From Speech to Personality: Mapping Voice Quality and Intonation into Personality Differences

ABSTRACT

From a cognitive point of view, personality perception corresponds to capturing individual differences and can be thought of as positioning the people around us in an ideal personality space. The more similar the personality of two individuals the closer their position in the space. This work shows that the mutual position of two individuals in the personality space can be inferred from prosodic features. The experiments, based on ordinal regression techniques, have been performed over a corpus of 640 speech samples comprising 322 individuals assessed in terms of personality traits by 11 human judges, which is the largest database of this type in the literature. The results show that the mutual position of two individuals can be predicted with up to 80% accuracy.

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

1. INTRODUCTION

Social cognition has shown that people attribute, spontaneously and unconsciously, a wide range of socially relevant characteristics to others [17]. Furthermore, the effect is so pervasive and ubiquitous that it takes place not only when people meet others in person, but also when others simply appear in audio and video recordings [14]. From a multimedia point of view, the main effect is that the perception of social and psychological phenomena taking place in the data influences significantly what we remember about the data we consume [3, 5].

This work considers one aspect of this phenomenon, namely the spontaneous attribution of personality traits to unacquainted speakers. In particular, the article proposes an approach for *Automatic Personality Perception* (APP) based on *prosody*, the combination of (i) intonation, namely the combination of loudness, pitch, and speaking rate that characterizes the *way* someone speaks and (ii) voice quality, which reflects the way energy distribution across the frequency spectrum affects speech.

The main motivation for this choice is that the influence of both intonation and voice quality on personality perception has been extensively investigated in human sciences (e.g., see [16]). 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, affective and psychological phenomena [18].

To date, only a few approaches for APP have been proposed in the computing literature (see, e.g., [6, 10, 11]). In contrast, the relationship between prosody and personality perception has been investigated for several decades in human sciences. The main findings can be summarized as follows: (i) high pitch variation tends to be perceived as higher competence and benevolence, and vice-versa [13], (ii) mean pitch tends to have negative correlation with respect to extraversion and dominance for females speakers, but positive correlation for male speakers [15], and (iii) speaking rate tends to be positively correlated with perceived competence [13]. In general, those findings suggest that prosody plays an important role in the way people perceive others.

The experiments of this work, performed over the largest database of speakers assessed in terms of perceived personality traits, show that it is possible to predict the mutual position of two speakers in the personality space with up to 80% accuracy. The proposed approach is based on Ordinal Regression, which is the most suitable methodology to classify ordinally labeled data. To the best of our knowledge, this is the first work that goes beyond the simple prediction of traits attributed to speakers by predicting differences between individuals, in line with the cognitive processes behind personality perception [4].

2. PERSONALITY: MODEL AND DATA

This section presents the personality model employed in this work and describes the data used in the experiments.

The "Big Five" Model.

Personality is the latent construct accounting for "*individuals' characteristic patterns of thought, emotion, and behavior together with the psychological mechanisms - hidden or not - behind those patterns*" [4]. The *Big Five* (BF) personality model is the most commonly applied and accepted personality model [19] and proposes a personality representation based on five traits that have been shown to account for most of the individual differences:

ACM-MM, Nara, Japan.

- *Extraversion*: Active, Assertive, etc.
- *Agreeableness*: Appreciative, Kind, etc.
- *Conscientiousness*: Efficient, Organized, etc.
- *Neuroticism*: Anxious, Self-pitying, etc.
- *Openness*: Artistic, Curious, etc.

The *BF* model represents personalities in terms of five scores (one for each of the traits above) that can be obtained with appropriate assessment questionnaires. The scores measure how well the adjectives accompanying the traits described a given individual. This work adopted the BFI-10 [12], a short version (see Table 1) of a longer questionnaire known as the *Big Five Inventory* (BFI) [12]. Each question in Table 1 is associated to a Likert scale including five points ranging from "*Strongly disagree*" to "*Strongly agree*" and mapped into the interval $[-2, 2]$. The scores corresponding to each trait are obtained by simple numerical calculations performed over the answers to the questionnaire (see [12] for more details). The main advantage of the BFI-10 is that it can be completed in less than a minute while still providing reliable results [12].

The Data.

The experiments of this study were carried out over a corpus of 640 10 seconds long speech clips randomly extracted from the 96 news bulletins that *Radio Suisse Romande*, the French speaking Swiss national broadcast service, has broadcast during February 2005. There is one speaker per clip and the total number of unique speakers is 322. The personality assessment pool included 11 judges that have listened to each clip of the corpus and, immediately after listening, have filled the BFI-10 questionnaire. The judges have never met one another and have worked independently without being co-located (the assessment was performed via an online application). The judges have worked no more than 60 minutes per day (split into two 30 minutes sessions) to avoid tiredness effects. The clips have been presented to each judge in a different and random order to cope with the reduction in attention observed towards the end of each session. The clips are in French and the 11 judges have signed a document where they state that they do not speak or understand such language. This ensures that the content of the clips influences the personality assessment process only to a minor extent.

