A Survey on Perceived Speaker Traits: Personality, Likability, Pathology, and the First Challenge

Björn Schuller[1,2], *Stefan Steidl*[3], *Anton Batliner*[2,3], *Elmar Nöth*[3,4], *Alessandro Vinciarelli*[5,6], *Felix Burkhardt*[7], *Rob van Son*[8,9], *Felix Weninger*[2], *Florian Eyben*[2], *Tobias Bocklet*[3], *Gelareh Mohammadi*[6], *Benjamin Weiss*[10]

[1] *Imperial College London, Department of Computing, England*
[2] *Technische Universität München, Machine Intelligence & Signal Processing Group, MMK, Germany*
[3] *Friedrich-Alexander University Erlangen-Nuremberg, Pattern Recognition Lab, Germany*
[4] *King Abdulaziz University, Jeddah, Saudi Arabia*
[5] *University of Glasgow, School of Computing Science, Scotland*
[6] *IDIAP Research Institute, Martigny, Switzerland*
[7] *Deutsche Telekom AG Laboratories, Berlin, Germany*
[8] *Netherlands Cancer Institute NKI-AVL, Amsterdam, The Netherlands*
[9] *University of Amsterdam, Phonetic Sciences, Amsterdam, The Netherlands*
[10] *Technische Universität Berlin, Quality & Usability Lab, Germany*
Corresponding Author: *Björn Schuller*
Imperial College London, Department of Computing, England
`bjoern.schuller@imperial.ac.uk`, *telephone: +44-207-59-48357*

Abstract

The INTERSPEECH 2012 Speaker Trait Challenge aimed at a unified test-bed for perceived speaker traits – the first challenge of this kind: personality in the five OCEAN personality dimensions, likability of speakers, and intelligibility of pathologic speakers. In the present article, we give a brief overview of the state-of-the-art in these three fields of research and describe the three sub-challenges in terms of the challenge conditions, the baseline results provided by the organisers, and a new openSMILE feature set, which has been used for computing the baselines and which has been provided to the participants. Furthermore, we summarise the approaches and the results presented by the participants to show the various techniques that are currently applied to solve these classification tasks.

Keywords: Computational Paralinguistics, Personality, Likability, Pathology, Survey, Challenge

1. Introduction

In 2009 to 2012, challenges (Schuller et al., 2009, 2010, 2011b, 2012, 2013a,b) were organised at the INTERSPEECH conferences that featured several different aspects of paralinguistics: topics of interest were not *what* the speaker said, i. e., word recognition, or the semantics behind word recognition, e. g., hot spots or ontologies, but *how* it was said; for that, pertinent information can either be found between words (vocal, non-verbal events), it can be modulated onto the word chain (typically supra-segmental phenomena such as prosody or voice quality), or it can be encoded in the (types of) words chosen and in the connotations of these words. Catalogues of (short-term) speaker states such as emotions and of (long-term) speaker traits such as gender or personality are given in (Schuller et al., 2013b; Schuller and Batliner, 2014). In the 2012 challenge and accordingly in the present article, we want to address speaker traits that were obtained by perceptual annotation and not by some 'objective' measurement such as placing subjects on a scale to find out about their weight, or simply by deciding between male or female.

There are different definitions for the field that deals with 'how' instead of 'what'; traditionally, *paralinguistics* is mostly conceived as dealing with the non-verbal, vocal aspects of communication, sometimes including, sometimes excluding multi-modal behaviour such as facial expression, hand gesture, gait, body

posture. Here, we follow the definition given in Schuller and Batliner (2014): paralinguistics is *"[...] the discipline dealing with those phenomena that are modulated onto or embedded into the verbal message, be this in acoustics (vocal, non-verbal phenomena) or in linguistics (connotations of single units or of bunches of units)."* Thus, we exclude multi-modality but include verbal phenomena: although most of the contributions to our challenges so far concentrated on acoustics, i. e. on vocal phenomena modulated onto or embedded into the verbal message, we do not want to exclude linguistic approaches such as the modelling of interjections, hesitations, part-of-speech, or n-grams.

Speech is produced by speakers, and when we aim at paralinguistics, then a specific type of speech (friendly speech, pathological speech) characterises a specific type of speakers – they display friendliness or pathological speech traits. Thus, we could subsume all these phenomena under *Speaker Characterisation* or *Speaker Classification* as was done by (Müller, 2007, V): *"[...] the term* speaker classification *is defined as assigning a given speech sample to a particular class of speakers. These classes could be Women vs. Men, Children vs. Adults, Natives vs. Foreigners, etc."*. Eventually, it is simply a matter of perspective whether we call the object of our investigation "type of speech" (indicated by specific speech characteristics) or "speaker traits" (indicated by specific speech characteristics extracted from the speech of specific speakers).

Irrespective of the term chosen, it is always about assigning one individual sample (speech or speaker) to $k = 1, ..., n$ groups (classes) of speakers; the larger n is, the more likely we may employ regression procedures instead of classification. Of course, it is always possible to map more or less continuous attributions such as rating scales onto a few classes. For challenges like the present one, we as organisers have to know which class a speaker in the test set belongs to. As mentioned above, this 'reference' (or 'ground truth', 'gold standard') can be obtained by (sort of) objective measures (for instance, speaker weight classes by following the 'body mass index') or by using perceptive evaluation. In this challenge on perceived speaker traits, we presented three sub-challenges where all speakers were assigned to (two) different classes, based on perceptive evaluation.

Perceptual judgements as basis for reference classes set specific edge conditions: basically, this mostly results in ranked/ordinal scales; however, often parametric procedures such as Pearson's correlation are used. Human annotators do not always agree; thus, we do need some measure for agreement, and some method for ending up with one 'unified' label per token. This is normally the mean of the rating scale scores of all annotators. If we aim at classes, we have to partition the scale at appropriate points (mean, median, etc.).

1.1. Why Such a Challenge: The Motivation Behind

When some of the authors started organising challenges back in 2009, the main motivation behind was to introduce a certain standard of comparability into the field of Computational Paralinguistics, by introducing concepts like

- a partitioning of the database into train, development, and test data; often, there were only train and test partitions, and researchers defined the partitions of the very same corpus in different ways
- a clearcut stratification of subjects for the partitions, if necessary and feasable, for instance, into male/female, old/young, etc.
- the 'open microphone setting' which means that all data recorded and available should be processed; this pertains especially realistic data that often were preselected, based on labeller agreement, quality of recordings, and alike
- adequate performance measures such as Unweighted Average (UA) Recall (UAR), that is, the unweighted (by number of instances in each class) mean of the percentage correctly classified in the diagonal of the confusion matrix; especially for more than two classes, this measure is more adequate than the usual Weighted Average Recall
- both feature extraction and machine learning procedures done with open source tools, to guarantee strict comparability (e. g. of different features, using exactly the same learning algorithm, and of various learning algorithms, using exactly the same features) and repeatability (ensuring, also by means of

software configuration management, that baseline results can be reproduced by anyone with access to the data and open source software, at any time)

- comparability between studies both within the setting of the challenge (this is easy to obtain because the organisers can define the settings in a strict way) and later on, after the challenge (this cannot be ascertained in a strict way, of course, but authors often refer to and apply the challenge settings)

In the later challenges 2010-2012, we basically kept these conditions, with slight modifications: we introduced some more performance measures (correlation and area under ROC (Receiver Operating Characteristic) curve (AUC)); we employed not only free interaction (as in our Speaker Personality Corpus, see Section 3.2) but controlled, prompted data as well (as in our likability and pathology corpus, see Sections 3.3 and 3.4); we implemented larger feature vectors, see Section 3.1.

1.2. Structure of the Article

The three speaker traits dealt with in the challenge are described in Section 2. Previous studies on these traits are summarised below in order to motivate research on their automatic recognition as well as demonstrate feasibility. In Section 3, after shortly presenting the challenge and the unified machine learning framework (feature vectors and learning algorithms employed for computing the baseline results), we introduce the three challenge corpora, together with baseline results. Section 4 presents the contributions to each of the three sub-challenges, and the winners – in contrast to the general literature review, this section serves to review state of the art methods in a comparable setting and to provide a form of quantitative meta-analysis. Section 5 aims at summarising what we have learnt from the challenge.

2. Three Speaker Traits

In this section, we want to give a short account of the state-of-the-art in research on perceived speaker traits within computational paralinguistics. The recognition of perceived speaker traits is exemplified by personality, likability, and pathology. These three traits have been chosen based on the quantity of available labelled data (a crucial prerequisite for meaningful machine learning experiments) and the existence of feasibility studies on automatic classification.

2.1. Speaker Personality

The key-idea of personality psychology is that there are stable individual characteristics that explain most observable differences between people, especially when it comes to overt behaviour. Therefore, the two main goals of personality research are, on the one hand, to identify such characteristics and, on the other hand, to establish a causal link between characteristics and behaviours (Matthews et al., 2009). For this reason, the individual characteristics adopted as a personality basis are one of the main aspects of every theory. Some of the most important personality models ground personality into unconscious (the *psychoanalytic* perspective), inner states (the *humanistic* perspective), mind (the *cognitive* perspective), environment (the *behaviorist* perspective), physiology (the *biological* perspective), etc.; see Funder (2001) for an extensive survey.

The success of a personality theory depends on how effectively it predicts measurable aspects of the life of people (e.g., professional success, amount and quality of social contacts, engagement in criminal activities, etc). From this point of view, the most successful personality theories are based on *traits*, broad dimensions capable of capturing most individual differences. The traits are typically identified by analysing the many adjectives people use to describe others and themeselves (roughly 18 000 in English). However different, such terms revolve around a few main dimensions – adopted as personality traits – that are likely to be the trace that salient psychological phenomena leave in language.

The main criticism against this type of models is that traits are purely descriptive and do not correspond to actual characteristics of individuals (Cloninger, 2009). However, several decades of experiments have shown that the same traits appear with surprising stability across cultures, individuals, situations, etc. This applies in particular to the *Big-Five* (BF) traits (Norman, 1963; McCrae and John, 1992) that are considered today "*the dominant paradigm in personality research, and one of the most influential models in all of psychology*" (McCrae, 2009). The BFs are as follows:

- ***O**penness*: artistic, curious, imaginative, insightful, original, wide interests, etc.
- ***C**onscientiousness*: efficient, organised, planful, reliable, responsible, thorough, etc.
- ***E**xtroversion*: active, assertive, energetic, outgoing, talkative, etc.
- ***A**greeableness*: appreciative, kind, generous, forgiving, sympathetic, trusting, etc.
- ***N**euroticism*: anxious, self-pitying, tense, touchy, unstable, worrying, etc.

In the *Big Five* framework, assessing the personality of an individual means to complete a questionnaire – see Rammstedt and John (2007) for an example – that "*measures*" how well the adjectives above describe the subject under examination. The same questionnaire can be used for self- or other assessment. In the first case, the scores are considered to represent the true personality of an individual (even though the literature shows that self-assessments are often biased towards the image people hope to convey). In the second case, the scores account for the way people are perceived by others. Correspondingly, computing approaches deal with personality under two main perspectives, namely *automatic personality recognition* (APR), aimed at mapping behaviour into self-assessed traits, and *automatic personality perception* (APP), aimed at mapping behaviour into traits assigned by others. Pianesi (2013) describes the main methodological differences between the two cases and shows how APP and APR match different aspects of the "*Brunswick's Lens*", one of the most effective models of human psychology.

Research on human-human communication shows that speech plays an important role for both recognition and perception of personality. On the one hand, according to Scherer (1979), "*personality dispositions are 'externalised' or expressed via indicator cues, i. e., objectively measured speech variables ('distal cues') [...]*". In other words, personality leaves markers in speech and this should make it possible to perform APR using spoken data. On the other hand, following Ekman et al. (1980), "*judgements made from speech alone rather consistently [have] the highest correlation with whole person judgements*". Therefore, speech data should allow one to perform APP reasonably well. While still being limited, the results proposed so far in the speech literature seem to confirm the indications above for both APR (Mairesse et al., 2007; Ivanov et al., 2011) and APP (Mairesse et al., 2007; Mohammadi and Vinciarelli, 2012; Polzehl et al., 2010; Valente et al., 2012; Nass and Min Lee, 2001; Schmitz et al., 2007; Trouvain et al., 2006).

In the work of Mairesse et al. (2007), APR experiments were conducted using the EAR corpus, a collection of random conversation snippets involving 96 subjects. The goal of the experiments was to predict whether each individual was in the upper or lower half of the scores observed for the *Big Five* traits. The performance was higher than chance to a statistically significant extent (65 % accuracy) only for Extroversion. The features were mean, extremes and standard deviation of pitch and intensity. In the same work, Mairesse et al. (2007) proposed APP experiments over the same data and using the same features. The speakers were assessed by 6 judges each and the goal of the tests was to predict the exact value of the personality scores. The best performance was a reduction by roughly 15 % of the error rate achieved when always predicting the average of the observed scores. Similar APR experiments (predicting whether people are above or below the median for each trait) were proposed by Ivanov et al. (2011) over the *PersIA* corpus, 119 conversations where 24 subjects took over the roles of tourist or tour operator. However, only the 12 subjects playing the latter role were used for the tests. The classification, performed by feeding 6 552 features available in openSMILE (Eyben et al., 2010) to BoosTexter (a boosting-based system for text categorisation), achieved an accuracy of 95 % for Extroversion and 63 % for Conscientiousness (no significant results for the other traits).

Similar experiments, but aimed at APP, were proposed by Mohammadi and Vinciarelli (2012). The tests were conducted using the *Speaker Personality Corpus* (see Section 3.2) and the proposed approach aimed at predicting whether speech samples, assessed by 11 judges, were above or below average with respect to each trait. The samples were represented with statistics of pitch, intensity, first two formants and length of voiced and unvoiced segments (minimum, maximum, mean, and relative entropy of differences between consecutive samples). The task was performed with a Logistic Regression and the accuracy was between 60 % and 75 % depending on the traits (best results for Extroversion and Conscientiousness). For their APP experiments, Polzehl et al. (2010) used 220 samples of one professional speaker acting 10 personality types. The features (1 450 in total, including Mel Frequency Cepstral Coefficients (MFCCs), Harmonic-to-Noise-Ratio (HNR),

Zero-Crossing-Rate, etc.) were fed to a support vector machine (SVM) and the accuracy was around 60 %. Valente et al. (2012) proposed experiments over 128 subjects of the AMI corpus, a collection of meetings where individuals are portrayed while interacting with others. The experiments aimed at predicting whether the subjects are above or below the median with respect to the *Big Five* traits and the features included speaking activity (e. g., total amount of speech for a given subject), prosody (speaking rate, mean, minimum, maximum, standard deviation and median of pitch, etc.), N-gram distributions and Dialogue Acts (e. g., questions, statements, etc.). Different types of features tend to predict better different traits – the accuracies range was between 50 % and 68 % – but the combination of all features produced statistically significant results only in the case of Agreeableness and Openness (accuracy around 55 % in both cases). As could be expected, the results improved significantly after excluding from the data subjects in the range $[m_t - 0.5, m_t + 0.5]$, where m is the median and t is one of the *Big Five* traits.

