Speaker Personality Classification Using Systems Based on Acoustic-Lexical Cues and an Optimal Tree-Structured Bayesian Network

Kartik Audhkhasi, Angeliki Metallinou, Ming Li, Shrikanth S. Narayanan

Signal Analysis and Interpretation Lab (SAIL), Electrical Engineering Department
University of Southern California, Los Angeles, CA, USA
{audhkhas, metallin, mingli}@sipi.usc.edu, shri@sipi.usc.edu

Abstract

Automatic classification of human personality along the Big Five dimensions is an interesting problem with several practical applications. This paper makes some contributions in this regard. First, we propose a few automatically-derived personality-discriminating lexical features which provide information complementary to the conventional acoustic-prosodic cues. We also design a frame-level Gaussian mixture model based system which adds complimentary information to the systems trained on global statistical functionals. Next, we note that the Big Five dimensions are correlated and thus model the dependency between these dimensions in the form of an optimal tree-structured Bayesian network. Our final sub-system consists of within class covariance normalization followed by L1-regularized logistic regression. Fusion of all these sub-systems achieves better classification performance than independently trained classifiers using just acoustic features.

Index Terms: Speaker Personality Classification, Bayesian Network Structure Learning, Gaussian Mixture Models, Within Class Covariance Normalization

1. Introduction

Many recent research efforts in speech processing and understanding have focussed on paralinguistic information in addition to the words spoken by the speaker. Well-known examples of such aspects include emotion [1], gender and age [2]. It is believed that paralinguistic information can enhance design of realistic automated agents, such as in interactive voice response systems.

Conventional paralinguistic aspects such as emotion and gender are not the only ones which have been addressed in the speech processing literature. Another important paralinguistic characterization of speakers is based on their personality. The Merriam-Webster dictionary defines personality as *"the complex of characteristics that distinguishes an individual or a nation or group; especially: the totality of an individual's behavioral and emotional characteristics"*. It is typically addressed in literature in terms of the Big Five dimensions - Openness, Conscientiousness, Extroversion, Agreeableness and Neuroticism, abbreviated as OCEAN [3]. Each of these dimensions is highly subjective. Thus, the individual or other human evaluators are asked to fill a questionnaire designed to contain discriminative information about the OCEAN dimensions. The long version of this questionnaire has 44 questions [4], while a more recent one-minute version (BFI-10) just has 10 items [5]. Scores from subsets of questions are averaged to arrive at personality scores for an individual.

The Speaker Trait Challenge at Interspeech 2012 has automatic personality classification as one of the sub-tasks. The Speaker Personality Corpus (SPC) distributed as part of the challenge contains 640 clips randomly extracted from the French news bulletins that Radio Suisse Romande broadcasted during February 2005. Each clip is roughly 10 seconds long. The labeling was performed by 11 judges using the BFI-10 questionnaire. The 11 ratings for a clip along each of the OCEAN dimensions are fused into a single binary label by deciding whether at least 6 judges assign it a score higher than their average for the particular dimension. We are omitting further details about the SPC since they can be found in the overview paper on the challenge [6].

Many previous works have tried to address the problem of automatic classification along the OCEAN dimensions. Mairesse et al. [7] present a survey of various approaches which utilize acoustic-prosodic, lexical and speech act-type features. They use various standard classifiers such as support vector machines and naive Bayes, and achieve significantly better performance than chance accuracy.

The present paper makes several contributions. First, we derive some personality specific rate and durations features using an automatic speech recognition (ASR) system in Section 2.2. We demonstrate that these features provide similar discrimination as conventional acoustic-prosodic features for classification along some of the OCEAN dimensions. We also design a Gaussian Mixture Model (GMM) based system using frame-level features instead of global functionals computed over the entire utterance (Section 3). Such systems have been shown to perform extremely well in another paralinguistic tasks, such as emotion recognition from speech. Next, we observe significant correlation between the OCEAN dimensions for samples in the SPC. This indicates that design of independent classifiers is sub-optimal. We thus propose the use of structured prediction by first learning an optimal tree-structured Bayesian network followed by classification on the lines of tree-augmented Naive Bayes [8] (Section 4). As the final system, we use within class covariance normalization [9] followed by L1-regularized logistic regression in Section 5. Experiments and results on the SPC are presented in Section 6 and we conclude the paper in Section 7.

