University of Glasgow

Tayarani, M., Esposito, A. and Vinciarelli, A. (2019) What an “ehm” leaks about you: mapping fillers into personality traits with quantum evolutionary feature selection algorithms. *IEEE Transactions on Affective Computing*, 13(1), pp. 108-121. (doi: 10.1109/TAFFC.2019.2930695).

This is the author’s final accepted version.

There may be differences between this version and the published version. You are advised to consult the publisher’s version if you wish to cite from it.

http://eprints.gla.ac.uk/202931/

Deposited on: 11 November 2019

Enlighten – Research publications by members of the University of Glasgow
http://eprints.gla.ac.uk

What an "Ehm" Leaks About You: Mapping Fillers into Personality Traits with Quantum Evolutionary Feature Selection Algorithms

Mohammad Tayarani, Anna Esposito and Alessandro Vinciarelli *Member, IEEE*

Abstract—This work shows that fillers - short utterances like "ehm" and "uhm" - allow one to predict whether someone is above median along the Big-Five personality traits. The experiments have been performed over a corpus of 2,988 fillers uttered by 120 different speakers in spontaneous conversations. The results show that the prediction accuracies range between 74% and 82% depending on the particular trait. The proposed approach includes a feature selection step - based on Quantum Evolutionary Algorithms - that has been used to detect the personality markers, i.e., the subset of the features that better account for the prediction outcomes and, indirectly, for the personality of the speakers. The results show that only a relatively few features tend to be consistently selected, thus acting as reliable personality markers.

Index Terms—Social Signal Processing, Personality Computing, Quantum Evolutionary Algorithms, Computational Paralinguistics.

1 Introduction

It was 1927 when Edward Sapir - widely known for the hypothesis of linguistic relativity - stated that "*[...] looking for the thing we call personality we have the right to attach importance to the thing we call voice [...] the nervous processes that control voice production must share in the individual traits of the nervous organization that condition the personality*" [1]. In the last decade, computing domains like Social Signal Processing [2] and Computational Paralinguistics [3] appeared to confirm such an early intuition by showing that, at least to a certain extent, it is possible to map features automatically extracted from speech into personality traits (see [4] for an extensive survey).

In line with the above, the goal of this work is to show that the *fillers* uttered during a conversation allow one to predict whether an individual is above median along the Big-Five personality traits [5]. The fillers are short vocalizations like "*ehm*" or "*uhm*" that "*are characteristically associated with planning problems [...] planned for, formulated, and produced [...] just as any word is*" [6]. In other words, the fillers are those vocalizations that speakers utter when they want to hold the floor, but they do not know what to say next. Fillers occur frequently during spontaneous conversations and, in particular, the analysis presented in [7] shows that the speakers involved in the experiments of this work utter, on average, one filler every 10.9 seconds.

- *Mohammad Tayarani and Alessandro Vinciarelli are with the University of Glasgow. E-mail: firstname.lastname@glasgow.ac.uk*
- *Anna Esposito is with the Università della Campania "Luigi Vanvitelli" (Italy). E-mail: anna.esposito@unicampania.it*

To the best of our knowledge, no theory explains why fillers should carry personality-relevant information and, according to the literature, "*[...] little research examines the correlation between self-report personality traits and filler words*" [8]. The assumption underlying this article is that individual differences captured through the Big-Five traits lead to different motivations behind the use of fillers and, hence, to different ways of uttering them. For example, when it comes to information seeking behaviour, people with high Openness tend to take into account more sources [9]. Given that fillers often correspond to planning problems (what to say next), it is possible that the tendency to consider more alternatives requires one to hold the floor for longer time and, correspondingly, to utter fillers in a different and more sustained way. A similar explanation applies to highly conscientious people that, in the case of planning problems, are likely to go more thoroughly through all possible alternatives and, hence, to utter the fillers in a way that allows one to hold the floor longer and to communicate high cognitive load (a tendency to use fillers more frequently when the education level is higher has been observed in [10]).

Extraversion has been shown to be a higher order trait encompassing dominance, the tendency to control people and the environment [11]. This suggests that Extraversion is associated to a tendency to control the floor and, hence, to utter the fillers in a way that ensures such a goal to be achieved. In the case of Agreeableness, the trait of those that tend to do what others need and like, fillers are probably used as a way to ensure smoother turn-taking. Finally, the anxiety associated to Neuroticism has been shown to increase the number of

speech disfluencies [12]. These include the fillers that, in this case, are uttered in a way that conveys negative emotions rather than planning problems.

The experiments have been performed over the $2,988$ fillers - uttered by 120 individuals - of the *SSPNet Vocalization Corpus*, a publicly available dataset that has been used for the *Interspeech 2013 Computational Paralinguistics Challenge* [13], an international benchmarking campaign aimed at the automatic detection of fillers in spontaneous speech streams. At the moment of the campaign, the personality assessments were not available and, hence, this is the first work that uses the data to perform *Automatic Personality Recognition* (APR) [4], i.e., the automatic inference of self-assessed traits from observable behavior.

The approach proposed in this work includes three main steps, namely *feature extraction*, *feature selection* and *classification*. The first step is performed using *openSMILE*, a feature extraction tool commonly adopted in experiments aimed at the inference of personality, emotions or other social and psychological constructs from speech [14], [15]. The tool provides a standard set of 384 features that cover a wide spectrum of acoustic properties. Given the dimensionality of the feature set, the approach includes a selection step based on Quantum Evolutionary Algorithms (QEA) [16], well known for their performance in combinatorial optimization problems. In particular, the QEA adopted in this work - the *Principal Component Analysis* QEA (PCA-QEA) - is original and it has been designed to concentrate the search efforts in those regions of the feature space where there is less certainty about whether the features should be selected or discarded. Finally, the classification is performed with eight standard classifiers.

Besides reducing the dimensionality of the feature vectors, the selection approach allows one to identify the features most likely to carry personality relevant information in the fillers. This is important because it provides insight about the relationship between personality and speech production hypothesized by Sapir and mentioned at the beginning of this section. In particular, Section 5.3 shows that a relatively small number of features (between 7 and 48 out of 384 depending on the traits) is selected at least 90% of the times during the multiple iterations of the feature selection approach used in the experiments. Compared to correlational analysis - the approach typically adopted in Psychology for such a purpose [17] - the main advantage is that the features are not considered individually, but as elements of subsets expected to maximize the classification performance. Thus, the selection approach provides better insights on how multiple speech characteristics jointly convey personality information.

The classification experiments have addressed two main problems. The first is to predict whether the speaker that has uttered *a given filler* is above median along the Big Five traits (accuracy up to 68.0% depending on the particular trait), the second is to predict whether the speaker that has uttered *a set of fillers* is above median along the same traits (accuracy up to 81.2% depending on the trait). These results seem to suggest that there is a relationship between personality and fillers.

To the best of our knowledge, the main novelties of this article are as follows:

- This is the first work showing that it is possible to infer the self-assessed personality of speakers from the way they utter fillers;
- This is the first work that identifies the physical characteristics of fillers that better account for the outcome of the classification approaches used in the experiments and, hence, account indirectly for the traits of the speakers;
- The classification approach includes an original feature selection methodology.

The rest of this work is organised as follows: Section 2 presents a survey of previous work, Section 3 describes the data used in the experiments, Section 4 illustrates the approach proposed in this article, Section 5 presents experiments and results and the final Section 6 draws some conclusions.

2 Survey of Previous Work

This section proposes a survey of previous work on the inference of personality traits and on the Quantum Evolutionary Algorithms aimed at feature selection.

2.1 Mapping Speech into Personality

According to the terminology proposed in [4], the approaches aimed at the inference of personality traits can be split into two major groups, namely those that address Automatic Personality Recognition (APR) - the inference of the traits that the speakers attribute to themselves - and those that address Automatic Personality Perception (APP) - the inference of the traits that the listeners attribute to the speakers. From a personality point of view, the main difference is that people self-assessing their own personality, unlike those that assess the personality of others, do not access only the information available in the speech signal, but also the rest of their experience, including aspects that are not directly accessible to the observation of others. The main consequence of such a difference is that, in general, the relationship between data and traits tends to be less consistent in APR than in APP. Hence, the performance achieved in the latter task tends to be higher than in the former one [4]. From a methodological point of view, APR and APP share the problem of inferring personality traits (self-assessed or assessed) from speech. However, there is a problem that must be addressed in APP and not in APR, namely the reliability of the assessments obtained through the involvement of multiple personality raters [18]. In particular, whenever the judgment of multiple raters is aggregated, it is necessary to ensure that they agree beyond chance.

Only a few APR works have used speech in a unimodal approach [19], [20]. In both articles, the goal of the experiments was to predict whether an individual is in the upper or lower half of the personality scores observed in the data, a binary task similar to the one performed in this article. The approach proposed in [20] combines both verbal and nonverbal aspects of speech, but the experiments, performed over the EAR Corpus, do not lead to accuracies higher than chance. In the case of [19], the experiments have been performed over the *PersIA* corpus (119 conversations involving 24 subjects). The features have been extracted with OpenSMILE [14] like in this work (see Section 5) and the accuracies are up to 95% in the case of Conscientiousness.

In other works [21], [22], [23], [24], [25], speech is combined with other behavioral cues and, in particular, with gestures detected automatically in videos. In these works, the APR task to be performed is a binary classification similar to the one proposed in this work and the features extracted from speech account for prosody (e.g., mean and standard deviation of formants, spectral entropy, autocorrelation peaks, energy, etc.) and speech activity (e.g., percentage of speaking time per subject, number and length of voiced segments, etc.). In the case of [23], [25], the data corresponds to 12 meeting recordings each including 4 different individuals. In the case of [22], the data is a collection of 89 self-presentations given via Skype. The accuracy achieved over the meetings goes up to 90% thanks to the large amount of information available in meeting recordings, but it is lower (65% to 75% depending on the trait) in the case of self-presentations.

