# Humans as Feature Extractors: Combining Prosody and Personality Perception for Improved Speaking Style Recognition

*Abstract*—**This paper presents experiments where natural and spontaneous cognitive processes, in particular those who lead to the attribution of personality traits to unacquainted people, are used as a natural form of feature extraction. In particular, personality assessments provided by human judges are used as features to distinguish between professional and non-professional speakers. The same task is performed with prosodic features extracted with a fully automatic process for comparison purposes. Furthermore both prosodic features and personality assessments are combined. The results show that the discrimination between professional and non-professional speaking styles can be performed with an accuracy of $87.2\%$ when using prosodic features, of $75.5\%$ when using personality assessments, and of $90.0\%$ when using the combination of the two.**

*Index Terms*—**Speaking Style, Social Signal Processing, Prosody, Personality Perception, Big Five Personality Model**

## I. Introduction

A large body of evidence shows that we are "flexible interpreters" [1], i.e. we spontaneously infer meaning from everything surrounds us, independently of an explicit goal or need for doing it. The phenomenon is particularly interesting when it concerns others: as soon as we enter in contact with other individuals, we spontaneously attribute to them a large number of socially relevant traits, including goals, beliefs, values, intentions, etc [2].

This work proposes experiments where the phenomenon above is exploited as a natural form of feature extraction, i.e. as a way to represent data (the voice of people talking on the radio in the case of this work) in a form suitable for automatic processing. More in particular, the experiments show that personality traits attributed by judges to speakers they do not understand (because they speak in a foreign language) and they are not acquainted with, can be used as features to distinguish between professional and non-professional speaking styles. Furthermore, the experiments show that the personality traits can be combined with automatically extracted prosodic features (pitch, energy, speaking rate, etc.) leading to statistically significant improvements. In other words, humans appear to be effective feature extractors not only when they act alone, but also when they are combined with machines. Such a result is interesting in the perspective of Implicit Human-Centered Tagging, the effort of using natural behavioral reactions for better indexing and understanding of multimedia data [3].

The experiments are performed over a dataset of $640$ speech clips split into two classes: professional speakers ($309$ samples) and non-professional speakers ($331$ speakers). For each clip, $10$ assessors have filled a questionnaire resulting into a personality assessment in terms of the Big-Five, the five broad personality dimensions that have been shown to capture most of the individual differences [4]. As the assessments are represented with five-dimensional vectors, the average of the $10$ assessments can be used as a feature vector for distinguishing between professional and non-professional speakers.

In parallel, the clips have been processed with a speech processing tool [5] allowing the extraction of features accounting for the speaking style, namely pitch, energy, formants, length of (un-)voiced segments and their respective statistics (minimum, maximum, mean and entropy of variation). This has led to another feature vector that has been used to perform the same classification as above. The two feature vectors (personality traits and prosodic features) have then been combined to verify whether human perception and automatic audio processing are *diverse*, i.e. account for different aspects of the same data (the speech signal in both cases).

The results show that the discrimination between professional and non-professional speaking styles can be performed with an accuracy of 87.2% when using prosodic features, of 75.5% when using personality assessments, and of 90.0% when using the combination of the two. In other words, the combination leads to a statistically significant improvement with respect to the best of the two feature sets. Hence, even though personality assessments have a lower performance, they are still diverse with respect to prosodic features and allow a performance improvement.

The rest of this paper is organized as follows: Section II provides a short introduction to the concept of personality and its measurement, Section III describes the approach used for the experiments, Section IV reports on experiments and results, and Section V draws some conclusions.

## II. Measuring Personality

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*" [6]. This work adopts the Big-Five (BF) model, the personality model most widely accepted and commonly applied [4]. The BF is based on five broad traits that have been shown to account for most of the individual differences (each trait is accompanied by some adjectives commonly associated to it) [4]:

- Extraversion (Active, Assertive, Energetic, etc.)
- Agreeableness (Appreciative, Generous, Kind, etc.)
- Conscientiousness (Efficient, Organized, Reliable, etc.)
- Neuroticism (Anxious, Tense, Touchy, etc.)
- Openness to experience (Artistic, Curious, Original, etc.)

The model represents personalities in terms of five scores corresponding to the above traits and obtained by filling appropriate questionnaires. This work adopts the BFI-10 [7], a questionnaire including ten items that, while needing less than a minute to be filled, it provides reliable personality assessments.

## III. The Approach

The approach proposed in this work includes two main steps, the first is the feature extraction and the second is the mapping of the feature vectors into one of the two classes represented in our data, namely professional and non-professional speakers. The feature extraction is performed with two different techniques, the first is the extraction of prosodic features and the second is the collection of personality assessments.