2. Feature Design

Many studies in psychology and affective sciences have explored the link between personality and acoustic-prosodic characteristics of the speaker. Scherer [10] was one of the pioneering researchers to study this link. He found out that several perceptual cues in speech such as resonance, breathiness, loudness etc. correlate significantly with some personality ratings.

Even though many of the perceptual characterizations considered in that paper were highly subjective (such as gloom), they are linked with simple acoustic-prosodic features (such as energy). The baseline feature set for the Interspeech challenge is motivated by this link.

2.1. Baseline Acoustic-Prosodic Features

The baseline feature set is constructed from a set of 64 low-level descriptors (LLDs) such as Mel Frequency Cepstral Coefficients (MFCCs), pitch, energy etc. computed at the frame level from the speech signal. Since the OCEAN labels are assigned over the entire utterance, the challenge feature set consists of statistical functionals (such as mean, standard deviation, percentiles etc.) of these LLDs computed over all frames in the audio clips. A listing of the LLDs and functionals is provided in Table 1 of the challenge paper [6]. This baseline feature set is 6125-dimensional.

Many researchers, most notably Pennebaker et al. [11, 12], have also studied text-based features, broadly termed as LIWC (Linguistic Inquiry and Word Count). These features show significant correlation with various personality dimensions, which motivated us to explore them further.

2.2. Lexical Rate and Duration Features

Since the SPC does not include the reference audio transcriptions, we used a large vocabulary ASR system for extracting the lexical features. The acoustic and language models trained by the speech group at LIUM, France were used with the Sphinx-3 decoder [13]. The MFCC-based acoustic model consists of 3-state, left-to-right Hidden Markov Models with 5725 tied states and 22 Gaussian mixture components per state. The vocabulary has approximately 64000 words and includes word fragments as well. The fillers include breathiness, laughter and music in addition to silence.

Each utterance was decoded by the ASR system and the top hypothesis with word boundary information was used to extract the rate and duration features. The former capture the rate of speech using various basic units such as words and phonemes. The latter set captures duration of speech and non-speech events such as words and filled pauses. The rate (number per sec) and total duration (normalized by utterance duration) of words, characters, phonemes, fillers, silence, filled pause (*euh*), breathiness and laughter were computed, resulting in a 16-dimensional feature vector for each utterance.

It must be noted that the many of the LIWC features [11, 12] can only be extracted based on word identity (e.g. part-of-speech tags). Since our ASR system output is extremely noisy, we did not extract these word identity-dependent features. Furthermore, the fact that the human evaluators in SPC could not understand French also made word identity-related information potentially redundant.

3. GMM-based Frame-Level System

Taking global statistical functionals of acoustic-prosodic features may smooth-out personality-specific local information. Thus, we train a GMM-based system with frame-level acoustic-prosodic features to capture this detailed information. The feature set for this system contains MFCCs, Mel filter-bank energies (MFBs), pitch, energy, loudness, perceptual linear prediction (PLP) features, zero crossing rate, and other LLDs from the challenge feature set. These features were extracted over 30 msec window with a 10 msec shift.

	O	C	E	A	N
O	-	**0.22**	**0.30**	-0.12	0.10
C	**0.22**	-	**0.37**	0.02	0.08
E	**0.30**	**0.37**	-	**-0.29**	**0.27**
A	-0.12	0.02	**-0.29**	-	**-0.43**
N	0.10	0.08	**0.27**	**-0.43**	-

Table 1: Pearson's correlation coefficient between pairs of OCEAN dimensions. Numbers in bold represent statistically significant correlation coefficients using the 2-sided t-test at the 5% significance level.

In order to reduce the dimensionality of the feature set, we explored principal component analysis and ranking based on Fisher's criterion [14], which gives high score to features that achieve small within-class and large between-class variability. We selected the top 20 features, and further reduced our feature set by excluding features which have a high correlation coefficient (greater than 0.8) with other features. When choosing between two competing, highly-correlated features, we pick the one that has the largest Fisher score. Finally, we append first time derivatives to our feature vector, resulting in a 24 to 40-dimensional vector depending on the personality dimension.

