Automatic Attribution of Personality Traits Based on Prosodic Features

Gelareh Mohammadi
Idiap Research Institute,1920 Martigny
Switzerland
École Polytechnique Fédérale de Lausanne
(EPFL), Switzerland
gelareh.mohammadi@epfl.ch

Alessandro Vinciarelli
University of Glasgow, G12 8QQ Glasgow,
United Kingdom
Idiap Research Institute,1920 Martigny
Switzerland
vincia@dcs.gla.ac.uk

ABSTRACT
This paper proposes an approach for Automatic Personality Perception, the task of predicting the personality traits attributed by human listeners to speakers they listen to for the first time and they are not acquainted with. The experiments are performed over a corpus of speech clips (330 individuals in total) assessed in personality terms by 11 human judges. The results show that it is possible to predict with an accuracy between 60 and 72 percent (depending on the trait) whether a person is perceived to be in the lower or upper part of the personality scales corresponding to the Big-Five, the five broad dimensions capable of explaining most of the individual differences.

Categories and Subject Descriptors: H.3.1 [Content Analysis and Indexing]. General Terms: Experimentation.
Keywords: Personality Assessment, Big Five Personality Mode, Social Signal Processing, Nonverbal Vocal Behavior

1. INTRODUCTION

Human sciences have shown that people infer unconsciously and spontaneously a wide range of socially relevant characteristics about others (e.g., attitudes, intention, beliefs, etc.), especially during social interactions [17]. This work considers one aspect of this phenomenon, namely the spontaneous attribution of personality traits to unacquainted speakers. In particular, this article proposes an approach for Automatic Personality Perception (APP) based on prosody, the combination of loudness, pitch and speaking rate characterizing the way someone speaks. The main reason behind this choice is that the effect of prosody on personality perception has been extensively investigated in human sciences (e.g., see [15]). Furthermore, domains like Social Signal Processing have shown that non-verbal behavioral cues (e.g., vocalizations, facial expressions, gestures, etc.) are a reliable evidence for machine understanding of social and affective phenomena [18].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
ACM MM’11, November 28 - December 1, Scottsdale, Arizona, USA.
Copyright 2011 ACM 978-1-60558-933-6/10/10 ...$10.00.
Big Five Inventory

a longer questionnaire known as the BFI-10 [12], a short version (see Table 1) of a personality assessment questionnaire. This work has adopted the BFI-10 [12], a short version (see Table 1) of a longer questionnaire known as the Big Five Inventory (BFI) [12]. The main advantage of the BFI-10 is that it can be completed in less than a minute while still providing reliable results.

3. THE APPROACH

The APP approach proposed in this work includes three main steps: extraction of short-term prosodic features, estimate of long-term statistical prosodic properties, and mapping of these latter into perceived personality traits.

3.1 Extraction of Short-Term Features

The approach starts with the extraction of the most important and most extensively investigated in psychology prosody features: pitch (vibration frequency of vocal cords), first two formants (resonance frequencies of vocal tract), voice energy and speaking rate (measured indirectly through voiced and unvoiced segments length). These features are extracted using PRAAT (version 5.1.15) [1] from 40 ms windows at regular time steps of 10 ms and reflect the short-term characteristics of vocal behavior.

3.2 Statistical Features Estimation

Personality perception is not influenced directly by the features above, but rather by their long-term properties. Hence, the approach estimates four statistical properties for each of the six short-term features: minimum, maximum, mean and relative entropy of the variation between values of the same feature extracted from two consecutive windows (see above). Minimum and maximum together account for the dynamic range, while the mean is a short-term characteristic of vocal behavior.

