Automatic Personality Perception: Prediction of Trait Attribution Based on Prosodic Features

Gelareh Mohammadi and Alessandro Vinciarelli, *Member*, *IEEE*

Abstract—Whenever we listen to a voice for the first time, we attribute personality traits to the speaker. The process takes place in a few seconds and it is spontaneous and unaware. While the process is not necessarily accurate (attributed traits do not necessarily correspond to the actual traits of the speaker), still it significantly influences our behavior toward others, especially when it comes to social interaction. This paper proposes an approach for the automatic prediction of the traits the listeners attribute to a speaker they never heard before. The experiments are performed over a corpus of 640 speech clips (322 identities in total) annotated in terms of personality traits by 11 assessors. The results show that it is possible to predict with high accuracy (more than 70 percent depending on the particular trait) whether a person is perceived to be in the upper or lower part of the scales corresponding to each of the Big Five, the personality dimensions known to capture most of the individual differences.

Index Terms—Personality traits, prosody, Big Five, social signal processing, automatic personality perception

1 Introduction

One of the main findings of social cognition is that spontaneous and unaware processes influence our behavior to a large extent, especially when it comes to social interactions [1]. In particular, there is a large body of evidence showing that "*people make social inferences without intentions, awareness, or effort, i.e., spontaneously*" [2]. Furthermore, the phenomenon is so pervasive that it has been observed not only when we meet other individuals in person, but also when we see them in pictures [3], we watch them in videos [4], or we listen to them in audio recordings [5].

This paper considers a facet of the phenomenon above, namely, the spontaneous attribution of personality traits to speakers we are not acquainted with. In particular, the paper proposes an approach for prosody-based *Automatic Personality Perception* (APP), i.e., for automatically mapping prosodic aspects of speech into personality traits attributed by human listeners. Unlike Automatic Personality Recognition (APR), the goal of APP is not to predict the *real* personality of an individual, but the personality as *perceived* by observers. In other words, APP is not expected to predict the real personality of a given person, but the personality that others attribute to her in a given situation.

There are at least three reasons to consider the APP problem important (especially in zero acquaintance scenarios): The first is that interactions with unknown individuals are frequent in our everyday life and include, e.g., phone conversations with call center operators, job interviews, meetings with new colleagues, etc. In all these cases, our behavior is not driven by the actual personality of the people we have around, but by the traits we spontaneously perceive and attribute to them [1], [2]. For example, the traits we attribute to politicians after having watched their picture for 100 ms predict whether we vote for them or not [3].

The second is that we attribute traits not only to people, but also to machines that exhibit human-like features and behaviors, including robots, embodied conversational agents, animated characters, etc. [4], [5]. In this case, only APP can be performed because machines do not have personality and APR is simply not possible. Furthermore, traits attributed to machines help to predict the attitude of the users. For example, extroverted people tend to spend more time with the robots they perceive to be extroverted than with those they perceive to be introverted [6].

The third is that perceived traits correlate with a wide spectrum of personal characteristics (e.g., professional choices, political orientations, well-being, etc.) better than self-assessed traits, typically considered as the actual personality of an individual [7]. The prediction of personal characteristics is one of the most important applications of personality theory [8] and APP approaches, aimed at predicting attributed traits, are likely to contribute to it. For example, Section 6.2 shows that perceived traits allow one to predict whether a person is a professional speaker or not.

The APP approach proposed in this work relies on prosody as a physical, machine detectable cue capable of explaining the traits perceived by human listeners. The choice is supported by extensive investigations in human sciences showing that nonverbal vocal behavior significantly influences personality perception [9] (see Section 4 for a short survey). Furthermore, domains like affective computing [10] and social signal processing [11] have shown that nonverbal behavioral cues (e.g., prosody, facial expressions, gestures, postures, etc.) work effectively as evidence for technologies dealing with emotional and social phenomena.

- *G. Mohammadi is with the IDIAP Research Institute, CP592, 1920 Martigny, Switzerland, and the Ecole Polytechnique Federale de Lausanne (EPFL). E-mail: gelareh.mohammadi@idiap.ch.*
- *A. Vinciarelli is with the University of Glasgow, Sir A. Williams Building, G128QQ Glasgow, United Kingdom, and the IDIAP Research Institute, CP592, 1920 Martigny, Switzerland. E-mail: vincia@dcs.gla.ac.uk.*

Manuscript received 20 Apr. 2011; revised 4 Feb. 2012; accepted 27 Feb. 2012; published online 13 Mar. 2012.
Recommended for acceptance by B. Schuller.
For information on obtaining reprints of this article, please send e-mail to: taffc@computer.org, and reference IEEECS Log Number TAFFC-2011-04-0035.
Digital Object Identifier no. 10.1109/T-AFFC.2012.5.

So far, only a few works have considered the APP problem in the literature (see Section 4 for a survey). This work contributes to the domain by addressing several issues that, to the best of our knowledge, are still open:

- This is the first work that measures quantitatively the effect of individual prosodic features on APP effectiveness.
- This is the first work that uses personality assessments as features for predicting personal characteristics.
- The dataset used in this work includes three times more individuals than any other APP experiment reported so far in the literature.
- This work considers nonverbal speech features neglected so far in both computing and psychological literature.

The first point is important for two main reasons: On one hand, it provides crucial information toward the development of better APP systems by identifying the most effective vocal cues. On the other hand, it suggests the characteristics that synthetic voices should have in order to elicit the perception of predefined traits. Previous work in psychology (see Section 3) has shown the impact of individual features in terms of correlation with personality traits, but no investigation has been made so far of how individual features contribute to an automatic prediction approach.

The second point is important because it shows that the ratings are coherent with respect to a variable collected independently of the assessments. Such a task-oriented methodology for assessing the reliability of ratings has never been used before.

The third point is important because it improves the statistical reliability of the results in a domain where the collection of data, especially when it comes to personality assessments, is expensive and time-consuming (every judge has to assess the entire dataset and this becomes difficult when the number of subjects and samples increases).

Finally, the fourth point is important because psychological studies can benefit from expertise on speech processing typically not available in the human sciences community. Hence, it is possible to consider other features than those considered so far and further deepen the study of the interplay between speech and personality.

In the long term, APP can be considered as a contribution to the efforts being done toward bridging the social intelligence gap between people and machines [11]. However, some early applications of APP have already been explored like the generation of synthetic voices eliciting desired social perceptions (see, e.g., [13]), the use of personality assessments in recommender systems [14], the interaction between humans and robots [6], or the indexing of multimedia data in terms of social and emotional user perceptions [15]. In this respect, the development of APP technologies is expected to be beneficial for several computing areas.

2 Personality and Its Assessment

Personality is the latent construct that accounts for "*individuals' characteristic patterns of thought, emotion, and behavior together with the psychological mechanisms—hidden or not—behind those patterns*" [16]. The literature proposes a large number of models (see [17] for an extensive survey), but the most common personality representation relies on the *Big Five* (BF), five broad dimensions that "*appear to provide a set of highly replicable dimensions that parsimoniously and comprehensively describe most phenotypic individual differences*" [18].

TABLE 1
The BFI-10 Questionnaire
Used in the Experiments of This Work

ID	Question
1	This person is reserved
2	This person is generally trusting
3	This person tends to be lazy
4	This person is relaxed, handles stress well
5	This person has few artistic interests
6	This person is outgoing, sociable
7	This person tends to find fault with others
8	This person does a thorough job
9	This person gets nervous easily
10	This person has an active imagination

The version reported here is the one that has been proposed in [12].

The BF have been identified by applying factor analysis to the large number of words describing personality in everyday language (around 18,000 in English [17]). Despite the wide variety of terms at disposition, personality descriptors tend to group into five major clusters corresponding to the BF:

- *Extroversion*: Active, Assertive, Energetic, etc.
- *Agreeableness*: Appreciative, Kind, Generous, etc.
- *Conscientiousness*: Efficient, Organized, Planful, etc.
- *Neuroticism*: Anxious, Self-pitying, Tense, etc.
- *Openness*: Artistic, Curious, Imaginative, etc.

In this perspective, the clusters are interpreted as the trace that salient psychological phenomena leave in language (the *lexical hypothesis* [18]), one of the main evidences supporting the actual existence of the BF [17].

