The INTERSPEECH 2012 Speaker Trait Challenge*

Björn Schuller[1], *Stefan Steidl*[2], *Anton Batliner*[2], *Elmar Nöth*[2],
Alessandro Vinciarelli[3,4], *Felix Burkhardt*[5], *Rob van Son*[6,7], *Felix Weninger*[1],
Florian Eyben[1], *Tobias Bocklet*[2], *Gelareh Mohammadi*[4], *Benjamin Weiss*[5]

[1]Technische Universität München, Institute for Human-Machine Communication, Germany
[2]FAU Erlangen-Nuremberg, Pattern Recognition Lab, Germany
[3]University of Glasgow, School of Computing Science, Scotland
[4]IDIAP Research Institute, Martigny, Switzerland
[5]Deutsche Telekom AG Laboratories, Berlin, Germany
[6]Netherlands Cancer Institute NKI-AVL, Amsterdam, The Netherlands
[7]University of Amsterdam, Phonetic Sciences, Amsterdam, The Netherlands
schuller@tum.de

Abstract

The INTERSPEECH 2012 Speaker Trait Challenge provides for the first time a unified test-bed for 'perceived' speaker traits: Personality in the five OCEAN personality dimensions, likability of speakers, and intelligibility of pathologic speakers. In this paper, we describe these three Sub-Challenges, Challenge conditions, baselines, and a new feature set by the openSMILE toolkit, provided to the participants.

Index Terms: Computational Paralinguistics, Speaker Traits, Personality, Likability, Pathology

1. Introduction

Whereas the first open comparative challenges in the field of paralinguistics targeted more 'conventional' phenomena such as emotion, age, and gender, there still exists a multiplicity of not yet covered, but highly relevant speaker states and traits. In the last instalment, we focused on speaker states, namely sleepiness and intoxication. Consequently, we now want to focus on speaker traits. In that regard, the INTERSPEECH 2012 Speaker Trait Challenge broadens the scope by addressing three less researched speaker traits – the computational analysis of personality, likability, and pathology in speech. Main applications are found in intelligent and socially competent dialogue systems, agents and robots [1], as well as in the medical domain. Three Sub-Challenges are addressed:

In the *Personality Sub-Challenge*, the personality of a speaker has to be determined based on acoustics but potentially including linguistics, for the OCEAN five personality dimensions [2], each mapped onto two classes.

In the *Likability Sub-Challenge*, the likability of a speaker's voice has to be determined by a learning algorithm and acoustic features. While the annotation provides likability in multiple levels, the classification task is binarised.

In the *Pathology Sub-Challenge*, the intelligibility of a speaker in a pathological condition has to be determined by a classification algorithm and acoustic features.

By that – as opposed to the INTERSPEECH 2010 Paralinguistic Challenge – we now deal with perceived speaker traits. The measure of competition will be Unweighted Average Recall of the two classes. Class labels of the train and development sets will be known. All Sub-Challenges allow contributors to find their own features with their own machine learning algorithm. However, a standard feature set will be provided for all Sub-Challenges. Participants will have to stick to the definition of training, development, and test sets. They are encouraged to report on results obtained on the development set, but have only five trials to upload their results on the test sets, whose labels are unknown to them. Each participation will be accompanied by a paper presenting the results that undergoes peer-review and has to be accepted for the conference in order to participate in the Challenge. The organisers preserve the right to re-evaluate the findings, but will not participate themselves in the Challenge. Participants are encouraged to compete in all Sub-Challenges.

In the following we introduce the Challenge corpora (Section 2), features (Section 3), and baselines (Section 4) before concluding (Section 5).