The APP problem was addressed in a larger number of works [20], [26], [27], [28], [29], [30] and was the subject of an international benchmarking campaign based on a corpus of video blogs [31]. All proposed approaches include a feature extraction step that typically represents speech samples in terms of the same characteristics as those used for APR (see above). The extracted features are then mapped into personality scores using standard machine learning algorithms such as, e.g., Support Vector Machines. In most cases [20], [26], [28], [30], the actual recognition task corresponds to a binary classification similar to the one proposed in this work.

Speech based APP was the subject of the *Interspeech 2012 Speaker Trait Challenge* [32], an international benchmarking campaign during which several groups have tested their models over the same data [33], [34], [35], [36], [37], [38], [39], [40], [41]. The experiments of the challenge were performed over the SSPNet Personality Corpus, a collection of 640 speech samples (322 subjects in total) rated in terms of the Big-Five by 11 assessors. Overall, the most successful APP approaches appear to be those that apply feature selection methodologies to identify the physical characteristics of speech that better explain the perception of the raters. The importance of feature selection in APR tasks is supported also by Personality Computing competitions [42]. However, the performance changes significantly from one trait to the other. In particular, while Extraversion and Conscientiousness are predicted to a satisfactory extent, the other traits are recognised beyond chance, but with limited accuracy. Like in the case of this work, the goal of the experiments was to predict whether people score above median or not with respect to the Big-Five traits.

Overall, the state-of-the-art shows that most of the works about APP and APR propose binary classification tasks like the one addressed in this article (see Table 1). Furthermore, it shows that when it has been possible to perform rigorous comparisons across multiple approaches, those that adopt feature selection approaches tend to perform comparatively better (such a result has been observed for APP, but such a problem is methodologically similar to APR).

2.2 Quantum Evolutionary Algorithms

By reducing the dimensionality of data, feature selection has an important role in the performance of machine learning algorithms [43]. The task is the optimization process of finding the optimal subset of features that offer the best performance for machine learning algorithms. A variety of optimization algorithms have been applied to feature selection, including complete search, greedy search, heuristic search, and random search [44], [45], [46], [47]. However, most of existing feature selection methods are prone to stagnation in local optima [48]. Because of their global search abilities, evolutionary algorithms have recently gained much attention [49].

The feature selection methodologies can be grouped into two major categories, namely *filter* and *wrapper* approaches. In the first case, the classifiers are not involved in the selection process and the focus is on the identification of features that are redundant with respect to the others through measures like, e.g., the correlation or the covariance. Filter approaches tend to be fast and to have a low computational burden, but they result in feature subsets that are not adapted to any classifier in particular. As a consequence, the performances tend to be lower, on average, than those achieved with wrapper methods. These latter use the performance that a classifier achieves using a feature subset as a criterion to retain or discard a feature. In this way, the subset of the selected features changes from one classifier to the other and, in general this leads to higher classification performances [50]. On the other hand, wrapper methods tend to be slower and computationally heavier than filter ones.

The main difficulty in feature selection is that the features interact with one another, a phenomenon called *epistasis* [51]. This means, for example, that a feature that is not discriminative individually can significantly improve its contribution to the the classification performance when it is used in conjunction with other features. Similarly, a feature that is discriminative individually, can become redundant when used jointly with other

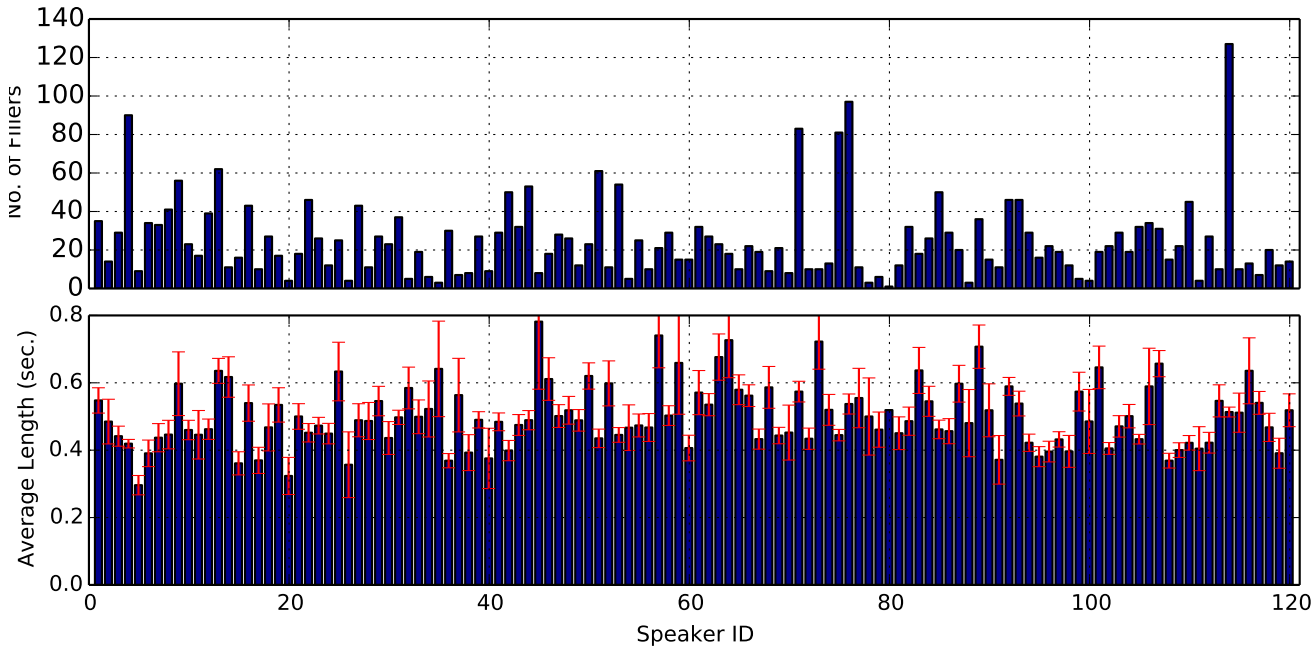

Fig. 1. The upper chart shows the number of fillers uttered by every speaker in the corpus. The lower chart shows the average length of the fillers uttered by every speaker (the error bars correspond to the standard errors).

Ref.	Subj.	Samples	Features	Task	Ext.	Agr.	Con.	Neu.	Ope.	Other
[19]	12	119 conversations	OpenSMILE speech features	C(2)	63.0 ACC	56.3 ACC	95.0 ACC	32.8 ACC	40.3 ACC	
[20]	96	96 conversation transcripts	prosody LIWC, MRC	C(2)	57.3 ACC	58.3 ACC	53.2 ACC	50.4 ACC	61.4 ACC	
[21]	43	4 collaborative tasks per subject	prosody, turn takings motion activity	C(2)	81.4 ACC	69.8 ACC	69.8 ACC	81.3 ACC	60.5 ACC	
[22]	89	89 self presentations	prosody, posture, face/hand/head movements	C(2)	70.8 ACC	65.2 ACC	73.0 ACC	76.4 ACC	66.3 ACC	
[25]	48	12 meetings of 4 persons	prosody, speech activity body movements	C(3) C(3)	94.4 85.0 ACC					ACC for LOC = 94.9 ACC for LOC = 86.0

TABLE 1

APR and nonverbal communication. The table (included from the survey in [4]) shows the main APR works based on speech and/or nonverbal communication. The columns contain, from left to right, the number of participants involved in the experiments, number and type of behavioural samples, main cues, type of task and performance over different traits. The column "*Other*" refers to works using models different from the Big-Five. C(n) for Classification with n classes, LOC for Locus of Control and ACC for accuracy (percentage of correctly classified samples).

features. Therefore, any method that evaluates features individually is very unlikely to find the optimal subset of features. This means that it is necessary to perform a global search if one intends to find the optimal subset of features. Note, however, that this optimal subset depends to a significant extent on the evaluation criterion and on the classification algorithm. Thus, any optimal subset that is found for a particular classifier works best only for that particular classifier and most probably is not the best subset for other classifiers.

Many selection approaches have used evolutionary algorithms, including Genetic Algorithms [52] and Genetic Programming [53], particle swarm optimisation [48] or ant colony [54]. Other global search algorithms recently used for feature selection include defferential evolution [55], memetic algorithms [56], learning classifier systems [57] and artificial immune systems [58].

3 The Data

The experiments of this work have been performed over $2,988$ fillers extracted from the *SSPNet Vocalization Corpus*, a publicly available dataset used for the *Interspeech Computational Paralinguistics Challenge* [13]. The benchmarking campaign was aimed at the automatic detection of vocalizations in a speech stream and the personality assessments were not available (see below for more details). Thus, this is the first work that uses the

data for Automatic Personality Recognition and, to the best of our knowledge, it is the first work that uses fillers to perform such a task. The extraction of the fillers has been performed manually. An annotator has identified the time boundaries of every filler and has provided the resulting audio segment to two other annotators. These have validated the segment or asked to change the boundaries depending on whether the filler was segmented correctly or not.

The fillers have been extracted from 60 dyadic conversations between unacquainted individuals (see [4], [7] for a full description of the data) for a total of 120 participants (63 female and 57 male), all native English speakers of British nationality. The conversations are based on the *Winter Survival Task* (WST) [59]: The participants are said to be part of a rescue team that will assist the survivors of a plane crash in a polar area. In particular, the participants are given a list of 12 items[1] that the survivors have found in the area of the accident and the goal of the conversation is to identify those that are most likely to be helpful while the survivors move from the place of the crash to a point where they can be rescued. The participants are asked to cooperate and provide their suggestions as quickly as possible because it is dangerous for the survivors to remain in the area of the crash.

The total number of fillers is $2,988$, corresponding to an average of 24.9 samples per subject. The average duration of the samples is 502 ms with a standard deviation of 262 ms. Figure 1 shows the distribution of the number of samples across the subjects and the average duration of the fillers for every subject. Overall, female and male subjects have uttered $1,297$ and $1,691$ fillers, respectively (the averages are 20.6 for female speakers and 29.7 for male ones.). According to a χ^2 test, the difference is statistically significant ($p < 10^{-12}$) meaning that the male subjects, on average, tend to utter fillers more frequently than the female ones.