In light of the above, the BF model represents a personality with five scores (one per trait) that can be thought of as the position on an ideal personality map. Thus, in the BF perspective, personality assessment means essentially to obtain those scores. As the BF account for "*phenotypic individual differences*" (see quote from [18] above), the main instruments for score assignment are questionnaires where a person is assessed in terms of *observable* behaviors and characteristics, i.e., in terms of what a person does or how a person appears to be.

Table 1 shows the *Big Five Inventory 10* (BFI-10), the questionnaire used in this work [12]. Each question is associated with a 5 point Likert scale (from "*Strongly Disagree*" to "*Strongly Agree*") mapped into the interval $[-2, 2]$. The BFI-10 includes the 10 items that better correlate with the assessments obtained using the full BFI (44 items). The personality scores can be obtained using the answers provided by the assessors (Q_i is the answer to item i):

- *Extroversion*: $Q_6 - Q_1$.
- *Agreeableness*: $Q_2 - Q_7$.
- *Conscientiousness*: $Q_8 - Q_3$.
- *Neuroticism*: $Q_9 - Q_4$.
- *Openness*: $Q_{10} - Q_5$.

The main advantage of the BFI-10 is that it takes less than 1 minute to complete.

3 Speech and Personality

It was around one century ago that the hypothesis of "*speech as a personality trait*" [19] was proposed for the first time. Since then a large number of studies have analyzed the effect of vocal behavior on personality perception, especially when it comes to prosody, voice quality, and fluency.

Prosodic cues mainly include pitch (oscillation frequency of glottal folds), energy (perceived as loudness), and speaking rate. The results of [20] show that males with higher pitch variation are perceived as more dynamic, feminine, and aesthetically inclined. In contrast, females with higher pitch variation are rated as more dynamic and extroverted. The work in [21] investigates the joint effect of pitch variation and speaking rate. High pitch variation combined with high speaking rate leads to perception of high competence and vice versa. Similar effects are observed with respect to benevolence: Low-pitch variation and high speaking rate lead to low benevolence ratings, the contrary of high-pitch variation and low energy. In the same vein, a negative correlation between mean pitch and both extraversion and dominance has been observed for American female speakers (in [22] as cited in [23]). The same study has shown that the correlation is positive for American male speakers. In the case of German speakers, higher pitch leads to low extroversion perception.

The correlation between speaking rate and competence has been consistently observed in a large number of studies [21], [24] (as cited in [25]), [26] (as cited in [21]), [25]. The tendency is to associate higher speaking rate to higher competence and vice versa, but some contradicting evidence has been found as well [25]. Another study [20] showed that faster speakers are perceived as more animated and extroverted. The relation between speaking rate and benevolence seems to be more controversial. Some works (e.g., [24], [27] as cited in [25]) suggest that average speaking rates lead to higher benevolence ratings (an invert U relation), while others indicate that these latter decrease when the speaking rate decreases as well (a direct proportionality relation) [25].

The effect of loudness has been examined in [9] and [21]. Findings of the first study report a positive correlation between mean and dynamic range of loudness on one side and emotional stability and extroversion on the other side. The other work [21] indicates that louder speakers are perceived as more competent and vice versa.

The effect of voice quality on perception of 40 personality-related adjectives is investigated in [20]: For male speakers, breathier voices are perceived as younger and more artistic; for female ones as prettier, more feminine, more sensitive, and richer in sense of humor. Earlier studies [28], [29] (as cited in [23]) show that breathy voices sound more introverted and neurotic.

Thinner female voices elicit the perception of immaturity at different levels (social, physical, emotional, and mental) [20]. For both genders, flat voices are perceived as more masculine, sluggish, and colder. Nasality in both genders was perceived as socially undesirable and the same applies to tenseness, perceived as an indicator of bad temper (male speakers) or youth, emotional instability, and low intelligence (female speakers). Throatiness in men was perceived as being older, realistic, mature, sophisticated, and well-adjusted, while in women it was perceived as being more selfish. Orotundity showed positive correlation with being energetic, healthy, artistic, sophisticated, proud, interesting, and enthusiastic (for male speakers). In the case of female voices, orotundity was perceived as higher in liveliness and aesthetic sensitivity, but too proud and humorless as well.

Fluency aspects of speech like silent and filled pauses have also been explored. The study in [30] has examined the effect of extroversion on pauses: Extroverted people speak with fewer filled pauses, fewer pauses longer than 2 seconds, shorter periods of silence, and lower number of silent hesitation pauses. However, in [31], extroverted German speakers are found to have more silent pauses. Other investigations have shown that anxious speakers speak with fewer short silent pauses but more frequent longer pauses [23]. However, the relationship between personality traits and pausing may be more complex because social psychological factors like social skills, self-presentation strategies, etc., have to be taken into account [31].

The works presented in this section investigate the personality perception problem from a psychological point of view and, unlike this paper, do not include any attempt to develop computational approaches capable of automatically predicting the traits perceived by human listeners. Computing-oriented works are surveyed in the next section.

4 Personality and Computing

Only a few approaches have been dedicated to personality in the computing community, mostly in domains like social signal processing [11] that aim at modeling, analysis, and synthesis of nonverbal communication. Table 2 is a synopsis of the main works presented so far in the literature. Voice and speech-related cues have been used in all of the approaches, while other forms of nonverbal behavior (e.g., the amount of energy associated with body gestures known as *fidgeting*) appear in only a few cases. The main reason is probably that vocal behavior has been shown to be significantly correlated (more than, e.g., facial expressions and body movements) to important personality aspects like inhibition, dominance, or pleasantness [32].

The main problem with the current state of the art seems to be the low number of individuals represented in the corpora used for experiments. The largest corpus includes 2,479 identities and the same number of samples, but it contains only written essays. Thus, it cannot be compared with the corpora used in works based on nonverbal behavior. In these latter, the largest dataset seems to be the one used in this work (322 identities for 640 samples), more than three times larger than the closest corpus (96 individuals for 96 samples).

The first computing approach dealing with personality was presented in [13]. This work shows not only that manipulating the prosody of synthetic voices (pitch, intensity, speaking rate) influences the perception of extroversion, but also that synthetic voices perceived as extroverted tend to be more appreciated by extroverted persons and vice versa.

TABLE 2
State of the Art

Author	No. of samples in the corpus	No. of Identities	Real/Acted/Synthetic voice[1]	Personality Traits[2]	Perceived or Self-assessed[3]	No. of Annotators per subject	Non-verbal Vocal features	Other Features
Nass & Lee [13]	10	-	S	Ext.	P	36	Statistics of: Pitch,Intensity, Speaking rate	-
Schmitz et al. [33], [34]	12	-	S	Sin. Exc. Com. Sop. Rug.	P	36	Pitch range, Pitch level, Tempo, Intensity	-
Krahmer et al. [35]	8	-	S	Ext.	P	24	Pitch range, Pitch variation, Speaking rate	eye movement, eye blinking, eyebrow movement
Tapus et al. [36]	12	-	S	Ext.	P	-	Loudness, Speaking rate	proxemics
Mairesse et al. [37], [38]	96	96	R	Big Five	S,P	6	Statistics of: Pitch, Intensity, Speaking rate	LIWC, MRC, Utterance
Mohammadi et al. [39]	640	322	R	Big Five	P	3	Statistics of: Pitch, Formants, Speaking rate, Voiced & unvoiced segment's length	-
Polzehl et al. [40]	30	1	A	Big Five	P	20	Statistics of: Pitch, Intensity, Formants, Loudness, Spectral Energy, MFCC, HNR, ZCR	-
Pianesi et al. [41]	48	48	R	Ext. LOC	S	-	Statistics of: Pitch, Formants, Spectral energy, Energy in frame, Speaking time, Speaking rate, Voiced Segment's length, etc.	Energy of: Head movement, Hand movement, Body movement
Olguin et al. [42]	67	67	R	Big Five	P	-	Speech volume, Speaking time, Voiced speaking, Time	Statistics of: Amount of physical activity, duration & No. of face-to-face interactions, Proximity to other people/ bed/phone, etc.
Zen et al. [43]	2	13	R	Ext., Neu.	S	-		Velocity, Proximity

The abbreviations are as follows: Real-voice (R), Acted-voice (A), Synthetic voice (S), Locus of control (LOC), Extraversion (Ext.), Neuroticism (Neu.), Sincerity (Sin.), Excitement (Exc.), Competence (Com.), Sophistication (Sop.), Ruggedness (Rug.), Perceived (P), Self-assessed (S).

