely used kernels are linear (inner product), polynomial, RBF and sigmoid. Related to the kernel used (RBF, polynomial, sigmoid) we have to determine its parametric coefficients. For instance, with the sigmoid kernel :

$$k(\mathbf{x}, \mathbf{y}) = \tanh(\gamma \mathbf{x}^T \mathbf{y} + c_0)$$

where $\mathbf{x}$ and $\mathbf{y}$ are input vectors and γ (slope) and c_0 (intercept) are parametric values related to the kernel.

We choose to use the distance between a new sample and the learned boundary as an anomaly score, instead of directly deciding which class the samples belong to. We then use a classification threshold based on this anomaly score to decide if a sample is normal or abnormal. This two steps strategy introduces some adjustable fuzziness around the boundary. This strategy corresponds to *a posteriori* modifying the role played by the slack-variable penalty in the first step.

3. Materials

The experiments are conducted over the SSPNET-Personality Corpus Mohammadi and Vinciarelli (2012). The corpus contains 640 audio clips divided into 307 audio clips of professional speakers and 333 audio clips of guests in French language. The duration of

each audio clip is about 10 seconds, based on the assumption that it takes short time to get an opinion about others personality.

3.1. Big Five features

To evaluate the personality score of an individual under a personality model, a bundle of different assessments are available in Boyle and Helmes (2009). In our case, we are focusing on the Big Five Inventory. It is a self-report inventory designed to measure the Big Five dimensions (that contains 44 personality items). It consists of short phrases with relatively accessible vocabulary John and Srivastava (1999). Even if the BFI is relatively simple to fill-in, it takes over five minutes to be completed. For the creation of a corpus, the time per sample is too long. A shorter version of the BFI, the BFI-10 Questionnaire Rammstedt and John (2007), has the advantage to be really quick to fill-in (about one minute per sample).

For each sample, 11 non-native french speakers have evaluated the BFI-10 Questionnaire Rammstedt and John (2007) from which a score is computed for each Big Five's scale. For our experiment, we consider two sets of features based on the big five model:

- *BigFive5*: for each sample and for each Big Five scale, we compute the mean of the 11 evaluations which leads to 5 features.
- *BigFive55*: for each sample, we concatenate the 11 evaluations given for each sample which leads to 55 features.

The purpose of this last set of features is to verify if the information contained in several distinct annotations is complementary or can simply be aggregated in a lower dimension feature vector, as in *BigFive5* feature vector. The main drawback of these two feature sets is that they are the result of a manual annotation, which is furthermore difficult to predict directly from the speech signal.

3.2. Prosodic Features

Prosodic features are commonly used to capture affect cues in a speech signal. Contrary to the Big Five traits, a large number of prosodic features are much easier to extract automatically from the speech signal. In this study, we adopt the 6 dimensional prosodic feature set as described in Mohammadi and Vinciarelli (2012). Using the Praat software Boersma and Weenink (2016), we extract the pitch, the first two formants, the energy and the duration of voiced/unvoiced segments with a sliding analysis window size of 40 ms with a step of 10 ms. From these low-level features, resulting from the extraction, we derive the final features that summarize their time evolution by computing mean, maximum, minimum and entropy values for each of the 6 features. Consequently, the final *Prosodic* feature set is composed of 24 features for each audio clip.

3.3. Pre-processing

Normalizing the features is an important pre-processing step for a lot of machine learning algorithms. In our case, OC-SVM could be affected by features that do not share the same range of values and the same variations. Therefore this step is particularly important in the case of *Prosodic* features, which are composed of different kind of features. We choose to

use a standard normalization by centering and reducing the features. For the iForest and GMM approaches, this pre-processing step will not affect significantly the final result.

4. Experimental Setup

The different experiments described in this section are carried out by following the same procedure. From the corpus, we build three sets (train, test, and validation sets), as described below. To increase the statistical confidence of our results, we run each experiment 60 times by distributing randomly the samples on the three sets.