Each of the 120 subjects included in the corpus has filled the *Big-Five Inventory 10* (BFI-10) [60], a 10-items questionnaire aimed at personality self-assessment in terms of the Big-Five traits [5] (see Table 2). As a result, it is possible to know, for every subject, the five scores corresponding to the Big-Five traits, namely *Openness* (the tendency to be intellectually curious and open), *Conscientiousness* (the tendency to be planful and reliable), *Extraversion* (the tendency to be socially active and assertive), *Agreeableness* (the tendency to do what others appreciate) and *Neuroticism* (the tendency to experience the negative side of life). The Big-Five is the most commonly applied personality model - both in computing [4] and psychology [61] - and it is particularly suitable for technology because it represents personality as a 5-dimensional vector, thus allowing the application of statistical approaches like those adopted in this work.

1. Steel wool, axe, pistol, butter can, newspaper, lighter without fuel, clothing, canvas, airmap, whisky, compass and chocolate.

ID	Trait	Question
1	Ext.	I am reserved
2	Agr.	I am generally trusting
3	Con.	I tend to be lazy
4	Neu.	I am relaxed, handle stress well
5	Ope.	I have few artistic interests
6	Ext.	I am outgoing, sociable
7	Agr.	I tend to find fault with others
8	Con.	I do a thorough job
9	Neu.	I get nervous easily
10	Ope.	I have an active imagination

TABLE 2
The BFI-10 questionnaire used in the experiments of this work. The version reported here is the one that has been proposed in [60].

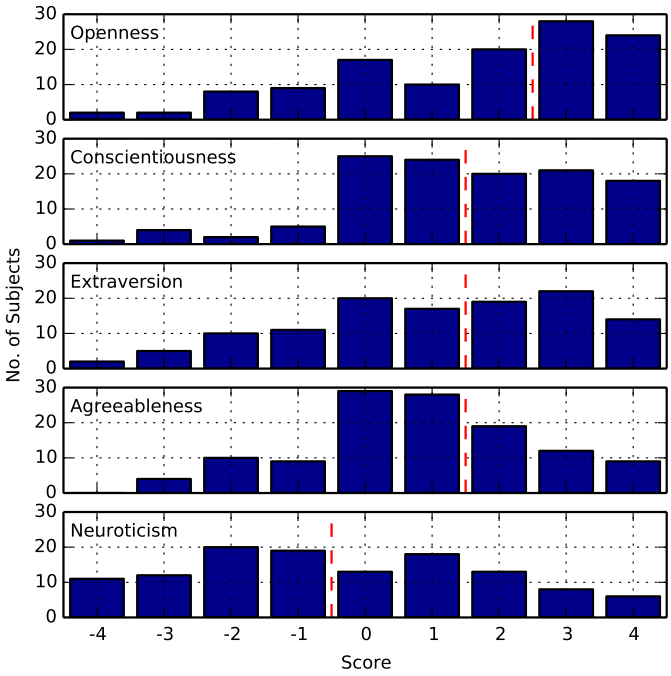

Fig. 2. The chart shows the distribution of the scores for each of the Big Five Traits. The vertical dashed line corresponds to the median and separates the two classes, namely *low* (left of the line) and *high* (right of the line).

Figure 2 shows the distribution of the trait scores across the 120 participants of the experiment.

4 The Proposed Approach

The proposed approach includes three main steps, namely *feature extraction* (see Section 4.1), *feature selection* (see Section 4.2) and *classification* (see Section 4.3).

4.1 Feature Extraction

The proposed approach extracts the features with *OpenSMILE* [14], [15], a publicly available tool commonly adopted for the inference of social and psychological constructs from speech. OpenSMILE applies the

methodologies typical of *computational paralinguistics* [3], namely it converts a speech sample into a sequence $\mathcal{Y} = (\vec{y}_1, \ldots, \vec{y}_T)$ of short-time feature vectors and, then, it estimates the statistical properties of the short-term features to build a vector $\vec{x}$ that represents the sample as a whole. The short-term feature vectors $\vec{y}_k$ are extracted from analysis windows that must be long enough to allow a reliable extraction of the features, but short enough to ensure that the signal properties are stable. The literature shows that, in the case of speech, windows of length between 20 and $40\ ms$ lead to satisfactory results (see [3], page 188). Thus, the approach proposed in this work adopts $25\ ms$ long windows. Similary, the literature suggests that a *frame rate* - number of short-term vectors $\vec{y}_k$ extracted per second - suitable for speech is 100, meaning that the windows must start at regular time steps of $10\ ms$ (*Ibidem*). Thus, the approach proposed in this work adopts such a rate for the experiments. The only feature that is extracted from the filler as a whole without passing through the process above is the duration. The reason is that such a measure is not a short-term property and can only be measured taking into account the whole sample.

The short-term feature vectors $\vec{y}_k$ include 16 features with their respective *delta regression coefficients* [14], [15], for a total of 32 features. The 16 features are the *Root Mean Square* (RMS) of the energy, the first 12 *Mel Frequency Cepstrum Coefficients* (MFCC), the *Zero Crossing Rate* (ZCR), the *Voicing Probability* (VP) and the *fundamental frequency* or *pitch* ($F0$). All features have been smoothed, meaning that the value of a feature extracted from window k is replaced with the average of the feature values extracted from windows $k-1$ to $k+1$ (the delta regression coefficients have been extracted after that the features have been smoothed).

The MFCCs have been included because they are well known to capture information about the energy (coefficient 1) as well as about the phonetic content of the signal (coefficients 2 to 12) [62], [3]. This allows one to investigate whether the particular type of filler being uttered - e.g., the use of different vowels like in "*eh*" or "*uh*" or the presence of a final consonant like in "*ehm*" or "*uhm*" - has a relationship with the speaker's traits. Root Mean Square of the energy, $F0$ and length of the filler account for the *Big Three* of prosody, namely loudness, pitch and tempo, respectively. Prosodic features have been widely applied in *Personality Computing* [4] and they have the advantage of being controlled - at least when it comes to loudness and tempo - by the speaker. Hence, they can provide information about the speaking style. The remaining features (ZCR and VP) provide information about the possible presence of unvoiced segments in the filler [63], [64].

For each of the 32 short-term features described above, the approach estimates 12 statistical properties, thus resulting into a 385-dimensional vector - the 385^{th} is the duration, for which no statistical properties are estimated because there is only one value. The statistical properties are minimum, maximum, range (difference between maximum and minimum), position of the window where the maximum value has been extracted, position of the window where the minimum value has been extracted, arithmetic mean, slope of the linear approximation of the contour, offset of the linear approximation of the contour, difference between linear approximation and actual contour, standard deviation, skewness (third order moment) and kurtosis (fourth order moment minus three).

z_i	b_i	$f(z) \geq f(b)$	$\Delta\theta$
0	0	true	0
0	1	true	0
1	0	true	0
1	1	true	0
0	0	false	0
0	1	false	$\beta\pi$
1	0	false	$-\beta\pi$
1	1	false	0

TABLE 3
Calculation of $\Delta\theta$. The z_i is the i-th element of a solution in Θ and b_i is the i-th element of the best solution until iteration t.

4.2 Feature Selection

The goal of the feature selection step is to identify a subset of the original feature set F - a *solution* hereafter - that allows one to achieve the highest possible performance while including the smallest possible number of features. A solution can be represented as a D-tuple $z = (z_1, \ldots, z_D)$ of binary numbers, where D is the dimension of the original feature vectors, $z_k = 1$ if the k^{th} feature has been retained and $z_k = 0$ otherwise.

The selection approach proposed in this work is based on *Quantum Evolutionary Algorithms* (QEA) [16], [65]. These represent all possible solutions with the help of two elements, namely a D-tuple $\theta = (\theta_1, \theta_2, \ldots, \theta_D)$ - called *quantum individual* - and an operator O - called the *Observation Operator*. The components $\theta_k \in [0, \pi/2]$ are angles such that $p(z_k = 1) = \cos^2\theta_k$, $\forall k \in [1, \ldots, D]$. The operator O, when applied to a quantum individual, produces a solution z by assigning every z_k value 1 or 0 with probabilities $\cos^2\theta_k$ and $\sin^2\theta_k$, respectively. In this way, a quantum individual is sufficient to generate all 2^D solutions that can result from a selection process.

The main characteristic of QEA based selection processes is that they can be modeled as a search through the space of quantum individuals. The main novelty of the approach proposed in this work is that it adopts the *Principal Component Analysis* (PCA) to identify the directions of such a space along which the search is more worth (see below for more details). Given the important role of the PCA, the approach has been called PCA-QEA.

The PCA-QEA is an iterative process and the first step, aimed at initialization, starts with the creation of a set

Algorithm 1 PCA Quantum-Evolutionary Algorithm

1: **procedure** PCA-QEA ITERATION
2: Generate Z^{t+1} by applying the O operator to the quantum individuals of Θ^t;
3: Evaluate Z^{t+1};
4: Store the best solution that each quantum individual has generated into B^{t+1};
5: Obtain the set $H^{t+1} = H^t \cup Z^{t+1}$;
6: Rank the solutions of H according to their classification performance and keep in H only the $M = 50$ top ranking solutions;
7: Apply PCA to the elements of H;
8: Update Θ^t taking into account the results of the PCA;
9: Go back to step 2 unless the termination condition has been reached.

$\Theta^{(0)} = \{\theta_1^{(0)}, \ldots, \theta_N^{(0)}\}$ in which the components of the quantum individuals $\theta_k^{(0)}$ are set to $\pi/4$ so that every feature is selected or discarded with probability 0.5 ($N = 20$ in the experiments of this work). The application of the operator O to each of the N quantum individuals generates a set $Z^{(0)} = \{z_1, \ldots, z_N\}$ of binary solutions - one per quantum individual - that can be *evaluated*, i.e., that can be used to perform a task with performances $f(z_1), \ldots, f(z_N)$, respectively. In the experiments of this work, the task is a classification and the performance is measured in terms of *accuracy* (the percentage of times that the classification is correct). The last two tasks of this step are the initialization of a set H - expected to contain the history of the best solutions - to the empty set $\emptyset$, and the creation of a set $B^{(0)} = Z^{(0)}$. At every step t, the set $B^{(t)}$ includes the best solutions - meaning that they lead to the highest performances - that every quantum individual has generated until step t. The best solution in $B^{(t)}$ is called b and, at the initialization step, it corresponds to the best solution in $B^{(0)}$.