2. Challenge Corpora

2.1. Speaker Personality Corpus (SPC)

In the *Personality Sub-Challenge* the "Speaker Personality Corpus" (SPC) serves for analyses and comparison [3]. The corpus includes 640 clips randomly extracted from the French news bulletins that Radio Suisse Romande, the Swiss national broadcast service, has transmitted during February 2005. There is only one person per clip and the total number of individuals is 322. The most frequent speaker appears in 16 clips, while 61.0 % of the people talk in one clip and 20.2 % in two. The length of the clips is, with a few exceptions, 10 seconds (roughly one hour and 40 minutes in total).

A pool of eleven judges performed the personality assessment. Each judge listened to all clips and, for each one of them, completed the BFI-10, a personality assessment questionnaire commonly applied in the literature [4] and aimed at calculating

* The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 289021 (ASC-Inclusion). The authors would further like to thank the sponsors of the challenge, the HUMAINE Association and Telekom Innovation Laboratories, and Catherine Middag for adding phoneme alignments for the Pathology Sub-Challenge. The responsibility lies with the authors.

Table 1: *Partitioning of Speaker Personality Corpus (X: high on trait X / NX: low on trait X, X ∈ { O, C, E, A, N }). #: number of instances per set and class.*

SPC Sub-Task	#	Train	Devel	Test	Σ
OPENNESS	O	97	70	80	247
	NO	159	113	121	393
CONSCIENTIOUS.	C	110	81	99	290
	NC	146	102	102	350
EXTRAVERSION	E	121	92	107	320
	NE	135	91	94	320
AGREEABLENESS	A	139	79	105	323
	NA	117	104	96	317
NEUROTICISM	N	140	88	90	318
	NN	116	95	111	322
Σ		256	183	201	640

a score for each of the Big-Five dimensions [2]: OPENNESS to experience (Artistic, Curious, Imaginative, Insightful, Original, Wide interests); CONSCIENTIOUSNESS (Efficient, Organized, Planful, Reliable, Responsible, Thorough); EXTRAVERSION (Active, Assertive, Energetic, Outgoing, Talkative); AGREEABLENESS (Appreciative, Forgiving, Generous, Kind, Sympathetic, Trusting); NEUROTICISM (Anxious, Self-pitying, Tense, Touchy, Unstable, Worrying). The BFI-10 was completed via an on-line application, thus the judges were never in direct contact with each other. The judges were allowed to work no more than 60 minutes per day (split into two 30 minutes long sessions) to ensure a proper level of concentration during the entire assessment. Furthermore, the clips were presented in a different order to each judge to avoid tiredness effects in the last clips of a session. The judges signed a formal declaration that they do not understand French, in order to ensure that only nonverbal cues are taken into account. Attention has been paid to avoid clips containing words that might be understood by non-French speakers (e. g., names of places or famous persons) and might have a priming effect. For a given judge, the assessment of each clip yields five scores corresponding to the OCEAN traits. Each clip is labelled to be above average (X) for a given trait X ∈ { O, C, E, A, N } if at least six judges (the majority) assign to it a score higher than their average for the same trait; otherwise, it is labelled NX. Training, development and test set are defined by speaker independent subdivision of the SPC, stratifying by speaker gender (Table 1).

2.2. Speaker Likability Database (SLD)

In the *Likability Sub-Challenge* the "Speaker Likability Database" (SLD) is used [5]. The SLD is a subset of the German Agender database [6], which was originally recorded to study automatic age and gender recognition from telephone speech. The speech is recorded over fixed and mobile telephone lines at a sample rate of 8 kHz. The database contains 18 utterance types taken from a set listed in detail in [6]. An age and gender balanced set of 800 speakers is selected. For each speaker, we used the longest sentence consisting of a command embedded in a free sentence, in order to keep the effort for judging the data by many listeners as low as possible.