The data is thus split into three folds as follows:

- Train set: 207 clips from professional speakers and a variation from 0 to 103 guest clips.
- Test set: 50 clips from professional speakers and 50 guests clips.
- Validation set: 50 clips from professional speakers and a variation from 0 to 50 guest clips.

The number of correlated features (Pearson correlation coefficient over 0.7) for each set of features is the following: 0 for BF5 , 4 for BF55 and 2 for Prosodic .

We use python 3.5 and scikit-learn Pedregosa et al. (2011) to conduct the different experiments.

4.1. Hyper-parameters tuning

According to the available data, since we have 207 samples in the training set, we need to be careful about the number of parameters in our model to avoid over-fitting.

In the case of the Gaussian Mixture approach, we need to determine two hyper-parameters: the covariance matrices type and the number of components. Using a diagonal covariance matrix seems to be a good compromise between complexity of the model, the size of training set and the few correlated features. Then, we need to choose the number of considered components (referred to as #cp) . To make this choice, a common method is to balance the complexity of the model (number of components) and the quality of the model on the training set. The AIC and BIC scores Steele and Raftery (2010) are well-known criteria to reach this goal. In our case, the number of features is low (see in section 3), so the BIC score is more likely to be reliable Schwarz and others (1978).

In the case of the OC-SVM approach, after testing different types of kernel in our settings, we have chosen to present the results with the sigmoid kernel which gives the best results. The parameter c_0 is fixed arbitrarily to 0. The two hyper-parameters ν and γ are chosen using a grid search approach to maximize the classification score.

In the case of the iForest approach, we need to find two hyper-parameters: ψ (sub-sampling size given as a percentage of the available data) and T (number of sub-estimators *i.e* iTrees). When using an iForest, each sample is associated to an anomaly score. In our case, we consider the 10 percents of the samples with the highest probability to be abnormal samples. With this assumption, we can use a classification score as we did previously.

Table 1: Hyper-parameters chosen for each AD

	GMM		iForest		OC-SVM		
	Co-var	#cp	T	ψ	Kernel	γ	ν
BF5	diagonal	4	17	0.29	sigmoid	0.016	0.7
BF55	diagonal	14	77	0.31	sigmoid	0.007	0.6
PROSODIC	diagonal	10	59	0.26	sigmoid	0.009	0.57

For each set of features, we considered a training set with all the available normal samples, to determine the hyper-parameters with respect to the aforementioned methods. Parameter tuning results are provided in Table 1.

4.2. Experiments

4.2.1. Comparing AD models on the different feature sets

To compare the different sets of features according to the chosen approach, we have carried out the following steps: for each set of values (*BigFive5*, *BigFive55*, *Prosodic*), we test each anomaly detection model after tuning the hyper-parameters as explained before (section 4.1). Then, we compare the different models by using a ROC (Receiver Operating Characteristic) curve. It means that for a given FPR (False Positive Rate), we search for the associated TPR (True Positive Rate). This sampled association is obtained by testing a range of possible threshold values. The ROC curve gives an information about the detector quality as a function of the threshold value. In this paper, an AUC score refers to the area under a ROC curve, and AUC score range is between 0 and 1. A 0.5 AUC value corresponds to a random binary classifier score.

A second step consists in estimating the robustness of each detector to a degradation of the training set. In this purpose, we introduce a certain percentage of abnormal samples into the training set (0% to 50%). We keep the different hyper-parameter values unchanged.

A last step consists in analyzing the anomaly score obtained for the different sets of features with the three approaches. By considering the hyper-parameters obtained previously, we trained the predictor with all the available normal samples. Then, we evaluated the anomaly score for each normal and abnormal sample separately.

4.2.2. Influence of the validation set purity on the threshold choice

In the previous experiments, we have used a range of possible values for the threshold which allow to separate the normal and abnormal classes. In this step, we seek to select the best threshold value. In this purpose, we design a validation set which contains 50 normal samples and a variable percentage of abnormal samples (variation of the percentage of abnormal samples, starting from 0, ending at 50 with a step of 5 percents). To evaluate the quality of each model, we use the F1-score (i.e. the harmonic mean between recall and precision) with normal instances as positives and abnormal instances as negatives. For each level of impurity, we determine the threshold value that maximizes a F1-score on the validation set. Then, we evaluate the F1-score on the test set with the threshold value obtained before.