The APP studies presented so far deal with real voices, but the literature shows that human listeners tend to assign personality traits to artificial speakers as well (Nass and Min Lee, 2001; Schmitz et al., 2007; Trouvain et al., 2006). Nass and Min Lee (2001) pioneered the manipulation of prosody (intensity, pitch, pitch range, and speaking rate) aiming at producing speech that sounds more or less extroverted. Their experiments showed not only that the personality ratings assigned by human judges are consistent with the manipulation of prosody, but also that listeners tend to prefer artificial voices that they perceive as being closer to themselves in terms of personality. Such an effect, known as *similarity-attraction*, takes place between humans as well (Aronson et al., 2009) and seems to suggest that people do not make differences, at an unconscious level, between real and artificial voices. In the same vein, the experiments proposed by Schmitz et al. (2007) and Trouvain et al. (2006) show that it is possible to elicit the attribution of predefined traits by changing the characteristics of artificial voices (36 assessors were involved in the tests).

Besides APP in the Big Five framework as defined above, there have been a handful of studies focussing on specific aspects of perceived personality. For instance, it has been shown that observer ratings of charisma are significantly correlated with prosodic features derived from F0, as well as speaking rate and lexical features such as phrase length and use of personal pronouns (Rosenberg and Hirschberg, 2009); however, this study does not report results of automatic classification. Regarding automatic prediction, it has been shown that traits of perceived leadership, comprising the archetypal personalities of charismatics, achievers and team-players, can be predicted with accuracy significantly above chance level (Weninger et al., 2012), reaching above 72 % accuracy in binary classification. In this study, a large set of acoustic features (Schuller et al., 2010) as well as lexical features derived from ASR were used to perform predictions on a database of speeches collected from YouTube, which were rated by observers in the above-mentioned categories. Coninciding with the findings of earlier studies on the correlation of personality and speech (e. g. Scherer (1979)), acoustic features derived from F0 and loudness were found to be most important, while lexical features from ASR performed considerably worse due to high word error rates, owing to the heterogeneity of the recordings.

2.2. Speaker Likability

How much we like a speaker, based on the sound of his/her voice and manner of speaking, is a fascinating yet complex topic. Likability is defined as the attitude of a person towards the speaker and thus represents an individual, subjective expression of valence within a specific relationship (Ferguson and Fukukura, 2012). This high-level evaluation has many facets and might be constituted or at least strongly influenced by varying factors ranging from sexual attraction to reciprocal liking (Aronson et al., 2009). For certain speaker roles and professional situations, likability does also relate to competence (Nesler et al., 1993). As we are dealing with a first impression in a passive rating scenario, this valence information can only be prospective and is likely to be based on vocal stereotypes and attributional information available (Ambady and Skowronski, 2008; Kreiman and Van Lancker Sidtis, 2011), e. g. a friendly tone or benevolent character as potential signal for reciprocal liking.

Typical items to assess likability are, e. g. 'I think he/she could be a friend of mine', 'I would like to have a friendly chat with her/him', 'It would be difficult to meet and talk with him/her', 'He/she just wouldn't fit into my circle of friends', 'We could never establish a personal friendship with each other', 'He/she would be pleasant to be with', see the revised version of McCroskey and McCain (1974); or adjectives such as 'likeable',

'friendly', 'socially attractive', 'relaxed' (Schweitzer and Lewandowski, 2013). Others just use a single scale (Kenny, 1994; Gravano et al., 2011).

With respect to the database used in the challenge, we left the exact definition open by deliberately not stating the nature of 'likability' that the listeners should judge. Instead, we used a single scale (the German antonyms *sympathisch–unsympathisch*, representing an unambiguous appraisal of valence). This scale has a strong correlation ($r = .86$) with the 'likability'-factor of a full-fledged questionnaire for a different database, but collected in a comparable fashion (20 speakers, 46 raters), see Weiss and Burkhardt (2010); Weiss and Möller (2011). The likeability factor of the questionnaire comprises the items likable, pleasant, friendly, and sympathetic (only positive poles listed). However, due to the laboratory set-up of the rating procedure, the telephone quality of the samples, and the technical content, a more factual and formal situation can be assumed for most raters, comparable, e. g. to telephone-based service communication.

Until recently, this aspect of speaker classification has not been studied very often, when compared with other paralinguistic phenomena, for example with the expression of affect. Therefore, we additionally cite studies on potentially related concepts in the following, like pleasant or attractive voices, or speaking proficiency; this is, however, not a complete overview of all these fields of research.

It is obvious that the impression a voice or manner of speaking has on us is highly idiosyncratic and differs between cultural contexts, similarity attraction being an important factor (Aronson et al., 2009). Prosodic features, especially pitch and duration, seem to be strongly correlated with the expertise in speaking (e. g. the acoustic measures F0, rate, pauses, but also fluency (Strangert and Gustafson, 2008; Moniz et al., 2008)); this has been more often addressed in the literature than the question 'what makes a good voice'. Also, differences in pronunciation can result in a negative evaluation (Eklund and Lindström, 2001; van Bezooijen, 2005). For the case of professional speakers, politicians' speech samples were analysed for charisma, using lexical and prosodic information, in (Rosenberg and Hirschberg, 2009). While this trait can be related to likability, the aforementioned study does not report results of automatic classification. A cross-cultural study on likability perception of speech has – to our knowledge – not been done yet. In a study contrasting Swedish and US-American speakers, Dahlbäck et al. (2007) found that participants' preferred speakers mirror their own accent in a tourist information system.

In order to automatically detect awkward, friendly, assertive, or flirtatious speaking style in cross-gender conversations, prosodic, lexical, and conversational features were analysed in Ranganath et al. (2013). Results are better than the baseline even for speaker-independent classification, except for female self-ratings of friendliness. A friendly stance is characterized by a conversational style including laughter and appreciations. These results are currently valid only for this special situation (speed-dating sessions). For everyday social interaction of one female Japanese speaker, voice quality measured by a normalized amplitude quotient revealed highest values (i.e. breathiest) for a careful style, mediocre values for a friendly style, and lowest values for a casual style (Campbell and Mokhtari, 2003). A study solely concerned with the prosodic features F0 and durations synthesised Chinese speech styles, based on findings from a rating experiment (Li and Wang, 2004): F0 was found to be the most important cue for friendliness, i. e. higher F0 for prosodic words, and higher ratings for interrogative sentences.

One of the most often studied aspects of likability is interpersonal attraction between men and women. In Collins (2000), men with voices with closely spaced, low-frequency harmonics were judged as being more attractive, older and heavier, more likely to have a hairy chest and being of a more muscular body type. This was confirmed by Hodges-Simeon et al. (2010), who found that ratings of attractiveness by women for men's voices were predicted by low mean F0, lower spacial distribution of formants and high intensity, across fertility cycle phase and mating context.

Ketzmerick studied acceptability of voices for e-learning applications. In her study, low pitch, "clear" voice, and constant voice "capacity" – assessed perceptually – were relevant for likable voices (Ketzmerick, 2007). From a different analysis she found correlations of these characteristics with acoustic parameters, especially with minimal F0 and F0 frequency distribution. When analysing single long vowels, a correlation between pleasantness of men's voices and pitch and the second formant frequency were found (Jürgens et al., 1996). The authors concluded in advising to choose low and throaty voices for generating text-to-speech systems. In Chattopadhyay et al. (2003), speech rate, pausing, and pitch were analysed as for their effects in advertising, showing that higher rate and lower pitch significantly correlate with a positive attitude towards

the advertisement. For clear speech, it could be shown that out of several descriptive factors especially *warm/relaxed* correlates significantly with likability, and with acoustic parameters of less pressed, more breathy voice quality and lower spectral centre of gravity (Weiss and Möller, 2011; Weiss and Burkhardt, 2010).

One of the few automatic classifications has been conducted by Pinto-Coelho et al. (2013), who found best performance for a combination of SVM with Gaussian Mixture Models (GMM)/Naive Bayes to binary classify voice pleasantness of 77 female speakers of European Portuguese. A related approach was also successful for female voices of different languages (Pinto-Coelho et al., 2011).

2.3. Speaker Pathology

Communication disorders can affect different levels of speech production: phonation, articulation, and language production (Ruben, 2000). These disorders can be congenital or can be the consequence of illness or diseases; they can distort speech severely and by that impair the professional and social lives of the affected speakers. Language disorders denote disorders where the production of language is impaired (Paul and Norbury, 2011). This can vary from language development disorders in children where the language, vocabulary or grammar of a child are not developed normally according to his/her age, to Dysphasia/Aphasia, which are caused by lesions of the left hemisphere, often provoked by a stroke. Phonation or voice disorders are disorders where the process of phonation is disturbed (Aronson and Bless, 2009). This can range from mild issues like hoarseness (Pretorius and Milford, 2008), over persons with vocal fold paralysis (MacGregor et al., 1994), to persons with laryngeal cancer and tracheoesophageal substitute voice (Brown et al., 2003). Speech disorders (Harrison, 2010) refer to disorders where problems exist within the articulation tract. These disorders can have several causes, e. g., cancer in the oral and pharyngeal cavity (Blot et al., 1988), loss of teeth (Tanaka, 1973), and cleft lip and palate (Tolarova and Cervenka, 1998). Dysarthria denotes the form of disorder where motor deficits lead to an dysfunctional speech control. Dysarthria counts as a sub-group of articulation disorders. An example is Parkinson's Disease, a degenerative disorder of the central nervous system where dopamine-generating cells die, which has an impact on the control of muscles (MacGregor et al., 1994). Disorders in written language and dysgraphia respectively dyslexia are also communication disorders but are not further addressed in this overview.

In the current trend towards evidence-based medicine, a reliable evaluation is required to support the effectiveness of treatments and therapies. Currently, the effectiveness of treatment regimes and therapies for speech disorders are mostly evaluated by human listeners; connected speech is preferred since this reflects best the everyday usage of patients. An evaluation based on sustained vowels or single syllables would allow a functional analysis but is often not meaningful for the evaluation of the treatment or the speech therapy. Perceptual evaluations are not only a logistic problem as there is a shortage of qualified therapists, and human evaluation is very expensive: the ratings of these therapists are also subjective and may be biased due to varying listening conditions (McColl, 2006; McColl et al., 1998), differences in the individual experience, varying personal conditions (Bunton et al., 2007; Sheard et al., 1991) or speaker characteristics such as gender (Eadie et al., 2008), and the common habituation to a speaker's idiosyncrasies. Thus, there is a strong need for automatic approaches in clinical routine.

Advances in speech and spoken language processing in the last decades open a new research/application field in pathologic speech and health application nowadays. These fields can be the detection of speech pathologies, i. e., whether a speaker is likely to suffer from a specific articulation disorder, or the assessment of speech pathologies. Assessment is the task of monitoring the severity of an articulation disorder or the impact of a disorder with respect to voice and speech.

Pathology detection is often addressed with systems known in the field of speaker identification/verification (Murillo et al., 2011; Bocklet et al., 2011b; Sturim et al., 2011). Automatic assessment techniques can vary from measuring quality issues like disfluency counts of stuttered speech (Nöth et al., 2000; Heeman et al., 2011) or nasality to non-measurement-based criteria with fuzzy definition. These fuzzy definitions are often assessed perceptually by naive listeners or speech experts/therapists. Connected speech is thereby preferred. Examples are vocal effort, voice quality, roughness, breathiness, hoarseness, or speech intelligibility. No consistent evaluation schemes can be found in the literature. However, in most studies the evaluation is either based on Likert scales or on visual analogue scales. Haderlein (2007) focuses on an objective automatic

intelligibility measurement of pathologic voices by automatic speech recognition (ASR). He showed that the error rate of the ASR system shows a strong correlation with speech intelligibility of speakers with laryngeal cancer on a read text. The underlying idea of automatic intelligibility assessment can be summarised as follows: If humans have difficulties to understand the sounds and words of speech, this will also increase the error rate of the ASR system. The correlation between the ASR error rate and the human evaluation of intelligibility can then be used to develop ASR as an objective measure of speech intelligibility. Systems have evolved from simply scoring word error rates in a read passage (Ferrier et al., 1995) over word-lists to phone and phonological feature recognition in full text reading (Middag et al., 2014). It also has been shown that other approaches based on phonological modelling (Middag et al., 2009), vocal fold modelling (Bocklet et al., 2011a), or acoustic modelling (Bocklet et al., 2012) can improve intelligibility measurement of speakers with pathologic voices. These approaches tend also to be language- and text-independent or 'ASR-free', see Middag et al. (2010, 2011). 'Text-independent' means that we are not confined to one specific text such as 'Northwind and sun', 'ASR-free' means that we do not employ any automatic speech recognition but only acoustic features.

The use of automatic speech processing techniques in order to allow an automatic speech intelligibility measurement has also been applied to articulation disorders diagnosed for patients with oral squamous cell carcinoma (Stelzle et al., 2011), speakers suffering from dysarthria (Vijayalakshmi et al., 2006; Hummel et al., 2011), or children with cleft lip and palate (CLP) (Maier et al., 2009).

3. The First Challenge on Perceived Speaker Traits: Personality, Likability, Pathology

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 previous 2011 challenge, we focused on medium-term speaker states, namely sleepiness and intoxication. Consequently, we now wanted to focus on long(er) term speaker traits. In that regard, the INTERSPEECH 2012 Speaker Trait Challenge broadened the scope by addressing three less researched speaker traits:

Personality Sub-Challenge: the personality of a speaker had to be determined based on acoustics, but potentially including linguistics, for the five OCEAN personality dimensions (Wiggins, 1996), each mapped onto two classes.

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

Pathology Sub-Challenge: based on machine learning and suited acoustic features, the intelligibility of a speaker in a pathological condition had to be predicted.

In order to keep the evaluation of the tasks comparable across the three sub-challenges, we decided in favour of two-class classification problems optimising the unweighted average recall of both classes in order to not prefer the most frequent class in case of unbalanced data[1] and providing an intuitive and transparent measure. Class labels of the training and development sets were known to the participants. All sub-challenges allowed contributors to find their own features with their own machine learning algorithm; however, a standard feature set was provided for all sub-challenges. Participants had to stick to the definition of training, development, and test sets. They were encouraged to report on results obtained on the development set, but had only five trials to upload their results on the test sets, whose labels were unknown to them. Each participation had to be accompanied by a paper presenting the results that underwent peer-review and had to be accepted for the conference in order to participate in the challenge. The organisers preserved the right to re-evaluate the findings, but did not participate themselves in the challenge. Participants were encouraged to compete in all sub-challenges.

[1] Conventional accuracy can be thought of as 'weighted average recall', with weights corresponding to the class priors.

Table 1: *64 low-level descriptors of the INTERSPEECH 2012 Speaker Trait Challenge baseline feature set.*

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, psycho-acoustic sharpness, harmonicity
6 voicing related LLD
F0 by SHS + Viterbi smoothing, probability of voicing
logarithmic HNR, jitter (local, delta), shimmer (local)

3.1. Challenge Features

So far, adding new features and feature types to the feature vectors of previous challenges proved to be beneficial for classification performance. Moreover, such features bundles established for computing the baselines are not necessarily thought of being an end in themselves; instead, they provide a comprehensive basis for feature evaluation and selection.