Likability ratings of the data were established by presenting the stimuli to 32 participants (17 male, 15 female, aged 20–42, mean=28.6, standard deviation=5.4). To control for effects of gender and age group on the likability ratings, the stimuli were presented in six blocks with a single gender / age group. To mitigate effects of fatigue or boredom, each of the 32 participants rated only three out of the six blocks in randomised order with a short break between each block. The order of stimuli within each block was randomised for each participant as well. The participants were instructed to rate the stimuli according to their likability, without taking into account sentence content or transmission quality. The rating was done on a seven point Likert scale. All participants were paid for their service. A preliminary analysis of the data shows no significant impact of participants' age or gender on the ratings, whereas the samples rated are significantly different (mixed effects model, $p < .0001$). To establish a consensus from the individual likability ratings (16 per instance), the evaluator weighted estimator (EWE) [7] was used. The EWE is a weighted mean, with weights corresponding to the 'reliability' of each rater, which is the cross-correlation of her/his rating with the mean rating (over all raters). For each rater, this cross-correlation is computed only on the block of stimuli which s(he) rated. In general, the raters exhibit varying 'reliability' ranging from a cross-correlation of .057 to .697.

Table 2: *Partitioning of Speaker Likability Database (L: likable / NL: non-likable).*

SLD #	Train	Devel	Test	Σ
L	189	92	119	400
NL	205	86	109	400
Σ	394	178	228	800

The EWE rating was discretised into the 'likable' (L) and 'non-likable' (NL) classes based on the median EWE rating of all stimuli in the SLD. For the Challenge, the data were partitioned into a training, development, and test based on the subdivision for the Interspeech 2010 Paralinguistic Challenge (Age and Gender Sub-Challenges). We 'shifted' roughly 30 % of the development speakers to the test set (in a stratified way), in order to increase its size. The resulting partitioning for this Challenge is shown in Table 2. While the Challenge task is classification, the EWE is provided for the training and development sets, and participants are encouraged to present regression results in their contributions.

2.3. NCSC

For the *Pathology Sub-Challenge* we selected the "NKI CCRT Speech Corpus" (NCSC) recorded at the Department of Head and Neck Oncology and Surgery of the Netherlands Cancer Institute as described in [8]. The corpus contains recordings and perceptual evaluations of 55 speakers (10 female, 45 male) who underwent concomitant chemo-radiation treatment (CCRT) for inoperable tumors of the head and neck. Recordings and evaluations in the corpus were made before and after CCRT: before CCRT (T0; 54 speakers), 10-weeks after CCRT (T1; 48 speakers) and 12-months after CCRT (T3; 39 speakers). Average speaker age was 57. Not all speakers were Dutch native speakers. All speakers read a Dutch text of neutral content.

Thirteen recently graduated or about to graduate speech pathologists (all female, native Dutch speakers, average age 23.7 years) evaluated the speech recordings in an online experiment on an intelligibility scale from 1 to 7. Participants were requested to complete the evaluations in a quiet environment. All participants completed an on-line familarisation module.

All samples were manually transcribed and an automatic phoneme alignment was generated by a speech recogniser

Table 3: *Partitioning of NKI CCRT Speech Corpus (segment level, I: intelligible / NL: non-intelligible).*

NCSC #	Train	Devel	Test	Σ
I	384	341	475	1 200
NI	517	405	264	1 186
Σ	901	746	739	2 386

Table 4: *64 provided low-level descriptors (LLD).*

4 energy related LLD
Sum of auditory spectrum (loudness)
Sum of RASTA-style filtered auditory spectrum
RMS Energy
Zero-Crossing Rate
54 spectral LLD
RASTA-style auditory spectrum, bands 1-26 (0–8 kHz)
MFCC 1–14
Spectral energy 250–650 Hz, 1 k–4 kHz
Spectral Roll Off Point 0.25, 0.50, 0.75, 0.90
Spectral Flux, Entropy, Variance, Skewness, Kurtosis, Slope, Psychoacoustic Sharpness, Harmonicity
6 voicing related LLD
F0 by SHS + Viterbi smoothing, Probability of voicing logarithmic HNR, Jitter (local, delta), Shimmer (local)