Table 2: Mean area under ROC curve (ROC-AUC score) and standard deviations, for the three approaches combined with the three feature sets.

	BF5	BF55	PROSODIC
GMM	0.759±0.024	0.961±0.020	0.910±0.031
OC-SVM	0.857±0.030	0.918±0.024	0.876±0.036
iForest	0.762±0.031	0.890±0.032	0.801±0.037

4.2.3. Importance of the features

Here we address the estimation of the contribution to the prediction brought by each feature f_i in the scope of a given feature set F. With this purpose, we run the feature selection algorithm 1 on $BF5$ and *Prosodic* feature sets.

Algorithm 1 Forward feature selection.

Require: Let AD be an anomaly method.
Require: Let $f_0, .., f_N$ be a feature set.
  $S_{Best} \leftarrow \varnothing$
  $S_{Open} \leftarrow f_0, .., f_N$
  repeat
    $S_{test} \leftarrow \varnothing$
    for each f_i in S_{Open} **do**
      $n \leftarrow 0$
      $Scores \leftarrow \varnothing$
      repeat
        Train AD with the feature set $S_{Best} \cup f_i$ on the training set.
        AUC $\leftarrow$ Compute the area under the roc curve obtained by AD on the test set.
        $Scores \leftarrow$ Scores $\cup \{AUC\}$
        $n \leftarrow n + 1$
      until $n = 60$
      $S_{test} \leftarrow S_{test} \cup \{mean(Scores), i\}$
    end for
    (AUC,i)$\leftarrow \arg\max_{AUC}(S_{test})$
    $S_{Best} \leftarrow S_{Best} \cup \{f_i\}$
    $S_{Open} \leftarrow S_{Open} \setminus \{f_i\}$
  until S_{Open} is $\varnothing$

5. Results

5.1. Comparing AD models on the different feature sets

Our first experiment (Table 2) consists in evaluating the quality of each feature set. On our data, the *BigFive55* feature vector leads to better AUC values regardless of the method used. For the OC-SVM approach, the *Prosodic* and *BigFive5* feature vectors get comparable results, but for the GMM approach, the *Prosodic* set reached a score closer to the one reached