A similar approach has been proposed in [33], [34] where the results show that prosodic features (pitch range, pitch level, tempo, and intensity) of brand-related synthetic voices have an impact on the perception of several traits (sincerity, excitement, competence, sophistication, and ruggedness). A wider spectrum of automatically generated nonverbal cues (pitch, speaking rate, gaze, and eyebrow movements) has been explored in [35]. The results show that all of the cues actually have an influence on how extroverted an embodied conversational agent is perceived to be. The work in [36] has shown that extroverted people tend to spend more time with robots simulating an extroverted personality (via speaking rate, loudness, and interpersonal distance) than with those simulating an introverted one and vice versa.

On the analysis side, the earliest approaches were proposed in [37], [38]. These works consider both personality perception and personality recognition and use written data as well as speech samples for the experiments. Both psycholinguistic, like Linguistic Inquiry and Word Count (LIWC) or MRC (see [37] for more details), and prosodic features (average, minimum, maximum, and standard deviation of pitch, intensity, voiced time, speech rate) have been used, separately and in combination. The recognition is performed using different statistical approaches (C4.5 decision tree learning, nearest neighbor, Naive Bayes, Ripper, Adaboost, and Support Vector Machines (SVM) with linear kernels). The results show that it is possible to predict whether a person is (or is perceived to be) below or above average along the Big Five dimensions with an accuracy between 60 and 75 percent, depending on the trait and on the features used.

Similar approaches have been proposed in [39] and [40]. In the first work, statistical functions (entropy, minimum, maximum, etc.) of the main prosodic features (pitch, energy, first two formants, length of voiced, and unvoiced segments) have been used to predict whether a speaker is perceived as above or below average along each of the Big Five dimensions. The prediction is performed with Support Vector Machines and the accuracies range between 60 and 75 percent depending on the trait. The other work [40]

applies a total of 1,450 features based on statistics (e.g., moments of the first four orders) of intensity, pitch, loudness, formants, spectral energy, and Mel Frequency Cepstral Coefficients. These are first submitted to a feature selection approach and then fed to Support Vector Machines to recognize 10 different personality types acted by the same speaker. The recognition rate is 60 percent.

As personality plays a major role in social interactions, some works have focused on scenarios where people participate in social exchanges [41], [42], [43]. The approach proposed in [41] focuses on one minute windows extracted from meeting recordings. The meeting participants are represented in terms of mean and standard deviation of vocal characteristics (e.g., energy, formants, largest autocorrelation peak, etc.) as well as fidgeting. These features are fed to a Support Vector Machine trained to recognize two personality traits (extroversion and locus of control) with an accuracy up to 95 percent (however, one of the three classes identified by the authors accounts for 66 percent of the test material).

The experiments in [42] use wearable devices to extract behavioral evidences related to speaking activity (speaking time, voiced time, loudness, etc.), movement (intensity, power, etc.), proximity (time in proximity of others, time in proximity of phones, etc.), face-to-face interactions (number of face-to-face interactions, etc.), and position in a social network (centrality, betweenness, etc.). The results consist of the correlation between the measures above and self-assessed personality traits. In some cases, the absolute value is higher than 0.4 (e.g., the correlation between speaking activity and agreeableness, and the correlation between social network features and openness). Proximity is used in [43] as well, where interpersonal distance features (e.g., the minimum distance with respect to others, the distribution of interpersonal distances across others, etc.) and velocity are used to predict self-assessed extroversion and neuroticism. The accuracy in predicting whether someone is above or below the median along a certain trait is 66 percent for extroversion and 75 percent for neuroticism. The main limitation of the work is that the number of subjects is low (13 individuals).

While at an early stage, the state of the art has covered a wide spectrum of behavioral cues (both verbal and non-verbal) and scenarios, but a number of issues still remain open. The first is the number of individuals involved in the experiments: The collection of personality assessments is expensive and time consuming, especially in APP experiments where the number of raters per subject must be higher than 10. As a consequence, the experiments typically consider only a few tens of subjects and the statistical reliability of the results is potentially limited (see Table 2 for the number of subjects involved in different works).

The second is the use of personality assessments to predict personal characteristics of individuals. Such a task is one of the most important applications of personality theory [8], but so far it has been largely, if not all, neglected by the computing community. The third problem applies in particular to the APP case and it is the low agreement between assessors that rate the personality of the same individual (the correlation tends to be low to moderate [44]). The phenomenon accounts for the inherent ambiguity of the personality perception problem and it is not evident how to deal with it. The fourth problem is that the Big Five model is a dimensional representation of personality, but both APP and APR approaches quantize the assessments in order to apply classifiers. This transforms the dimensional representation into a categorical one, but the categories typically have no psychological motivation (the most common approach is to consider assessments below and above average as two classes). Last, but not least, personality perception is culture dependent (listeners belonging to different cultures tend to assign different traits to the same speaker) [45], but such an effect has never been taken into account in computing approaches. In other words, automatic systems tend to either neglect the problem (the culture of both subjects and assessors is simply not taken into account) or to limit the investigation to one culture only to avoid multicultural effects.

5 The Approach

The APP approach proposed in this work includes three main steps: extraction of low-level short-term prosodic features from the speech signal, estimate of statistical features accounting for long-term prosodic characteristics of speech, and mapping of statistical features into attributed traits.

5.1 Low-Level Feature Extraction

The low-level features extracted in this work are pitch (number of vibrations per second produced by the vocal cords, the main acoustic correlate of tone and intonation), first two formants (resonant frequencies of the vocal tract), energy of the speech signal, and length of voiced and unvoiced segments (an indirect measure of speaking rate). The *rationale* behind the choice is not only that pitch, rate and energy, often called the *Big Three*, are the most important aspects of prosody, but also that they are the most investigated cues when it comes to the relation between speech and personality perception (see Section 3). Furthermore, the formants can capture possible gender and verbal content effects [5].

The extraction has been performed with Praat (version 5.1.15), one of the most widely applied speech processing tools [46]. The features are extracted from 40 ms long analysis windows at regular time steps of 10 ms and reflect short-term characteristics of vocal behavior. As the extraction is performed every 10 ms, the process converts a speech clip into a sequence of frame feature vectors $F = (\vec{f}_1, \ldots, \vec{f}_N)$, where the components $f_i^{(j)}$ of each $\vec{f}_i$ correspond to the six low-level features mentioned above. After having been extracted, the features are transformed into z-scores using their mean and standard deviation as estimated in the training set.

5.2 Estimation of Statistical Features

In this work, four statistical properties are estimated: minimum, maximum, mean, and relative entropy of the differences between low-level feature values extracted from consecutive analysis windows. Minimum and maximum are used because together they account for the dynamic range. The entropy of the differences between consecutive feature values accounts for the predictability of a given prosodic

characteristic: the higher the entropy, the more the difference between consecutive values is uncertain, i.e., the more difficult it is to predict the next value given the current one (the approach is inspired by Song et al. [47]). If $\Delta f_i^{(j)} = f_i^{(j)} - f_{i-1}^{(j)}$ is the difference between two consecutive values of the jth low-level feature and $\mathcal{Y}^{(j)} = \{y_1^{(j)}, y_2^{(j)}, \ldots, y_{|\mathcal{Y}^{(j)}|}^{(j)}\}$ is the set of the values that $\Delta f_i^{(j)}$ can take, then the entropy H for the jth low level feature is

$$H\left(\Delta f_i^{(j)}\right) = \frac{-\sum_{k=1}^{|\mathcal{Y}^{(j)}|} p\left(y_k^{(j)}\right) \log p\left(y_k^{(j)}\right)}{\log\left(|\mathcal{Y}^{(j)}|\right)}, \quad (1)$$

where $p(y_k^{(j)})$ is the probability of $\Delta f_i^{(j)} = y_k^{(j)}$ (estimated with the fraction of times the value $y_k^{(j)}$ is actually observed) and $|\mathcal{Y}^{(j)}|$ is the cardinality of $\mathcal{Y}^{(j)}$ (number of elements in $\mathcal{Y}^{(j)}$). The term $\log |\mathcal{Y}^{(j)}|$ works as a normalization factor; the upper bound ($H = 1$) is reached when the distribution is uniform (maximum uncertainty). When the entropy is higher, it means there is higher uncertainty and the feature is less predictable.