Table 5: *Applied functionals.* ^{1}*: arithmetic mean of LLD / positive* Δ *LLD.* ^{2}*: only applied to voice related LLD.* ^{3}*: not applied to voice related LLD except F0.* ^{4}*: only applied to F0.*

Functionals applied to LLD / Δ LLD
quartiles 1–3, 3 inter-quartile ranges
1 % percentile (≈ min), 99 % percentile (≈ max)
position of min / max
percentile range 1 %–99 %
arithmetic mean^1, root quadratic mean
contour centroid, flatness
standard deviation, skewness, kurtosis
rel. duration LLD is above / below 25 / 50 / 75 / 90% range
rel. duration LLD is rising / falling
rel. duration LLD has positive / negative curvature^2
gain of linear prediction (LP), LP Coefficients 1–5
mean, max, min, std. dev. of segment length^3
Functionals applied to LLD only
mean of peak distances
standard deviation of peak distances
mean value of peaks
mean value of peaks – arithmetic mean
mean / std.dev. of rising / falling slopes
mean / std.dev. of inter maxima distances
amplitude mean of maxima / minima
amplitude range of maxima
linear regression slope, offset, quadratic error
quadratic regression a, b, offset, quadratic error
percentage of non-zero frames^4

trained on Dutch speech using the Spoken Dutch Corpus (CGN). Transcription and phonemisation are provided for the participants. For the Challenge, the original samples were segmented at the sentence boundaries. The training, development, and test partitions were obtained by stratifying according to age, gender and nativeness of the speakers, roughly following a 40 % / 30 % / 30 % partitioning (cf. Table 3). The average rank correlation (Spearman's rho) of the individual ratings with the mean rating is .783. In accordance with the Likability Sub-Challenge, the EWE was calculated and discretised into binary class labels (intelligible, non-intelligible), dividing at the median of the distribution. Note that the class labels of the speech segments are not exactly balanced (1 200 / 1 186) since the median was taken from the ratings of the non-segmented original speech. As for likability, we provide the EWE for the training and development sets.

3. Challenge Features

For the baseline acoustic feature set used in this Challenge, we unify the acoustic feature sets used for the INTERSPEECH 2010 Paralinguistic Challenge dealing with ground truth ('non-perceived') speaker traits (age and gender) with the new acoustic features introduced for the INTERSPEECH 2011 Speaker State (SSC) and Audio-Visual Emotion Challenges (AVEC) aiming at the assessment of perceived speaker states. Again, we use TUM's open-source openSMILE feature extractor [9] and provide extracted feature sets on a per-chunk level and a configuration file to allow for additional frame-level feature extraction. The general strategy was to preserve the high-dimensional 2011 SSC feature set including energy, spectral and voicing related low-level descriptors (LLDs); a few LLDs are added including logarithmic harmonic-to-noise ratio (HNR), spectral harmonicity, and psychoacoustic spectral sharpness, as in the AVEC 2011 set. The chosen set of LLDs is shown in Table 4. Regarding functionals, on the one hand, we aim at a more careful selection—for example, from delta regression coefficients we do not compute the simple arithmetic mean as in the 2011 SSC set, but rather the mean of positive values, and the utterance duration is not considered as a useful feature, in contrast to the assessment of speaker states. On the other hand, we added a variety of functionals related to local extrema, such as mean and standard deviation of inter-maxima distances, as in the AVEC 2011 feature set. Furthermore, to compute the location of these extrema, we use a refined peak picking algorithm with respect to the 2011 SSC. The set of applied functionals is given in detail in Table 5. Altogether, the 2012 Speaker Trait Challenge feature set contains 6 125 features, which is roughly a 40 % increase over previous year's feature set.