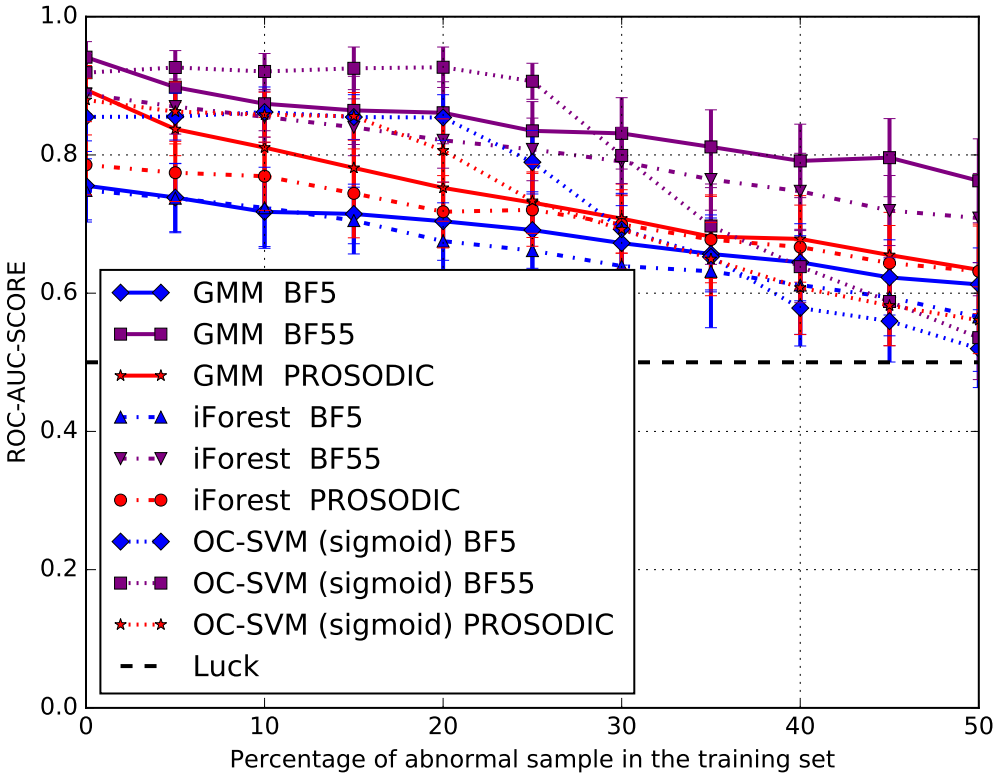

Figure 1: Robustness of the models to the introduction of abnormal samples in the training set.

by the *BF55* set than the *BF5* set. For the iForest approach and the GMM approach, *BigFive5* feature vector achieves the worst AUC values.The *BigFive5* feature vector has the lowest AUC results in general, thus showing that averaging the individual annotations induces a significant information loss.

In addition, the GMM approach performs better than the two other methods for the *Prosodic* and *BF55* feature sets.

Our second experiment (Figure 1) consists in evaluating the quality of each AD model when the training set is degraded, *i.e.* when a certain percentage of abnormal samples are included in it. The OC-SVM approach shows a degradation in two phases: before a certain percentage of contamination it is resilient to degradation and after this limit its AUC scores decrease rapidly down to around 0.5 (random detector) for 50% of degradation. Isolation Forest and Gaussian Mixture approaches show a regular low slope linear decrease along the degradation axis.

For the OC-SVM approach, the *BigFive55* feature vector seems to maintain stable results for approximately 25% of degradation. The other two feature vectors start to have a decrease of their AUC value for less than 15% of degradation. For the iForest approach, the *prosodic* feature vector seems to be more robust to degradation than the *BigFive5*.