The transition between iterations $t-1$ and t takes place by evaluating the solutions in $Z^{(t-1)}$ and by changing the components of the quantum individuals in $\Theta^{(t-1)}$ by a value $\Delta\theta$ calculated according to the rules in Table 3, thus resulting into the set $\Theta^{(t)}$. The rationale behind the rules of Table 3 is that the quantum individuals generating solutions that perform better than b must not be changed (see upper part of the table), while the others have to be changed so that they are more likely to generate solutions similar to b (see lower part of the table).

Once the set $\Theta^{(t)}$ is available, the iteration follows the steps of the pseudocode described in Algorithm 1. The application of O allows one to generate a set $Z^{(t)}$ of solutions that can be evaluated. These can then be used to update $B^{(t-1)}$ to $B^{(t)}$ and to set $H = H \cup Z^{(t)}$. The subset of the best M solutions in H ($M = 50$ in the experiments of this work) is then analyzed with PCA to find the directions that account for the largest variance. When the variance concentrates on a relatively small number of Principal Components, it means that most components in the best solutions tend to be stable. In other words, the selection approach has reached a local minimum where most features tend to be always retained or always discarded [66], [67]. In such a condition, further exploration of the solutions' space can be effective - meaning that the uncertainty above can be removed - only if another local minimum can be reached. For this reason, the core-idea of the PCA-QEA, is to ensure that the search through the solutions' space tends to explore to a finer scale the directions that the PCA shows to carry most of the variance [68], [69].

In this respect, the most important difference with the standard QEA is that the PCA-QEA does not apply the rules of Table 3 with the same $\Delta\theta$ for all components, but takes into account the results of the PCA to explore the directions along which there is most variance. In particular, the PCA-QEA randomly selects one of the Principal Components according to the following probability distribution:

$$p(\epsilon_k) = \frac{\epsilon_k}{\sum_{j=1}^{D} \epsilon_j}, \tag{1}$$

where ϵ_k is the eigenvalue associated to the k^{th} Principal Component ν_k, i.e., the data variance along the direction identified by ν_k. Following such a distribution, the eigenvectors along which the variance is larger tend to be chosen more frequently. Once a Principal Component ν_i has been selected, the k^{th} components of the vectors in Θ^t are modified by a value $\Delta\omega_k$ calculated as follows:

$$\Delta\omega_k = \Delta\theta \times (1 - u_k) + \frac{\Delta\theta}{2}, \tag{2}$$

where the value u_k corresponds to the following:

$$u_k = \frac{|\nu_{ik}| - \min_{j=1}^{D} |\nu_{ij}|}{\max_{j=1}^{D} |\nu_{ij}| - \min_{j=1}^{D} |\nu_{ij}|}, \tag{3}$$

where $|\nu_{ij}|$ is the absolute value of the j^{th} component in ν_i. The use of $1 - u_k$ in Equation (2) ensures that the update is smaller in those directions along which the components of the eigenvectors are larger. The reason for such a choice is that it is important to explore at a finer scale those directions along which the variance is larger, because these are the directions corresponding to features about which the algorithm is uncertain. Hence, it is along these directions that it is possible to find the minima where there is no uncertainty about the corresponding features.

4.3 Classification

The classification step is based on eight classifiers, namely the *Cascade Forward Neural Network* (CFNN) [70], the *Feed Forward Neural Networks* (FFNN) [71], the *Fuzzy Neural Networks* (FNN) [72], the *Generalized Regression Neural Networks* (GRNN) [73], the *k Nearest Neighbors* (kNN) [74], the *Linear Discriminant Function* [74], the *Naïve Bayes Classifier* (NB) [74], and the *Support Vector Machines* (SMV) [75]. While the goal of the work is to predict whether a person is above median or not along the Big-Five traits, the actual classification takes place

Trait	Ope.	Con.	Ext.	Agr.	Neu.
$p(low)$ fillers	56.7%	54.2%	51.7%	66.7%	50.8%
$p(high)$ fillers	43.3%	45.8%	48.3%	33.3%	49.2%
$\hat{\alpha}$ fillers	50.9%	50.3%	50.0%	55.5%	50.0%
$p(low)$ subjs.	56.7%	50.8%	54.2%	66.7%	51.7%
$p(high)$ subjs.	43.3%	49.2%	45.8%	33.3%	48.3%
$\hat{\alpha}$ subjects	50.9%	50.1%	50.3%	55.5%	50.0%

TABLE 4
Class Distribution. The table shows the a-priori probabilities of class high and low for different traits, both for individual fillers and subjects. Furthermore, the table shows the accuracy of a baseline classifier that assigns a sample (filler or subject) to a class according to the a-priori probabilities.

at the level of individual fillers. The main advantage of such an approach is that the number of samples at disposition for performing feature selection and training is significantly larger. In fact, the number of subjects is 120, but the number of fillers is $2,988$ (see Section 3).

If $\Phi = \{\phi_1, \ldots, \phi_L\}$ is the set of the L fillers uttered by a given speaker, then the classification at the level of the subject is done through a majority vote (a subject is assigned to the class that her or his fillers are more frequently assigned to):

$$c^* = \arg\max_{c \in \mathcal{C}} |\{\phi_k : c(\phi_k) = c\}|, \tag{4}$$

where c^* is the class actually assigned to the subject, $\mathcal{C}$ is the set of all predefined classes (above median and below or equal to median in the experiments of this work), $|.|$ is the cardinality of a set, and $c(.)$ is the classifier mapping the fillers into one of the classes $c \in \mathcal{C}$. If there is a tie, the subject is assigned to one of the classes according to the respective a-priori probabilities.

5 Experiments and Results

Every speaker is either above median (class *high*) or not (class *low*) along the Big-Five traits. Therefore, it is possible to perform, for each trait, the following two main tasks:

- Filler Classification: to infer the class of a speaker from one individual filler (see Section 5.1);
- Speaker Classification: to infer the class of a speaker from the set of all the fillers that she or he has uttered (see Section 5.2).

Both tasks can be considered a form of Automatic Personality Recognition because they both allow one to infer information about the traits of a speaker. In addition to the tasks above, the selection approach allows one to identify the features most likely to increase the classification accuracy and, hence, most likely to carry personality relevant information (see Section 5.3).

Table 4 shows the distribution over the classes for both fillers and speakers. Correspondingly, the table shows the accuracy $\hat{\alpha}$ (percentage of times that the classification is correct) of a random classifier that assigns a sample to class c with probability p_c (p_c is the *a-priori* probability of c):

$$\hat{\alpha} = \sum_{c \in \mathcal{C}} p_c^2, \tag{5}$$

where $\mathcal{C}$ is the set of the predefined classes. Such an accuracy is used as a baseline to test whether the proposed approach performs better than chance.

5.1 Filler Classification Results

The fillers adopted in the experiments have been uttered by 120 individuals involved in 60 dyadic conversations (see Section 3). This allows the adoption of a *leave-one-conversation-out* experimental setup: the fillers uttered by the two subjects involved in a given conversation are used as a test set while all of the others are used to train the classifiers and to perform the feature selection. The process is iterated 60 times and, at each iteration, a different conversation is left out as a test set. The main advantage of such a setup, inspired by the leave-one-out approach, is that it allows one to perform tests over the whole dataset at disposition while still keeping a rigorous separation between training and test sets. Furthermore, in the case of the experiments of this work, the setup has the advantage of being *speaker independent*, meaning that none of the subjects is represented in both training and test set. This ensures that the approach recognizes the personality traits and not the voice of the speakers.

Table 5 shows the results that have been obtained over the $2,988$ fillers both with and without feature selection. The maximum accuracy is above 60% for all traits except Agreeableness where it is 59.3% (a possible reason is that the distribution of the samples over the two classes is more unbalanced for such a trait than for the others). Furthermore, the accuracy α of the systems that include the feature selection step is always higher, to a statistically significant extent, than the corresponding $\hat{\alpha}$ values in Table 4. Therefore, it is possible to say that the relationship between the physical characteristics of the fillers and the personality traits is consistent enough to allow the automatic inference of the latter from the former (to an extent that is better than chance to a statistically significant extent). A possible interpretation of such an observation is that people with different personality traits tend to utter the fillers in a different way and the difference is sufficiently consistent to allow the inference of the traits above chance.

The application of the feature selection step reduces to a statistically significant extent the accuracy of a system using the full feature set only in one case (FNN for Openness). In contrast, there is a statistically significant improvement in 17 cases out of 40 (see Table 5). These observations confirm that the feature selection approach is effective in discarding the features that do not carry relevant information while retaining those that allow the

Classifier	O(S)	O(F)	C(S)	C(F)	E(S)	E(F)	A(S)	A(F)	N(S)	N(F)
CFNN	54.7*	50.8	62.8**	57.8	59.5	56.3	58.1**	53.4	57.6**	52.9
FFNN	56.6**	49.9	62.8**	55.0	60.9**	51.6	57.0**	51.7	58.7	57.6
FNN	52.2	55.7*	60.4	59.7	60.1	58.0	57.7**	51.7	60.9	57.9
GRNN	55.6*	52.1	57.8	54.8	56.0	54.4	58.7*	54.9	57.4	54.1
KNN	55.7	54.3	58.7	57.3	57.2	54.8	**59.3**	56.6	57.0	55.0
LDF	**63.4****	50.8	**67.8****	60.3	**64.8****	56.3	58.9**	54.0	**63.3**	**60.2**
NB	59.4	**57.2**	62.9	**62.1**	54.1**	48.8	59.1	56.6	58.5	57.1
SVM	56.1	54.8	62.2**	57.5	62.1	**59.4**	58.0	**57.3**	59.7	57.1
$\hat{\alpha}$	50.9%	50.9%	50.3%	50.3%	50.0%	50.0%	55.5%	55.5%	50.0%	50.0%

TABLE 5

Filler classification. The table reports the accuracies obtained over individual fillers. In the column titles, the letter "S" stands for "selection" (the results have been obtained by applying the PCA-QEA) and the letter "F" stands for "full feature set" (the results have been obtained without applying the feature selection). The acronyms of the first column stand for Cascade Forward Neural Network (CFNN), Feed Forward Neural Networks (FFNN), Fuzzy Neural Networks (FNN), Generalized Regression Neural Networks (GRNN), k Nearest Neighbors (kNN), Linear Discriminant Function (LDF), Naïve Bayes Classifier (NB) , and Support Vector Machines (SVM). The double and single stars mean that the accuracy after the selection is higher than the accuracy without feature selection with 99% and 95% confidence level, respectively (according to a two-tailed t-test with Bonferroni correction). The t-tests have been performed according to the approach proposed in [76] to take into account the dependence across the the multiple fillers uttered by the same subject. The accuracy written in bold is the highest in the column. The last row shows the baseline accuracy $\hat{\alpha}$ (see Table 4).