4. Challenge Baselines

As evaluation measure, we retain the choice of unweighted average (UA) recall as used since the first Challenge held in 2009 [10]. In the given case of two classes ('X' and 'NX'), it is calculated as (Recall(X)+Recall(NX))/2, i. e., the number of instances per class is ignored by intention. The motivation to consider unweighted average recall rather than weighted average (WA) recall ('conventional' accuracy, additionally given for reference) is that it is also meaningful for highly unbalanced distributions of instances among classes, as given in former Challenges, and for more than two classes. In the case of equal distribution, UA and WA naturally resemble each other. In related disciplines of spoken language technology, evaluation often makes use of the Detection Error Trade-off (DET, False Negative Rate vs. False Positive Rate) curve, which is an alternative to the Receiver Operating Characteristic (ROC, True Positive Rate vs. False Positive Rate). As additional measure we thus provide the Area Under

Table 6: Personality, Likability, *and* Pathology Sub-Challenge *baseline results by linear SVM and random forests (ensembles of unpruned REPTrees trained on random feature sub-spaces) by unweighted and weighted average (UA/WA) recall in percent (weighting by number of instances per class). C: complexity parameter; $N \times P$: # of trees / sub-space size; S_{opt}: optimal random seed on Devel; mean ± standard deviation across random seeds for RF. Competition measure is UA.*

Task	*SVM*					*Random Forests*					
		Devel		**Test**			**Devel**			**Test**	
	C	**UA** (WA)	AUC	**UA** (WA)	AUC	$N \times P$	**UA** (WA)	AUC	S_{opt}	**UA** (WA)	AUC
Personality Sub-Challenge											
(N)O	10^{-3}	60.4 (62.8)	67.6	57.8 (59.7)	62.9	100×.1	57.7 ± 2.3 (64.4)	67.0	15	59.0 (63.7)	67.4
(N)C	10^{-2}	74.5 (74.9)	80.0	80.1 (80.1)	84.5	1 000×.02	74.9 ± 0.9 (74.8)	81.2	25	79.1 (79.1)	83.7
(N)E	10^{-2}	80.9 (80.9)	90.5	76.2 (76.6)	84.1	1 000×.01	82.8 ± 0.9 (82.8)	92.0	28	75.3 (75.6)	85.2
(N)A	10^{-3}	67.6 (65.6)	71.1	60.2 (60.2)	62.1	1 000×.01	67.2 ± 1.4 (64.6)	71.6	5	64.2 (64.2)	66.7
(N)N	10^{-2}	68.0 (68.3)	71.9	65.9 (65.7)	71.8	1 000×.05	68.9 ± 0.6 (68.9)	73.5	10	64.0 (63.7)	71.6
Mean		70.3 (70.5)	76.2	68.0 (68.5)	73.1		70.3 (71.1)	77.1		**68.3** (69.3)	74.9
Likability Sub-Challenge											
(N)L	10^{-4}	58.5 (58.4)	60.8	55.9 (56.1)	61.1	1 000×.02	57.6 ± 1.4 (57.5)	57.0	26	**59.0** (59.2)	64.7
Pathology Sub-Challenge											
(N)I	10^{-3}	61.1 (61.0)	63.9	68.0 (66.2)	76.6	1 000×.02	64.8 ± 0.5 (64.8)	69.9	8	**68.9** (67.5)	75.0

the Curve (AUC) as given by the WEKA toolkit that reduces the curve to a single measure. Note, however, that this measure is not compliant with the principle of result calculation in this series of Challenges: It demands for multiple evaluations of the learning algorithm's model or knowledge of confidences per instance; however, participants are allowed to submit only five uploads of their predictions on unlabelled test data and are not required to provide learnt models or confidences.