At the end of the assessment process, each clip is assigned five scores corresponding to the BFs. Each score is the average of the 11 scores assigned individually by the assessors. The average scores for each trait were then converted into N ordinal categories so that they represented a "degree" associated each personality trait. This was achieved by ordering the samples according to the corresponding score and then by splitting the resulting ranking into N equally sized groups.

3. THE APPROACH

The proposed APP approach comprises three main steps: (i) extraction of short-term speech features, (ii) estimation of long-term statistical features, and (iii) mapping of those features into ordinal categories.

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

Table 1: The BFI-10 questionnaire used in the experiments (as proposed in [12]).

Short-Term Feature Extraction.

The feature extraction process starts by identifying the *syllable* boundaries in the speech signal. The reason for doing so is that most relevant prosodic features can be reliably extracted only from *syllable nuclei*, namely the speech segments that most likely correspond to vowels. Hence, the feature extraction process starts with the syllable segmentation approach proposed in [9]. Briefly, such an approach identifies syllables as speech segments enclosed between consecutive energy minima. Although this algorithm has been tested on Italian and English only, the principle is generally valid to identify syllable-like segments, sometimes called *phonetic syllables*. The nuclei are then identified as regions that lie within -3 dB from the energy peak that characterizes each syllable.

The short-term features can be split into two main groups, depending on whether they account for *intonation* or *voice quality*. Intonation features describe prosodic strategies realized in the time domain, whereas voice quality features describe energy distribution across the frequency spectrum (this a very rough distintion as there are many overlaps between the two groups, but it is useful for our purposes). The first group includes *energy*, *pitch* (the perceived fundamental frequency of the utterance) and *syllable length* related measures. Pitch and energy are extracted from 40 ms long windows at regular time steps of 10 ms. The average of the multiple values extracted from the same nucleus are used as pitch and energy of the nucleus.

Voice quality features are based upon the Long Term Average Spectrum (LTAS) of the whole nucleus. The *Harmonicity* measures the ratio between the energy in the periodic part of the speech signal and the noise; in this work it is computed with the method proposed in [2]. The *spectral centroid* is the mean of the frequencies in the spectrum weighted by their magnitude and it describes the energy distribution over the spectrum. The *spectral skewness* is the difference between the energy distributed above and below the spectral centroid. The *spectral kurtosis* measures how much the energy distribution differs from a Gaussian centered around the spectral centroid. The *spectral slope* is the inclination of the line fitting the frequency bins, i.e., the amounts of energy distributed between 0 and 10^3 Hz, 10^3 and 2×10^3 Hz and so on until 4×10^3 Hz (these settings are commonly used in the literature).

The feature set is completed by *jitter*, *shimmer* and *glissando likelihood*. The first two account for the variations of pitch and energy, respectively. The two features are measured every 10 ms in the nuclei. If F_i is the i^{th} measurement of the pitch, the jitter is the average of $|F_i^{-1} - F_{i-1}^{-1}|$. If A_i

	$N=3$			$N=4$			$N=5$			$N=6$		
ρ	0%	50%	80%	0%	50%	80%	0%	50%	80%	0%	50%	80%
Ext.	78.6%	84.2%	88.8%	76.1%	79.5%	81.2%	75.0%	77.3%	76.8%	74.9%	76.9%	81.4%
Agr.	65.8%	69.0%	74.7%	63.6%	67.8%	76.9%	64.6%	67.5%	70.5%	64.1%	67.0%	70.6%
Con.	70.8%	74.8%	76.4%	69.4%	73.9%	81.7%	68.9%	73.6%	74.8%	68.2%	71.3%	75.6%
Neu.	72.0%	75.7%	77.8%	70.4%	74.2%	76.2 %	69.9%	73.4%	73.4%	69.0%	71.3%	69.4%
Ope.	63.9%	70.1%	69.3%	61.3%	64.7%	62.4%	61.6%	66.0%	69.6%	61.3%	65.6%	66.1%

Table 2: Pairwise ranking results. The table reports the accuracy in predicting, for each trait, the speaker that has been scored higher by the assessors. The results were obtained for different numbers N of ordinal categories and different values ρ of rejection rate.

is the i^{th} amplitude measurement, the shimmer is the average of $|A_i - A_{i-1}|$ divided by the average of A_i as per estimated over the five samples from A_{i-2} to A_{i+2}. The glissando likelihood is the ratio between the actual rate of change of the pitch movement crossing the syllable nucleus and the glissando perception threshold that was empirically found in [8]. The value of the glissando likelihood saturates to 1 and gives an account of how likely it is that the pitch movement crossing the syllable nucleus will be perceived as a dynamic rather than a static tone.