For the baseline acoustic feature set used in this challenge, we therefore unified the acoustic features 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 Challenge (SSC) and the Audio-Visual Emotion Challenges (AVEC) aiming at the assessment of perceived speaker states. Again, we used TUM's open-source openSMILE feature extractor (Eyben et al., 2010, 2013; Eyben, 2014) and provided 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 were added including logarithmic HNR, spectral harmonicity, and psycho-acoustic spectral sharpness, as in the AVEC 2011 set. The chosen set of LLDs is shown in Table 1, where RMS stands for Root Mean Square, and SHS for Sub-Harmonic Summation (Hermes, 1988). Regarding functionals, on the one hand, we aimed at a more careful selection – for example, from delta regression coefficients we did not compute the simple arithmetic mean as in the 2011 SSC set, but rather the mean of positive values, and the utterance duration was 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 used a peak picking algorithm, refined with respect to the 2011 SSC. The set of applied functionals is given in detail in Table 2. Altogether, the INTERSPEECH 2012 Speaker Trait Challenge feature set contained 6 125 features, which is roughly a 40 % increase over previous year's feature set.

The latest version of openSMILE is described in Eyben et al. (2013). Details can be found in the openSMILE handbook, source code, and configuration file[2], and in Eyben (2014).

3.2. Speaker Personality Corpus (SPC)

In the Personality Sub-Challenge, the Speaker Personality Corpus (SPC) served for analyses and comparison (Mohammadi and Vinciarelli, 2012). The corpus includes 640 clips randomly extracted from the French

[2]URL: http://opensmile.sourceforge.net/; name of configuration file: IS12_SpeakerTrait.conf

Table 2: *Functionals of the INTERSPEECH 2012 Speaker Trait Challenge baseline feature set.*

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 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]

[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

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 subjects talk in one clip only and 20.2 % in two clips. 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 (Rammstedt and John, 2007) and aimed at calculating a score for each of the *Big-Five* dimensions described in Section 2 (Wiggins, 1996). 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 linguistic cues were not 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 for a given trait $X \in \{\mathbf{O}, \mathbf{C}, \mathbf{E}, \mathbf{A}, \mathbf{N}\}$ if at least six judges (the majority) assign to it a score higher than their average for the same trait; otherwise, it is labelled $\neg X$. Training, development, and test set are defined by speaker-independent subdivision of the SPC, stratified by speaker gender (Table 3).

Table 3: *Partitioning of the Speaker Personality Corpus (X: high on trait X / ¬NX: low on trait X, $X \in \{\boldsymbol{O}, \boldsymbol{C}, \boldsymbol{E}, \boldsymbol{A}, \boldsymbol{N}\}$). #: number of instances per set and class.*

SPC Sub-Task	**#**	**Train**	**Devel**	**Test**	**Σ**
Openness	O	97	70	80	247
	¬O	159	113	121	393
Conscientiousness	C	110	81	99	290
	¬C	146	102	102	350
Extraversion	E	121	92	107	320
	¬E	135	91	94	320
Agreeableness	A	139	79	105	323
	¬A	117	104	96	317
Neuroticism	N	140	88	90	318
	¬N	116	95	111	322
Σ		256	183	201	640
Duration [min]		44.0	31.6	33.4	109.0

3.3. Speaker Likability Database (SLD)

In the Likability Sub-Challenge, the Speaker Likability Database (SLD) was used (Burkhardt et al., 2011). The SLD is a subset of the German Agender database (Burkhardt et al., 2010), 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 Burkhardt et al. (2010). The topics of these were *command words*, *embedded commands*, *month*, *days of the week*, *relative time description*, *public holiday*, *birth date*, *time*, *date*, *telephone number*, *postal code*, *first name*, *last name*, *yes/no* with free or preset inventory and 'eliciting' questions such as *"Please tell us any date, for example the birthday of a family member"*.

The age groups in the database (Children: 7–14, Youth: 15–24, Adults: 25–54, Seniors: 55–80 years) are of equal size. All age groups have equal gender distribution and contain at least 100 German speakers from all German federal states. For the experiments described in this paper, we excluded children because we found it hard to judge likability for children's voices; the resulting set of 800 speakers was almost balanced as for age and gender (Youth female/male: 121/112, Adults: 135/130, Seniors: 147/155). For each speaker, we selected the longest sentence available consisting of a command embedded in a free sentence, based on the number of word tokens. This resulted in sentences with a maximal length of eight words (4.4 tokens on average). Typical sentences include "*mach weiter mit der Liste*" ("*continue with the list*") or "*ich hätte gerne die Vermittlung bitte*" ("*I'd like to talk to an operator, please*").

Likability ratings of the data were established by presenting the stimuli to 32 participants (17 male, 15 female, aged 20–42, average age: 28.6 years, standard deviation: 5.4 years). 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; see as well above Section 2.2. The rating was done on a seven point Likert scale. All participants were paid for their service.

A preliminary analysis of the data showed no significant impact of participants' age or gender on the ratings, whereas the samples rated were significantly different (mixed effects model, $p < .0001$). Also, there was no significant effect of the gender of the speakers, nor of the interaction between gender of the raters and speakers. However, age groups were rated differently (age $F(2, 12765) = 3.49$, $p < 0.05$). Speakers from the younger group were more positively assessed than those from the older group ($alpha = .05$), although the effect was very small (effect size $part.\eta^2 = 0.039$). This effect may be specific to this group of participants, as the majority of them would belong to the age group of adults.

Table 4: *Partitioning of Speaker Likability Database (L: likable / ¬L: non-likable).*

SLD #	**Train**	**Devel**	**Test**	**Σ**
L	189	92	119	400
¬L	205	86	109	400
Σ	394	178	228	800
Duration [min]	21.2	9.6	12.5	43.3

To establish a consensus from the individual likability ratings (16 per instance), the evaluator weighted estimator (EWE) by Grimm and Kroschel (2005) was used. The EWE is a weighted mean with weights corresponding to the 'reliability' of each rater, which is the cross-correlation of his/her rating with the mean rating (over all raters). For each rater, this cross-correlation was computed only on the block of stimuli which s(he) rated. In general, the raters exhibited varying 'reliability' ranging from a cross-correlation of .057 to .697. The EWE rating was discretised into the 'likable' (L) and 'non-likable' (¬L) classes based on the median EWE rating of all stimuli in the SLD. Even this binary classification was conceived as challenging because the distribution of EWE ratings is roughly normal and quite symmetric.

For the challenge, the data were partitioned into a training, development, and test set 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 sub-challenge is shown in Table 4. While the challenge task is classification, the EWE is provided for the training and development sets, and participants were encouraged to present regression results in their contributions.

3.4. The Pathology Corpus – The NKI CCRT Speech Corpus (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 van der Molen et al. (2009) and Clapham et al. (2012). The corpus contains recordings and perceptual evaluations of 55 speakers (10 female, 45 male) who underwent concomitant chemo-radiation treatment (CCRT) over a period of seven weeks for stage III-IV inoperable tumours located from the oral cavity down to the hypopharynx. For the NCSC, recordings were selected made before CCRT (T0; 54 speakers), ten weeks after CCRT (T1; 48 speakers) and twelve months after CCRT (T3; 39 speakers). The average speaker age was 57 (32–79 years). All speakers read a Dutch text of neutral content. The native language of the speakers had not been registered but it was clear that not all speakers were Dutch native speakers. One of the authors (RvS) evaluated the recordings to classify speakers as probable non-native Dutch speakers. Recordings were made in a sound-treated room with a Sennheiser MD421 Dynamic Microphone and a portable 24-bit digital wave recorder (Edirol Roland R-1). The sampling frequency was 44.1 kHz; the distance between mouth and microphone was 30 cm.

Thirteen recently graduated or about to graduate speech pathologists (all female, native Dutch speakers, average age 23.7 years, range 21–27) evaluated the speech recordings in an on-line experiment on an intelligibility scale from 1 to 7: "Speech intelligibility was defined as the difficulty/ease with which the listener decodes the speech signal. Listeners were instructed to try to ignore aspects of voice acceptability, reading fluency and any interrupting noises in the files." (Clapham et al., 2012). Participants were requested to complete the evaluations in a quiet environment.

The recorded text was divided into three fragments based on natural breaks in the text (fragment 1: 70 words, fragment 2: 68 words, fragment 3: 51 words). Not all Dutch phonemes were present in the text and phoneme frequencies were not balanced between fragments; for instance, the phoneme `/f/` only occurs in fragment 1. The intensity of the fragments was normalised to 70 dB; more details are given in Clapham et al. (2012). The first two fragments were rated, fragment 3 was used as practise stimuli. Fragment 1 contains 49 unique words and fragment 2 50 unique words. The corpus contains 141 recordings of fragment 1 and one less of fragment 2 (one speaker only read fragment 1). The stimuli were presented in a randomised order

Table 5: *Partitioning of NKI CCRT Speech Corpus (segment level, I: intelligible / ¬I: non-intelligible).*

NCSC #	**Train**	**Devel**	**Test**	**Σ**
I	384	341	475	1 200
¬I	517	405	264	1 186
Σ	901	746	739	2 386
Duration [min]	51.4	39.4	41.4	132.2

Table 6: *Reliability of observer annotations: Mean CC of individual raters with average rating, and mean Cohen's κ of binary individual ratings with majority vote among binary ratings (high on trait / low on trait).*

Response	**CC**	**κ**
Speaker Personality Corpus		
Openness	.307	.231
Conscientiousness	.439	.317
Extraversion	.584	.438
Agreeableness	.449	.334
Neuroticism	.453	.340
Speaker Likability Database		
Likability	.503	.365
NKI CCRT Speech Corpus		
Intelligibility	.837	.674

over three presentation blocks. The listeners could replay each stimulus as often as they wished. On average, it took 70 minutes to complete a block. All participants completed an on-line familiarisation module. The module contained examples of 'good', 'reasonable', and 'poor' speech intelligibility. Participants used their own anchors and received no feedback on performance.

All samples were orthographically transcribed and an automatic phoneme alignment was generated by Catherine Middag (Ghent) using a speech recogniser trained on Dutch speech using the Spoken Dutch Corpus (CGN) based on Middag et al. (2014). Interfering sounds and disturbances were manually annotated. The participants were provided with the transcription and phonemisation. 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 5). The average rank correlation (Spearman's ρ) of the individual ratings with the mean rating is .78. 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 only approximately 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 continuous EWE value for the training and development sets.

3.5. Annotator Agreement

Table 6 shows the reliability of the observer annotations of the challenge corpora (in case of the SPC, results are subdivided by personality trait). On the one hand, the correlation coefficient (CC) of the individual raters' scores with the mean rating is computed. On the other hand, a 'discrete' measure as for the recognition tasks of the challenge is introduced by mapping the individual ratings to a binary label for 'high' / 'low' whenever the rater's score is above / below the mean of the rater's scores across all instances; then, Cohen's κ agreement coefficient is computed between the resulting binary labels of each rater and the majority vote among the binary ratings. Finally, the average of the CC and κ is computed across all raters to obtain a measure for the whole corpus. The κ figures display high agreement for intelligibility ($\kappa = .674$) while the

agreement for personality and likability is considerably lower. It is interesting that the agreement on likability is in the range of the agreement on personality, since the former could be intuitively regarded as a much more subjective trait. In the case of the personality assessments, the average correlation between raters ranges between 0.12 and 0.28 depending on the particular trait. While weak, such correlations are statistically significant and correspond to the typical values observed in zero acquaintance scenarios like those adopted for the data collection (Biesanz and West, 2000). Similar considerations probably apply to likeability as well.

3.6. Baseline Results

As evaluation measure, we retain the choice of unweighted average (UA) recall as used since the first challenge held in 2009 (Schuller et al., 2011a). In the given case of two classes ('X' and '$\neg X$'), it is calculated as $(\text{recall}(X) + \text{recall}(\neg X))/2$, 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 the ROC curve (AUC). However, this measure was not considered for the official evaluation since it would have restricted potential submissions to those systems providing class posteriors ('confidences') per test instance.

For the baselines we exclusively exploit acoustic feature information. As a first, simple baseline, we use logistic functions of the form

$$d_i(x_i) = \frac{1}{1 + \exp(-(a_i x_i + b_i))} \tag{1}$$

where x_i is the value of feature i. For each feature and binary recognition task, the parameters a_i and b_i are fitted to the training set by the least squares method, modelling class X as positive and class $\neg X$ as negative outcome of a Bernoulli trial. The implementation in R, a language and environment for statistical computing, is used for reproducibility. A decision for the positive class is taken whenever $d_i > .5$. Among the functions for the individual features, the one that achieves highest UA on the development set is chosen for the experiments on the test set. This baseline serves to verify the acoustic feature extraction procedure and acts as a reference for the results obtained with more sophisticated machine learning algorithms.

First, we integrate linear SVMs by considering logistic functions that 'convert' SVM hyperplane distances to class pseudo-posteriors,

$$d_{\text{SVM}}(\boldsymbol{x}) = \frac{1}{1 + \exp(-(a(\boldsymbol{w}^T \boldsymbol{x} + b_1) + b_2))}, \tag{2}$$

where $\boldsymbol{w}$ is the normal vector of the SVM hyperplane, $\boldsymbol{x}$ is an acoustic feature vector, b_1 is the SVM bias and a and b_2 are parameters of the logistic function, which are fitted to the SVM outputs on the training set in analogy to the method described above. As per the definition of SVMs, the vector $\boldsymbol{w}$ is sparse, increasing the robustness against overfitting in high dimensional feature spaces. The sparsity is controlled by the complexity parameter C which weighs the trade-off between classification error and the norm of $\boldsymbol{w}$. In our experiments, we choose the value of $C \in \{10^{-4}, 10^{-3}, \ldots, 1\}$ that achieves the best UA recall on the development set. The weight vector $\boldsymbol{w}$ is determined with sequential minimal optimisation (SMO).

Second, we evaluate random forests (RF), which avoid the curse of dimensionality by constructing ensembles of REPTrees trained on random feature subspaces as proposed by Ho (1998). On the development set, we determine the optimal feature subspace size $P \in \{.01, .02, .05, .1\}$ and the 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 $S_{\text{opt}} \in \{1, \ldots, 30\}$ 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

Table 7: *Personality, Likability, and Pathology Sub-Challenge baseline results by logistic regression on single features. Best single features by UA achieved in classification of development set data using logistic functions optimised on training set feature-label mappings; two-task problem, X vs. ¬X: (¬)X; IQR: interquartile range.*

Task	**Feature**	**Devel**		**Test**	
		UA (WA)	AUC	**UA** (WA)	AUC
Personality Sub-Challenge					
(¬)O	IQR 2–3 of Δ RASTA band 3.4–4.0 kHz	64.0 (68.3)	72.2	56.8 (60.7)	58.7
(¬)C	IQR 2–3 of Δ RASTA band 3.4–4.0 kHz	72.6 (72.1)	78.3	70.6 (70.6)	79.8
(¬)E	IQR 1–3 of Δ Loudness	79.3 (79.2)	86.0	69.1 (68.7)	76.6
(¬)A	IQR 1–3 of Δ Energy 1.4–4.0 kHz	67.4 (63.9)	72.4	63.0 (63.7)	68.5
(¬)N	99-percentile of Δ log. HNR	69.3 (69.4)	71.5	67.7 (67.7)	72.9
Mean		70.5 (70.6)	76.1	65.4 (66.3)	71.3
Likability Sub-Challenge					
(¬)L	99-percentile of RASTA band 1.2–1.5 kHz	63.9 (63.5)	63.2	46.3 (46.5)	47.2
Pathology Sub-Challenge					
(¬)I	Mean peak distance of Loudness	62.8 (63.7)	67.9	63.1 (59.7)	71.8

the extended training set. For transparency and reproducibility, we use the open source SVM and RF implementations from the WEKA data mining toolkit (Hall et al., 2009).