Classifier	O(S)	O(F)	C(S)	C(F)	E(S)	E(F)	A(S)	A(F)	N(S)	N(F)
CFNN	58.3**	42.5	**78.3****	59.2	68.3	65.8	55.8	51.7	67.5*	55.0
FFNN	64.2*	51.7	73.3**	54.2	69.2**	51.7	63.3**	45.0	60.0	62.5
FNN	47.5	57.5	69.2	65.8	65.0	63.3	65.0**	45.8	**70.8**	62.5
GRNN	60.8*	48.3	64.2	53.3	62.5	57.5	**71.7***	60.0	61.7	58.3
KNN	60.0	56.7	65.0	64.2	62.5	53.3	61.7	59.2	60.0	51.7
LDF	**73.3****	42.5	77.5	76.7	**75.0***	61.7	66.7**	52.2	70.0	70.0
NB	70.0	64.2	67.5	70.0	55.8*	43.3	65.8	58.3	65.0	61.7
SVM	62.5	58.3	70.8	61.7	72.5	65.8	65.0	61.7	62.5	67.5
$\hat{\alpha}$	50.9%	50.9%	50.1%	50.1%	50.3%	50.3%	55.5%	55.5%	50.0%	50.0%

TABLE 6

Speaker Classification. The table reports the accuracies obtained over the 120 speakers by applying a majority vote over all fillers they uttered. In the column titles, the letter "S" stands for "selection" (the results have been obtained by applying the PCA-QEA) and the letter "F" stands for "full feature set" (the results have been obtained without applying the feature selection). The acronyms of the first column stand for Cascade Forward Neural Network (CFNN), Feed Forward Neural Networks (FFNN), Fuzzy Neural Networks (FNN), Generalized Regression Neural Networks (GRNN), k Nearest Neighbors (kNN), Linear Discriminant Function (LDF), Naïve Bayes Classifier (NB) , and Support Vector Machines (SVM). The double and single stars mean that the accuracy after the selection is higher than the accuracy without feature selection with 99% and 95% confidence level, respectively (according to a two-tailed t-test with Bonferroni correction). The accuracy written in bold is the highest in the column. The last row shows the baseline accuracy $\hat{\alpha}$ (see Table 4).

classifiers to perform better than chance. Therefore, the features that are selected more frequently can reliably be considered as personality *markers*, i.e., as physical and machine detectable externalizations of the personality.

The Linear Discriminant Function is the classifier that, in combination with the feature selection approach, appears to consistently outperform the others for all traits. In the case of Agreeableness, where the top performing classifier is kNN, the accuracy difference with respect to LDF is not statistically significant ($p > 0.05$ according to a two tailed t-test). One possible interpretation of such an observation is that the LDF is more deterministic than the others. This allows the selection process to measure the fitness of the solutions more accurately and, correspondingly, to limit the noise that can decrease the effectiveness of the search process.

Section 3 shows that there are two orders of magnitude between the minimum and the maximum number of fillers uttered by a given speaker in the corpus. Given a particular classifier and a particular trait, it is possible to obtain $L = 120$ pairs (n_i, α_i), where n_i is the number of fillers that speaker i has uttered, α_i is the accuracy

achieved over the fillers of speaker i and L is the total number of speaker. The Spearman correlation coefficient - more robust to the outliers than the Pearson coefficient - has been estimated for each of the combinations between a classifier and a trait that have been tested in Table 5. The results show that the coefficient is statistically significant ($p < 0.05$) only in 2.5% of the cases (after Bonferroni correction). This seems to suggest that, overall, the accuracy does not depend on the number of fillers at disposition for a given speaker. This is important in view of the application of a majority vote aimed at classifying the speakers rather than the individual fillers (see Section 5.2).

5.2 Speaker Classification Results

Table 6 shows the results that can be obtained by applying a strict majority rule (when there is a tie, a subject is considered to be wrongly classified). Like in the case of the individual fillers, the feature selection never reduces the accuracy of the approach to a statistically significant extent. The LDF accuracy is the highest for two traits (Openness and Extraversion), while it is within a statistical fluctuation from the highest accuracy for the other traits. Therefore, it is possible to say that the combination between the selection approach and the LDF remains the most effective at the level of the speakers as well.

The accuracy is above 70% for all traits, thus confirming that, unlike most previous approaches in the literature, the inference of the traits from the fillers leads to satisfactory results along all the traits rather than along only some of them. This appears to confirm that the fillers carry personality relevant information and can act as reliable markers, at least in the conversation scenario targeted in the experiments. When it comes to the traits, the highest accuracies are observed, like in the case of the individual fillers, for Openness and Conscientiousness. However, in the case of the subjects, the difference between the highest and lowest accuracies - 78.3% and 70.8%, respectively - is not statistically significant.

The results of Table 6 have been obtained by applying a strict majority rule, i.e., by making a decision only when there is not a tie between the number of fillers assigned to class low and the number of fillers assigned to class high. However, when there is a tie, it is still possible to assign the subjects to one of the two classes - high or low - according to the a-priori probabilities of Table 4 (estimated over the training set). In this respect, the results of Table 6 can be considered as a lower bound of the accuracy that can be achieved through the aggregation of the decisions made at the level of the individual fillers.

Table 7 shows the average accuracies obtained after 100 repetitions of the experiment (given that there is a stochastic component, the accuracy changes from one repetition to another). The results are similar to those of Table 6: for every trait, the LDF accuracy is either the highest or it is within a statistical fluctuation with respect to the highest accuracy. Furthermore, the accuracy is below 75% only in the case of Agreeableness, the trait for which the class distribution is the most unbalanced (see Table 4).

Section 3 shows that there is a difference between female and male subjects in terms of number of fillers (female subjects tend to utter less fillers). According to a t-test with Bonferroni correction, there is only one case (Neuroticism with FFNN) in which the difference is statistically significant. This seems to suggest that the approach performs in the same way over both female and male subjects and the observable differences between the fillers uttered by subjects of different gender (in particular the length that tends to be lower for male subjects) do not play a role in APR.

In the majority vote, every speaker is assigned to the class that her or his fillers are most frequently assigned to. Therefore, it is possible to measure the correlation between the percentage of fillers such a class is assigned to and the trait scores of the speakers before the binarization (see Figure 2). In this way, it is possible to test whether the fillers of the speakers that are at the extreme of the scales tend to be assigned more consistently than the others to the winning class. The results show that, for any trait and any classifier, the correlation is not statistically significant. This seems to suggest that the effectiveness of the majority vote does not change with the trait scores. In other words, the speakers that are at the extremes of the scales are not classified with an effectiveness different from the others.

5.3 Feature Selection and Personality Markers

On average, the selection process retains half of the original features and this suggests that the speakers externalize their personality through large numbers of markers. However, there are features that are selected - for a given trait - with higher probability across the multiple classifiers and this suggests that they are more likely to act as personality markers. For this reason Figure 3 shows, for every feature and trait, how frequently a feature is selected during the application of the PCA-QEA. The main pattern that can be observed is that the delta regression coefficients (the features with the suffix "*de*" in the figure) tend to be selected less frequently than the others. One possible explanation is that these features are expected to capture temporal variations, but the fillers tend to be uttered as prolonged vowels in which the speech properties remain stable and, hence, no major variations are observed. The main exceptions with respect to such a general pattern can be observed for Extraversion, where the delta regression coefficients appear to be selected more frequently in the case of energy ("*pcm-RMSEnergy*" in the figure), voicing probability ("*voiceProb*") and Fundamental frequency ("*F0*"). One possible explanation is that the speakers externalize their Extraversion through the variability along such dimensions - correlational analysis suggests that the most

Classifier	O(S)	O(F)	C(S)	C(F)	E(S)	E(F)	A(S)	A(F)	N(S)	N(F)
CFNN	62.0**	45.8	**81.2****	62.9	70.1	70.0	58.7	55.9	70.8**	56.7
FFNN	65.9*	53.8	77.0**	57.5	70.9**	53.8	66.6**	47.5	62.1	66.2
FNN	52.1	59.6	72.5	67.5	67.9	64.5	67.9**	49.1	72.9	66.2
GRNN	65.4*	52.9	67.1	57.5	65.0	60.8	**73.4****	63.7	64.6	59.5
KNN	63.3	58.8	66.7	65.4	64.6	57.0	65.4	62.1	63.3	54.6
LDF	**76.2****	47.9	79.9	78.8	**75.4***	64.2	68.8**	55.5	**75.0**	72.5
NB	73.3	65.9	71.2	73.7	58.3**	45.8	67.5*	62.5	68.7	64.2
SVM	64.6	60.8	72.9	65.0	**75.4**	67.9	67.1	63.8	65.0	69.6
$\hat{\alpha}$	50.9%	50.9%	50.1%	50.1%	50.3%	50.3%	55.5%	55.5%	50.0%	50.0%

TABLE 7

Classification Results. The table reports the average accuracies obtained by assigning the subjects for which there is a tie to one of the two classes according to their a-priori probabilities. In the column titles, the letter "S" stands for "selection" (the results have been obtained by applying the PCA-QEA) and the letter "F" stands for "full feature set" (the results have been obtained without applying the feature selection). The double and single stars mean that the accuracy after the selection is higher than the accuracy without feature selection with 99% and 95% confidence level, respectively (according to a two-tailed t-test with Bonferroni correction). The accuracy written in bold is the highest in the column. The last row shows the baseline accuracy $\hat{\alpha}$ (see Table 4).