For each personality dimension $Y \in \{O, C, E, A, N\}$, we train GMMs for Y and NY (Λ^1 and Λ^0 respectively) denoting presence and absence of the trait. Given a new test utterance with feature vector $\mathbf{x}$, we use the following maximum a posteriori (MAP) rule for making a decision:

$$P(\mathbf{x}|\Lambda^1)P(1) \underset{NY}{\overset{Y}{\gtrless}} P(\mathbf{x}|\Lambda^0)P(0) \quad (1)$$

4. Tree-Structured Bayesian Network Structure Learning

The baseline system presented in [6] performs classification independently along each of the OCEAN dimensions. However, many pairs of dimensions have appreciable correlation coefficients, as indicated by Table 1. Thus, design of independent classifiers for each of the dimensions makes an unrealistic assumption. Instead, we propose to learn a classifier which jointly predicts all the 5 OCEAN labels for a given utterance. One possibility is to perform multi-class classification over each of the 32 unique label combinations. Another is to construct a fully connected Bayesian network. Both these methods require extremely large amounts of training data, which is not the case in the SPC. Thus, we constrain the Bayesian network to have a tree structure.

The Chow-Liu algorithm [15] was used to infer the optimal tree from the training labels in the maximum likelihood sense. Another interpretation of this algorithm is that it tries to minimize the Kullback-Liebler (KL) divergence between the actual joint distribution of the N random variables (the five OCEAN dimensions in our case) and a second order product approximation. This is equivalent to finding the tree with maximum sum of pairwise mutual informations, which can be done by finding the maximum weight tree from a fully connected graph with mutual information as edge weights. We used the Prim-Jarnik algorithm [16] for this purpose. Fig. 1 shows the optimal tree found by using this algorithm on the training set OCEAN labels. The root node was arbitrarily selected to be neuroticism.

Once this tree is constructed, we can connect the feature vector node ($\mathbf{x}$) to all the other nodes. Ideally, we should have

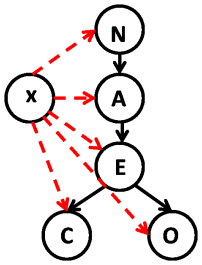

Figure 1: Optimal tree-structured Bayesian network for personality classification. The root node was arbitrarily picked to be neuroticism after structure learning. **x** denotes a feature vector.

learnt these additional links as well. But estimation of mutual information with continuous-valued random variables is difficult, especially with limited training data.

After structure learning, we need to learn the following conditional probability distribution functions (CPDFs) from the training data: $P(N|\mathbf{x}), P(A|N = 0, \mathbf{x}), P(A|N = 1, \mathbf{x}), P(E|A = 0, \mathbf{x}), P(E|A = 1, \mathbf{x}), P(C|E = 0, \mathbf{x}), P(C|E = 1, \mathbf{x}), P(O|E = 0, \mathbf{x}), P(O|E = 1, \mathbf{x})$ where 1 and 0 denote the presence and absence of the trait respectively. Any classifier can be used to estimate these conditional distributions. We used linear SVM with a logistic regression model trained on its scores for this purpose.

During test, the maximum a posteriori (MAP) 5-tuple of OCEAN labels can be found out as:

$$(o, c, e, a, n)_{MAP} = \arg \max_{(o,c,e,a,n)\in\{0,1\}^5} \Big[P(N = n|\mathbf{x}) \\ P(A = a|N = n, \mathbf{x})P(E = e|A = a, \mathbf{x}) \\ P(C = c|E = e, \mathbf{x})P(O = o|E = e, \mathbf{x})\Big] \quad (2)$$

4.1. Smoothing Conditional Probability Distributions

Even though the second order CPDFs require fewer instances to train than the full joint PDF, there is a risk of overfitting in case of the SPC because the training set only has 256 instances, which will further be partitioned into two sets conditioned on the parent random variable being 0 or 1. We thus interpolate all the CPDFs with the corresponding unconditional distribution prior to the test phase as follows:

$$P_{int}(L_2 = l_2|L_1 = l_1, \mathbf{x}) = \alpha P(L_2 = l_2|L_1 = l_1, \mathbf{x}) \\ + (1 - \alpha)P(L_2 = l_2|\mathbf{x}) \quad (3)$$

where L_1 and L_2 are two labels from OCEAN and $\alpha \in [0, 1]$ is the global interpolation weight. Setting $\alpha = 0$ is equivalent to performing independent classification along each of the OCEAN dimensions, i.e. a Bayesian network with no edges. $\alpha = 1$ corresponds to the Bayesian network in Fig. 1. This hyperparameter is tuned based on classification performance on 10-fold cross-validation for each of the five dimensions.