### A. Extraction of Prosodic Features

The prosody features employed in this work are pitch, first and second formant, energy and speaking rate (measured indirectly through the length of voiced and unvoiced segments). These are not only the most important prosodic features but also the most commonly explored in *speech based personality personality perception* (see [8] for an extensive survey). In the experiments, the features are estimated on $40$ $ms$ windows at regular time steps of $10$ $ms$ using PRAAT [5]. These low level features reflect only short-term characteristics of vocal behavior whereas speaking style recognition is affected by long-term characteristics of vocal behavior. Thus it is necessary to estimate the statistical properties of the low-level features over an entire speech clip.

In this work, four statistical measures have been estimated for each of the primary low-level features: minimum and maximum (together indicate the dynamic range of the vocal features), mean and entropy of feature variation. Entropy measures the uncertainty of a random variable: If $X$ is a discrete random variable with the set of possible values $\mathcal{X} = \{x_1, x_2, ..., x_{|\mathcal{X}|}\}$ then the entropy ($H$) of its distribution $P(X)$ is:

$$H(X) = \frac{-\sum_{i=1}^{|\mathcal{X}|} P(x_i)\log(P(x_i))}{\log(|\mathcal{X}|)} \quad (1)$$

in which $P(x_i)$ is the probability of $X = x_i$ (estimated with the observed frequency of $x_i$) and $|\mathcal{X}|$ is the cardinality of $\mathcal{X}$. The term $\log|\mathcal{X}|$ is a normalization factor, the upper bound $H(X) = 1$ is reached when the distribution is uniform. In this work, this measure has been applied to the first derivative of low-level features. The first derivative accounts for variation during time, so the entropy measures the predictability of feature variation.

### B. Personality Assessment

In this work, $10$ judges have filled the BFI-10 questionnaire for each of the $640$ clips used in the experiments. For each clip, the resulting assessment is the average of the $10$ individual assessments. The judges have filled the questionnaires through an on-line system that they have accessed in a place of their own choice. In this way, they have been working without being physically co-located and any mutual influence has been avoided. The judges do not understand the clips so that they are influenced only (or at least mostly) by nonverbal communication. In order to avoid tiredness effects, the clips have been assessed in a different, random order for each of the judges. Furthermore, the assessments have been done over a period of several weeks and each judge has never worked more than one hour per day.

### C. Speaking Style Recognition

The classification of a given feature vector (including prosodic features, personality assessments, or the concatenation of the two) in terms of speaking style (professional vs non-professional) has been performed with a logistic function. This model estimates the probability of a vector $\vec{f}$ belonging to class $C$ as follows:

$$P(C|\vec{f}) = \frac{1}{1 + \exp(\theta_0 - \sum\limits_{i=1}^{D} \theta_i f_i)} \quad (2)$$

where $D$ is the dimension of $\vec{f}$ and the $\theta_i$ are the model parameters. The advantage of such a model is that the weights give an indication of the contribution of each feature in the classification task. Furthermore, the model does not make any assumption about the distribution of the data [1]. As there are two classes, $\vec{f}$ is assigned to $C$ if $P(C|\vec{f}) \geq 0.5$.

The experimental setup is based on a $k$-fold cross-validation method: The entire dataset is split into $k$ equal size subsets, $k-1$ parts are used for training and the remaining one for testing. This procedure is repeated $k$ times (each time a different subset is used for testing) and the average performance of all $k$ runs will be reported as a performance measure. In our experiments, $k = 15$.

## IV. Experiments and Results

Three experiments have been performed: In the first experiment, the feature vector $\vec{f}$ includes only prosodic features, in the second experiment it includes only personality scores, in the third it includes both prosodic features and personality scores. The rest of this section presents the results that have been obtained.

### A. The data

The corpus used for the experiments includes $640$ speech clips for a total of 330 individuals. Each clip is $10$ seconds long and it has been extracted randomly from a collection of 96 news bulletins broadcast by Radio Suisse Romande, the Swiss