Table 7 shows that the logistic regression single feature baselines perform considerably well for personality and pathology recognition. Looking at the selected features for personality, we observe that the most relevant ones are derived delta regression coefficients, pointing to the importance of temporal dynamics. For Openness and Conscientiousness, energy changes in the high frequency bands seem to be particularly of relevance. Extroversion appears to be mostly correlated to loudness while high Agreeableness is apparently characterised by low variation in the speech energy. Finally, Neuroticism manifests in high maximum change in HNR, which is correlated to speech rate. For likability, the maximum energy in a middle frequency band is selected as most important feature; while being hard to interpret, this could point at 'loud' persons being perceived as less likable. However, the selection of this feature does not generalise to the test set, indicating a severe mismatch in the data sets, because its performance on the test set stays below chance. Finally, it is interesting that a rather simple, purely acoustic feature, namely the mean peak distance of loudness, delivers around 63 % UA on both development and test sets of the Pathology Sub-Challenge. However, this feature is arguably related to speech rate, not intelligibility as such – obviously, intelligible persons also speak faster (95 % confidence interval of difference between I and ¬I: $[-.050, -0.38]$, $p \ll .001$); this might simply be due to them having less difficulties in production and articulation.

In Table 8, the official baselines employing the full feature set are shown. Regarding the two other classifiers, Table 8 shows that RF deliver a 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, as will be used in the ongoing). Furthermore, all results with RF on the test set are significantly above chance level ($p < .05$). Of the tasks investigated, the recognition of Conscientiousness (80.1 % UA recall on test using RF), Extroversion (75.3 %) and intelligibility (68.6 %) can be performed most robustly. Generally, the choice of optimal C parameters for the SVMs reveals that sparse weight vectors deliver optimal performance (on the development set), which is in accordance with the somewhat impressive performance of the single feature baselines on the development set. On the Personality Sub-Challenge test set, the RF and SVM results are only slightly ($<$ 3 % mean UA, $p > .05$) above the single feature baselines. In contrast, on the likability test set, the single feature baseline stays below chance level while RF deliver 59.0 % UA (significantly above chance level, cf. above). RF also outperform the single feature baseline for the Pathology test set ($p < .05$). This points at a more multi-faceted nature of these tasks compared to personality recognition, i. e., to the recognition of single traits out of the *Big Five* inventory; for likability and pathology, the observer ratings are apparently tied to multiple acoustic features.

Table 8: *Personality, Likability, and Pathology Sub-Challenge baseline results on all features by linear SVMs and random forests (ensembles of unpruned REPTrees trained on random feature sub-spaces) by unweighted and weighted average (UA/WA) recall in percent. 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; two-task problem, X vs. ¬X: (¬)X.*

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											
(¬)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
(¬)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
(¬)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
(¬)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	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											
(¬)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											
(¬)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

4. Challenge Results

One of the requirements for participation in the challenge was the submission and acceptance of a paper to the INTERSPEECH 2012 Speaker Trait Challenge, which was organised as a special event at the INTERSPEECH conference. Overall, 52 research groups registered for the challenge, and finally, 18 papers were accepted for presentation in the regular review process of the conference. All participants were encouraged to compete in all three sub-challenges. Table 9 shows how many participants took part in which sub-challenge. Ten groups participated only in one of the three sub-challenges: four groups participated only in the Personality Sub-Challenge, two groups only in the Likability Sub-Challenge, and four teams only in the Pathology Sub-Challenge. Six teams took part in all three sub-challenges. One group participated in the Personality and the Likability Sub-Challenge, and one team in the Likability and the Pathology Sub-Challenge. In the following sub-sections we briefly summarise the approaches of the participants. If a paper describes the contribution to more than one sub-challenge, it is mentioned in the section of that sub-challenge where the best results have been achieved.

4.1. Contributions to the Personality Sub-Challenge

Figure 1a shows the official results of the eleven participants in the Personality Sub-Challenge. Obviously, the baseline results provided by the organisers are very competitive: only two teams performed better on the official test set. However, the purpose of the challenge was not just to identify the most successful approach; it is way more than only comparing evaluation scores. The procedures employed by the participants show a large variety of various kinds of ways to address these speaker trait classification problems and by that, give an overview of the state-of-the-art in this field of research. Furthermore, the majority vote of the n best systems shows what can be achieved if the approaches of different teams are combined. Although many teams already built various systems and fused them, the performances of the winning team can still be improved. Figure 1b shows the results of the majority vote for a varying number n of fused systems. In general, better results are obtained if the best three to ten systems (with the exception of the fusion of the best four systems) are combined. The fusion of the best five results yields the best performance. Note that picking the best value for n is some sort of optimisation on the test set; yet, it is obvious that the fusion of various contributions is beneficial. Due to the small size of the test set, the improvements have to be rather large to be significant. To judge the significance, Figure 2a shows what absolute improvements in percent are needed for different significance levels. The plots are based on a one-tailed statistical significance

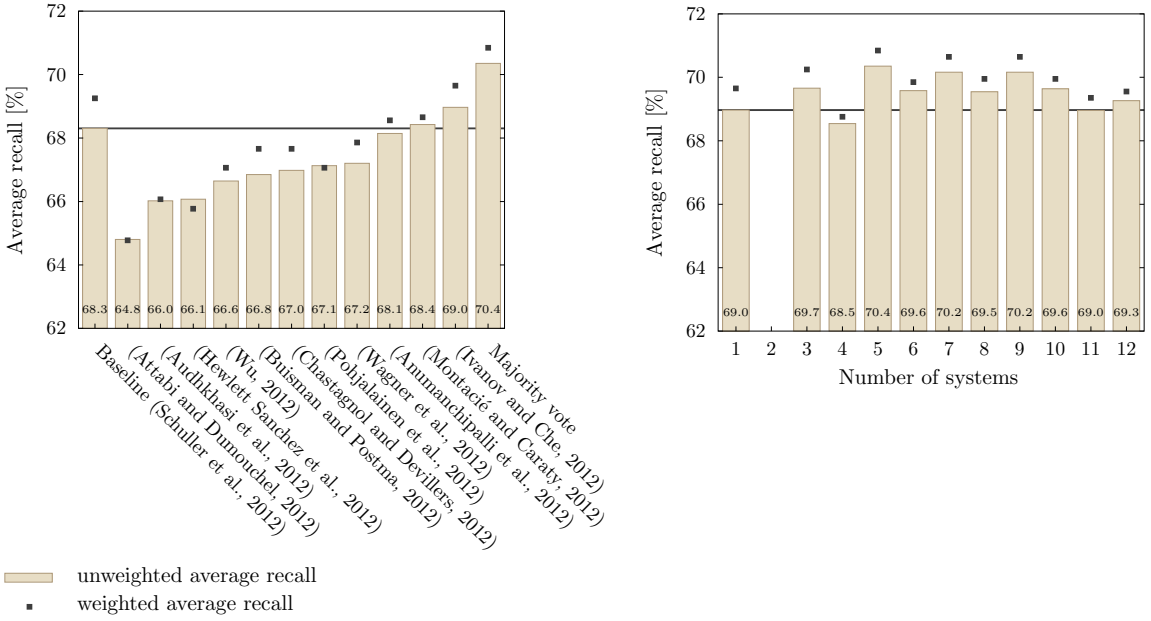
Table 9: Participants of the INTERSPEECH 2012 Speaker Trait Challenge

Sub-Challenge			Participants	
Personality	**Likability**	**Pathology**	**Count**	**Related papers**
✓			4	Audhkhasi et al. (2012) Chastagnol and Devillers (2012) Ivanov and Che (2012) Wagner et al. (2012)
	✓		2	Brueckner and Schuller (2012) Cummins et al. (2012)
		✓	4	Huang et al. (2012) Kim et al. (2012) Stark et al. (2012) Zhou et al. (2012)
✓	✓		1	Buisman and Postma (2012)
	✓	✓	1	Lu and Sha (2012)
✓	✓	✓	6	Attabi and Dumouchel (2012) Anumanchipalli et al. (2012) Hewlett Sanchez et al. (2012) Montacié and Caraty (2012) Pohjalainen et al. (2012) Wu (2012)
11	10	11	18	

z-test assuming that the accuracy of the second experiment is better than the one of the first experiment. Below, the approaches of the participants who only or mainly took part in the Personality Sub-Challenge, are summarised.

Personality-specific local information might be smoothed out if global statistics are computed over the whole utterance. Hence, Audhkhasi et al. (2012) built a GMM-based system using the LLDs of the official baseline feature set at the frame level. Principal component analysis (PCA) and ranking based on Fisher's criterion were used to reduce the dimensionality of the LLDs, resulting in 24 to 40 features depending on the personality dimension. As the authors observed an appreciable correlation between the five personality dimensions, their second system was a tree-structured Bayesian network in order to jointly predict the five OCEAN labels. The conditional probability distribution functions were learnt with linear SVMs with a logistic regression model and smoothed to avoid overfitting due to the small number of samples. The third classification approach was based on a binary logistic regression classifier with L_1 regularisation to obtain a sparse weight vector for a better generalisation. Within class covariance normalisation (WCCN) was used to reduce speaker variability. In addition to these systems based on acoustic features, the authors proposed three systems based on 16 lexical features, modelling rate, and total duration of speech and non-speech events. These features were computed on the output of a large vocabulary ASR system. Again, the authors used the tree-structured Bayesian network, the WCCN system, and a SVM classifier instead of the GMM system. The Bayesian network and the WCCN approach performed better than the GMM-based system for most dimensions; the acoustic features were better than the lexical features; this might be due to the high ASR error rate and the short duration of each audio clip. The selection of the best classifier for each dimension was better than majority vote-based fusion. Although the results were promising on the development set, the proposed systems did not generalise as expected.

Chastagnol and Devillers (2012) addressed the problem of identifying relevant features for the Personality Sub-Challenge. They proposed a modification of the sequential floating forward search (SFFS) algorithm. Wrapper feature selection techniques require the training and evaluation of a classifier for each subset of features being tested. To speed up this process, the authors suggested to estimate the relevance of a feature if added to an existing subset by the relevance of this feature if added to the most similar subset that has



(a) Results of the participants of the Personality Sub-Challenge.

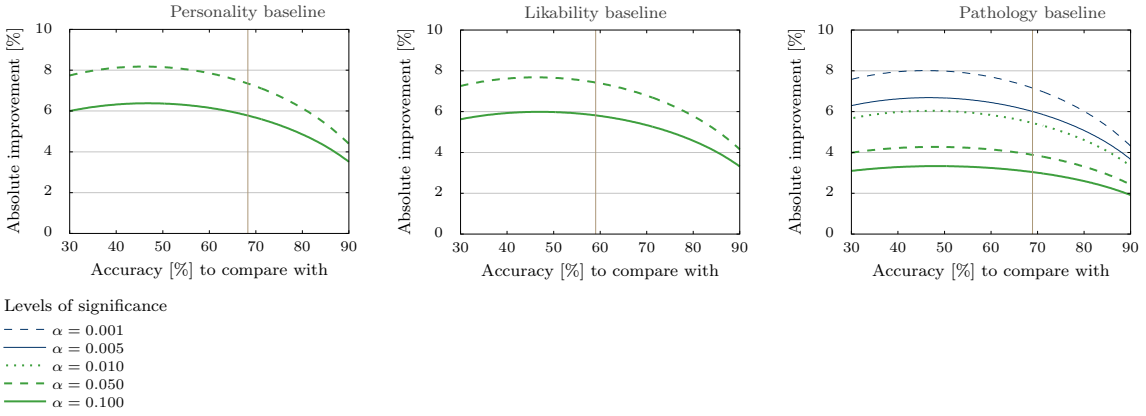
(b) Results of the fusion of the n best systems.

Figure 1: Results of the Personality Sub-Challenge

been computed previously. The Jaccard index was used as a measure for the similarity of two sets: in the forward step, the most promising features are tried to be added first, in the backward step, the least relevant features are tried to be removed first. The machine classifiers were SVMs with radial basis function (RBF) kernel; parameters were optimised in a 5-fold cross-validation on the training set at each step, and a grid search for parameter optimisation was parallelised. Limited by the available computation time, the authors reduced the dimensionality of the 6 125 features of the official baseline feature set depending on the personality dimension, ranging from 534 for Extroversion to 1 014 for Neuroticism. Although the authors obtained absolute improvements of up to 13.7 % for all five OCEAN dimensions on the development set, the method seems to lack in generalisation capability: the average over all five dimensions on the test set stayed below the baseline result. The most relevant families of features were low-level descriptors related to energy and spectral characteristics.

Wagner et al. (2012) addressed the problem that speaker traits are prominent only in some regions of the speech signal and are surrounded by non-meaningful information or noise. They clustered speech frames based on the LLDs of the official baseline feature set reduced by linear discriminant analysis (LDA), and kept only those frames that belong to homogeneous clusters; a cluster is homogeneous if the proportion of observations linked to the respective class is above a certain threshold. The authors applied a smoothing in order to avoid fragmentation of the frames. After identifying the meaningful frames, the authors computed the official features and classified them with SVMs with a polynomial kernel. The authors evaluated their approach on the test set of the Personality Sub-Challenge but could not improve the baseline result.

Ivanov and Che (2012) achieved the best results (official results on the test set: 69.0 %) for the Personality Sub-Challenge. They added four streams of varying fast Fourier transform (FFT) size (16, 32, 64, and 128 samples for both spectrum computations) of modulation spectrum analysis features to the set of baseline features, resulting in a set of 21 760 additional features. The large number of features was reduced based on the Kolmogorov-Smirnov statistical test. Features were selected based on the dissimilarity of the feature distributions for different class labels, and on the similarity of feature distributions over the training and the development set in order to get representative features. The size of the feature set differed for the five personality traits, ranging from 6 719 features for Openness to 13 425 features for extroversion. Classification



(a) Personality Sub-Challenge (b) Likability Sub-Challenge (c) Pathology Sub-Challenge

Figure 2: Significance of improvements

UA [%]	Baseline	Contributions		
Trait		Mean/std.dev.	Min.	Max.
O	59.0	58.3 ± 2.6	54.5	62.5
C	79.1	77.2 ± 1.7	74.6	80.1
E	75.3	73.9 ± 2.9	68.9	79.2
A	64.2	60.0 ± 3.6	56.0	66.8
N	64.0	65.7 ± 2.7	60.5	69.2

Table 10: Personality Sub-Challenge: Performance (UA) of the baseline, mean and standard deviation over participants' performance (UA), and minimum/maximum UA achieved in participants' contributions, by Big Five trait (O, C, E, A, N).