For the baselines we exclusively exploit acoustic feature information. For transparency and reproducibility, we use open-source classifier implementations from the WEKA data mining toolkit [11]. As classifiers, we first use linear Support Vector Machines (SVM) trained with Sequential Minimal Optimisation (SMO), as they are robust against overfitting in high dimensional feature spaces. We choose the complexity parameter $C \in \{10^{-4}, 10^{-3}, \dots, 1\}$ for the SMO algorithm that achieves best UA recall on the development set. Logistic models were fitted to the SVM outputs to calculate meaningful AUC values. Second, we evaluate Random Forests (RF), which avoid the curse of dimensionality by constructing ensembles of REPTrees trained on random feature subspaces as proposed in [12]. On the development set, we determine an optimal feature subspace size $P \in \{.01, .02, .05, .1\}$ and number of trees $N \in \{100, 200, 500, 1\,000\}$. To allow for robust parameter selection, the parameters N and P yielding the best average UA recall across random seeds 1–30 on the development set are selected. While we display mean and standard deviation of UA recall for the optimal values of N and P, the official Challenge baseline on the test set is the (single) result achieved by choosing N, P and the random seed $S_{\mathrm{opt}} \in \{1, \dots, 30\}$ that is optimal on the development set. For evaluation on the test set, we re-train the models using the training and development set for evaluation on the test set. Parameter selection was found to generalise well to the extended training set.

Table 6 shows that RF deliver slightly better UA recall on the test set than SVM, for all three Sub-Challenges; however, the difference is not significant ($p > .05$ according to a z-test). Furthermore, all results with RF on the test set are significantly above chance level UA ($p < .05$). Of the tasks investigated, the recognition of conscientiousness (80.1 % UA recall on test using RF), extraversion (75.3 %) and intelligibility (68.6 %) can be performed most robustly.

5. Conclusion

We introduced the INTERSPEECH 2012 Speaker Trait Challenge that concentrated on perceived speaker traits. As for previous Challenges, we focused on realistic settings including radio broadcast, mobile phone, and genuine pathologic speech – the baseline results show the difficulty of the investigated automatic recognition tasks. We have provided a baseline using a rather 'brute force' feature extraction and classification approach for the sake of consistency across the Sub-Challenges; particularly, for the *Pathology Sub-Challenge*, no information on the phonetic content is used or assessed in the baseline. Hence, it will be of interest to see the performance of methods that are more tailored to peculiarities of the presented tasks.

6. References

[1] F. Metze, A. Black, and T. Polzehl, "A review of personality in voice-based man machine interaction," in *Proc. HCI International*, vol. 2. Orlando, FL: Springer, 2011, pp. 358–367.

[2] J. Wiggins (ed.), *The Five-Factor Model of Personality*. Guilford, 1996.

[3] G. Mohammadi, M. Mortillaro, and A. Vinciarelli, "The voice of personality: Mapping nonverbal vocal behavior into trait attributions," in *Proc. International Workshop on Social Signal Processing*, Florence, Italy, 2010, pp. 17–20.

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

[5] F. Burkhardt, B. Schuller, B. Weiss, and F. Weninger, "'Would You Buy A Car From Me?'—On the Likability of Telephone Voices," in *Proc. of Interspeech*. ISCA, 2011, pp. 1557–1560.

[6] F. Burkhardt, M. Eckert, W. Johannsen, and J. Stegmann, "A database of age and gender annotated telephone speech," in *Proc. International Conference on Language Resources and Evaluation (LREC)*. ELRA, 2010, pp. 1562–1565.

[7] M. Grimm and K. Kroschel, "Evaluation of natural emotions using self assessment manikins," in *Proc. of ASRU*. IEEE, 2005, pp. 381–385.

[8] L. van der Molen, M. van Rossum, A. Ackerstaff, L. Smeele, C. Rasch, and F. Hilgers, "Pretreatment organ function in patients with advanced head and neck cancer: clinical outcome measures and patients' views," *BMC Ear Nose Throat Disorders*, vol. 9, no. 10, 2009.

[9] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proc. ACM Multimedia*. Florence, Italy: ACM, 2010, pp. 1459–1462.

[10] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 9/10, pp. 1062–1087, 2011.

[11] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The WEKA Data Mining Software: An Update," *SIGKDD Explorations*, vol. 11, 2009.

[12] T. K. Ho, "The Random Subspace Method for Constructing Decision Forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 832–844, 1998.