The extraction of the low-level features is performed every 10 ms and a clip of length T seconds results into $T \times 100 - 4$ values for each low-level feature. (The last 40 ms are occupied by one analysis window only.) Most of the clips of the database are 10 s long (593 out of 640) and this corresponds to 996 observations. Such a number is sufficient to avoid the effect of possible outliers on mean and entropy. As there are six low-level features and four statistical properties, the resulting feature vector $\vec{x}$ for a speech clip has 24 dimensions.

5.3 Recognition

The last step aims at assigning $\vec{x}$ to one of the two classes associated with each personality trait, namely, *High* or *Low* (see Section 6.3).

The classification is performed with a Logistic Regression, a binary classifier expressing the probability of a feature vector belonging to class C as follows:

$$p(C|\vec{x}) = \frac{\exp\left(\sum_{i=1}^{D} \theta_i x_i - \theta_0\right)}{1 + \exp\left(\sum_{i=1}^{D} \theta_i x_i - \theta_0\right)}, \quad (2)$$

where D is the dimension of the feature vectors and the θ_i are the parameters of the model. As the problem is binary, $\vec{x}$ is assigned to C if $p(C|\vec{x}) \geq 0.5$. This value might be changed to take into account, e.g., different class distributions, but the experiments do not take into account this possibility for the sake of simplicity. The model is trained by maximizing the entropy over the training set with the Limited Memory BFGS method [48].[1]

This classifier has several advantages: The first is that it is discriminative and it does not make any assumption about the distribution of the feature vectors $\vec{x}$. The second is that the θ_i parameters weight the features x_i according to their influence on the classification result. This is important in a problem like APP where it is necessary not only to achieve good accuracy, but also to explain what the features are that most influence the perception of the listeners. Since the logistic regression might result in low accuracies, the experiments are completed by using a SVM with Gaussian kernel. While having a lower explanatory power, the SVM might provide better accuracies.

1. See www.cs.grinnell.edu/ weinman/code/index.shtml for implementation details.

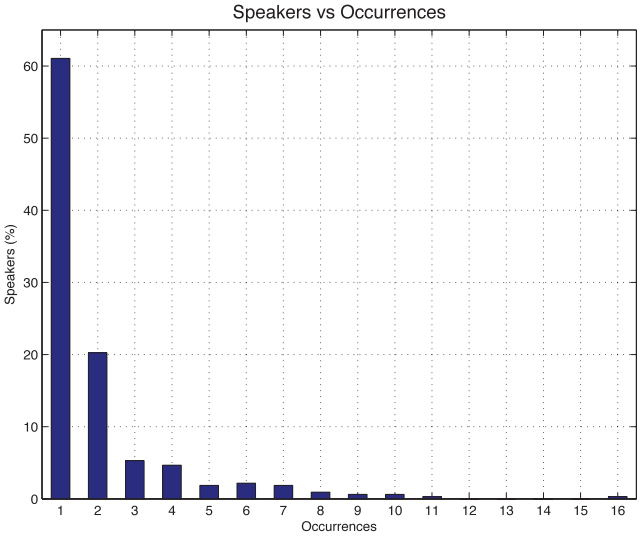

Fig. 1. Distribution of speaker occurrences. The chart shows the percentage of speakers appearing a given number of times. Roughly two-thirds of the individuals represented in the corpus appear only once.

The experimental setup is based on the k-fold cross-validation method [49]: The entire dataset is split into k equal size subsets, $k-1$ are used for training and the remaining one for testing. The procedure is repeated k times and each time a different subset is left out for testing (in the experiments of this work, $k = 15$). This allows one to test the approach over the entire corpus at disposition while keeping a rigorous separation between training and test set. The folds have been obtained with a random process with the only constraint of keeping all samples of a given speaker in the same fold (see Fig. 1 for the distribution of the number of occurrences per speaker). In this way, the task is speaker independent and each fold reproduces, in the limits of statistical fluctuations, gender, and speaker category distribution of the entire dataset (see Section 6.1 for more details about speaker categories).

The performance is expressed in terms of accuracy, percentage of samples correctly classified in the test set. The overall accuracy is the average of the accuracies obtained over the k partitions mentioned above. The statistical significance of differences observed when comparing classifiers is assessed with the t-test. An accuracy difference is considered significant when the p-value (the probability of observing at least such a difference in the hypothesis that the two accuracies result from the same underlying distribution) is lower than 5 percent.

6 Experiments and Results

This section describes experiments and results performed in this work.

6.1 The Data

The corpus used for the experiments contains 640 speech clips where a total of 322 individuals are represented (see Fig. 1 for the distribution of the number of samples per

speaker). The clips have been randomly extracted from the 96 news bulletins broadcast by Radio Suisse Romande, the French speaking Swiss national broadcast service, during February 2005. In 593 cases the length of the clips is exactly 10 seconds, in the remaining 47 samples the length is lower because the randomly extracted segments included more than one voice. The clips are emotionally neutral and they do not contain words that might be easily understood by individuals who do not speak French (e.g., names of places or well-known people). As the judges (see below) do not speak such a language, the personality assessments should be influenced mainly by nonverbal behavior. Furthermore, there is only one speaker per clip to avoid effects due to conversational behavior on personality perception. In any case, since the experiments focus on perceived traits and not on real personality, potential effects of transient states (e.g., emotions) do not represent a major problem. The use of short clips is motivated not only by the social cognition literature (see, e.g., [1], [2]), but also by social psychology observations showing that thin slices of behavior are sufficient to make reasonable guesses about the people we have around [50], [51].

The speakers can be grouped into two major categories: *professional* (307 samples) and *nonprofessional* (333 samples). The former includes journalists that work for Radio Suisse Romande and talk regularly on the radio. The latter includes people who happen to talk on the radio because they are involved in a specific event but do not appear regularly on the media. The assessors (see below) are not aware of the categories.

The personality assessments have been performed by 11 judges who have filled in the BFI-10 questionnaire for each of the clips in the corpus [12]. The assessments have been done using an online system, asking each judge to first listen to a clip and then fill in, immediately after, the questionnaire. It was not possible to move from one clip to the next one before having completed the questionnaire and, once the questionnaire for a given clip was completed, it was not possible to go back and edit the assessments.

The judges do not know one another and they have performed the assessments in different places and at different moments. Hence, the assessments can be considered fully independent of one another (no influences between judges). The average of $|\rho|$ (absolute value of the correlation between assessors) ranges between 0.12 and 0.28 depending on the trait. While weak, such a correlation is statistically significant and it corresponds to the values typically observed in psychological experiments on personality perception [44].

The clips have not been assessed all at once, but in separate sessions of length between 30 and 60 minutes. In any case, the judges have never worked more than 1 hour per day. Furthermore, the clips have been presented in a different (random) order to each judge. In this way, no clips have been assessed systematically at the beginning or at the end of a given session, when tiredness conditions of the assessors can be very different.

The final personality assessments for each clip are obtained by averaging over the scores assigned by each of the judges separately. The results are 5D vectors distributed in a space where each component accounts for a personality trait. The application of Principal Component Analysis to these vectors allows one to project the personality assessments over the bidimensional plan where each point accounts for the personality attributed to a specific speaker (see Fig. 2). The first two principal components account for roughly 70 percent of the variance and at least four components are needed to go beyond 90 percent of the variance. Thus, the BF confirms to be a parsimonious representation where all the components are actually necessary and none of them can be discarded without loosing significant information [18].

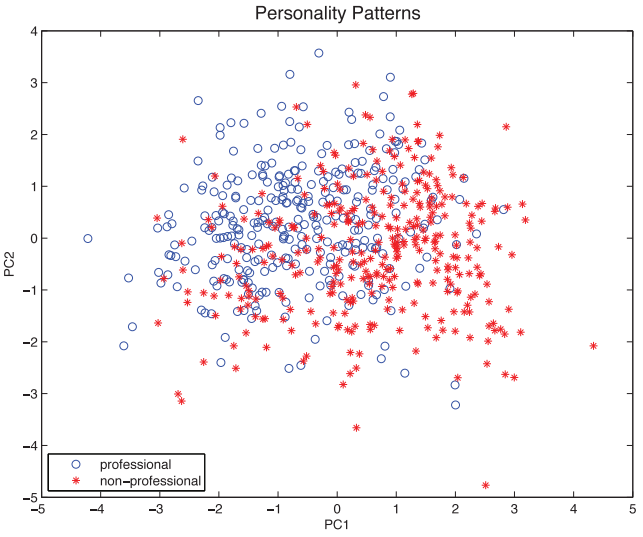

Fig. 2. Personality patterns. The coordinates of each point are the projections of a personality assessment over the first two principal components ($PC1$ and $PC2$).