3.3 Recognition

The last step consists in predicting whether a speaker is perceived to be above or below median with respect to a given trait. The reported performance is based on the 24 dimensional feature vectors described above. As there are four statistical features for each of the six low-level features, the total number of features extracted from each clip is 24.

\[ p(C|\vec{x}) = \frac{1}{1 + \exp\{-\sum_{i=1}^{D} \theta_i x_i - \theta_0\}\} \]

where the \( \theta_i \) are the model parameters obtained by maximizing the entropy of the model over a training set. The advantage of such a model is that the weights indicate the contribution of each feature in the classification task (the higher the absolute value of the weight, the higher the contribution). This is important in APP because it can explain what are the nonverbal cues actually influencing the perception of personality. Furthermore, the model is discriminative, thus it does not make any assumption about the distribution of \( \vec{x} \).

The experimental setup is based on the k-fold cross validation method [5]: The entire dataset is split into \( k \) equal size subsets and \( k-1 \) are used for training while the remaining one for testing. This procedure is iterated \( k \) times and each time a different subset is left out for testing (\( k = 15 \) in the experiments of this work). The reported performance (see below) is the average percentage of samples correctly classified over all test subsets.

4. EXPERIMENTS AND RESULTS

The experiments of this study were carried out over a corpus of 640 speech clips, 10 seconds long each, recorded from 96 news bulletins of Radio Suisse Romande, the French speaking Swiss national broadcast service, during February 2005. There is only one speaker per clip and there is a total number of 330 people in the corpus;

\[ \text{See } \text{www.cs.grinnell.edu/~weinman/code/index.shtml } \text{for implementation details.} \]

<table>
<thead>
<tr>
<th>Trait Description</th>
<th>Agree strongly</th>
<th>Agree a little</th>
<th>Neither agree nor disagree</th>
<th>Disagree a little</th>
<th>Disagree strongly</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. This person is reserved</td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
</tr>
<tr>
<td>2. This person is generally trusting</td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
</tr>
<tr>
<td>3. This person tends to be lazy</td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
</tr>
<tr>
<td>4. This person is relaxed, handles stress well</td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
</tr>
<tr>
<td>5. This person has few artistic interests</td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
</tr>
<tr>
<td>6. This person is outgoing, sociable</td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
</tr>
<tr>
<td>7. This person tends to find fault with others</td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
</tr>
<tr>
<td>8. This person does a thorough job</td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
</tr>
<tr>
<td>9. This person gets nervous easily</td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
</tr>
<tr>
<td>10. This person has an active imagination</td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
<td>(5)</td>
</tr>
</tbody>
</table>

Table 1: The BFI-10 questionnaire used in the experiments (as proposed in [12]).
Table 2: Performance as a function of the agreement between assessors. The number in parenthesis is the percentage of the corpus for which at least \( n \) judges agree on the same label for a given trait. For \( n \geq 9 \) the number of samples retained is too small to perform meaningful experiments.

<table>
<thead>
<tr>
<th>Trait</th>
<th>( n \geq 6 )</th>
<th>( n \geq 7 )</th>
<th>( n \geq 8 )</th>
<th>( n \geq 9 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extraversion</td>
<td>73.0 (100.0)</td>
<td>76.3 (77.6)</td>
<td>78.7 (57.3)</td>
<td>84.8 (36.1)</td>
</tr>
<tr>
<td>Agreeableness</td>
<td>62.7 (100.0)</td>
<td>63.7 (67.2)</td>
<td>69.5 (40.9)</td>
<td>73.1 (18.6)</td>
</tr>
<tr>
<td>Conscientiousness</td>
<td>72.7 (100.0)</td>
<td>78.8 (67.0)</td>
<td>82.4 (38.1)</td>
<td>90.9 (14.5)</td>
</tr>
<tr>
<td>Neuroticism</td>
<td>67.7 (100.0)</td>
<td>70.3 (68.4)</td>
<td>74.3 (38.9)</td>
<td>78.9 (20.8)</td>
</tr>
<tr>
<td>Openness</td>
<td>59.8 (100.0)</td>
<td>68.8 (55.1)</td>
<td>74.0 (22.8)</td>
<td>80.0 (6.2)</td>
</tr>
</tbody>
</table>

As the goal of this work is to study the effect of non-verbal vocal behavior on the perception of personality, the judges who assess the clips do not understand French; to avoid effects of verbal and emotional content, the clips are emotionally neutral and do not contain words that might be understood by non-French speakers (e.g., names of people and places). The judges (11 in total) have filled the BFI-10 questionnaire immediately after listening to each clip of the corpus. The clips were played in different random order for each judge in order to avoid tiredness effects (the judges have completed their work over a period of several weeks and they have never worked more than 30-60 minutes per day).