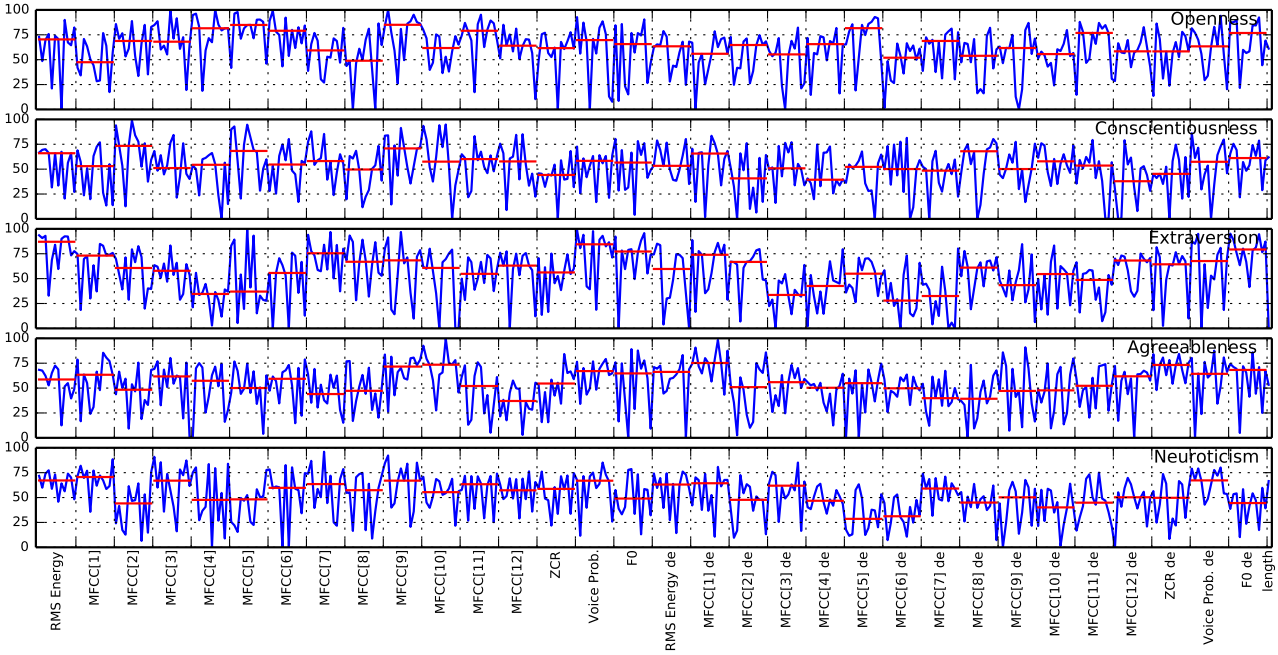

Fig. 3. Selection probability. The plots show the probability of every feature to be selected. The red horizontal lines show the median of the selection probabilities corresponding to the 12 statisticals associated to every feature extracted from the speech signal.

extraverted subjects tend to display more variability - but the same variability does not change consistently with the other traits (the relationship between Extraversion and energy has actually been observed earlier in the literature [77]).

Another observable pattern is that the first two MFCCs tend to be selected more frequently for Conscientiousness and Neuroticism than for the other traits. This seems to suggest that energy (related to the firts MFCC) and the particular vowel a filler corresponds to - the second to twelfth MFCCs change with the phonemes being uttered (see [3], page 198) - act as a personality marker.

For what concerns the Big Three of prosody - pitch ("*F0*" in the figure), loudness (corresponding to "*pcm-RMSenergy*" in the figure) and tempo ("*length*" in the figure) - Figure 3 shows that the first plays an important role in the case of Conscientiousness, Extraversion and Neuroticism, the second interplays significantly with Extraversion, Agreeableness and Neuroticism, while the third is not selected frequently for any of the traits. This is in line with the previous results of the literature

(see [4] for an extensive survey) where features related to prosody are often successfully applied in Automatic Personality Recognition. One possible explanation is that prosodic features can be controlled, up to a certain extent, by the speakers and then they are more likely to act as personality markers than other features that depend on anatomy and therefore less dependent on the speaking style of the speaker.

The more frequently a feature is selected, the more it is likely to carry personality relevant information that allows the classifiers to achieve a high accuracy. In other words, the subset of the features that are selected more frequently is likely to include the most reliable personality markers. Therefore, in addition to the general patterns outlined above, it is possible to analyze what are the most frequently selected features for all traits. In particular, the subset of the features that are retained at least 90% of the times is appears to be different for the various traits, both in terms of size and of elements (48 for Openness, 13 for Conscientiousness, 39 for Extraversion, 13 for Agreeableness and 7 for Conscientiousness).

In the case of Openness, the features that are selected at least 90% of the times include mainly the statisticals of the MFCCs between 2 and 11 (41 out of the 48 elements of the subset). This seems to suggest that the phonetic content of the fillers is the main marker for the trait. The other features of the subset correspond to the voicing probability, the fundamental frequency and their respective delta regression coefficients. This suggests that variations of the intonation and the voice emission are an externalization of Openness. In the case of Conscientiousness, the subset of the features selected at least 90% of the times includes only the statisticals of the MFFCs coefficients between 2 and 12. Therefore, the phonetic content of the fillers seems to be the main cue adopted to manifest the trait.

For Extraversion, 26 out of the 39 features selected at least 90% of the times account for the energy (how loud the fillers are uttered), fundamental frequency (and its delta regression coefficients) and voicing probability (and its delta regression coefficients). In particular, the correlational analysis suggests that more extraverted people tend to utter fillers more loudly, to have higher and more variable pitch and, finally to have higher variability in the voicing probability. For the last two traits (Agreeableness and Neuroticism), the features in the subset are the MFCCs between 2 and 12, meaning that it is the phonetic content of the fillers that plays the role of the marker.

6 Discussion and Conclusions

This work has shown that it is possible to predict whether a person is above median along the Big-Five traits using the fillers that she or he utters during a spontaneous conversation. The results show that the accuracy - percentage of times the proposed approach makes the right decision about a filler - is close to or above 60% for all the traits. Furthermore, the results show that the application of a majority vote over the fillers uttered by a given speaker, allows one to predict whether this latter is above median along the traits with an accuracy around 75% for all traits. To the best of our knowledge, these performances are in line with the previous results observed for the same task in the literature (though a rigorous comparison is not possible because the experiments have not been performed over the same data).

The results above are interesting from at least two points of view. The first is that they further confirm the relationship between speech and personality traits, while still being innovative because, to the best of our knowledge, the fillers have never been used before for Automatic Personality Recognition from speech [4]. The second is that the fillers can provide a more honest evidence of personality traits with respect to self's assessment psychometric instruments known to be affected by social desirability biases - people may bias their answers in order to provide a positive view of themselves [78]. Furthermore, the approach presented in this article can be of help to other technologies. For example, interactive artificial agents such as social robots [79], companions [80] or Embodied Conversational Agents [81] can infer the personality traits of their users from the fillers these utter and adapt their behavior correspondingly (see, e.g., [82] for the benefits resulting from matching the personality of the users). Finally, fillers can be synthesized to make artificial voices more effective in conveying personality traits [83].

The literature shows that the personality traits tend to be distributed differently across persons of different gender [84]. However, in the data of this work, the Kullback-Leibler Divergence between the trait distributions of female and male participants is, within a statistical fluctuation, null (according to a one-sample t-test). In other words, the traits appear to be equally distributed over female and male participants. Such a peculiarity makes it unnecessary to use gender normalized scores and, as a confirmation, the results show that there is no statistically significant difference between the accuracies achieved over participants of different gender (see Section 5). The collection of further data in which the distributions are different, in the same way as it happens in the general population, can possibly show whether the use of gender-dependent normalizations can further improve the performance of the proposed approach.

The personality questionnaire used in this work [60] includes only 10 items (see Table 2). This reduces the time needed to obtain a personality self-assessment, but it lowers the granularity of the scores. In particular, the BFI-10 approximates the personality traits, that are continuous variables, with an integer score that can assume only 9 different values. The main consequence is that most of the participants tend to concentrate around the median and to form two classes rather than to distribute along a personality dimension. In the case

of this work, the percentage of participants that fall within 2 points from the median is 68% for Openness, 75% for Conscientiousness, 65% for Extraversion, 73% for Agreeableness and 58% for Neuroticism. Such a situation makes it possible to perform effectively the binary classification presented in this article and in most APR works presented in the literature (see Section 2), but does not allow one to apply regression approaches capable to better account for the natural variance in the data.

In addition, while being widely used, the BFI-10 questionnaire is affected by the problems typical of short questionnaires that "*[...] may measure only some subdimension of a trait [...] leading to either regression dilution or overestimation of the association between a trait and a criterion measure*" [85]. In addition, it has been shown that short questionnaires can increase both Type 1 and Type 2 errors, thus increasing the chances of overestimating or missing the relationship between personality and other observable variables [86]. This suggests that the use of questionnaires including more items (see [4] for a survey of the main instruments) is a necessary step to improve the state-of-the-art in APR.

Section 5 shows that there is no statistically significant correlation between the number of fillers available for a given person and the performance of the approach in inferring her personality. Such an observation suggests that a few fillers can be sufficient to reliably predict whether a person is above median along the traits. This is important because it means that it is not necessary to collect large amounts of data about a person and, hence, the time necessary to collect a sufficient number of fillers remains comparable to - if not lower than - the time required to fill a questionnaire (the most popular self-assessment instruments include several tens of items and take up to one hour to be filled).