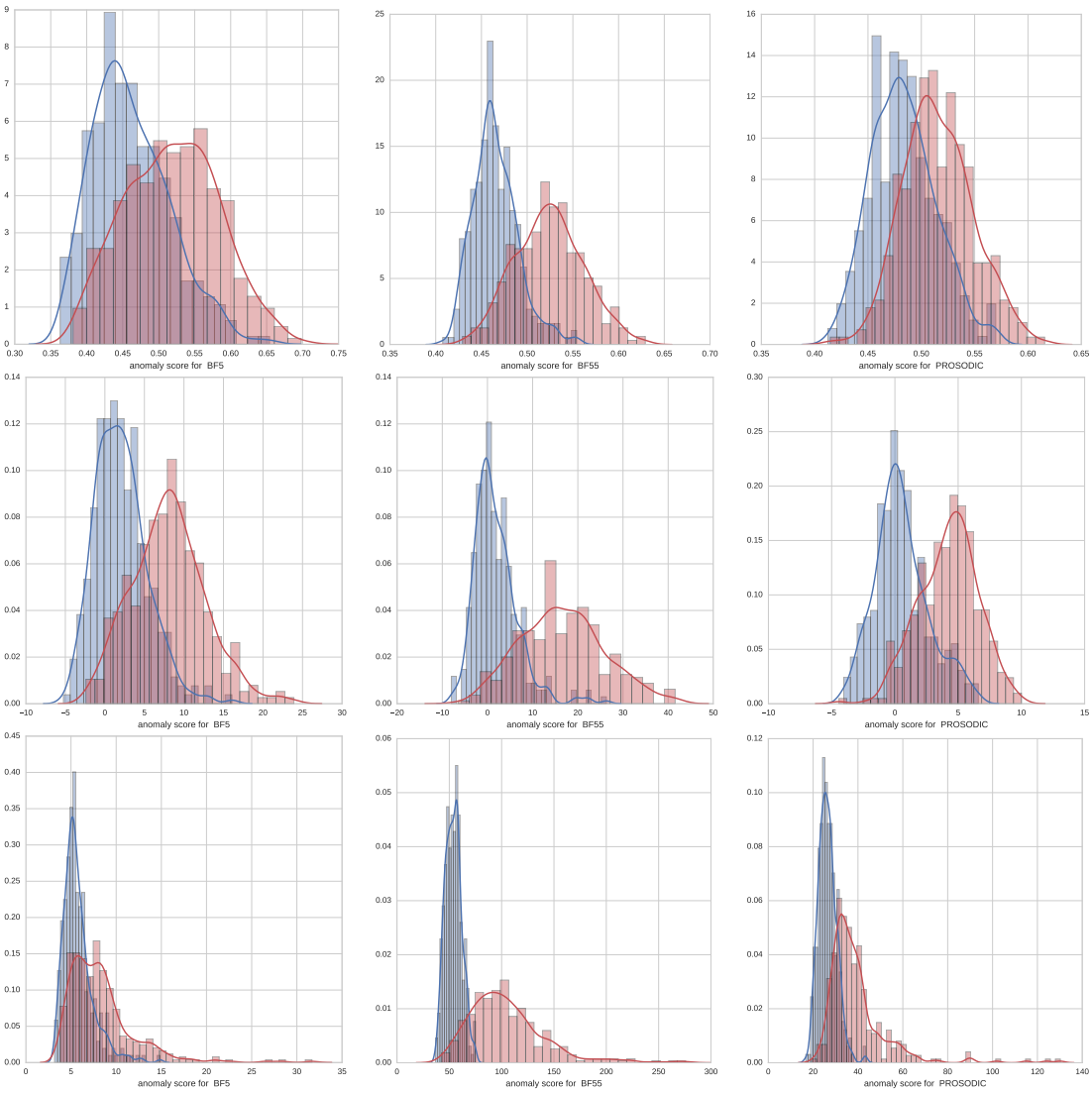

Figure 2: Anomaly scores (red (right) for abnormal samples and blue (left) for normal samples) obtained for each feature BF5 (left), BF55 (middle), Prosodic (right) with the three approaches iForest (top), OC-SVM (middle) and GMM (bottom).

The iForest and GMM approaches share the same trends to degradation, but the GMM approach reaches a higher AUC score compared to iForest. In any case, the *BigFive55* feature vector provides a more robust separation between normal and abnormal samples.

In our final experiment (Figure 2), we analyze the anomaly score obtained for all available samples in the corpus. Figure 2 presents the histograms of anomaly scores obtained for normal samples (blue) and another one for abnormal samples (red) by the three tested AD methods (iForest (top), OC-SVM (middle) and GMM (bottom)) and the three feature vectors (BF5 (left), BF55 (middle), Prosodic (right)).

For the three approaches, the intersection area between the abnormal and normal curves is greater for BF5 than for Prosodic, and greater for Prosodic than for BF55. Indeed, by aggregating the annotations of BF55 to get BF5, we reduce the ability to discriminate between normal and abnormal samples as already mentioned.

It is noticeable that the iForest achieves the lowest performances compared to the OC-SVM or the GMM approaches. This suggests that, in our context, modeling the normality for detecting an abnormal sample is better than considering that an anomaly is characterized by very different feature values (*i.e.* that corresponds to isolated points in the feature space). Furthermore, the GMM approach seems to provide a better separation between normal and abnormal samples compared to the OC-SVM approach.

5.2. Influence of the validation set purity on the threshold choice

To conduct this experiment, we considered the threshold values obtained through the ROC curve computation as explained in section 4.2.2. For the three approaches, the F1-score curves obtained (Figure 3) for a feature vector seem to share a same pattern.

The *BigFive55* feature vector curves reach an asymptotic value for at least 10% of abnormal samples on the validation set (around 6 samples), and at least 30% (21 samples) for the two other feature vectors.

The asymptotic F1-score reached with *BigFive55* feature vector and GMM approach is above 0.8, and the F1-score obtained with *Prosodic* feature vector and OC-SVM approach corresponds to the one (an F1-measure around 0.75) obtained in a supervised way with SVM Mohammadi and Vinciarelli (2012). The variation of the F1-score in Figure 3 shows the difficulty to reach a good and stable threshold value.