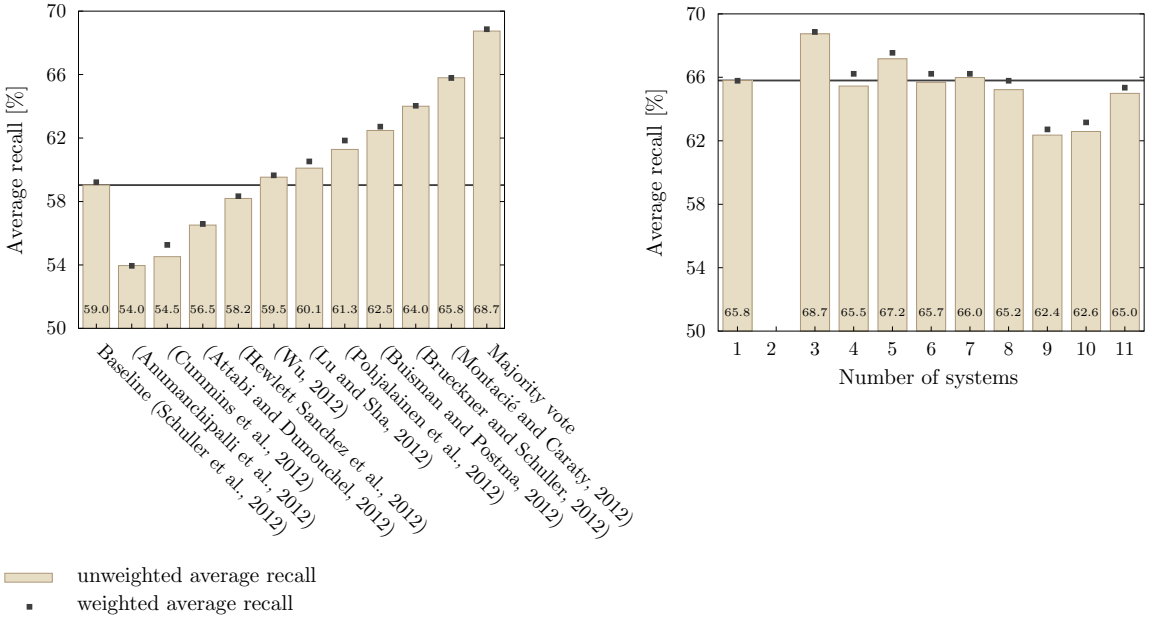
was performed using AdaBoost.

Table 10 shows the performance achieved in the Personality Sub-Challenge for each of the Big Five personality traits. It can be seen that for all of the traits, there is at least one contribution that improves over the SVM/RF baseline performance in terms of UA. Regarding the mean performance of the participants on the individual traits, we observe similar tendencies as in the baseline – the overall best performance is achieved for the Conscientiousness trait, while Openness seems to be the hardest task. The observed 'difficulty' of the recognition of the traits is also in accordance with observers' agreement (cf. Table 6). It is notable that on average the participants do not improve over the baseline except for the Neuroticism trait. In the latter case though, it has to be taken into account that the average participants' performance (65.7 % UA) is still below the single feature baseline (67.7 % UA, cf. Table 7).

4.2. Contributions to the Likability Sub-Challenge

Figure 3a shows the results of the ten participants in the Likability Sub-Challenge on the official test set. Six groups were able to outperform the baseline result of the organisers. Again, the best single performance could be improved by fusing the best three contributions. Figure 3b shows the performance for a varying number n of fused systems. As the size of the test set is only slightly larger than that of the Personality Sub-Challenge, the absolute improvements have to be rather large again as shown in Figure 2b, in order to be statistically significant at the α-level chosen. The following is a summary of the approaches of the participants in the Likability Sub-Challenge.

Cummins et al. (2012) evaluated the performance of single prosodic, voice quality, and spectral features, as well as multidimensional cepstral and spectral features (MFCC, perceptual linear prediction coefficients (PLP),



(a) Results of the participants of the Likability Sub-Challenge.

(b) Results of the fusion of the n best systems.

Figure 3: Results of the Likability Sub-Challenge

line spectral pairs (LSP), linear prediction cepstral coefficients (LPCC), and the spectral centroid frequencies and amplitudes). The authors trained gender-dependent models to cope with the gender-specific characteristics of likability. A GMM universal background model (UBM) approach with maximum a posteriori adaptation (MAP) was used to model one- and multidimensional features directly at the frame level. At utterance level, global statistics were computed for the one-dimensional features. For the multidimensional features, GMM supervectors were used. In both cases, these features were fed to SVMs and sparse representation classifiers (SRC). Furthermore, the authors evaluated the effect of the amount of data used for training the UBM, and used two other data sets (the data of the Personality Sub-Challenge and the NIST 2004 database) for training. The best performance on the test set of the Likability Sub-Challenge was achieved with MFCC supervectors and SRC. However, a clear lack of generalisation could be observed between the development and the test set.

Wu (2012) addressed the curse of dimensionality and proposed a new feature selection scheme. First, the 6 125 acoustic features of the official baseline feature set were ranked with respect to the Fisher information metric and the best 2 000 features were selected. Then, a second feature selection was applied based on a genetic algorithm. Five feature sets with a varying number of features N were generated: $N \in \{100, 200, 300, 400, 500\}$. SVMs were trained for each set and a majority vote was used to fuse the systems. Although the dimensionality of the feature space was reduced drastically compared to the official baseline feature set, the result on the test set of the Likability Sub-Challenge was only slightly better than the SVM baseline result. In contrast, the results on the test sets of the other two sub-challenges did not reach the baseline.

Pohjalainen et al. (2012) applied two feature selection methods (and a combination of both) to the feature set provided by the challenge organisers. For the first method, the authors treated the feature selection problem as a set covering problem with equal costs: the set of audio files has to be covered, i. e., it has to be classified correctly, using the minimum number of features. Well-known techniques are available to solve these kinds of integer linear programming problems. For classification of audio files based on a single feature, unidimensional Gaussian mixture models for each class were applied. The authors investigated a supervised and a partially unsupervised approach. The second method selected those features that display the highest

statistical dependency between a quantised version of the features (65 bins, all bins contain the same amount of samples) and the class label. The classification was based on a k nearest neighbour classifier. The number of selected features depended on the classification task and varied between 124 and 411 on the test set. The same holds for the number of nearest neighbours: k varied between 15 and 137. The authors took part in all three sub-challenges but could only achieve better results than the baseline for the Likability Sub-Challenge.

Buisman and Postma (2012) applied the log-Gabor transform, a main transform for feature extraction in image analysis, to spectrograms: the large number of raw features is normalised and reduced by PCA before normalising them again and classifying them with SVMs with a RBF kernel. The authors trained gender-dependent as well as gender-independent models and took part in the Personality and the Likability Sub-Challenges. On average, the gender-dependent models gave better results. However, the results were not clear-cut for the five classification tasks of the Personality Sub-Challenge. In the Likability Sub-Challenge, the authors were able to outperform the baseline systems on the test set using only their features.

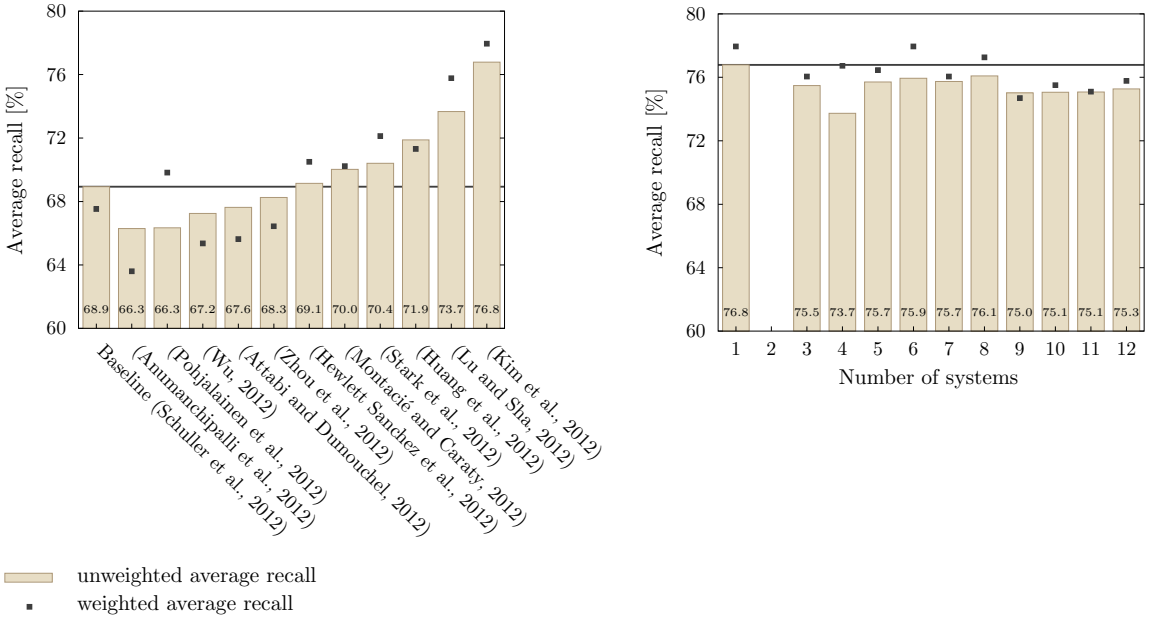
Brueckner and Schuller (2012) evaluated Gaussian-Bernoulli restricted Boltzmann machines (GBRBM) for predicting likability. For comparability with the baseline results, the set of acoustic features was the same as the official one provided by the organisers. The authors used these GBRBMs as a first stage of a two-layered neural network and added a logistic regression network with one output node as an additional layer on top in order to predict the binary target labels. Furthermore, the authors evaluated deep belief networks, which can be constructed by stacking several restricted Boltzmann machines of the same layer size, as well as bottleneck architectures with a decreasing layer size towards the output layer. However, the two-layered neural networks outperformed the more complex architectures: the UA on the test set could be improved from 59.0 % to 64.0 %. Note that due to their affiliation with the organisers of the challenge, the authors did not officially take part in the challenge, although they strictly followed the challenge rules, i. e., they only used the information available to all competitors and had the same number of trials.

Montacié and Caraty (2012) studied pitch and intonation features. They defined three sets of features based on F0 and ΔF0 computed with openSMILE, which differ in the number of functionals and whether Viterbi smoothing is applied or not. The three sets consisted of 97, 134, and 134 features, respectively. Furthermore, the authors defined 34 features characterising a stylised pitch curve based on modelling melody (MOMEL) and the international transcription system for intonation (INTSINT). The classification results obtained with a leave-one-speaker-out cross-validation on the training and development set outperformed in some cases the results of the baseline results although the number of features is drastically smaller. Finally, the authors added the baseline feature set (excluding the pitch features) and defined an extended feature set of 10 172 features obtained by adding more functionals. The final system fused up to three sub-systems based on various combinations of the proposed feature sets and backward best fit feature selection. The authors took part in all three sub-challenges and outperformed all three baseline systems. They obtained the largest improvement for the Likability Sub-Challenge that they won (65.8 % UA).

4.3. Contributions to the Pathology Sub-Challenge

Figure 4a shows the results of the eleven participants in the Pathology Sub-Challenge on the official test set. Again, six groups were able to outperform the baseline result of the organisers, some of them clearly. Surprisingly, the best single performance could not be improved by a majority vote of the best contributions. This might be due to the clear superiority of the winner. Figure 3b shows the performance of the majority vote for a varying number n of fused systems. As the size of the test set is clearly larger than the one of the test sets of the other two sub-challenges, smaller absolute improvements are significant as shown in Figure 2c. The following is a summary of the approaches of the participants in the Pathology Sub-Challenge.

Anumanchipalli et al. (2012) exploited the fact that the pathological voice recordings are read speech and the underlying texts are known. Based on the phonetic alignments of the intelligible utterances, they built a statistical parametric speech synthesis model representing an intelligible Dutch speaker and computed the 'distance' in terms of the Mel cepstral distortion between a test utterance and the corresponding synthetic utterance generated by the model. Furthermore, they used the speech rate, the error between the phoneme duration predicted by the model and the actual duration in the alignment, the F0 prediction error, and the L_2 distance between the baseline openSMILE feature vector of the test utterance and the average vector for intelligible speakers. Besides these five 'text' features, the authors used the following acoustic features:



(a) Results of the participants of the Pathology Sub-Challenge.

(b) Results of the fusion of the n best systems.

Figure 4: Results of the Pathology Sub-Challenge

the standard 6 125 openSMILE features, the 128 LLDs of the official feature set at the frame level, and 26-dimensional PLP features (of order 12, plus energy, and their deltas). Various classification systems were built: linear SVMs, classification trees, and RFs for the complete set of openSMILE features, an SVM system for the LLDs at the frame level, a multi-layer perceptron for the PLP features, and an SVM and an RF system for the five text features. The authors focused on the Pathology Sub-Challenge, since the new text-dependent features could only be computed for this sub-challenge. However, results were reported for the other two sub-challenges and the systems based on acoustic features. The text-dependent features outperformed the baseline system on the development set of the Pathology Sub-Challenge by about 20 % absolute, but unfortunately did not generalise on the test set.

Attabi and Dumouchel (2012) solved the two-class classification problems of the three sub-challenges using anchor models. For each class, a Gaussian mixture model was trained using one of the two standard approaches: maximum likelihood estimation (MLE) and the UBM-MAP adaptation approach. The anchor model representing one class was defined as a vector of the log-likelihood scores of the training data of this class produced by the class-dependent GMM. Interestingly, the anchor model space was spanned by the 14 models of the seven two-class classification problems from all three sub-challenges. A test utterance was assigned to the class with the smallest Euclidean or cosine metric among those classes belonging to the respective classification task. The authors applied WCCN with full and diagonal covariance matrices on the log-likelihood scores of the anchor vector. Although the results were promising on the development set, the random forest baseline results generalised better on the test sets of all three Sub-Challenges.

Zhou et al. (2012) focused on auditory-inspired spectro-temporal modulation features. Their large number of 7 680 features was reduced by PCA to 140 and in a second step by a modified LDA (MLDA) to only 60 features. The features were either classified with SVMs (linear or RBF kernel) or GMMs with 64 or 128 components. The SVM system with RBF kernel and 140-dimensional features performed best; the results were almost identical with the linear SVM baseline results, demonstrating the effectiveness of this kind of features. In addition to the classification, the authors compared intelligible and non-intelligible speech and observed that the highest amplitude peaks for non-intelligible speech are at a lower rate than for intelligible speech. The authors explain this fact with a slower speaking rate and a more discontinuous speech of the

non-intelligible subjects.

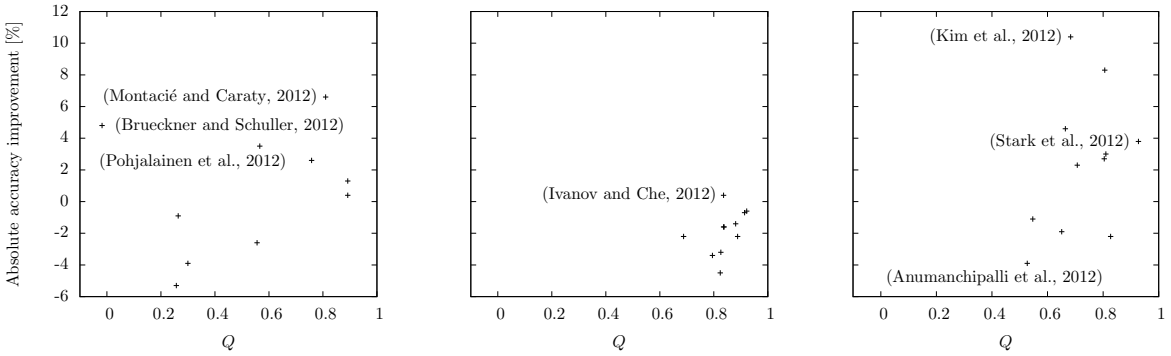
Hewlett Sanchez et al. (2012) built seven different systems, which they finally fused using logistic regression. The individual systems used different feature sets as well as different classifiers. Besides the official feature set, the authors used basic prosodic features computed with their own software, features modelling pitch, energy, and spectral tilt in regions of 200 ms with Legendre polynomial regression of order 5, MFCC features (and their first, second, and third derivatives), and shifted delta cepstrum (SDC) features. Two systems were based on feature selection using those features that performed best when a classifier was trained with just one feature. For classification, the authors used SVMs, GMMs with MAP adaptation of a UBM (UBM-GMM MAP) trained on English speech, a UBM-GMM nuisance compensation system, and an eigenchannel SDC system. The authors participated in all three sub-challenges. Although their results were promising on the development set for all three sub-challenges, the best fusion system could only outperform the baseline system slightly on the Pathology test set.