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

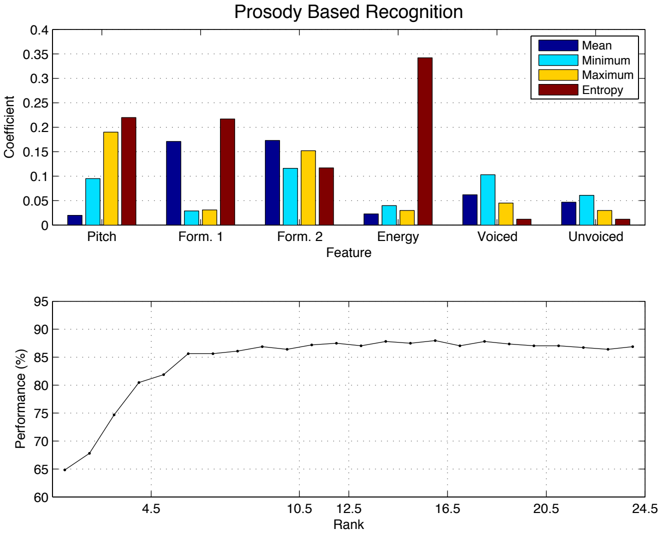

Fig. 1. The upper chart shows the absolute values $|\theta|$ of the coefficients for the different prosodic features. The lower plot shows how the performance changes when using only the feature corresponding to the highest $|\theta|$, only the two features corresponding the two highest $|\theta|$, and so on.

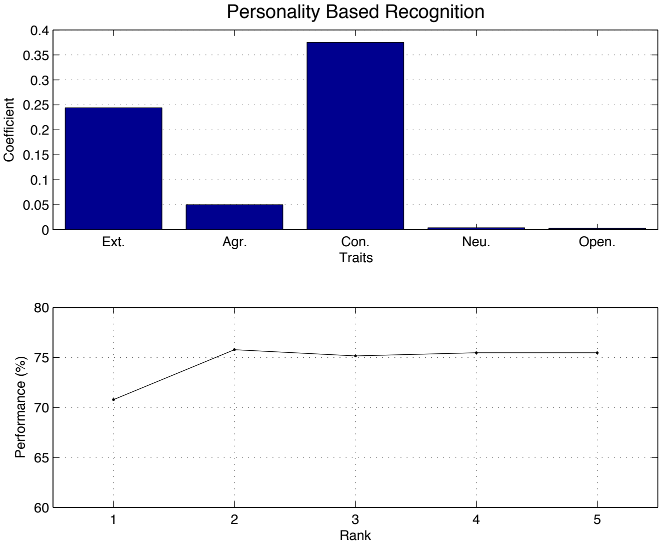

Fig. 2. The upper chart shows the absolute values $|\theta|$ of the coefficients for the different personality traits. The lower plot shows how the performance changes when using only the traits corresponding to the highest $|\theta|$, only the two traits corresponding the two highest $|\theta|$, and so on.

national broadcast service, during February 2005. The clips portray both professional (309) and non-professional (331) speakers. To avoid the effect of verbal content and emotion on personality assessments, the clips are emotionally neutral and 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 do not speak French, the personality assessments should be influenced mainly by nonverbal behavior.

| Experiments | total | "Prof." | "Non-Prof." |
|---|---|---|---|
| Prosody-based | 87.2% | 88.0% | 86.5% |
| Personality-based | 75.5% | 76.2% | 73.8% |
| Combination | 90.00 % | 89.9% | 90.1% |

TABLE I
PERFORMANCE OF PROFESSIONAL/NON-PROFESSIONAL CLASSIFICATION.

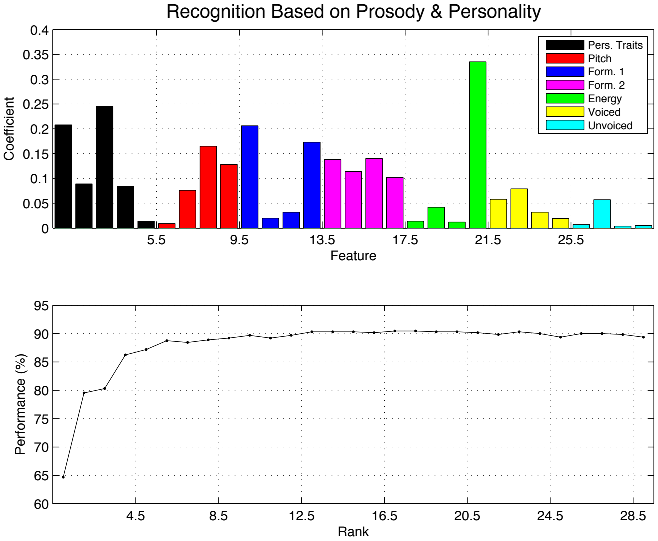

Fig. 3. Upper and lower plots have the same structure as those of the other figures. In the upper plot, the personality traits have the same order as in Figure 2 and the bars associated to each prosodic feature correspond to mean, minimum, maximum and entropy.