The main limitation of the current approach is that the fillers have been extracted manually from the speech stream. The application of an automatic filler extraction methodology is likely to introduce noise in the data and, hence, to reduce the performances observed in this work. For this reason, the future work will focus on the analysis of the interplay between the errors resulting from the automatic analysis of the fillers and the accuracy of the APR approach. Given that a few fillers are sufficient to achieve a good performance, the manual extraction can still be an option, but the possibility of a fully automatic approach that takes as input spoken data and gives as output an assessment of the personality of the speakers can allow the use of the methodologies proposed in this work in applications like, e.g., implicit tagging [87], personality based recommender systems [88] and the indexing of large-scale collections of multimedia recordings [89], [90].

The application of a feature selection approach has allowed the identification of the features - physical measurements automatically extracted from the data - that appear to maximize the accuracy of the classifiers. In particular, the adoption of a feature selection approach allows one to identify patterns - subsets of features that possibly interact with one another - rather than individual features. This is an advantage with respect to most psychological works that tend to work on the correlation between individual features and constructs of interest, thus missing that "*most biological and behavioral phenomena are the products of patterns of conditions [...] investigators have to gather patterns of measures in order to differentiate among the varied sequences that can give rise to the same outcome*" [91]. In this respect, this article contributes to shed further light on the interplay between speech and personality originally hypothesized by Edward Sapir (see beginning of Section 1) [1].

ACKNOWLEDGMENTS

The work in this article has been supported by the United Kingdom Engineering and Physical Sciences Research Council (EPSRC) through the projects "Socially Competent Robots" (EP/N035305/1), "School Attachment Monitor" (EP/M025055/1) and "UKRI Centre for Doctoral Training in Socially Intelligent Artificial Agents" (EP/S02266X/1).

REFERENCES

[1] E. Sapir, "Speech as a personality trait," *The American Journal of Sociology*, vol. 32, no. 6, pp. 892–905, 1927.
[2] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social Signal Processing: Survey of an emerging domain," *Image and Vision Computing Journal*, vol. 27, no. 12, pp. 1743–1759, 2009.
[3] B. Schuller and A. Batliner, *Computational paralinguistics: emotion, affect and personality in speech and language processing*. John Wiley & Sons, 2014.
[4] A. Vinciarelli and G. Mohammadi, "A survey of Personality Computing," *IEEE Transaction on Affective Computing*, vol. 5, no. 3, pp. 273–291, 2014.
[5] G. Saucier and L. Goldberg, "The language of personality: Lexical perspectives on the five-factor model," in *The Five-Factor Model of Personality*, J. Wiggins, Ed. Guilford Press, 1996, pp. 21–50.
[6] H. Clark and J. Fox Tree, "Using "uh" and "um" in spontaneous speaking," *Cognition*, vol. 84, no. 1, pp. 73–111, 2002.
[7] A. Vinciarelli, P. Chatziioannou, and A. Esposito, "When the words are not everything: the use of laughter, fillers, back-channel, silence and overlapping speech in phone calls," *Frontiers in ICT*, vol. 2, no. 4, 2015.
[8] C. Laserna, Y. Seih, and J. Pennebaker, "Um ... who like says you know: Filler word use as a function of age, gender, and personality," *Journal of Language and Social Psychology*, vol. 33, no. 3, pp. 328–338, 2014.
[9] J. Heinström, "Five personality dimensions and their influence on information behaviour," *Information Research*, vol. 9, no. 1, pp. 9–1, 2003.
[10] G. Tottie, "Uh and um as sociolinguistic markers in British English," *International Journal of Corpus Linguistics*, vol. 16, no. 2, pp. 173–197, 2011.
[11] D. Watson and L. Clark, "Extraversion and its positive emotional core," in *Handbook of Personality Psychology*, R. Hogan, J. Johnson, and S. Briggs, Eds. Elsevier, 1997, pp. 767–793.
[12] T. Buchanan, J. Laures-Gore, and M. Duff, "Acute stress reduces speech fluency," *Biological Psychology*, vol. 97, pp. 60–66, 2014.
[13] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 Computational Paralinguistics Challenge: social signals, conflict, emotion, autism," in *Proceedings of Interspeech*, 2013.

[14] F. Eyben, M. Woellmer, and B. Schueller, "OpenSMILE: the Munich versatile and fast open-source audio feature extractor," in *Proceedings of ACM International Conference on Multimedia*, 2010, pp. 1459–1462.

[15] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in OpenSMILE, the Munich open-source multimedia feature extractor," in *Proceedings of the ACM International Conference on Multimedia*, 2013, pp. 835–838.

[16] K.-H. Han and J.-H. Kim, "Quantum-inspired Evolutionary Algorithm for a class of combinatorial optimization," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 6, pp. 580–593, 2002.

[17] L. Wilkinson, "Statistical methods in psychology journals: Guidelines and explanations," *American Psychologist*, vol. 54, no. 8, pp. 594–604, 1999.

[18] R. Rosenthal, "Conducting judgment studies: Some methodological issues," in *The New Handbook of Methods in Nonverbal Behavior Research*, J. Harrigan, R. Rosenthal, and K. Scherer, Eds. Oxford University Press Oxford, 2005, pp. 199–234.

[19] A. V. Ivanov, G. Riccardi, A. J. Sporka, and J. Franc, "Recognition of personality traits from human spoken conversations," in *Proceedings of Interspeech*, 2011, pp. 1549–1552.

[20] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text," *Journal of Artificial Intelligence Research*, vol. 30, pp. 457–500, 2007.

[21] L. Batrinca, B. Lepri, N. Mana, and F. Pianesi, "Multimodal recognition of personality traits in human-computer collaborative tasks," in *Proceedings of International Conference on Multimodal Interaction*, 2012, pp. 39–46.

[22] L. Batrinca, N. Mana, B. Lepri, F. Pianesi, and N. Sebe, "Please, tell me about yourself : Automatic personality assessment using short self-presentations," in *Proceedings of the International Conference on Multimodal Interfaces*, 2011, pp. 255–262.

[23] B. Lepri, R. Subramanian, K. Kalimeri, J. Staiano, F. Pianesi, and N. Sebe, "Connecting meeting behavior with extraversion - a systematic study," *IEEE Transactions on Affective Computing*, vol. 3, no. 4, pp. 443–455, 2012.

[24] F. Pianesi, M. Zancanaro, B. Lepri, and A. Cappelletti, "A multimodal annotated corpus of consensus decision making meetings," *Language Resources and Evaluation*, vol. 41, no. 3, pp. 409–429, 2007.

[25] F. Pianesi, N. Mana, A. Cappelletti, B. Lepri, and M. Zancanaro, "Multimodal recognition of personality traits in social interactions," in *Proceedings of the International Conference on Multimodal Interfaces*, 2008, pp. 53–60.

[26] F. Mairesse and M. Walker, "Words mark the nerds: Computational models of personality recognition through language," in *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, 2006, pp. 543–548.

[27] G. Mohammadi, A. Origlia, M. Filippone, and A.Vinciarelli, "From speech to personality: Mapping voice quality and intonation into personality differences," in *Proceedings of ACM International Conference on Multimedia*, 2012, pp. 789–792.

[28] G. Mohammadi and A. Vinciarelli, "Automatic personality perception: Prediction of trait attribution based on prosodic features," *IEEE Transactions on Affective Computing*, vol. 3, no. 3, pp. 273–278, 2012.

[29] K. Polzehl, T. ans Schoenenberg, S. Moller, F. Metze, G. Mohammadi, and A. Vinciarelli, "On speaker-independent personality perception and prediction from speech," in *Proceedings of Interspeech*, 2012.

[30] F. Valente, S. Kim, and P. Motlicek, "Annotation and recognition of personality traits in spoken conversations from the AMI meetings corpus," in *Proceedings of Interspeech*, 2012.

[31] F. Celli, B. Lepri, J.-I. Biel, D. Gatica-Perez, G. Riccardi, and F. Pianesi, "The workshop on computational personality recognition 2014," in *Proceedings of the ACM international conference on Multimedia*, 2014, pp. 1245–1246.

[32] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, "A survey on perceived speaker traits: Personality, likability, pathology and the first challenge," *Computer Speech and Language*, vol. 29, no. 1, pp. 100–131, 2015.

[33] Y. Attabi and P. Dumouchel, "Anchor models and WCCN normalization for speaker trait classification," in *Proceedings of Interspeech*, 2012.

[34] K. Audhkhasi, A. Metallinou, M. Li, and S. Narayanan, "Speaker personality classification using system based on acoustic-lexical cues and an optimal tree-structured bayesian network," in *Proceedings of Interspeech*, 2012.

[35] H. Buisman and E. Postma, "The log-gabor method: Speech classification using spectrogram image analysis," in *Proceedings of Interspeech*, 2012.

[36] C. Chastagnol and L. Devillers, "Personality traits detection using a parallelized modified SFFS algorithm," in *Proceedings of Interspeech*, 2012.

[37] A. V. Ivanov and X. Chen, "Modulation spectrum analysis for speaker personality trait recogniton," in *Proceedings of Interspeech*, 2012.

[38] C. Montacie and M. Caraty, "Pitch and intonation contribution to speakers' traits classification," in *Proceedings of Interspeech*, 2012.

[39] J. Pohjalainen, S. Kadioglu, and O. Rasanen, "Feature selection for speaker traits," in *Proceedings of Interspeech*, 2012.

[40] M. H. Sanchez, A. Lawson, D. Vergyri, and H. Bratt, "Multi-system fusion of extended context prosodic and cepstral features for paralinguistic speaker trait classification," in *Proceedings of Interspeech*, 2012.

[41] J. Wagner, F. Lingenfelser, and E. Andre, "A frame pruning approach for paralinguistic recognition tasks," in *Proceedings of Interspeech*, 2012.

[42] F. Celli, F. Pianesi, D. Stillwell, and M. Kosinski, "Workshop on Computational Personality Recognition (shared task)," in *Proceedings of the Workshop on Computational Personality Recognition*, 2013.

[43] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[44] M. Dash and H. Liu, "Feature selection for classification," *Intelligent Data Analysis*, vol. 1, no. 3, pp. 131–156, 1997.

[45] Y. Liu, F. Tang, and Z. Zeng, "Feature selection based on dependency margin," *IEEE Transactions on Cybernetics*, vol. 45, no. 6, pp. 1209–1221, 2015.