5.2.1. Importance of the features

The figure 4 corresponds to the results (in term of AUC score) obtained by using the feature selection algorithm described in 4.2.3 for the three AD approaches considered in this paper.

For the $BF5$ feature set, the first two features selected over ten runs are Extroversion and Conscientiousness, regardless to the AD-method used. Taken individually, we obtain an AUC score over 0.7 for both features while the other features obtain individually an AUC score between 0.5 to 0.6. Moreover, for GMM and iForest approaches, the results show that considering all the traits penalizes the overall score compared to the score obtained by the feature set composed of Extroversion and Conscientiousness only.

For the *Prosodic* feature set, in contrary to $BF5$, we do not observe a stable or partially stable feature set, for ten runs. We notice that the features which have a high AUC score in the first step, don't systematically appear as the first features selected. This means that they are highly correlated with previously selected features. One can observe that with only a subset of five features we can obtain a good score, for OC-SVM and iForest. For the GMM approach the score keep growing until ten features.

For the two sets of features $BF5$ and *Prosodic*, considering all the available features seems not being the best strategy. For $BF5$, GMM and iForest approaches seem to be impacted by the size of the feature vector, while OC-SVM seems to be less sensitive to it. For *Prosodic*, the iForest approach is the most impacted by the size of the feature vector, OC-SVM seems less impacted and GMM is stable.

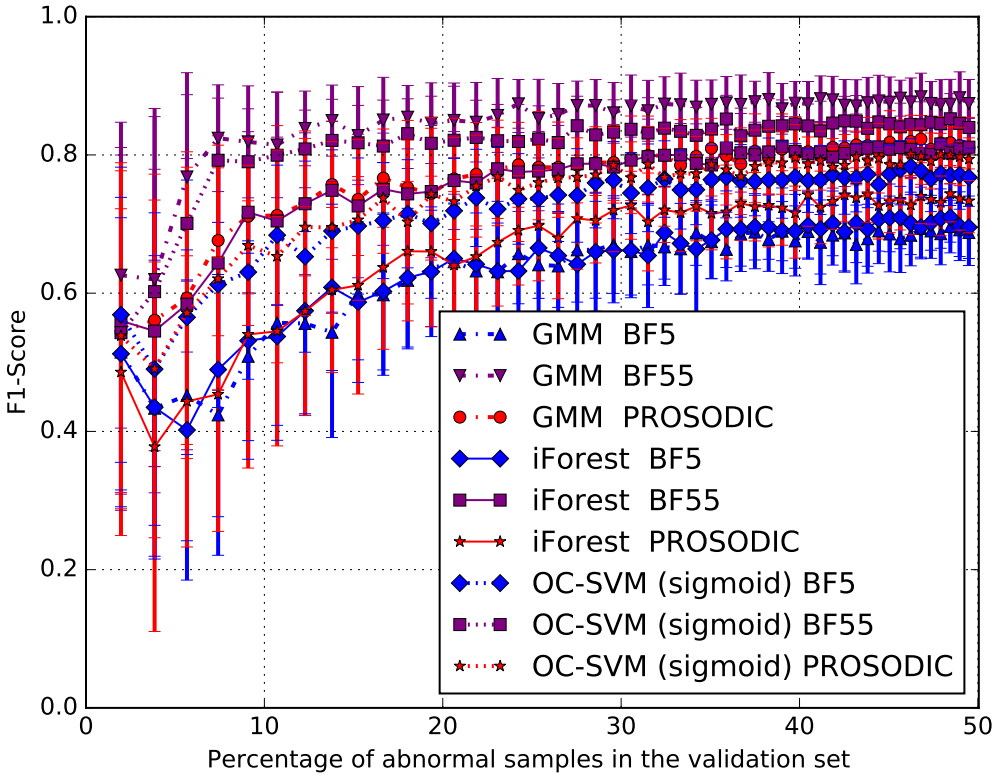

Figure 3: Variation of the F1-Score in function of the percentage of abnormal samples in the validation set.