6.2 Perceived Personality as a Predictor

Consider the set $\Pi = \{\vec{\pi}_1, \ldots, \vec{\pi}_N\}$ of the personality assessments (each $\vec{\pi}_i$ is a 5D vector where the components correspond to the Big Five). Fig. 2 shows the projection of the $\vec{\pi}$ vectors onto the first two principal components [52] extracted from Π (roughly 70 percent of the total variance). Assessments corresponding to professional speakers, 307 samples, and nonprofessional ones, 333 samples, are not completely overlapping (details about the categories are provided in Section 6.1). Each speaker of the corpus belongs to one of the two categories, but the judges are not aware that these exist. Furthermore, none of the items of the BFI-10 (Table 1) is specifically related to one of the two categories.

The partial separation between categories is important because it shows that the assessments are not random, but actually capture meaningful differences between speakers. Furthermore, it suggests that the $\vec{\pi}$'s can be used as feature vectors to automatically classify speakers as professional or nonprofessional. Since the prediction of personal characteristics is one of the most important applications of personality theory [8], the classification of personality assessments can be used as a test to verify whether the ratings are actually coherent or not. This can provide indications that are more useful than those obtained with the simple measurement of the correlation between assessors, typically weak in personality perception experiments [44].

The upper plot of Fig. 3 shows the θ coefficients of a logistic function trained to actually map the π vectors into

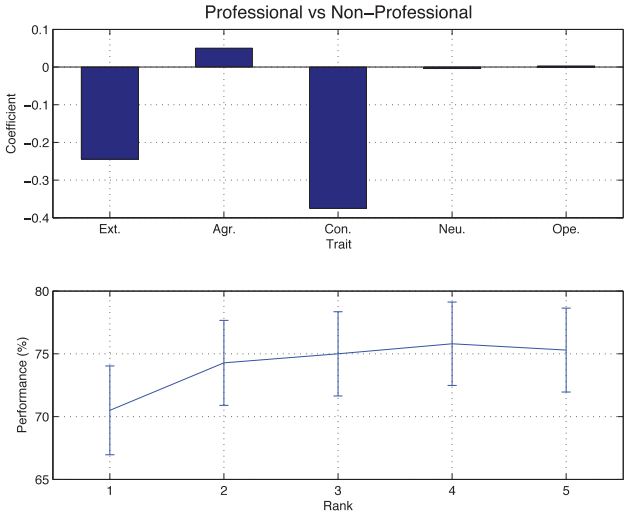

Fig. 3. Categorization. The upper plot shows the weights of the traits used as features to distinguish between professional and nonprofessional speakers. The lower plot shows how the accuracy changes when using an increasing number of traits ordered by their weight (the error bars account for the 95 percent confidence interval).

one of the two classes. The highest coefficients correspond to extroversion and conscientiousness, well known to be the most important traits from a social point of view, especially in zero acquaintance scenarios like the one considered in this work [53]. The probable reason for conscientiousness corresponding to the highest coefficient is that individuals higher in this trait are typically described as "*organized*," "*knowledgeable*," "*thorough*," "*reliable*," etc. [18]. In principle, these are exactly the characteristics that radio speakers try to convey when they talk on the radio.

The θ coefficients allow one to rank the features (traits in this case) according to the absolute value of the corresponding θ_i. The lower plot of Fig. 3 shows how the accuracy changes when using only the conscientiousness score (highest θ_i), when using both conscientiousness and extroversion (the two highest θ_i coefficients), and so on. Using only conscientiousness, the accuracy is around 70 percent (difference with respect to chance statistically significant). By using conscientiousness and extroversion, the accuracy increases to roughly 74 percent (the improvement is not statistically significant). By adding the other traits (ordered by θ coefficient), the accuracy does not increase any more and the recognition rate when using all of the traits is 75.3 percent. These results show that the judges can be considered effective "feature extractors" (or "flexible interpreters" following the definition of [1]) and, overall, the assessments collected in this work are reliable.

6.3 Prosody-Based Personality Perception

Each judge fills out a personality questionnaire for all clips in the corpus. Hence, for a given trait and a given judge, a clip will be either in the upper half of the scores assigned by the judge or in the lower one. This allows one to label a clip as *High* if it is in the upper half for the majority of the judges, or *Low* otherwise. As the number of judges is 11, the majority always includes at least 6 of them. However, the experiments can be restricted to those clips for which the number n of judges in the majority is higher. The prosodic features are extracted using all the material available for each speaker (on average, 40 seconds). This allows a more reliable estimate of the statistical features.

One of the main assumptions behind the use of logistic regression is that the features are not correlated. In the case of our data, the average of $|\rho|$, absolute value of the correlation, is 0.2 (4 percent of the variance in common). Such a value is considered weak and in only two cases does $|\rho| \geq 0.8$, a threshold above which the correlation is considered strong. The first case corresponds to maxima of the first and second formant, the second to a group of four features, including the entropies of the first formant, second formant, length of voiced segments, and length of unvoiced segments. This means that four features are likely to be redundant and therefore they have not been used in the rest of the experiments (maximum of the first formant, entropy of the second formant, entropy of voiced and unvoiced segments length).

Tables 3 and 4 report the accuracy (percentage of correctly labeled clips) of logistic regression and SVM, respectively, as a function of n. No substantial differences can be observed between the two classifiers used to perform APP. The numbers in parentheses correspond to the fraction of clips for which at least n judges actually agree on the same label. The results are compared with a baseline B that corresponds to the performance obtained when predicting always the class with the highest a priori probability. The only trait for which the difference with respect to B is not significant is openness.

The first column of the table reports the results for $n \geq 6$, when the experiments involve the entire corpus. Extroversion and conscientiousness tend to be recognized better than the other traits. In the case of extroversion, this is not surprising because such a trait is typically perceived by people more quickly and accurately [53]. In contrast, the high accuracy on conscientiousness is peculiar to this work. The probable reason is that such a trait is one that better accounts for the difference between professional and nonprofessional speakers, the two categories of individuals represented in the corpus (see Section 6.1). In this respect,

TABLE 3
Logistic Regression Accuracy as a Function of the Agreement between Assessors (Including 95 Percent Confidence Interval)

Trait	$n \geq 6$	B	$n \geq 7$	B	$n \geq 8$	B	$n \geq 9$	B
Extraversion	71.4 ± 3.5 (100.0)	50.0	75.5 ± 3.8 (77.6)	50.0	79.0 ± 4.1 (57.3)	52.0	85.3 ± 4.6 (36.1)	53.0
Agreeableness	58.8 ± 3.8 (100.0)	50.0	61.6 ± 4.6 (67.2)	51.0	67.6 ± 5.7 (40.9)	55.0	63.0 ± 8.7 (18.6)	59.0
Conscientiousness	72.5 ± 3.5 (100.0)	55.0	79.0 ± 3.8 (67.0)	57.0	82.0 ± 4.8 (38.1)	56.0	86.0 ± 7.0 (14.5)	62.0
Neuroticism	66.1 ± 3.7 (100.0)	50.0	69.6 ± 4.3 (68.4)	51.0	72.7 ± 5.5 (38.9)	51.0	74.4 ± 7.4 (20.8)	52.0
Openness	58.6 ± 3.8 (100.0)	61.0	65.7 ± 4.9 (55.1)	65.0	70.6 ± 7.4 (22.8)	69.0	80.0 ± 12.4 (6.2)	67.0

The number in parentheses is the percentage of the corpus for which at least n judges agree on the same label for a given trait. The B columns show the baseline performance, namely, the accuracy of a system that always gives as output the class with the highest a priori probability.