Statistical Features Estimation.

At the end of the short-term feature extraction process, each nucleus is represented by a vector where each component corresponds to one of the features above. Statistics estimated over the feature values extracted from each nucleus individually are then used to represented a speech sample. In particular, the mean is computed for all features, the standard deviation is computed for nuclei and syllable length, pitch, energy, spectral slope, harmonicity, and spectral centroid, the entropy is estimated for nuclei and syllable length, pitch, energy, spectral slope, spectral centroid, and glissando likelihood. Mean and bandwidth of the first three formants are also extracted from each syllable nucleus The feature set is completed by the minimum of the pitch and the maximum of the energy. Overall, the number of features is 35.

Ordinal Regression.

Personality perception is about capturing phenotypic differences between individuals. Hence, the last step of the approach consists in automatically ranking people according to the personality traits attributed by human assessors. The most suitable method for such a purpose is *Ordinal Regression* (OR) [7]. In OR, samples $\mathbf{x}_i$ are assigned to ordinal labels y_i belonging to the ordered set $C = (1, 2, \ldots, N)$. This work employs a linear probabilistic approach to OR as in [7]. The assumptions are that (i) the observed ordinal class labels are conditionally independent given the probabilities $\pi_h(\mathbf{x})$ of a sample $\mathbf{x}$ to belong to the h-th ordinal class, and (ii) the following linear model for $h = 1, \ldots, N-1$ holds:

$$\log\left[\frac{p(y \leq h|\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})}{p(y > h|\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})}\right] = \alpha_h + \mathbf{x}^T\boldsymbol{\beta}. \quad (1)$$

In other words, this OR model has parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{N-1})$ and $\boldsymbol{\beta}$ and models the logarithm of the ratio $p(y \leq h|\mathbf{x})/p(y > h|\mathbf{x})$ by a linear combination of the features with a bias term depending on h. Simple calculations show that the above assumption is equivalent to assuming

$$p(y \leq h|\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{1 + \exp[-(\alpha_h + \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})]} = l(\alpha_h + \mathbf{x}^T\boldsymbol{\beta})$$

where $l(z)$ is the *logistic function*. This allows one to obtain the probabilities for the observed y_i as

$$\pi_{y_i}(\mathbf{x}_i) = l(\alpha_{y_i} + \mathbf{x}_i^T\boldsymbol{\beta}) - l(\alpha_{y_i-1} + \mathbf{x}_i^T\boldsymbol{\beta}) \quad (2)$$

for $y_i > 1$ and $\pi_{y_i}(\mathbf{x}_i) = l(\alpha_1 + \mathbf{x}_i^T\boldsymbol{\beta})$ for $y_i = 1$. Based on the conditional independence assumption, the likelihood of n observed samples is readily available as:

$$\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = p(y_1, \ldots, y_n|\mathbf{x}_1, \ldots, \mathbf{x}_n, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{i=1}^{n} \pi_{y_i}(\mathbf{x}_i) \quad (3)$$

The parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are estimated by maximizing the (logarithm of) $\mathcal{L}$ with the *Newton-Raphson* method [1, 7].

4. EXPERIMENTS AND RESULTS

The goal of APP is to rank people according to the personality traits attributed to them by human assessors. One way to evaluate the predictive power of an APP approach is to test its ability to rank correctly all possible pairs of test samples. In order to do so, consider a pair of test samples $\mathbf{x}_i, \mathbf{x}_j$ such that the corresponding labels for a given personality trait satisfy, say, $y_i > y_j$. The performance score is simply the average number of times that the APP approach predicts a label for $\mathbf{x}_i$ that is greater than the label predicted for $\mathbf{x}_j$ over the entire test set (the probability of being correct by chance is 50%). Given the probabilistic nature of the proposed APP approach, predicting the most likely ranking for a pair of test samples $\mathbf{x}_i$ and $\mathbf{x}_j$, with corresponding predictive probabilities $p(h_i|\mathbf{x}_i)$ and $p(h_j|\mathbf{x}_j)$, becomes

$$\underset{(h_i,h_j)\in\mathcal{C}\times\mathcal{C}}{\arg\max} \{p(h_i|\mathbf{x}_i)p(h_j|\mathbf{x}_j)\}, \quad (4)$$

with the constraint that $h_i \neq h_j$. In this application, the number of ordinal categories ranges from 3 to 6, and so the solution to eq. 4 is found by enumerating all possible rankings. Another advantage of taking a probabilistic approach, is that it is possible to reject the percentage ρ of samples where the ordinal regression approach is less confident about the prediction, as illustrated next.