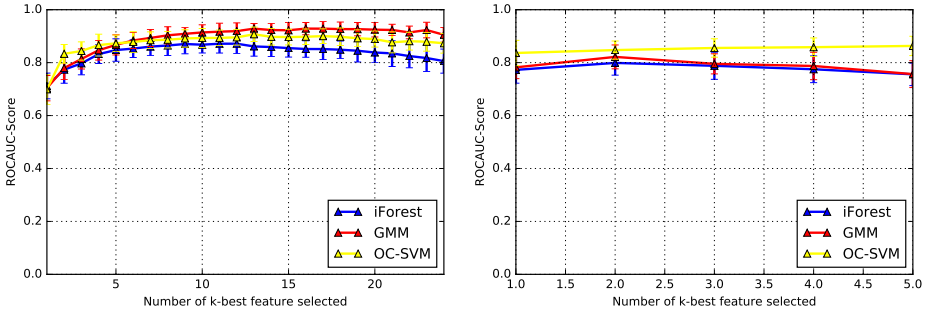

Figure 4: ROC-scores obtained at each step of the feature selection algorithm 1 for the best sub-set of features for the BF5 feature set (right) and Prosodic feature set (left)

The *Prosodic* feature set is composed of 24 features which are the result of 6 low level features (pitch, F1, F2, energy, voicedLength, unvoicedLength) aggregated by 4 hierarchical functions (maximum, minimum, mean, entropy). When we apply the feature selection algorithm (Algorithm 1), we observe that the maximum, minimum and mean have a same importance in the first step, and the entropy seems to be the less important of the fourth

high level method. Moreover, if we consider the six feature groups as described above, we obtain that the Pitch group (maxPitch, minPitch, meanPitch, entropyPitch) seems to be the less important feature group compared to the others.

6. Conclusion

The main objective of this paper was to compare the use of prosodic cues and the Big Five annotation traits as feature sets in an anomaly detection task. We have conducted experiments with the SSPNET-Personality Corpus using professional speakers as normal samples and guest speakers as abnormal samples. This choice was motivated by Mohammadi *et al.* work Mohammadi and Vinciarelli (2012) that demonstrates the effectiveness of using the Big Five features to train a supervised classifier able to separate these two categories of speakers. We built three sets of features (*BigFive5*, *BigFive55* and *Prosodic*) based on the speech signal and a psychological evaluation (Big Five model) available for the dataset. We have used three different unsupervised machine learning methods (iForest, OC-SVM, GMM) to build our anomaly predictors. Based on our results, the good performance of the *BigFive55* feature set over the *Prosodic* feature set indicates that features based on psychological information can bring more discriminant information than audio features only for the considerd task. However, the prosodic features are easy to extract from the speech signal and thus seem to be the best compromise between ease of extraction and performance. The better result obtained with OC-SVM or GMM compared to iForest seems to indicate that an anomaly in our context is better identified through the learning of a normality model from the available data than searching for samples which are really different from the others in the feature space. A natural follow-up is to test other feature sets: for instance one can increase the number of features to reach the same order of dimensions as the *BigFive55* feature vector. Another perspective is to build a meta anomaly detector using bagging or boosting approaches to aggregate the three proposed AD models that rely on quite different conceptual ways to separate anomaly from normal data. Finally, conducting these experiments on other audio corpora, or trying to generalize our results on multimedia data are part of our future work.

7. Acknowledgments

This research has been financially supported by the French Ministry of Defense - Direction Générale pour l'Armement and the région Bretagne (ARED).

References

Aisha Abdallah, Mohd Aizaini Maarof, and Anazida Zainal. Fraud detection system: A survey. Journal of Network and Computer Applications, 68:90–113, 2016.

Jesús B Alonso, Fernando Díaz-de María, Carlos M Travieso, and Miguel Angel Ferrer. Using nonlinear features for voice disorder detection. In ISCA tutorial and research workshop (ITRW) on non-linear speech processing, 2005.