### B. *Prosody Based Recognition*

In the first experiment, $\vec{f}$ includes only the prosodic features. Table I reports a recognition rate significantly higher than chance for both classes. The parameters of the logistic function allow one to rank the features according to the influence they have on the classification. Figure 1 shows the absolute value of $(\theta_i)$ for each feature. The lower plot of Figure 1 shows how the performance changes using only the feature with the highest $|\theta|$ value, only the two features corresponding to the two highest $|\theta|$ values, and so on. The plot clearly shows how the six most important features (corresponding to the entropies of pitch energy,first formants, maximum pitch and mean of formants) allow one to reach the same performance as the entire feature set.

### C. *Personality Based Recognition*

In the second experiment, $\vec{f}$ includes only the personality scores obtained during the collection of the assessments. This experiment shows whether there is a difference between professional and non-professional speakers in terms of perceived personality. Furthermore, it shows whether perceived personality can be used as an evidence for discriminating between the two categories of speakers above.

The results are reported in Table I and Figure 2. This latter shows the absolute values of the $\theta$ coefficients for each trait of the *Big Five* model. Not surprisingly, Extraversion and Conscientiousness are the most influential traits (they are well known to be those who are most quickly and accurately perceived in zero acquaintance scenarios [9]). The higher $|\theta|$ for Conscientiousness seems to suggest that such a trait explains most of the difference between professional and non-professional speakers. The lower plot of Figure 2 shows how performance changes when using the traits corresponding to the top $N$ absolute values of $|\theta|$. Conscientiousness and extraversion alone lead to a $74\%$ recognition rate. The performance is lower than in the case of prosodic features, but the model seems to capture correctly the way people perceive, in terms of personality, the difference between professional and non-professional speakers.

### *D. Combination of Prosody and Personality Features*

In the third experiment, $\vec{f}$ includes both prosodic features and personality assessments. The goal is to verify whether the two feature sets are diverse and, if yes, whether their combination can lead to statistically significant improvements. The results are reported in Table I and Figure 3. The recognition rate is higher than the best individual feature set ($p$-value $< 0.05$). The $\theta$ coefficients confirm that the entropies of energy and first formant, mean of first formant, maximum pitch , Conscientiousness and Extraversion are the most important factors influencing the discrimination between professional and non-professional speaking styles. In other words, none of the feature sets prevails on the other and they both carry different information so that the combination can actually be beneficial.

## V. Conclusion

This paper has presented experiments where personality perception, an unconscious process that takes place each time humans enter in contact with an unacquainted person, is used as a natural form of feature extraction in order to distinguish automatically between professional and non-professional speakers. The results show that personality assessments collected in a zero-acquaintance scenario (i.e., in a condition where the assessors do not know the persons they assess) achieve a satisfactory performance for the discrimination between the two categories of people mentioned above. Furthermore, the experiments show that the assessments can improve, to a statistically significant extent, the performance of prosodic features (more effective than personality assessments when used alone).

The results are of interest under two main perspectives. The first is Implicit Human-Centered Tagging, the new domain aimed at using spontaneous cognitive and behavioral processes to extract information from multimedia data, especially when it comes to indexing and content analysis [3]. The second is crowdsourcing [10], the new technique for gathering information from large pools of assessors. In both cases, personality perception might become a technique to model data where people play an important role (e.g., broadcast material, home-videos, video-lectures, etc.).

## References

[1] J. S. Uleman, L. S. Newman, and G. B. Moskowitz, *People as flexible interpreters: Evidence and issues from spontaneous trait inference*. Elsevier, 1996, vol. 28, pp. 211–279.

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

[3] M. Pantic and A. Vinciarelli, “Implicit human-centered tagging,” *IEEE Signal Processing Magazine*, vol. 26, no. 6, pp. 173–180, 2009.

[4] J. Wiggins, Ed., *The Five-Factor Model of Personality*. Guildfor Press, 1996.

[5] P. Boersma, “Praat, a system for doing phonetics by computer,” *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2002.

[6] D. Funder, “Personality,” *Annual Reviews of Psychology*, vol. 52, pp. 197–221, 2001.

[7] 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,” *in Journal of Research in Personality*, vol. 41, pp. 203–212, 2007.

[8] C. Nass and S. Brave, *Wired for Speech*. The MIT Press, 2005.

[9] C. Judd, L. James-Hawkins, V. Yzerbyt, and Y. Kashima, “Fundamental dimensions of social judgment: Unrdestanding the relations bteween judgments of competence and warmth,” *Journal of Personality and Social Psychology*, vol. 89, no. 6, pp. 899–913, 2005.

[10] J. Howe, *Crowdsourcing: Why the power of the crowd is driving the future of business*. Three Rivers Press, 2009.