In order to test the approach over the entire corpus while keeping a rigorous separation between training and test set, the experiments were performed using a K-fold validation procedure ($K = 15$) as follows. The corpus was split into K subsets of which $K - 1$ were used for training and one for testing. The folds were obtained randomly, but it was

ensured that the same person did not appear in both training and test set. Performance were evaluated leaving one of the K folds out at each time and averaging the results obtained.

APP Results.

Table 2 reports the results obtained using models with $N = 3, 4, 5$ and 6 (and $\rho = 0\%, 50\%$ and 80%.). The higher the number of ordinal categories N, the higher the resolution at which it is possible to discriminate between people. The performance difference with respect to chance is always statistically significant with p-value $p < 5\%$. The results suggest that the approach is robust with respect to the number of ordinal categories as no major performance losses are observed when going from $N = 3$ to $N = 6$. The influence of ρ depends on the particular trait, but the general trend is of an increase by roughly 5% when going from no rejection to $\rho = 50\%$, and by another 5% when further increasing ρ to 80%.

5. CONCLUSIONS

This work proposes an APP approach to map prosodic features into personality differences. The key elements of the proposed approach are (i) the use of features extracted from intonation and voice quality, (ii) the use of a probabilistic approach to map such features into the personality space, and (iii) a thorough evaluation based on the largest database of personality assessments from radio broadcasts available in the literature. The results show that it is possible to automatically rank people with different degrees of personality traits with an accuracy around 80%. One direction of current investigation is the possibility to employ a methods to better characterize the uncertainty in the predictions by integrating out the parameters of the ordinal regression model, rather than optimizing the likelihood. Also, an extended version of this work will report the analysis of the parameters $\boldsymbol{\beta}$ that allow a direct interpretation of the importance of different features in mapping samples in the personality space.

6. REFERENCES

[1] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer Verlag, 2007.

[2] P. Boersma. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *Proc. of IFA*, volume 17, pages 97–110, 1993.

[3] S. Dumais, E. Cutrell, J. J. Cadiz, G. Jancke, R. Sarin, and D. C. Robbins. Stuff i've seen: a system for personal information retrieval and re-use. In *Proc. of ACM Intl. Conf. on Research and Development in Information Retrieval*, pages 72–79, 2003.

[4] D.C. Funder. Personality. *Annual Reviews of Psychology*, 52:197–221, 2001.

[5] A. Jaimes, N. Sebe, and D. Gatica-Perez. Human-centered computing: a multimedia perspective. In *Proceedings of the ACM Intl. Conf. on Multimedia*, pages 855–864, 2006.

[6] 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*, 30:457–500, 2007.

[7] P. McCullagh. Regression models for ordinal data. *Journal Royal Statistical Society B*, 42:109–142, 1980.

[8] P. Mertens. The prosogram: Semi-automatic transcription of prosody based on a tonal perception model. In *Proc. of Speech Prosody*, 2004.

[9] M. Petrillo and F. Cutugno. A syllable segmentation algorithm for english and italian. In *Proc. of Eurospeech*, pages 2913–2916, 2003.

[10] F. Pianesi, N. Mana, and A. Cappelletti. Multimodal recognition of personality traits in social interactions. In *In Proc. of the Intl. Conf. on Multimodal Interfaces*, pages 53–60, 2008.

[11] T. Polzehl, S. Moller, and F. Metze. Automatically assessing personality from speech. *In Proceedings of IEEE Intl. Conf. on Semantic Computing*, pages 134–140, 2010.

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

[13] G. B. Ray. Vocally cued personality prototypes: An implicit personality theory approach. *Journal of Communication Monographs*, 53(3):266–276, 1986.

[14] B. Reeves and C. Nass. *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press, 1996.

[15] K. R. Scherer. Effect of stress on fundamental frequency of the voice. *in Journal of Acoustical Society of America*, 62(S1):25–26, 1977.

[16] K. R. Scherer. Personality inference from voice quality: The loud voice of extroversion. *European Journal of Social Psychology*, 8:467–487, 1978.

[17] J. S. Uleman, S. A. Saribay, and C. M. Gonzalez. Spontaneous inferences, implicit impressions, and implicit theories. *Annual Reviews of Psychology*, 59:329–360, 2008.

[18] A. Vinciarelli, M. Pantic, and H. Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing Journal*, 27(12):1743–1759, 2009.

[19] J.S. Wiggins, editor. *The Five-Factor Model of Personality*. Guildfor Press, 1996.