TABLE 4
SVM Accuracy as a Function of the Agreement between Assessors (Including 95 Percent Confidence Interval)

Trait	$n \geq 6$	B	$n \geq 7$	B	$n \geq 8$	B	$n \geq 9$	B
Extraversion	73.5 ± 3.4 (100.0)	50.0	77.9 ± 3.6 (77.6)	50.0	78.5 ± 4.2 (57.3)	52.0	85.3 ± 4.6 (36.1)	53.0
Agreeableness	63.1 ± 3.7 (100.0)	50.0	62.3 ± 4.6 (67.2)	51.0	64.5 ± 5.7 (40.9)	55.0	73.1 ± 8.0 (18.6)	59.0
Conscientiousness	71.3 ± 3.5 (100.0)	55.0	81.1 ± 3.7 (67.0)	57.0	84.4 ± 4.5 (38.1)	56.0	85.0 ± 7.3 (14.5)	62.0
Neuroticism	65.9 ± 3.7 (100.0)	50.0	69.0 ± 4.3 (68.4)	51.0	72.3 ± 5.6 (38.9)	51.0	72.9 ± 7.5 (20.8)	52.0
Openness	60.1 ± 3.8 (100.0)	61.0	67.7 ± 4.9 (55.1)	65.0	71.2 ± 7.3 (22.8)	69.0	75.00 ± 13.4 (6.2)	67.0

The number in parentheses is the percentage of the corpus for which at least n judges agree on the same label for a given trait. The B columns show the baseline performance, namely, the accuracy of a system that always gives as output the class with the highest a priori probability.

the result further confirms the findings of Section 6.2, where conscientiousness is shown to discriminate between the two classes of speakers. The confusion matrices for both Logistic Regression and SVM are reported in Table 5. On average, the performance is roughly the same for both high and low classes. The only exception is openness, but the performance of the classifier is not significantly different from chance for this trait.

One of the main difficulties in APP is that the agreement between raters tends to be low. In the case of this work, the average absolute value of the correlation between assessors ranges between 0.12 and 0.28, depending on the trait, in line with the values typically observed in the psychological literature [44]. Such an effect depends on the inherent ambiguity of the task, especially when it comes to zero acquaintance scenarios like the one considered in the experiments. The influence of the phenomenon above on the accuracy has been assessed by performing tests on subsets of the corpus for which n is higher, i.e., for which there is higher agreement between assessors.

When n increases to 7, the accuracy improves to a statistically significant extent for some traits, but one-third of the data must be eliminated from the corpus, on average. When $n \geq 8$, the fraction of data that can be retained decreases and the result consequently becomes less reliable from a statistical point of view. Furthermore, it becomes more difficult to train the system. However, the trends observed when passing from $n \geq 6$ to $n \geq 7$ seem to be confirmed. The effect of n on the results suggests that the variability of judgment across different assessors is one of the main sources of error in APP. In principle, the only way to address the problem is to remove the samples for which there is no consensus, but such an approach is not correct because there is no "right" or "wrong" perception. In other words, the variability of the ratings does not come from errors, but from the inherent ambiguity of the problem.

One of the main reasons for using the logistic function is that the parameter vector $\vec{\theta}$ provides indications about the features that most influence the outcome of the classifier and, correspondingly, the perception of the assessors. Fig. 4 shows, for each trait, what the θ coefficients associated with the different features are and, in parallel, how the accuracy changes when the number of features increases (the first point corresponds to the use of the only feature corresponding to the highest absolute value $|\theta_i|$, the second point corresponds to the use of the two features corresponding to the two highest absolute values $|\theta_i|$, and so on). The rest of this section shows how such information can be used in the case of $n \geq 6$.

For extroversion, the pitch entropy appears to be the most influential cue, in line with the results of the psychological literature described in Section 3 and showing that higher pitch variability leads to higher extroversion ratings [21]. The same applies to the mean of the unvoiced segments length, a cue related to the length of pauses. The corresponding coefficient is negative because the longer the pauses, the less extroverted a speaker sounds, exactly as observed in [30]. The first two formants appear to play an important role and might account for both gender effects (women tend to have higher formants) and influence of the words being spoken (though the assessors do not understand what the subjects say).

In the case of conscientiousness, the highest coefficients correspond to the entropies of pitch, first formant, and energy, suggesting that greater variety in the way of speaking tends to be perceived as a sign of competence (see [54] and references therein for a confirmation in the psychological literature). The negative coefficients for the mean of the first formant and the minimum of the second one might to correspond to a gender effect known as "*benevolent stereotype*" [53]: Women tend to be perceived as higher in extroversion, but lower in conscientiousness. However, it could be the effect of the words being uttered as well. The negative

TABLE 5
Confusion Matrices for Logistic Regression (Upper Part) and SVM (Lower Part)

	Ext.		Agr.		Con.		Neu.		Ope.	
Class	Low	High	Low	High	Low	High	Low	High	Low	High
Low	76.3	23.7	63.5	36.5	72.8	27.2	62.0	38.0	18.6	81.4
High	33.4	66.6	46.0	54.0	27.7	72.3	29.8	70.2	16.3	83.7

SVM

	Ext.		Agr.		Con.		Neu.		Ope.	
Class	Low	High	Low	High	Low	High	Low	High	Low	High
Low	79.4	20.6	64.7	35.3	69.0	31.0	62.6	37.4	21.5	78.5
High	32.5	67.5	38.5	61.5	26.8	73.2	30.7	69.3	15.5	84.5

The rows correspond to the actual label of the samples, while the columns correspond to the label assigned by the approach. The element (i, j) of each matrix is the percentage of samples belonging to class i that have been assigned to class j.

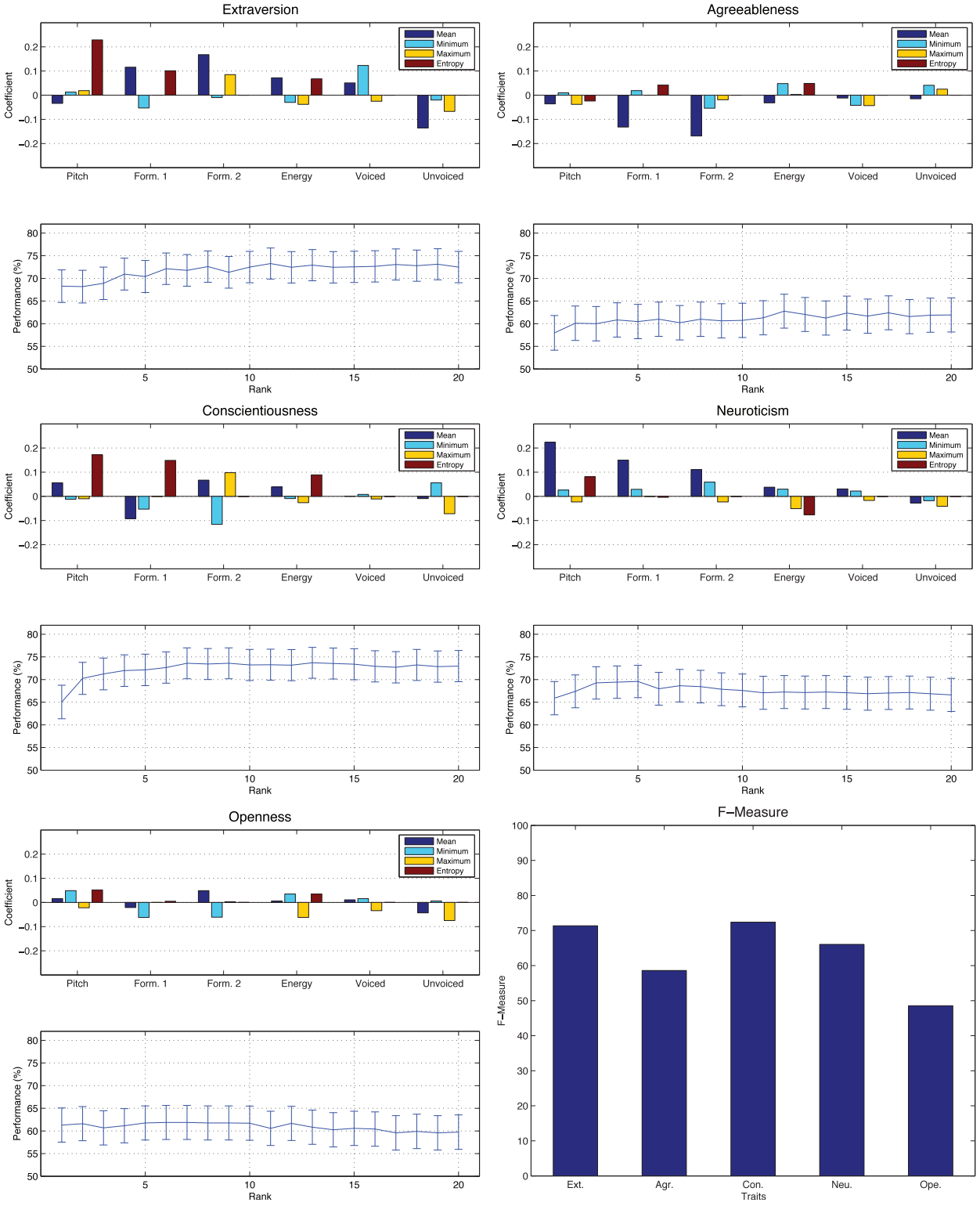

Fig. 4. For each trait, the upper chart shows the θ coefficients associated to the different features. For each cue (e.g., pitch), there are four statistics, namely, mean, minimum, maximum, and entropy. The lower plot of each trait shows the accuracy achieved when using only the N top ranking features (in terms of absolute values $|\theta|$ of the coefficients). The error bars correspond to the 95 percent confidence interval. The last plot shows the F-measures obtained when using all features. All plots correspond to $n \geq 6$.

coefficient for the maximum length of unvoiced segments seems to suggest that people using too many pauses appear to be less competent.