Stark et al. (2012) performed speaker-dependent as well as speaker-independent cross-validation experiments. They showed that a classifier trained on the complete official feature set learns a lot of speaker characteristics if the partitioning ignores the speaker overlap between the folds. If the dimension of the features is reduced drastically with PCA, much of the speaker information gets lost, while the intelligibility information is still preserved, as experiments with speaker-independent folds showed. In a second thread of experiments, the authors trained separate linear SVMs for each of the 17 sentences and observed significant differences in terms of the classification accuracy, which might be due to a varying proportion of voiced/unvoiced speech, varying phonetic content, different utterance length, or some other factor. The classification based on these 17 classifiers outperformed the baseline result on the test set.

As the number of features in the official feature set is much larger than the number of observations in the Pathology Sub-Challenge, Huang et al. (2012) addressed this problem by applying asymmetric sparse partial least squares regression to estimate the optimal decision boundary. The authors split the official feature set into three subsets (voice quality features, spectral, and harmonicity features, and hierarchical features) and classified them separately. In a second step, the classification results were fused using three fusion methods (FoCal fusion, AdaBoost fusion, and simple fusion). Best results on the test set of the Pathology Sub-Challenge were obtain by simple fusion and outperformed the baseline.

Lu and Sha (2012) explored Gaussian processes for classification and regression using the standard openSMILE feature set provided by the organisers of the challenge. They split the feature set into 13 groups and trained separate Gaussian RBF kernels. In order to deal with the high-dimensional feature space, the authors introduced sparse Gaussian processes based on a greedy feature selection method. The learnt kernels were used for kernel PCA to transform the features into a lower space. For the Likability Sub-Challenge, the authors evaluated gender-specific models: higher classification performances were observed for male than for female speakers. The authors outperformed the baseline systems in both the Likability and the Pathology Sub-Challenge and placed second in the latter one.

The winners of the Pathology Sub-Challenge were Kim et al. (2012). They proposed three sub-systems that were fused on the decision level using naïve Bayes and noisy majority models. The first sub-system captured variations of the acoustic properties of the Dutch speech signals by computing phoneme likelihoods produced by a Czech, a Hungarian, and a Russian phoneme recogniser. The second system modelled prosody and intonation. The third system was based on voice quality (HNR, jitter, and shimmer) and pronunciation features (39 MFCCs, formants 2–4 in vowel regions, and some temporal features). The best result was obtained by the prosodic and intonational sub-system; its performance could not be improved by fusion with the other sub-systems. However, results could be improved by adding five additional features of the baseline feature set that were chosen by a brute-force forward feature selection, and a joint classification approach where samples were clustered based on their acoustic similarity. The underlying assumption was that two utterances with very similar speech characteristics are annotated with the same label. All test samples of the same cluster were assigned the majority classification decision. By that, the unweighted average recall was raised to 76.8 % on the test set.



(a) Likability (b) Personality (c) Pathology

Figure 5: Q-statistic vs. baseline against accuracy improvement over the baseline for participants' systems. Higher Q means higher similarity of the systems' errors compared to the baseline. Points labelled with citations are referred to the text.

4.4. Winners of the INTERSPEECH 2012 Speaker Trait Challenge

The winners of the Personality Sub-Challenges were A. V. Ivanov and Xin Che with their paper "*Modulation Spectrum Analysis for Speaker Personality Trait Recognition*" (Ivanov and Che (2012)). C. Montacié and M.-J. Caraty won the Likability Sub-Challenge with their paper "*Pitch and Intonation Contribution to Speakers' Traits Classification*" (Montacié and Caraty (2012)). In the Pathology Sub-Challenge, the best performance was achieved by J. Kim, N. Kumar, A. Tsiartas, Ming Li, and S. Narayanan with their paper "*Intelligibility Classification of Pathological Speech Using Fusion of Multiple Subsystems*" (Kim et al. (2012)). The approaches of the winning teams are described in the sections above.

4.5. Analysing Participants' Systems: Accuracy vs. Complementarity

Let us now conclude our analysis of the participants' systems from another perspective, one that is not motivated by the goal of reaching utmost performance. A characteristic of the series of INTERSPEECH Challenges is that a state-of-the-art baseline approach is provided to the participants that can be reproduced using a set of features given to the participants, and open-source software. Hence, in terms of improving performance, there is a strong incentive to start from the baseline and continue with incremental improvements, such as modifications to the feature set or classifier. In contrast, novel approaches to the problem built 'from scratch' may perform worse – yet if the goal is to gain new insights into the paralinguistic classification problems, these approaches might be highly interesting to study further.

As an indicator of how much the participants' systems differ from the baseline, we compute the Q-statistic (Yule, 1900; Afifi and Azen, 1979) with the baseline. Informally, $Q_B(A)$ measures whether the system A commits the same errors on the evaluation set as the baseline B, information which is not contained in simple accuracy comparisons. Interestingly, there is evidence that the Q-statistic is also indicative of performance in fusion of classifier predictions (Kuncheva and Whitaker, 2003) – informally, systems that are far from each other in the errors they make do possess complementary strengths. More precisely, $Q_B(A)$ is defined by

$$Q_B(A) = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}}, \tag{3}$$

where N^{11} and N^{00} are the numbers of instances where the predictions of A and B are both correct or incorrect, respectively, and N^{01} and N^{10} are the numbers of instances where only A or B commit an error.

We now set the Q-statistics in relation to the accuracy improvements, both compared to the baseline. Since the Q-statistic does not take into account class imbalance, it is set in relation to the improvement in WA (not UA). The results are shown in Figure 5.

In the Likability Sub-Challenge (Figure 5a), the winner (Montacié and Caraty, 2012) shows strong agreement with the baseline ($Q = .81$) while improving its accuracy significantly (by 6.6 % absolute WA). This makes sense, as their approach combines the baseline feature set with more advanced pitch features and adds additional functionals. The second ranked DNN approach by Brueckner and Schuller (2012) is an outlier in the graph, because it improves over the SVM baseline by 4.8 % absolute while not agreeing at all with it ($Q = -0.02$). The third ranked approach by Pohjalainen et al. (2012), which is based on feature selection on the baseline feature set and k-NN classification, is again closer to the baseline in performance while moderately agreeing with it. Since the DNN used by Brueckner and Schuller (2012) takes the baseline feature set as input, but performs non-linear transformations on it, it can be argued that these non-linearities make its predictions complementary to the baseline based on random decision forests. The rank correlation of the Q-statistic and the improvement over the baseline is statistically significant ($\rho = .68$, $p < .05$) if and only if we exclude the outlying DNN approach. We conclude that for the vast majority (9/10) of the approaches submitted, improvement over the baseline is obtained when the predictions are not too different from it.

In the Personality Sub-Challenge (Figure 5b), we observe that only one system (Ivanov and Che, 2012) can slightly improve over the baseline (+ 0.4 % WA) at $Q = .84$. While this system is based on modulation spectrum features, it is similar to the baseline in that it over-generates a large feature space followed by subsequent feature selection (which is done implicitly in the random forest baseline). Overall, the errors of the systems are all somewhat similar to the ones of the baseline ($Q \geq .69$), and as in the Intelligibility Sub-Challenge, we observe significant correlation between the Q-statistic and the performance improvement ($\rho = .68$, $p < .05$).

In the Pathology Sub-Challenge (Figure 5c), the winning system by Kim et al. (2012) is considerably different from the baseline ($Q = .683$) while delivering a vast performance improvement (+ 10.4 % WA). It is still interesting that the Q-statistic for this system is rather high, since only five of the baseline features are used and most of the system's performance stems from improved prosody and intonation features, as well as acoustic clustering. The last ranked system by Anumanchipalli et al. (2012), which employs a novel approach for intelligibility based on the comparison with text-to-speech (TTS) systems has at the same time the lowest Q-statistic of all systems ($Q = .53$), while unfortunately not generalising to the test set (3.9 % decrease in WA from the baseline). The system by Stark et al. (2012) is representative of a system close to the baseline in terms of Q ($Q = .92$) but providing significant performance improvent (+3.8 % WA, ranked third in terms of UA and fourth in terms of WA). Their approach is a traditional engineering one: Based on observed deficiencies in the training algorithm for the baseline system, specific improvements were implemented. Overall, in the Pathology Sub-Challenge, there is no significant correlation between the Q-statistic and the performance improvement ($\rho = .33$, $p > .1$). In a way, this indicates that there are many multi-faceted and often well-motivated approaches to the intelligibility classification problem, but what works in practice has most likely to be determined by trial and error.

5. Concluding Remarks

5.1. Summary: Challenge Setup and Results

In this INTERSPEECH 2012 Speaker Trait Challenge, we focused on perceived speaker traits, i. e., on traits that have to be annotated by humans. The recording settings were realistic with respect to specific applications: radio broadcast in the case of personality, mobile and landline phone in the case of likability, and office environment in the case of pathological speech. The type of data was spontaneous, prompted, and read speech. Annotation was made using rating scales.

To keep the conditions constant and simple across experiments, we decided in favour of unweighted average (UA) recall, i. e., the rating scales were mapped onto a binary classification task. For the baseline features – enriched with respect to previous challenge feature sets – we employed a brute force approach, focusing only on acoustic features to give room for improvement. The data were clearly partitioned into a train, devel, and test set. This overall setting guarantees straightforward computing and strict comparability; thus it was not especially tailored for prospective applications – in such a case, we might have chosen correlation/regression approaches, or for utmost performance – in such a case, we might have chosen leave-one-speaker-out cross-validation instead of our partitioning.

The baseline performance was, for all three sub-challenges, competitive but could be surpassed by participants. The impact of single features or feature types was – this could be expected for such a challenge – less in the focus of the participants than classification performance. To this aim, (combinations of) new features – alone or with the baseline features provided by the organisers – and (combinations of) algorithms were employed. Early fusion on the feature level seems to pay off to some lesser extent than late fusion on the classifier level. Even if some participants themselves have already combined the output of classifiers yielding higher performance, combining the classifier output of different participants again resulted in some higher performance in the case of personality and likability. This outcome mirrors RFs and other 'combination approaches'. Adding new information to the feature set by adding new (types of) features obviously still can improve performance. However, it is not easy to find out the specific contribution of these new features. Most prominent is as detailed above speech tempo as single relevant feature for pathological read speech – a result that has been obtained for non-native read speech as well (Hönig et al., 2012). Good speakers are good at reading as well. It has to be validated if this conjecture holds for non-read speech up to the same extent as well.

Comparing the results obtained on the three traits considered in the Challenge, we find that participants could surpass the baseline mostly in the Pathology Sub-Challenge, followed by the Likability Sub-Challenge, while there was no significant improvement over the baseline in the Personality Sub-Challenge, and at the same time a large overlap to the baseline methodology. This is rather interesting, since from our literature review it seems that the former ones can be regarded as less established fields in speech research compared to the latter one. Hence, the observation could be attributed to convergence of the field, where participants have tend to submit 'tried and true' approaches as opposed to the rather 'creative' techniques applied to the likability and pathology tasks. However, it is also notable that personality is the only of the three Sub-Challenges where the simplistic single feature logistic regression baseline provides rather competitive results. This indicates a clear deficiency in the current machine learning approaches to the problem – whether this can be remedied by collecting more data, or whether it needs a fundamental algorithmic improvement, will be an interesting question for future research.

5.2. Why Such a Challenge: Prospect and Limitations

Such a challenge can make the community aware of new/other fields of research within Computational Paralinguistics and of the possibilities that strict comparability and competition provide. As such, it constitutes a first step towards higher and common standards; note that it does not necessarily mirror different standards in different fields. For instance, in research of pathological speech, binary judgements and by that, classification procedures are less common than rating scales and correlation measures. However, we do not know yet to which extent this is simply due to research traditions or because these approaches are more adequate for the specific tasks. It might be advisable to collect as much information as possible and to keep this information during the chain of processing as long as possible. This is an argument in favour of rating scales and correlation/regression procedures. At the end of the day, however, we often will have to decide in a binary fashion if it comes to applications: whether the manifestation of one of the personality traits of the *Big Five* or of all combined in an appropriate way, or whether the degree of likability or of pathological speech, will trigger some specific decision: e. g. do we hire this speaker for broadcasting or advertisement, or do we decide on continuing or not speech therapy. Repeated speaker-specific evaluations (the same speaker is evaluated several times, to check the benefits of treatment) and speaker-independent screenings (specific (age) groups are evaluated as for specific difficulties or pathologies such as hearing impairement or articulation problems) might need different types of measurements, and hopefully, when we can overcome the sparse data problem in some way, we will not only have to rely on brute force approaches but end up with some diversification of (acoustic and/or linguistic) main features that makes us a bit wiser.

We have to keep in mind the specifities of such challenges when evaluating our expectations (*what do we want to get*), the outcome (*what did we get*), and our hopes (*what will we get in the future*). Incentives (to be the winner of one of the challenges) and time constraints favour the participation of sites that are used to implement machine learning algorithms and adapt them to specific problems in a rather short time. These sites are not necessarily too strongly interested in interpreting specific features. Moreover, sites with some critical mass – meaning that colleagues do have different, complementing expertise in machine learning –

have higher chances to couple promising approaches into one approach, e. g. late fusion. So far, this is the most likely outcome of these challenges. Of course, we want to get wiser with respect to other aspects as well; the quest for the 'most important features' is as notorious as the quest for the holy Grail has been. This will take longer. Either we have to sort of 'outsource' this task to other endeavours – summer schools, for example, where different groups concentrate for several weeks on such a problem; or simply to the time to come, with more time for collecting additional data and evidence, and for addressing such questions anew with by then hopefully well-established reference databases, namely the ones that have been used in these challenges.

6. Acknowledgement

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007–2013) under grant agreements No. 289021 (ASC-Inclusion) and No. 338164 (ERC Starting Grant iHEARu), and from the German Research Foundation (DFG grant WE 5050/1-1). 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.

References

Afifi, A., Azen, S., 1979. Statistical Analysis. A Computer Oriented Approach. Academic Press, New York.

Ambady, N., Skowronski, J. J. (Eds.), 2008. First Impressions. Guilford Press, New York.

Anumanchipalli, G. K., Meinedo, H., Bugalho, M., Trancoso, I., Oliveira, L. C., Black, A. W., 2012. Text-dependent pathological voice detection. In: Proc. of Interspeech, Portland, Oregon, U. S. A. pp. 530–533.

Aronson, A. E., Bless, D., 2009. Clinical Voice Disorders, 3rd Edition. Thieme, New York, NY.

Aronson, E., Wilson, T., Akert, R. M., 2009. Social Psychology, 7th Edition. Prentice Hall.

Attabi, Y., Dumouchel, P., 2012. Anchor models and WCCN normalization for speaker trait classification. In: Proc. of Interspeech, Portland, Oregon, U. S. A. pp. 522–525.