5. Within Class Covariance Normalization Based L1-Regularized Logistic Regression

For our final system, we adopt a binary logistic regression classifier with L1 regularization in LIBLINEAR toolkit [17]. The L1-norm constraint encourages the weight vector to be sparse, which in turn leads to better generalization performance. Each feature dimension was standardized over the training set. Furthermore, there are multiple utterance samples from different speakers for each class in our binary classification task. In order to achieve robust performance, we employed the within class covariance normalization (WCCN) approach [9] to reduce the speaker variability before feeding the features to the logistic regression classifier. Due to the limited number of samples for training, we only perform WCCN after PCA. This also makes the inverse of the average within-class covariance matrix numerically stable. The weight of the L1 term in the log-likelihood objective function is a hyperparameter to be tuned.

6. Experiments and Results

The baseline feature set of 6125 features is too large considering that the number of training instances is only 256. To prevent overfitting, we removed all global functionals except mean and standard deviation. This resulted in a 384-dimensional acoustic-prosodic feature vector. Next, we performed principal component analysis (PCA) over this vector. The top 22 principal components contain 90% of the total variance, and were retained for all future experiments. We noticed that the performance of an SVM classifier along each of the OCEAN dimensions using this low-dimensional feature vector is close to the result using the entire 6125-dimensional feature vector.

Table 2 shows the development set unweighted average recall (UAR) for the multiple systems. The first column shows the baseline using the full feature set. All columns with "AP" denotes that the low-dimensional acoustic-prosodic feature set was used. "Lex" indicates that the 16-dimensional lexical feature set was used. The GMM-based sub-system uses only acoustic-prosodic features since frame-level lexical features are tough to extract. Within each row corresponding to a dimension (e.g. O), we present two results. The top UAR score is for the case when all hyperparameters are tuned on the development set. The bottom one is when the tuning is done using 10-fold cross-validation over the training set. The latter helps us predict the generalization accuracy, while the former is directly comparable to the development set performance given in the challenge paper. SVM's complexity constant is varied over $\{1e^{-5}, 1e^{-4}, 1e^{-3}, 1e^{-2}, 1e^{-1}, 1\}$, and the geometric mean of the best values over the 10-folds is taken. All numbers in bold represent an improvement over the baseline system. A star indicates that the sub-system performed the best for the particular OCEAN dimension. We present four types of fusion results. The first three are achieved by majority vote over all, acoustic-prosodic feature-based and lexical feature-based subsystems respectively. The final column shows the results for the best sub-system (the starred performance in each row).

We can make multiple key observations from the results. First, the Bayesian network and WCCN-based classifiers perform the best for most dimensions, followed by the GMM-based classifier. Second, acoustic-prosodic features give significantly better performance than the automatically derived lexical features in general. This could be attributed to several factors, such as high error rate of the 1-best hypothesis of the ASR system, and short duration of each clip which makes estimation of reliable lexical statistics difficult. Finally, we note that selection of the best classifier for each dimension gives better performance than fusion based on majority vote. This motivates the need for a weighted majority voting scheme which gives higher weight to more accurate systems without completely removing information from the weaker ones.

We next used the sub-systems corresponding to the final ("Best") column to evaluate performance on the test set. All hyperparameters were tuned by 10-fold cross-validation over the

System → Feature →	Baseline	GMM (AP)	BN (AP)	WCCN (AP)	1 SVM (Lex)	BN (Lex)	WCCN (Lex)	Fusion (All)	Fusion (AP)	Fusion (Lex)	Best
O	60.40	**65.36**	**60.78**	**67.57***	56.69	**62.07**	**63.42**	**63.87**	**62.03**	**62.72**	**67.57**
	59.70	**63.02**	**60.34**	**66.96***	57.14	**60.20**	**62.81**	**63.43**	59.62	59.65	**66.96**
C	74.50	73.31	74.29*	73.58	68.37	70.22	69.17	73.58	73.93	68.26	**74.29**
	72.80	70.23	**73.18**	**73.22***	62.35	62.28	68.06	72.48	**73.92**	66.28	**73.22**
E	80.90	79.23	**83.62***	79.75	75.94	77.58	78.11	80.80	80.86	77.58	**83.62**
	81.43	77.63	80.88*	79.75	71.02	75.94	78.10	79.71	80.86	77.03	80.88
A	67.60	65.88	**70.34***	**68.49**	53.99	58.21	56.67	**68.19**	**69.07**	57.71	**70.34**
	64.92	63.66	**69.02***	**66.59**	53.35	56.62	52.37	**66.60**	**68.77**	52.78	**69.02**
N	68.00	66.26	**69.31**	**71.17***	55.88	62.26	58.10	58.97	**71.02**	53.65	**71.17**
	68.62	64.68	68.30	**71.12***	54.93	60.20	55.02	60.00	**71.02**	54.70	**71.12**
Avg	70.30	70.00	69.31	**72.11***	62.17	66.06	65.09	69.08	**71.38**	63.98	**73.40**
	69.49	67.84	**70.34**	**71.53***	59.80	63.05	63.27	68.44	**70.84**	62.09	**72.53**