The remaining three traits have not been investigated as thoroughly as the above two in the psychological literature. However, the experiments still propose indications about the cues affecting listener perceptions. The mean of the formants appears to be the only important cue in the case of agreeableness. This suggests that voices with higher formants tend to be perceived as less agreeable. A similar situation is observed for neuroticism, where the means of pitch and first two formants appear to be the most

important cues. In the case of openness, the performance is not significantly better than the baseline B. Hence, the indications of the coefficients cannot be considered reliable. The main reason is probably that this trait is difficult to perceive in the particular setting of the experiments.

7 Conclusions

This work has presented experiments on prosody-based Automatic Personality Perception, i.e., on automatic prediction of personality traits attributed by human listeners to unknown speakers. The experiments of this work have been performed over the largest database used so far for this task in terms of both number of samples and individuals (see Section 4). Furthermore, the experiments propose a first attempt to use pattern recognition approaches to obtain indications about the behavioral cues affecting personality perception in the assessors. Whenever possible, the results in this respect have been matched with the findings of the related psychological literature.

The APP results show an accuracy ranging between 60 and 72 percent (depending on the trait) in predicting whether a speaker is perceived to be high or low with respect to a given trait (see Section 6.3). The accuracy tends to be higher for extroversion and conscientiousness, the two traits people tend to perceive with higher consensus in zero acquaintance scenarios. The accuracy for the latter trait is particularly high with respect to the other works of the literature (see Section 4). The most probable reason is that the corpus includes two categories of speakers (professional and nonprofessional ones) that differ in terms of characteristics typically related to the trait (e.g., thoroughness, reliability, efficiency, etc.).

Since the experiments focus on personality perception (how a person appears to be and not how he/she actually is), the agreement between assessors tends to be low [44]. This seems to be the main source of error in APP, given that the accuracy of both SVM and logistic regression improves when focusing on data for which the agreement is higher. Since there is no "right" or "wrong" perception, the problem above appears to be ineludible in APP. Probably the only solution is to design scenarios where different judges are more likely to agree on attributed traits.

Possible directions for future work have been outlined at the end of Section 4 and include the prediction of personal characteristics and behavior based on automatically perceived or recognized personality traits, the modeling of dimensional personality representations, or the inclusion of cultural effects in both APP and APR. In all cases, the collection of corpora of sufficient size will be one of the main obstacles because gathering personality assessments is an expensive and time-consuming task. This applies in particular to the APP problem, where each subject must be assessed by a sufficient number of judges (at least 10) and, in principle, the judges should be the same for all samples.

Personality significantly influences the existence of individuals in terms of life quality (e.g., professional success, development of stable relationships, etc.) as well as of interactions with others, with machines, and even with the data we consume during a significant fraction of our daily life (television programs, synthetic voices, etc.). Thus, the development of technologies dealing with personality appears to be an important step toward the development of socially intelligent machines capable of dealing with people in the same way that people do.

Acknowledgments

The research that led to this work was supported in part by the European Community Seventh Framework Programme (FP7/2007-2013), under grant agreement no. 231287 (SSPNet), in part by the Swiss National Science Foundation via the National Centre of Competence in Research IM2 (Information Multimodal Information management) and Indo-Swiss Joint Research Project CCPP (Cross-Cultural Personality Perception). The authors wish to thank Marcello Mortillaro for his help on the psychological aspects of this work and Olivier Bornet for the technical support.

References

[1] J.S. Uleman, L.S. Newman, and G.B. Moskowitz, "People as Flexible Interpreters: Evidence and Issues from Spontaneous Trait Inference," *Advances in Experimental Social Psychology,* M.P. Zanna, ed., vol. 28, pp. 211-279, Academic Press, 1996.

[2] J.S. Uleman, S.A. Saribay, and C.M. Gonzalez, "Spontaneous Inferences, Implicit Impressions, and Implicit Theories," *Ann. Rev. of Psychology,* vol. 59, pp. 329-360, 2008.

[3] C. Olivola and A. Todorov, "Elected in 100 Milliseconds: Appearance-Based Trait Inferences and Voting," *J. Nonverbal Behavior,* vol. 34, no. 2, pp. 83-110, 2010.

[4] B. Reeves and C. Nass, *The Media Equation: How People Treat Computers, Television, and New Media like Real People and Places.* Cambridge Univ. Press, 1996.

[5] C. Nass and S. Brave, *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship.* The MIT Press, 2005.

[6] A. Tapus and M. Mataric, "Socially Assistive Robots: The Link between Personality, Empathy, Physiological Signals, and Task Performance," *Proc. Assoc. for Advancement of Artificial Intelligence Spring Symp.,* 2008.

[7] D. Kolar, D. Funder, and C. Colvin, "Comparing the Accuracy of Personality Judgments by the Self and Knowledgeable Others," *J. Personality,* vol. 64, no. 2, pp. 311-337, 1996.

[8] D. Ozer and V. Benet-Martinez, "Personality and the Prediction of Consequential Outcomes," *Ann. Rev. of Psychology,* vol. 57, pp. 401-421, 2006.

[9] K.R. Scherer, "Personality Inference from Voice Quality: The Loud Voice of Extroversion," *European J. Social Psychology,* vol. 8, pp. 467-487, 1978.

[10] R. Picard, *Affective Computing.* The MIT Press, 2000.

[11] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social Signal Processing: Survey of an Emerging Domain," *Image and Vision Computing J.,* vol. 27, no. 12, pp. 1743-1759, 2009.

[12] 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," *J. Research in Personality,* vol. 41, no. 1, pp. 203-212, 2007.

[13] C. Nass and K.M. Lee, "Does Computer-Synthesized Speech Manifest Personality? Experimental Tests of Recognition, Similarity-Attraction and Consistency-Attraction," *J. Experimental Psychology: Applied,* vol. 7, no. 3, pp. 171-181, 2001.

[14] M. Tkalcic, T. Tasic, and J. Kosir, "Emotive and Personality Parameters in Multimedia Recommender Systems," *Proc. IEEE Int'l Conf. Affective Computing and Intelligent Interaction,* 2009.

[15] M. Pantic and A. Vinciarelli, "Implicit Human-Centered Tagging," *IEEE Signal Processing Magazine,* vol. 26, no. 6, pp. 173-180, Nov. 2009.

[16] D. Funder, "Personality," *Ann. Rev. of Psychology,* vol. 52, pp. 197-221, 2001.

[17] G. Matthews, I. Deary, and M. Whiteman, *Personality Traits.* Cambridge Univ. Press, 2003.

[18] G. Saucier and L. Goldberg, "The Language of Personality: Lexical Perspectives on the Five-Factor Model," *The Five-Factor Model of Personality,* J. Wiggins, ed., Guilford Press, 1996.

[19] E. Sapir, "Speech as a Personality Trait," *The Am. J. Sociology,* vol. 32, no. 6, pp. 892-905, 1927.

[20] D.W. Addington, "The Relationship of Selected Vocal Characteristics to Personality Perception," *J. Speech Monographs,* vol. 35, no. 4, pp. 492-503, 1968.

[21] G.B. Ray, "Vocally Cued Personality Prototypes: An Implicit Personality Theory Approach," *J. Comm. Monographs,* vol. 53, no. 3, pp. 266-276, 1986.