Audhkhasi, K., Metallinou, A., Li, M., Narayanan, S., 2012. Speaker personality classification using systems based on acoustic-lexical cues and an optimal tree-structures bayesian network. In: Proc. of Interspeech Portland, Oregon, U. S. A. pp. 262–265.

Biesanz, J., West, S., 2000. Personality coherence: Moderating self-other profile agreement and profile consensus. Journal of Personality and Social Psychology 79 (3), 425–437.

Blot, W. J., McLaughlin, J. K., Winn, D. M., Austin, D. F., Greenberg, R. S., Preston-Martin, S., Bernstein, L., Schoenberg, J. B., Stemhagen, A., Fraumeni, J. F., 1988. Smoking and drinking in relation to oral and pharyngeal cancer. Cancer Research 48, 3282–3287.

Bocklet, T., Nöth, E., Stemmer, G., 2011a. Voice assessment of speakers with laryngeal cancer by glottal excitation modeling based on a 2-mass model. In: Proc. of TSD 2011 – 14th International Conference on Text, Speech and Dialogue, September 1-5, Pilsen, Czech Republic. pp. 348–55.

Bocklet, T., Nöth, E., Stemmer, G., Ruzickova, H., Rusz, J., 2011b. Detection of persons with Parkinson's disease by acoustic, vocal, and prosodic analysis. In: Proc. of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) 2011, December 11-15, Big Island, Hawaii, U. S. A. pp. 478–483.

Bocklet, T., Riedhammer, K., Nöth, E., Eysholdt, U., Haderlein, T., 2012. Automatic intelligibility assessment of speakers after laryngeal cancer by means of acoustic modeling. Journal of Voice 26 (3), 390–397.

Brown, D. H., Hilgers, F. J. M., Irish, J. C., Balm, A. J. M., 2003. Postlaryngectomy voice rehabilitation: state of the art at the millennium. World Journal of Surgery 27, 824—831.

Brueckner, R., Schuller, B., 2012. Likability classification – A not so deep neural network approach. In: Proc. of Interspeech, Portland, Oregon, U. S. A. pp. 290–293.

Buisman, H., Postma, E., 2012. The log-Gabor method: Speech classification using spectrogram image analysis. In: Proc. of Interspeech, Portland, Oregon, U. S. A. pp. 518–521.

Bunton, K., Kent, R. D., Duffy, J. R., Rosenbek, J. C., Kent, J. F., 2007. Listener agreement for auditory-perceptual ratings of dysarthria. J Speech Lang Hear Res 50, 1481–1495.

Burkhardt, F., Eckert, M., Johannsen, W., Stegmann, J., 2010. A database of age and gender annotated telephone speech. In: Proc. of LREC 2010, 7th International Conference of Language Resources and Evaluation, May 19-21, 2010, Malta. pp. 1562–1565.

Burkhardt, F., Schuller, B., Weiss, B., Weninger, F., 2011. 'Would you buy a car from me?' – On the likability of telephone voices. In: Proc. of Interspeech, Florence, Italy. pp. 1557–1560.

Campbell, N., Mokhtari, P., 2003. Voice quality: the 4th prosodic dimension. In: Proc. of ICPhS. Barcelona, pp. 2417–2420.

Chastagnol, C., Devillers, L., 2012. Personality traits detection using a parallelized modified SFFS algorithm. In: Proc. of Interspeech, Portland, Oregon, U. S. A. pp. 266–269.

Chattopadhyay, A., Dahl, D. W., Ritchie, R. J. B., Shahin, K. N., 2003. Hearing voices: The impact of announcer speech characteristics on consumer response to broadcast advertising. Journal of Consumer Psychology 13 (3), 198–204.
Clapham, R. P., van der Molen, L., van Son, R. J. J. H., van den Brekel, M., Hilgers, F. J. M., 2012. NKI-CCRT Corpus – speech intelligibility before and after advanced head and neck cancer treated with concomitant chemoradiotherapy. In: Proc. of LREC 2012 – 8th International Conference on Language Resources and Evaluation, May 23-25, Istanbul, Turkey. pp. 3350–3355.
Cloninger, S., 2009. Conceptual issues in personality theory. In: Corr, P., Matthews, G. (Eds.), The Cambridge handbook of personality psychology. Cambridge University Press, pp. 3–26.
Collins, S. A., 2000. Men's voices and women's choices. Animal Behaviour 60, 773–780.
Cummins, N., Epps, J., Kua, J. M. K., 2012. A comparison of classification paradigms for speaker likeability determination. In: Proc. of Interspeech, Portland, Oregon, U. S. A. pp. 282–285.
Dahlbäck, N., Wang, Q., Nass, C., Alwin, J., 2007. Similarity is more important than expertise: Accent effects in speech interfaces. In: Proc. of CHI 2007 – Conference on Human Factors in Computing Systems, April 28 - May 3, San José, CA, U. S. A. . pp. 1553–1556.
Eadie, T. L., Doyle, P. C., Hansen, K., Beaudin, P. G., 2008. Influence of speaker gender on listener judgments of tracheoesophageal speech. Journal of Voice 22 (1), 43–57.
Eklund, R., Lindström, A., 2001. Xenophones: An investigation of phone set expansion in Swedish and implications for speech recognition and speech synthesis. Speech Communication 35 (2), 81–102.
Ekman, P., Friesen, W. V., O'Sullivan, M., Scherer, K., 1980. Relative importance of face, body, and speech in judgments of personality and affect. Journal of Personality and Social Psychology 38, 270–277.
Eyben, F., 2014. Real-time speech and music classification by large audio feature space extraction. Ph.D. thesis, Dissertation, Technische Universität München, forthcoming.
Eyben, F., Weninger, F., Gross, F., Schuller, B., 2013. Recent Developments in openSMILE, the Munich Open-source Multimedia Feature Extractor. In: Proc. of the 21st ACM International Conference on Multimedia. MM '13. New York, NY, USA, pp. 835–838.
Eyben, F., Wöllmer, M., Schuller, B. W., 2010. openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor . In: Proc. of MM 2010 – International Conference on Multimedia, October 25-29, Firenze, Italy. pp. 1459–1462.
Ferguson, M. J., Fukukura, J., 2012. Likes and dislikes: A social cognitive perspective on attitudes. In: Fiske, S. T., Macrae, C. N. (Eds.), The SAGE Handbook of Social Cognition. SAGE Publications Ltd, London, pp. 165–186.
Ferrier, L., Shane, H., Ballard, H., Carpenter, T., Benoit, A., 1995. Dysarthric speakers' intelligibility and speech characteristics in relation to computer speech recognition. Augmentative and Alternative Communication 11, 165–175.
Funder, D., 2001. Personality. Annual Reviews of Psychology 52, 197–221.
Gravano, A., Levitan, R., Willson, L., Beňuš, Š., Hirschberg, J., Nenkova, A., 2011. Acoustic and prosodic correlates of social behavior. In: Proc. of Interspeech. pp. 97–100.
Grimm, M., Kroschel, K., 2005. Evaluation of natural emotions using self assessment manikins. In: Proc. of ASRU 2005 – Automatic Speech Recognition and Understanding Workshop, Cancun, Mexico. pp. 381–385.
Haderlein, T., 2007. Automatic Evaluation of Tracheoesophageal Substitute Voices. Logos Verlag.
Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H., 2009. The WEKA data mining software: An update. SIGKDD Explorations Newsletter 11, 10–18.
Harrison, A. E., 2010. Speech Disorders: Causes, Treatment and Social Effects. Nova Science Publishers Inc.
Heeman, P. A., McMillin, A., Yaruss, J. S., 2011. Computer-assisted disfluency counts for stuttered speech. In: Proc. of Interspeech, Florence, Italy. pp. 3013–3016.
Hermes, D., 1988. Measurement of pitch by subharmonic summation. Journal of the Acoustical Society of America 83, 257–264.
Hewlett Sanchez, M., Lawson, A., Vergyri, D., Bratt, H., 2012. Multi-system fusion of extended context prosodic and cepstral features. In: Proc. of Interspeech, Portland, Oregon, U. S. A. pp. 514–517.
Ho, T. K., 1998. The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence 20, 832–844.
Hodges-Simeon, C. R., Gaulin, S. J. C., Puts, D. A., 2010. Different vocal parameters predict perceptions of dominance and attractiveness. Human Nature (Hawthorne, N. Y.) 21, 406–427.
Hönig, F., Batliner, A., Nöth, E., 2012. Automatic assessment of non-native prosody annotation, modelling and evaluation. In: Proc. of ISADEPT – International Symposium on Automatic Detection of Errors in Pronunciation Training, June 6-8, Stockholm, Sweden. pp. 21–30.
Huang, D.-Y., Yongwei Zhu, D. W., Yu, R., 2012. Detecting intelligibility by linear dimensionality reduction and normalized voice quality hierarchical features. In: Proc. of Interspeech, Portland, Oregon, U. S. A. pp. 546–549.
Hummel, R., Chan, W.-Y., Falk, T. H., 2011. Spectral features for automatic blind intelligibility estimation of spastic dysarthric speech. In: Proc. of Interspeech, Florence, Italy. pp. 3017–3020.
Ivanov, A. V., Che, X., 2012. Modulation spectrum analysis for speaker personality trait recognition. In: Proc. of Interspeech, Portland, Oregon, U. S. A. pp. 278–281.
Ivanov, A. V., Riccardi, G., Sporka, A. J., Franc, J., 2011. Recognition of personality traits from human spoken conversations. In: Proc. of Interspeech, Florence, Italy. pp. 1549–1552.
Jürgens, C., Johannsen, W., Fellbaum, K., 1996. Zur Eignung von Sprechern für die Lautelemente-Bibliothek eines Sprachsynthesesystems. In: Proc. of ITG Fachtagung Sprachkommunikation, September 17-18, Frankfurt am Main, Germany. pp. 101–104.
Kenny, D. A., 1994. Interpersonal Perception: A social relations analysis. Guilford Press, New York.
Ketzmerick, B., 2007. Zur auditiven und apparativen Charakterisierung von Stimmen. TUDpress, Dresden.

Kim, J., Kumar, N., Tsiartas, A., Li, M., Narayanan, S., 2012. Intelligibility classification of pathological speech using fusion of multiple subsystems. In: Proc. of Interspeech, Portland, Oregon, U. S. A. pp. 534–537.
Kreiman, J., Van Lancker Sidtis, D., 2011. Foundations of Voice Studies: An Interdisciplinary Approach to Voice Production and Perception. Wiley-Blackwell, Chichester.
Kuncheva, L. I., Whitaker, C. J., 2003. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Machine Learning 51 (2), 181–207.
Li, A., Wang, H., 2004. Friendly speech analysis and perception in standard Chinese. In: Proc. of Interspeech. Jeju Island, Korea, pp. 897–900.
Lu, D., Sha, F., 2012. Predicting likability of speakers with Gaussian processes. In: Proc. of Interspeech, Portland, Oregon, U. S. A. pp. 286–289.
MacGregor, F. B., Roberts, D. N., Howard, D. J., Phelps, P. D., 1994. Vocal fold palsy: a re-evaluation of investigations. Journal of Laryngology & Otology 108, 193–196.
Maier, A., Haderlein, T., Eysholdt, U., Rosanowski, F., Batliner, A., Schuster, M., Nöth, E., 2009. PEAKS – a system for the automatic evaluation of voice and speech disorders. Speech Communication 51, 425–437.
Mairesse, F., Walker, M. A., Mehl, M. R., Moore, R. K., 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. Journal of Artificial Intelligence Research 30, 457–500.
Matthews, G., Deary, I., Whiteman, M., 2009. Personality Traits. Cambridge University Press.
McColl, D., Fucci, D., Petrosino, L., Martin, D. E., McCaffrey, P., 1998. Listener ratings of the intelligibility of tracheoesophageal speech in noise. Journal of Communication Disorders 31, 279–289.
McColl, D. A., 2006. Intelligibility of tracheoesophageal speech in noise. Journal of Voice 20 (4), 605–615.
McCrae, R., 2009. The Five-Factor Model of personality. In: Corr, P., Matthews, G. (Eds.), The Cambridge handbook of personality psychology. Cambridge University Press, pp. 148–161.
McCrae, R. R., John, O. P., 1992. An introduction to the Five-Factor Model and its applications. Journal of Personality 60, 175–215.
McCroskey, J., McCain, T., 1974. The measurement of interpersonal attraction. Speech Monographs 41, 261–266.
Middag, C., Bocklet, T., Martens, J.-P., Nöth, E., 2011. Combining phonological and acoustic ASR-free features for pathological speech intelligibility assessment. In: Proc. of Interspeech, Florence, Italy. pp. 3005–3008.
Middag, C., Clapham, R., van Son, R., Martens, J.-P., 2014. Robust automatic intelligibility assessment techniques evaluated on speakers treated for head and neck cancer. Computer Speech & Language, 467–482.
Middag, C., Martens, J.-P., Van Nuffelen, G., De Bodt, M., 2009. Automated intelligibility assessment of pathological speech using phonological features. EURASIP Journal on Advances in Signal Processing 2009, 1–9.
Middag, C., Saeys, Y., Martens, J.-P., 2010. Towards an ASR-free objective analysis of pathological speech. In: Proc. of Interspeech, Makuhari, Japan. pp. 294–297.
Mohammadi, G., Vinciarelli, A., 2012. Automatic personality perception: Prediction of trait attribution based on prosodic features. IEEE Transactions on Affective Computing 3, 273–284.
Moniz, H., Mata, A. I., Trancoso, I., Viana, M. C., 2008. How can you use disfluencies and still sound as a good speaker? In: Proc. of Interspeech, Brisbane, Australia. p. 1687.
Montacié, C., Caraty, M.-J., 2012. Pitch and intonation contribution to speakers' traits classification. In: Proc. of Interspeech, Portland, Oregon, U. S. A. pp. 526–529.
Müller, C. (Ed.), 2007. Speaker Classification I and II. Springer, Heidelberg, New York, Berlin.
Murillo, J. L. B., Pozo, R. F., Toledano, D. T., Caminero, J., López, E., 2011. Analyzing training dependencies and posterior fusion in discriminant classification of apnea patients based on sustained and connected speech. In: Proc. of Interspeech, Florence, Italy. pp. 3033–3036.
Nass, C., Min Lee, K., 2001. Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. Journal of Experimental Psychology: Applied 7, 171–181.
Nesler, M. S., Storr, D. M., Tedeschi, J. T., 1993. The interpersonal judgment scale: A measure of liking or respect? The Journal of Social Psychology 133, 237–242.
Norman, W., 1963. Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. Journal of Abnormal and Social Psychology 66 (6), 574–583.
Nöth, E., Niemann, H., Haderlein, T., Decher, M., Eysholdt, U., Rosanowski, F., Wittenberg, T., 2000. Automatic stuttering recognition using hidden Markov models. In: Proc. of Interspeech, Beijing, China. pp. 65–68.
Paul, R., Norbury, C. F., 2011. Language Disorders from Infancy Through Adolscence: Listening, Speaking, Reading, Writing, and Communicating, 4th Edition. Elsevier Health Sciences, New York, NY.
Pianesi, F., 2013. Searching for personality. IEEE Signal Processing Magazine 30, 146–158.
Pinto-Coelho, L., Braga, D., Sales-Dias, M., Garcia-Mateo, C., 2011. An automatic voice pleasantness classification system based on prosodic and acoustic patterns of voice preference. In: Proc. of Interspeech, Florence, Italy. pp. 2457–2460.
Pinto-Coelho, L., Braga, D., Sales-Dias, M., Garcia-Mateo, C., 2013. On the development of an automatic voice pleasantness classification and intensity estimation system. Computer Speech and Language 27, 75–88.
Pohjalainen, J., Kadioglu, S., Räsänen, O., 2012. Feature selection for speaker traits. In: Proc. of Interspeech, Portland, Oregon, U. S. A. pp. 270–273.
Polzehl, T., Möller, S., Metze, F., 2010. Automatically assessing personality from speech. In: Proc. of ICSC 2010 – 4th International Conference on Semantic Computing, September 22-24, Pittsburgh, PA, U. S. A. pp. 134–140.
Pretorius, P. M., Milford, C. A., 2008. Investigating the hoarse voice. British Medical Journal 337, 1165–1168.
Rammstedt, B., John, O. P., 2007. 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, 203–212.