The assessment scores from each judge were grouped into two classes: \textit{High} (above or equal to the median) and \textit{Low} (below the median). The final label of each clip (\textit{High} or \textit{Low}) is based on a majority votes among the all 11 judges.

### 4.1 Automatic Attribution of Personality

The clips of the corpus are labeled as \textit{High} or \textit{Low} based on majority votes among the judges. As the number of judges is 11, there is always a majority of at least 6 judges “voting” for one of the two classes. However, it is possible to restrict the experiments to clips where the number \( n \) of judges voting for the same class is higher. Table 2 reports the performance as a function of \( n \).

The first column of the table (\( n \geq 6 \)) uses the entire corpus. The highest performance has been obtained on \textit{Extraversion}. This is not surprising as this trait is usually being perceived more quickly and accurately by people [3]. In contrast, the high performance on Conscientiousness obtained in this experiment, is not a common finding in the literature. One possible explanation is that this trait might explain the difference between \textit{professional} and \textit{non-professional} speakers, the two main categories of people present in the database.

When \( n \) increases to 7, the performance significantly improves for all traits, but at least one third of clips has to be excluded. This suggests that the ambiguity of the task is the main source of error and clips for which the consensus is higher can be better modeled from an APP point of view.

Figure 1 shows the \( \theta \) parameters corresponding to the two traits corresponding to the highest performances, i.e., Extraversion and Conscientiousness. For \textit{Extraversion}, the pitch entropy has received the highest weight, in line with human sciences investigations showing that higher pitch variability is positively correlated with Extraversion. The mean of unvoiced segments length appears to have high influence as well, meaning that longer pauses and slow rhythm elicit the perception of low Extraversion (as the \( \theta \) parameter is nega-
For Conscientiousness, entropies of pitch, first formant and energy have gained higher coefficients, suggesting that increasing the variety in the way of speaking leads to being perceived as more conscientious, a finding well documented in the psychological literature (see [4] and references therein).

5. CONCLUSION

This paper has presented an approach for Automatic Personality Perception, the automatic prediction of personality traits attributed by human listeners to speakers they are not acquainted with. The experiments have been performed over a corpus of 640 clips (330 individuals in total), to the best of our knowledge, the largest corpus used so far for this type of experiments. The results show that it is possible to predict with an accuracy up to 75 percent (depending on the particular trait), whether a speaker is perceived to be below or above the median with respect to each of the Big Five, the five broad personality dimensions known to account for most of the individual differences.

The results can be considered satisfactory for at least Extraversion and Conscientiousness, the two traits that typically people perceive most quickly and accurately in zero acquaintance scenarios like those considered in this work. However, there are several directions for future improvement. The first is to develop APP approaches capable of predicting the actual assessments, i.e. the continuous values obtained by averaging over the scores assigned by the judges. This would help to obtain more accurate and psychologically meaningful results. The second can be the extension of approaches like this to multimodal data, so that it would be possible to study the effect of different modalities. Furthermore, the indications obtained in this work about the features influencing personality perception should be used to synthesize personality coloured voices.

6. ACKNOWLEDGMENT

This work is supported by the European Community’s Seventh Framework Program (FP7/2007-2013), under grant agreement no.231287 (SSPNet). Furthermore, it is supported by the Swiss National Science Foundation through the Indo-Swiss Joint Research Project “Cross-Cultural Personality Perception” (Grant 122936), and through the National Centre of Competence in Research on Interactive Multimodal Information Management (IM2).

The authors would like to thank Olivier Bornet for technical support and Marcello Mortillaro for scientific advice.

7. REFERENCES