[46] H. Liu and Z. Zhao, "Manipulating data and dimension reduction methods: Feature selection," in *Encyclopedia of Complexity and Systems Science*. Springer, 2009, pp. 5348–5359.

[47] H. Liu, H. Motoda, R. Setiono, and Z. Zhao, "Feature selection: An ever evolving frontier in data mining," in *Feature Selection in Data Mining*, 2010, pp. 4–13.

[48] A. Unler and A. Murat, "A discrete particle swarm optimization method for feature selection in binary classification problems," *European Journal of Operational Research*, vol. 206, no. 3, pp. 528–539, 2010.

[49] B. Xue, M. Zhang, W. N. Browne, and X. Yao, "A survey on evolutionary computation approaches to feature selection," *IEEE Transactions on Evolutionary Computation*, vol. 20, no. 4, pp. 606–626, 2016.

[50] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.

[51] M. Mitchell, *An introduction to genetic algorithms*. MIT press, 1998.

[52] W. Siedlecki and J. Sklansky, "A note on genetic algorithms for large-scale feature selection," in *Handbook Of Pattern Recognition And Computer Vision*. World Scientific, 1993, pp. 88–107.

[53] D. P. Muni, N. R. Pal, and J. Das, "Genetic programming for simultaneous feature selection and classifier design," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 36, no. 1, pp. 106–117, 2006.

[54] Z. Yan and C. Yuan, "Ant colony optimization for feature selection in face recognition," in *Biometric Authentication*. Springer, 2004, pp. 221–226.

[55] R. N. Khushaba, A. Al-Ani, A. AlSukker, and A. Al-Jumaily, "A combined ant colony and differential evolution feature selection algorithm," in *Proceedings of the International Conference on Ant Colony Optimization and Swarm Intelligence*. Springer, 2008, pp. 1–12.

[56] Z. Zhu, Y.-S. Ong, and M. Dash, "Markov blanket-embedded genetic algorithm for gene selection," *Pattern Recognition*, vol. 40, no. 11, pp. 3236–3248, 2007.

[57] M. Iqbal, S. Naqvi, W. Browne, C. Hollitt, and M. Zhang, "Salient object detection using learning classifier systems that compute action mappings," in *Proceedings of the Annual Conference on Genetic and Evolutionary Computation*, 2014, pp. 525–532.

[58] M. Marinaki and Y. Marinakis, "A bumble bees mating optimization algorithm for the feature selection problem," *International Journal of Machine Learning and Cybernetics*, vol. 7, no. 4, pp. 519–538, 2016.

[59] M. Joshi, E. Davis, R. Kathuria, and C. Weidner, "Experiential learning process: Exploring teaching and learning of strategic management framework through the winter survival exercise," *Journal of Management Education*, vol. 29, no. 5, pp. 672–695, 2005.
[60] B. Rammstedt and O. 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*, vol. 41, no. 1, pp. 203–212, 2007.
[61] G. Matthews, I. Deary, and M. Whiteman, *Personality Traits*. Cambridge University Press, 2009.
[62] F. Al-Anzi and D. AbuZeina, "The capacity of Mel Frequency Cepstral Coefficients for speech recognition," *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, vol. 11, no. 10, pp. 1094–1098, 2017.
[63] B. Atal and L. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 3, pp. 201–212, 1976.
[64] M. Lee, J. Van Santen, B. Mobius, and J. Olive, "Formant tracking using context-dependent phonemic information," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 741–750, 2005.
[65] M. Tayarani-N and M. Akbarzadeh-T, "Improvement of the performance of the Quantum-inspired Evolutionary Algorithms: structures, population, operators," *Evolutionary Intelligence*, vol. 7, no. 4, pp. 219–239, 2014.
[66] M.-H. Tayarani-N and A. Prugel-Bennett, "On the landscape of combinatorial optimisation problems," *IEEE Transactions on Evolutionary Computation*, vol. 18, no. 3, pp. 420–434, 2014.
[67] M.-H. Tayarani-N. and A. Prügel-Bennett, "Quadratic assignment problem: a landscape analysis," *Evolutionary Intelligence*, vol. 8, no. 4, pp. 165–184, 2015.
[68] ——, "Anatomy of the fitness landscape for dense graph-colouring problem," *Swarm and Evolutionary Computation*, vol. 22, pp. 47 – 65, 2015.
[69] M.-H. Tayarani-N and A. Prügel-Bennett, "An analysis of the fitness landscape of travelling salesman problem," *Evolutionary Computation*, vol. 24, no. 2, pp. 347–384, 2016.
[70] S. Fahlman and C. Lebiere, "The cascade-correlation learning architecture," in *Advances in Neural Information Processing Systems*, 1990, pp. 524–532.
[71] A. Jain, J. Mao, and K. Mohiuddin, "Artificial neural networks: A tutorial," *IEEE Computer*, vol. 29, no. 3, pp. 31–44, 1996.
[72] J. Buckley and Y. Hayashi, "Fuzzy Neural Networks: A survey," *Fuzzy Sets and Systems*, vol. 66, no. 1, pp. 1–13, 1994.
[73] D. Specht, "A general regression neural network," *IEEE Transactions on Neural Networks*, vol. 2, no. 6, pp. 568–576, 1991.
[74] C. Bishop, *Pattern Recognition and Machine Learning*. Springer Verlag, 2006.
[75] M. Hearst, S. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support Vector Machines," *IEEE Intelligent Systems and Their Applications*, vol. 13, no. 4, pp. 18–28, 1998.
[76] L. Hedges, "Correcting a significance test for clustering," *Journal of Educational and Behavioral Statistics*, vol. 32, no. 2, pp. 151–179, 2007.
[77] K. R. Scherer, "Personality inference from voice quality: The loud voice of extroversion," *European Journal of Social Psychology*, vol. 8, pp. 467–487, 1978.
[78] T. Van de Mortel, "Faking it: social desirability response bias in self-report research," *Australian Journal of Advanced Nursing*, vol. 25, no. 4, pp. 40–48, 2008.
[79] C. Breazeal, *Designing sociable robots*. MIT Press, 2004.
[80] S. Biundo and A. Wendemuth, "Companion-technology for cognitive technical systems," *Künstliche Intelligenz*, vol. 30, no. 1, pp. 71–75, 2016.
[81] J. Cassell, J. Sullivan, E. Churchill, and S. Prevost, *Embodied Conversational Agents*. MIT Press, 2000.
[82] A. Tapus, C. Ţăpuş, and M. Matarić, "User-robot personality matching and assistive robot behavior adaptation for post-stroke rehabilitation therapy," *Intelligent Service Robotics*, vol. 1, no. 2, p. 169, 2008.
[83] M. Aylett, A. Vinciarelli, and M. Wester, "Speech synthesis for the generation of artificial personality," *IEEE Transactions on Affective Computing (accepted for publication)*, 2019.
[84] A. Feingold, "Gender differences in personality: A meta-analysis," *Psychological Bulletin*, vol. 116, no. 3, p. 429, 1994.
[85] B. Bakker and Y. Lelkes, "Selling ourselves short? How abbreviated measures of personality change the way we think about personality and politics," *The Journal of Politics*, vol. 80, no. 4, pp. 1311–1325, 2018.
[86] M. Credé, P. Harms, S. Niehorster, and A. Gaye-Valentine, "An evaluation of the consequences of using short measures of the Big Five personality traits," *Journal of Personality and Social Psychology*, vol. 102, no. 4, pp. 874–888, 2012.
[87] M. Pantic and A. Vinciarelli, "Implicit human-centered tagging," *IEEE Signal Processing Magazine*, vol. 26, no. 6, pp. 173–180, 2009.
[88] M. Tkalčič, U. Burnik, and A. Košir, "Using affective parameters in a content-based recommender system for images," *User Modeling and User-Adapted Interaction*, vol. 20, no. 4, pp. 279–311, 2010.
[89] S. Wang and Q. Ji, "Video affective content analysis: a survey of state-of-the-art methods," *IEEE Transactions on Affective Computing*, vol. 6, no. 4, pp. 410–430, 2015.
[90] M. Tkalcic, A. Odic, A. Kosir, and J. Tasic, "Affective labeling in a content-based recommender system for images," *IEEE transactions on Multimedia*, vol. 15, no. 2, pp. 391–400, 2013.
[91] J. Kagan, *Five Constraints on Predicting Behavior*. MIT Press, 2017.

Mohammad-H. Tayarani-N. received his Ph.D. degree from the University of Southampton, Southampton, U.K, in 2013. Then worked as research fellow at the University of Birmingham, Birmingham, UK. He is currently a research assistant at the University of Glasgow, Glasgow, UK. His main research interests include evolutionary algorithms, machine learning, and fractal image compression.

Anna Esposito received the PhD Degree in Applied Mathematics and Computer Science from the University of Naples "Federico II". She is currently Associate Professor in the Department of Psychology, Università della Campania "Luigi Vanvitelli" (Italy) and director of the Behavioural Cognitive Systems (BeCogSys) laboratory. She is research affiliate at the International Institute for Advanced Scientific Studies (IIASS). She has published 170+ peer reviewed papers and is edited/coedited 24+ international books.

Alessandro Vinciarelli is with the University of Glasgow where he is Full Professor at the School of Computing Science and Associate Academic at the Institute of Neuroscience and Psychology (`http://vinciarelli.net`). His main research interest is in Social Signal Processing, the domain aimed at modeling analysis and synthesis of nonverbal behavior in social interactions. Overall, he has published more than 130 works, including one authored book, and 35 journal papers. He has been General Chair of the IEEE International Conference on Social Computing in 2012 and of the ACM International Conference on Multimodal Interaction in 2017. Alessandro is or has been Principal Investigator of several national and international projects, including the UKRI Centre for Doctoral Training in Socially Intelligent Artificial Agents (`http://www.social-cdt.org`), a European Network of Excellence (the SSPNet, `www.sspnet.eu`), and more than 10 projects funded by the Swiss National Science Foundation end the UK Engineering and Physical Sciences Research Council. Last, but not least, Alessandro is co-founder of Klewel (`www.klewel.com`), a knowledge management company recognized with national and international awards.