Table 2: Unweighted average recall (UAR) for personality classification. AP and Lex stand for acoustic-prosodic and lexical cues respectively. For each OCEAN dimension, the number in the top row represents UAR when all hyper-parameters are tuned on the development set. The bottom row gives the UAR when the tuning is done in 10-fold cross-validation on the training set.

training and development set combines. The resulting system achieved an average UAR of 66.02, which is 1.78% less that the baseline of 68.0 using linear SVMs [6]. We hypothesize that this might be due to a mismatch between the test set and the rest of the database, which makes generalization difficult. Limited amount of training data further compounds this issue.

7. Conclusion and Future Work

This paper presents multiple approaches for automatically classifying human personality along the OCEAN dimensions using acoustic-prosodic and lexical cues. The presented systems utilize the characteristics of the challenge corpus in different ways. The GMM-based approach is motivated by the fact that global statistical functional of acoustic-prosodic features loose out on local frame-level information. The Bayesian network system exploits the correlations between the OCEAN dimensions and performs structured prediction. WCCN is motivated by the need to normalize the features in a class-dependent way by modifying the quadratic kernel of a support vector machine. While many of these systems surpass the challenge baselines on the development set, we observed that the overall performance drops slightly below the baseline for the test set. This indicates poor generalization of the trained systems in case of mismatched datasets and limited training data.

The SPC is a challenging database due to the small number of training instances. A major direction of future work would focus on improving the generalizability of all systems, for example by better smoothing and Bayesian techniques. Better feature design is another important area for additional work. Finally, we would also like to explore advanced label fusion techniques for combining the outputs from multiple systems.

8. References

[1] C.C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Communication*, 2011.

[2] M. Li, K. J. Han, and S. Narayanan, "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," *Computer, Speech, and Language*, 2012.

[3] W. T. Norman, "Towards an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality rating," *Journal of Abnormal and Social Psychology*, vol. 66, pp. 574–583, 1963.

[4] O.P. John, EM Donahue, and R. Kentle, "The big five inventory," *Berkeley Institute of Personality and Social Research, University of California*, 1991.

[5] 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*, vol. 41, no. 1, pp. 203–212, 2007.

[6] B. Schuller et al, "The Interspeech 2012 Speaker Trait Challenge," in *Interspeech*, 2012.

[7] F. Mairesse, M.A. Walker, M.R. Mehl, and R.K. Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text," *Journal of Artificial Intelligence Research*, vol. 30, no. 1, pp. 457–500, 2007.

[8] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Machine learning*, vol. 29, no. 2, pp. 131–163, 1997.

[9] A.O. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in *ICSLP*, 2006.

[10] K.R. Scherer, "Personality inference from voice quality: The loud voice of extroversion," *European Journal of Social Psychology*, vol. 8, no. 4, pp. 467–487, 1978.

[11] J.W. Pennebaker, M.E. Francis, and R.J. Booth, "Linguistic inquiry and word count: LIWC 2001," *Mahway: Lawrence Erlbaum Associates*, 2001.

[12] J.W. Pennebaker and L.A. King, "Linguistic styles: Language use as an individual difference.," *Journal of personality and social psychology*, vol. 77, no. 6, pp. 1296, 1999.

[13] Carnegie Mellon University, "Sphinx-3," *Pittsburgh, USA*.

[14] R.O. Duda, P.E. Hart, and D.G. Stork, "Pattern classification and scene analysis," 1995.

[15] C. Chow and C. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Transactions on Information Theory*, vol. 14, no. 3, pp. 462–467, 1968.

[16] V. Jarník, "O jistém problému minimálním," *Práca Moravské Prírodovedecké Spolecnosti*, vol. 6, pp. 57–63, 1930.

[17] R.E. Fan, K.W. Chang, C.J. Hsieh, X.R. Wang, and C.J. Lin, "LIBLINEAR: A library for large linear classification," *The Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.