Stefan Axelsson. Intrusion detection systems: A survey and taxonomy. Technical report, Chalmers University of Technology, Göteborg, Sweden, 2000.

Paul Boersma and David Weenink. Praat: doing phonetics by computer, 2016. URL http://www.praat.org/.

Gregory J Boyle and Edward Helmes. Methods of personality assessment. Cambridge University Press, 2009.

Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. ACM computing surveys (CSUR), 41(3):15, 2009.

Renee Peje Clapham, Lisette van der Molen, R. J. J. H. van Son, Michiel W. M. van den Brekel, and Frans J. M. Hilgers. NKI-CCRT corpus - speech intelligibility before and after advanced head and neck cancer treated with concomitant chemoradiotherapy. In Proceedings of the 8th international conference on Language Resources and Evaluation (LREC), 2012.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society. Series B (methodological), pages 1–38, 1977.

Lewis R Goldberg. Language and individual differences: The search for universals in personality lexicons. Review of personality and social psychology, 2(1):141–165, 1981.

Michael Gurven, Christopher von Rueden, Maxim Massenkoff, Hillard Kaplan, and Marino Lero Vie. How universal is the Big Five ? Testing the five-factor model of personality variation among foragerfarmers in the bolivian amazon. Journal of Personality and Social Psychology, 104(2):354–370, 2013.

Ling He, Margaret Lech, Namunu C. Maddage, and Nicholas Allen. Stress detection using speech spectrograms and sigma-pi neuron units. In Fifth International Conference on Natural Computation ICNC'09, volume 2, pages 260–264. IEEE, 2009.

Oliver P John and Sanjay Srivastava. The Big Five trait taxonomy: History, measurement, and theoretical perspectives, volume 2, pages 102–138. Guilford, 1999.

Lishuai Li, Maxime Gariel, R John Hansman, and Rafael Palacios. Anomaly detection in onboard-recorded flight data using cluster analysis. In IEEE/AIAA 30th Digital Avionics Systems Conference, pages 4A4–1–4A4–11. IEEE, 2011.

F. T. Liu, K. M. Ting, and Z. H. Zhou. Isolation forest. In Eighth IEEE International Conference on Data Mining, pages 413–422, 2008.

G. Mohammadi and A. Vinciarelli. Automatic personality perception: Prediction of trait attribution based on prosodic features. IEEE Transactions on Affective Computing, 3(3): 273–284, 2012.

Kyungseo Park, Yong Lin, Vangelis Metsis, Zhengyi Le, and Fillia Makedon. Abnormal human behavioral pattern detection in assisted living environments. In Proceedings of the 3rd International Conference on PErvasive Technologies Related to Assistive Environments, page 9, 2010.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in python. Journal of Machine Learning Research, 12:2825–2830, 2011.

Beatrice Rammstedt and Oliver P. John. Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. Journal of Research in Personality, 41(1):203–212, 2007.

Bjrn Schuller, Stefan Steidl, Anton Batliner, Elmar Nth, Alessandro Vinciarelli, Felix Burkhardt, Rob van Son, Felix Weninger, Florian Eyben, Tobias Bocklet, Gelareh Mohammadi, and Benjamin Weiss. A survey on perceived speaker traits: Personality, likability, pathology, and the first challenge. Computer Speech & Language, 29(1):100–131, 2015.

Gideon Schwarz and others. Estimating the dimension of a model. The annals of statistics, 6(2):461–464, 1978.

Bernhard Schlkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. Neural computation, 13(7):1443–1471, 2001.

Russell J Steele and Adrian E Raftery. Performance of bayesian model selection criteria for gaussian mixture models. Frontiers of Statistical Decision Making and Bayesian Analysis, 2:113–130, 2010.

Michel Valstar, Jonathan Gratch, Bjrn Schuller, Fabien Ringeval, Dennis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. AVEC 2016: Depression, mood, and emotion recognition workshop and challenge. In Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, pages 3–10. ACM, 2016.

A. Vinciarelli and G. Mohammadi. A survey of personality computing. IEEE Transactions on Affective Computing, 5(3):273–291, 2014.