Ranganath, R., Jurafsky, D., McFarland, D. A., 2013. Detecting friendly, flirtatious, awkward, and assertive speech in speed-dates. Computer Speech and Language 27, 89–115.
Rosenberg, A., Hirschberg, J., 2009. Charisma perception from text and speech. Speech Communication 51, 640–655.
Ruben, R. J., 2000. Redefining the survival of the fittest: Communication disorders in the 21st century. The Laryngoscope 110, 241–245.
Scherer, K. R., 1979. Personality markers in speech. In: Scherer, K. R., Giles, H. (Eds.), Social Markers in Speech. Cambridge University Press, Cambridge, pp. 147–209.
Schmitz, M., Krüger, A., Schmidt, S., 2007. Modeling personality in voice of talking products through prosodic parameters. In: Proc. of IUI 2007 – 12th International Conference on Intelligent User Interfaces, January 28-31, Honolulu, Hawaii, U. S. A. pp. 313–316.
Schuller, B., Batliner, A., 2014. Computational Paralinguistics – Emotion, Affect, and Personality in Speech and Language Processing. Wiley, Chichester, UK.
Schuller, B., Batliner, A., Steidl, S., Seppi, D., 2011a. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech Communication, Special Issue on "Sensing Emotion and Affect – Facing Realism in Speech Processing" 53, 1062–1087.
Schuller, B., Steidl, S., Batliner, A., 2009. The INTERSPEECH 2009 Emotion Challenge. In: Proc. of Interspeech, Brighton, UK. pp. 312–315.
Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., Narayanan, S., 2010. The INTERSPEECH 2010 Paralinguistic Challenge – Age, Gender, and Affect. In: Proc. of Interspeech, Makuhari, Japan. pp. 2794–2797.
Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., Narayanan, S., 2013a. Paralinguistics in speech and language – state-of-the-art and the challenge. Computer, Speech and Language, Special Issue on "Paralinguistics in Naturalistic Speech and Language" 27, 4–39.
Schuller, B., Steidl, S., Batliner, A., Nöth, E., Vinciarelli, A., Burkhardt, F., van Son, R., Weninger, F., Eyben, F., Bocklet, T., Mohammadi, G., Weiss, B., 2012. The INTERSPEECH 2012 Speaker Trait Challenge. In: Proc. of Interspeech, Portland, Oregon, U. S. A. pp. 254–257.
Schuller, B., Steidl, S., Batliner, A., Schiel, F., Krajewski, J., 2011b. The INTERSPEECH 2011 Speaker State Challenge. In: Proc. of Interspeech, Florence, Italy. pp. 3201–3204.
Schuller, B., Steidl, S., Batliner, A., Schiel, F., Krajewski, J., Weninger, F., Eyben, F., 2013b. Medium-term speaker states – a review on intoxication, sleepiness and the first challenge. Computer, Speech and Language, Special Issue on "Broadening the View on Speaker Analysis"To appear.
Schweitzer, A., Lewandowski, N., 2013. Convergence of articulation rate in spontaneous speech. In: Proc. of Interspeech. pp. 525–529.
Sheard, C., Adams, R. D., Davis, P. J., 1991. Reliability and agreement of ratings of ataxic dysarthric speech samples with varying intelligibility. Journal of Speech and Hearing Research 34, 285–293.
Stark, A., Bayestehtashk, A., Asgari, M., Shafran, I., 2012. INTERSPEECH Pathology Challenge: Investigations into speaker and sentence specific. In: Proc. of Interspeech, Portland, Oregon, U. S. A. pp. 538–541.
Stelzle, F., Maier, A., Nöth, E., Bocklet, T., Knipfer, C., Schuster, M., Neukam, F. W., Nkenke, E., 2011. Automatic quantification of speech intelligibility in patients after treatment for oral squamous cell carcinoma. Journal of Oral and Maxillofacial Surgery 69, 1493–1500.
Strangert, E., Gustafson, J., 2008. What makes a good speaker? Subjective ratings, acoustic measurements and perceptual evaluation. In: Proc. of Interspeech, Brisbane, Australia. pp. 1688–1691.
Sturim, D. E., Torres-Carrasquillo, P. A., Quatieri, T. F., Malyska, N., McCree, A., 2011. Automatic detection of depression in speech using Gaussian mixture modeling with factor analysis. In: Proc. of Interspeech, Florence, Italy. pp. 2981–2984.
Tanaka, H., 1973. Speech patterns of edentulous patients and morphology of the palate in relation to phonetics. Journal of Prosthetic Dentistry 29, 16–28.
Tolarova, M., Cervenka, J., 1998. Classification and birth prevalence of orofacial clefts. American Journal of Medical Genetics 75, 126–137.
Trouvain, J., Schmidt, S., Schröder, M., Schmitz, M., Barry, W. J., 2006. Modeling personality features by changing prosody in synthetic speech. In: Proc. of Speech Prodody 2006, May 2-5, 2006, Dresden, Germany. No pagination.
Valente, F., Kim, S., P. Motlicek, P., 2012. Annotation and recognition of personality traits in spoken conversations from the AMI meetings corpus. In: Proc. of Interspeech, Portland, Oregon, U. S. A. pp. 1183–1186.
van Bezooijen, R., 2005. Approximant /r/ in Dutch: Routes and feelings. Speech Communication 47, 15–31.
van der Molen, L., van Rossum, M. A., Ackerstaff, A. H., Smeele, L. E., Rasch, C. R. N., Hilgers, F. J. M., 2009. Pretreatment organ function in patients with advanced head and neck cancer: clinical outcome measures and patients' views. BMC Ear, Nose and Throat Disorders 9 (10), 1–9.
Vijayalakshmi, P., Reddy, M. R., O'Shaughnessy, D. D., 2006. Assessment of articulatory sub-systems of dysarthric speech using an isolated-style phoneme recognition system. In: Proc. of Interspeech, Pittsburgh, U. S. A. pp. 981–984.
Wagner, J., Lingenfelser, F., André, E., 2012. A frame pruning approach for paralinguistic recognition tasks. In: Proc. of Interspeech, Portland, Oregon, U. S. A. pp. 274–277.
Weiss, B., Burkhardt, F., 2010. Voice attributes affecting likability perception. In: Proc. of Interspeech, Makuhari, Japan. pp. 1485–1488.
Weiss, B., Möller, S., 2011. Wahrnehmungsdimensionen von Stimme und Sprechweise. In: Proc. of ESSV 2011 – 22. Konferenz Elektronische Sprachsignalverarbeitung, September 28-30, Aachen, Germany. pp. 261–268.
Weninger, F., Krajewski, J., Batliner, A., Schuller, B., 2012. The Voice of Leadership: Models and Performances of Automatic Analysis in On-Line Speeches. IEEE Transactions on Affective Computing 3 (4), 496–508.

Wiggins, J. S. (Ed.), 1996. The Five-Factor Model of Personality: The Theoretical Perspectives. The Guilford Press, New York, NY.
Wu, D., 2012. Genetic algorithm based feature selection for speaker trait classification. In: Proc. of Interspeech, Portland, Oregon, U. S. A. pp. 294–297.
Yule, G., 1900. On the association of attributes in statistics. Philosophical Transactions of the Royal Society A 194, 257–319.
Zhou, X., Garcia-Romero, D., Mesgarani, N., Stone, M., Espy-Wilson, C., Shamma, S., 2012. Automatic intelligibility assessment of pathologic speech in head and neck cancer based on auditory-inspired spectro-temporal modulations. In: Proc. of Interspeech, Portland, Oregon, U. S. A. pp. 542–545.

Vitae

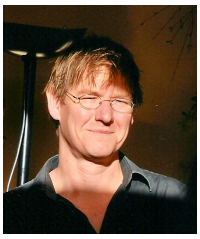
Björn Schuller received his diploma in 1999, his doctoral degree in 2006, and his habilitation in 2012, all in electrical engineering and information technology from TUM in Munich/Germany where he is tenured heading the Machine Intelligence & Signal Processing Group. He is further a Senior Lecturer at the Imperial College London/U. K. In 2013 he also chaired the Institute for Sensor Systems at the University of Passau/Germany. From 2009 to 2010 he was with the CNRS-LIMSI in Orsay/France and a visiting scientist in the Imperial College London. In 2012 he was with Joanneum Research in Graz/Austria, and in 2013 Visiting Professor of the Harbin Institute of Technology in Harbin/P. R. China and of the University of Geneva/Switzerland. Dr. Schuller is president of the Association for the Advancement of Affective Computing (AAAC), elected member of the IEEE SLTC, member of the ACM, IEEE, and ISCA and (co-)authored more than 390 peer reviewed publications leading to more than 6 100 citations – his current h-index equals 39.

Stefan Steidl received his diploma degree in Computer Science in 2002 from Friedrich-Alexander University Erlangen-Nuremberg in Germany (FAU). In 2009, he received his doctoral degree from FAU for his work on Vocal Emotion Recognition. He is currently a member of the research staff of ICSI in Berkeley/USA and the Pattern Recognition Lab of FAU. His primary research interests are the classification of naturally occurring emotion-related states and of atypical speech (children's speech, speech of elderly people, pathological voices). He has (co-)authored more than 40 publications in journals and peer reviewed conference proceedings and been a member of the Network-of-Excellence HUMAINE.

Anton Batliner received his M.A. degree in Scandinavian Languages and his doctoral degree in phonetics in 1978, both at LMU Munich/Germany. He has been a member of the research staff of the Institute for Pattern Recognition at FAU Erlangen/Germany since 1997. He is co-editor of one book and author/co-author of more than 200 technical articles, with a current h-index of 37 and more than 5 000 citations. His research interests are all aspects of prosody and paralinguistics in speech processing. Dr. Batliner repeatedly served as Workshop/Session (co)-organiser and has been Associate Editor for the IEEE Transactions on Affective Computing.

Elmar Nöth is a professor for Applied Computer Science at the University of Erlangen-Nuremberg. He studied in Erlangen and at M.I.T. and received the Dipl.-Inf. (univ.) degree and the Dr.-Ing. degree from the University of Erlangen-Nuremberg in 1985 and 1990, respectively. Since 1990 he was an assistant professor at the Institute for Pattern Recognition in Erlangen Since 2008 he is a full professor at the same institute and head of the speech group. Since 2013 he is Adjunct Professor at the King Abdulaziz University in Saudi Arabia. He is on the editorial board of Speech Communication and EURASIP Journal on Audio, Speech, and Music Processing. His current interests are prosody, analysis of pathologic speech, computer aided language learning and emotion analysis.

Alessandro Vinciarelli is Lecturer at the University of Glasgow (UK) and Senior Researcher at the Idiap Research Institute (Switzerland). His main research interest is Social Signal Processing, the domain aimed at modelling analysis and synthesis of nonverbal behaviour in social interactions. He has published more than 80 works (1700+ citations, h-index 23), organized the IEEE International Conference on Social Computing, and chaired twenty international scientific events. Furthermore, he is or has been Principal Investigator of several national and international projects, including a European Network of Excellence (the SSPNet, www.sspnet.eu). Last, but not least, Alessandro is co-founder of Klewel (www.klewel.com), a knowledge management company recognised with several awards.


Felix Burkhardt does tutoring, consulting, research and development in the working fields human-machine dialogue systems, Text-to-Speech synthesis, speaker classification, ontology based natural language modelling and emotional human-machine interfaces. Originally an expert of Speech Synthesis at the Technical University of Berlin, he wrote his PhD thesis on the simulation of emotional speech by machines, recorded the Berlin acted emotions database, EmoDB, and maintains the open source emotional speech synthesiser Emofilt. He has been working for the Deutsche Telekom AG since 2000, currently for the Telekom Innovation Laboratories in Berlin.


Rob van Son received a masters degree from the Radboud University in Nijmegen and a PhD in Phonetics from the University of Amsterdam. He has worked for the Amsterdam Center for Language and Communication (ACLC, University of Amsterdam) and the NKI-AVL in Amsterdam on a number of projects in the field of phonetics, psycholinguistics, and speech technology.


Felix Weninger received his diploma in computer science (Dipl.-Inf. degree) from TUM in 2009. He is currently pursuing his PhD degree as a researcher in the Machine Intelligence & Signal Processing Group at TUM's Institute for Human-Machine Communication. He has (co-)authored more than 60 publications in peer-reviewed books, journals and conference proceedings covering the fields of robust audio analysis, computational paralinguistics and medical informatics. Mr. Weninger serves as a reviewer for the IEEE Transactions on Audio, Speech and Language Processing, IEEE Transactions on Affective Computing and other high-profile journals and international conferences.


Florian Eyben obtained his diploma in Information Technology from TUM. He is currently pursuing his PhD degree in the Machine Intelligence & Signal Processing Group. His research interests include large

scale hierarchical audio feature extraction and evaluation, automatic emotion recognition from the speech signal, recognition of non-linguistic vocalisations, automatic large vocabulary continuous speech recognition, statistical and context-dependent language models, and Music Information Retrieval. He has over 90 publications in peer-reviewed books, journals, and conference proceedings covering many of his areas of research, leading to over 1 900 citations and an h-index of 23.

Tobias Bocklet received his diploma degree in computer science in 2007 and his PhD in 2012 both from the University of Erlangen-Nuremberg. In 2008 he was with the speech group at SRI International working on automatic speaker identification. From 2009–2013 he was a member of the research staff of the Institute of Pattern Recognition at the University of Erlangen-Nuremberg and the Department of Phoniatrics and Pedaudiology of the University Clinics Erlangen. In his work he focused on the assessment of speech and language development and pathologies. Tobias is now a researcher at Intel Corporation.

Gelareh Mohammadi is postdoctoral researcher at Idiap Research Institute, Martigny, Switzerland. Her work investigates the effect of nonverbal vocal behaviour on personality perception. She received her BSc in Biomedical Engineering from Amirkabir University of Technology, Iran, in 2003, her MSc in Electrical Engineering from Sharif University of Technology, Iran, in 2006, and her PhD in Electrical Engineering from EPFL in 2013. Her research interests include social signal processing, machine learning and pattern recognition.

Benjamin Weiss received his PhD in Linguistics in 2008 from Humboldt University of Berlin, doing his dissertation about speech tempo and pronunciation. In the same year, he evaluated embodied conversational agents as Visiting Fellow at the MARCS Auditory Laboratories, University of Western Sydney. Currently, he is working on likability of voices and multimodal Human-Computer Interaction at the Telekom Innovation Laboratories of TU Berlin.