[22] K.R. Scherer, "Effect of Stress on Fundamental Frequency of the Voice," *J. Acoustical Soc. of Am.,* vol. 62, no. S1, pp. 25-26, 1977.

[23] K.R. Scherer and U. Scherer, "Speech Behavior and Personality," *Speech Evaluation in Psychiatry,* pp. 115-135, Grune & Stratton, 1981.

[24] B.L. Smith, B.L. Brown, W.J. Strong, and A.C. Rencher, "Effect of Speech Rate on Personality Perception," *J. Language and Speech,* vol. 18, pp. 146-152, 1975.

[25] B.L. Brown, H. Giles, and J.N. Thakerar, "Speaker Evaluation as a Function of Speech Rate, Accent and Context," *J. Language and Comm.,* vol. 5, no. 3, pp. 207-220, 1985.

[26] R.L. Street and B.L. Brady, "Speech Rate Acceptance Ranges as a Function of Evaluative Domain, Listener Speech Rate and Communication Context," *J. Comm. Monographs,* vol. 49, pp. 290-308, 1982.

[27] M.A. Stewart, B.L. Brown, and S. Stewart, "A Comparison of Computer Manipulated Speech Rate with Subjectively Manipulated Speech Rate in Effects upon Personality Attributions," unpublished manuscript, 1984.

[28] G.E. Moore, "Personality Traits and Voice Quality Deficiencies," *J. Speech Disorders,* vol. 4, pp. 33-36, 1939.

[29] C.F. Diehl, R. White, and K.W. Burk, "Voice Quality and Anxiety," *J. Speech and Hearing Research,* vol. 2, pp. 282-285, 1959.

[30] A.W. Siegman and B. Pope, "Effects of Question Specificity and Anxiety Producing Messages on Verbal Fluency in the Initial Interview," *J. Personality and Social Psychology,* vol. 2, pp. 522-530, 1965.

[31] K. Scherer, "Personality Markers in Speech," *Social Markers in Speech,* pp. 147-209, Cambridge Univ. Press, 1979.

[32] P. Ekman, W. Friesen, M. O'Sullivan, and K. Scherer, "Relative Importance of Face, Body, and Speech in Judgments of Personality and Affect," *J. Personality and Social Psychology,* vol. 38, no. 2, pp. 270-277, 1980.

[33] M. Schmitz, A. Krüger, and S. Schmidt, "Modelling Personality in Voices of Talking Products through Prosodic Parameters," *Proc. 12th Int'l Conf. Intelligent User Interfaces,* pp. 313-316, 2007.

[34] J. Trouvain, S. Schmidt, M. Schroder, M. Schmitz, and W.J. Barry, "Modeling Personality Features by Changing Prosody in Synthetic Speech," *Proc. Third Int'l Conf. Speech Prosody,* 2006.

[35] E. Krahmer, S. Van Buuren, and W. Wesselink, "Audio-Visual Personality Cues for Embodied Agents: An Experimental Evaluation," *Proc. Workshop Embodied Conversational Characters as Individuals,* 2003.

[36] A. Tapus, C. Ţăpuş, and M. Mataric, "User-Robot Personality Matching and Assistive Robot Behavior Adaptation for Post-Stroke Rehabilitation Therapy," *Intelligent Service Robotics,* vol. 1, no. 2, pp. 169-183, 2008.

[37] 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," *J. Artificial Intelligence Research,* vol. 30, pp. 457-500, 2007.

[38] F. Mairesse and M. Walker, "Words Mark the Nerds: Computational Models of Personality Recognition through Language," *Proc. 28th Ann. Conf. Cognitive Science Soc.,* pp. 543-548, 2006.

[39] G. Mohammadi, A. Vinciarelli, and M. Mortillaro, "The Voice of Personality: Mapping Nonverbal Vocal Behavior into Trait Attributions," *Proc. Second Int'l Workshop Social Signal Proc.,* pp. 17-20, 2010.

[40] T. Polzehl, S. Moller, and F. Metze, "Automatically Assessing Personality from Speech," *Proc. Fourth IEEE Int'l Conf. Semantic Computing,* pp. 134-140, 2010.

[41] F. Pianesi, N. Mana, and A. Cappelletti, "Multimodal Recognition of Personality Traits in Social Interactions," *Proc. 10th Int'l Conf. Multimodal Interfaces,* pp. 53-60, 2008.

[42] D.O. Olguin, P.A. Gloor, and A. Pentland, "Capturing Individual and Group Behavior with Wearable Sensors," *Proc. Assoc. for Advancement of Artificial Intelligence Spring Symp.,* 2009.

[43] G. Zen, B. Lepri, E. Ricci, and O. Lanz, "Space Speaks: Towards Socially and Personality Aware Visual Surveillance," *Proc. ACM Int'l Workshop Multimodal Pervasive Video Analysis,* pp. 37-42, 2010.

[44] J. Biesanz and S. West, "Personality Coherence: Moderating Self-Other Profile Agreement and Profile Consensus," *J. Personality and Social Psychology,* vol. 79, no. 3, pp. 425-437, 2000.

[45] R. Hogan and M. Harris-Bond, "Culture and Personality," *The Cambridge Handbook of Personality Psychology,* P. Corr and G. Matthews, eds., pp. 577-588, Cambridge Univ. Press, 2009.

[46] P. Boersma, "Praat, a System for Doing Phonetics by Computer," *Glot Int'l,* vol. 5, nos. 9/10, pp. 341-345, 2002.

[47] C. Song, Z. Qu, N. Blumm, and A. Barabási, "Limits of Predictability in Human Mobility," *Science,* vol. 327, no. 5968, pp. 1018-1020, 2010.

[48] D. Liu and J. Nocedal, "On the Limited Memory BFGS Method for Large Scale Optimization," *Math. Programming,* vol. 45, no. 1, pp. 503-528, 1989.

[49] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," *Proc. Int'l Joint Conf. Artificial Intelligence,* pp. 1137-1145, 1995.

[50] N. Ambady, F. Bernieri, and J. Richeson, "Towards a Histology of Social Behavior: Judgmental Accuracy from Thin Slices of Behavior," *Advances in Experimental Social Psychology,* M. Zanna, ed., pp. 201-272, Academic Press, 2000.

[51] N. Ambady and R. Rosenthal, "Thin Slices of Expressive Behavior as Predictors of Interpersonal Consequences: A Meta-Analysis," *Psychological Bull.,* vol. 111, no. 2, pp. 256-274, 1992.

[52] C. Bishop, *Pattern Recognition and Machine Learning.* Springer, 2006.

[53] C. Judd, L. James-Hawkins, V. Yzerbyt, and Y. Kashima, "Fundamental Dimensions of Social Judgment: Understanding the Relations between Judgments of Competence and Warmth," *J. Personality and Social Psychology,* vol. 89, no. 6, pp. 899-913, 2005.

[54] S. Ketrow, "Attributes of a Telemarketer's Voice Persuasiveness," *J. Direct Marketing,* vol. 4, no. 3, pp. 8-21, 1990.

Gelareh Mohammadi received the BS degree in biomedical engineering from the Amirkabir University of Technology, Iran, in 2003 and the MS degree in electrical engineering from the Sharif University of Technology, Iran, in 2006. She is a third year PhD student at the Swiss Federal Institute of Technology, Lausanne (EPFL) and a research assistant at the Idiap Research Institute, Martigny, Switzerland. Her doctoral work investigates the effect of nonverbal vocal behavior on personality perception. Her research interests include social signal processing, image processing, and pattern recognition.

Alessandro Vinciarelli is a lecturer at the University of Glasgow and a senior researcher at the Idiap Research Institute, Switzerland. His main research interest is social signal processing, the new domain aimed at bringing social intelligence in computers. He is the coordinator of the FP7 Network of Excellence SSPNet (www.sspnet.eu), and is, or has been, principal investigator of several national and international projects. He has authored and coauthored more than 60 publications, including one book and 21 journal papers. He has initiated and organized a large number of international workshops (International Workshop on Socially Intelligent Surveillance and Monitoring, International Workshop on Social Signal Processing, etc.). He is a cochair of the IEEE Technical Committee on SSP and an associate editor of the *IEEE Signal Processing Magazine* for the social sciences. Furthermore, he is a founder of a knowledge management company (Klewel) recognized with several national and international prizes (www.klewel.com). He is a member of the IEEE.