Paralinguistics in Speech and Language — State-of-the-Art and the Challenge

Björn Schuller[1], Stefan Steidl[2,3], Anton Batliner[3], Felix Burkhardt[4], Laurence Devillers[1], Christian Müller[5], Shrikanth Narayanan[6]

[1] *CNRS-LIMSI, Spoken Language Processing Group, Orsay, France*
[2] *International Computer Science Institute (ICSI), Berkeley, CA, U. S. A.*
[3] *Friedrich-Alexander University Erlangen-Nuremberg, Pattern Recognition Lab, Germany*
[4] *Deutsche Telekom AG, Telekom Innovation Laboratories, Berlin, Germany*
[5] *German Research Center for Artificial Intelligence (DFKI), Saarbrücken, Germany*
[6] *University of Southern California, SAIL, Los Angeles, CA, U. S. A.*
Corresponding Author: *Björn Schuller*
CNRS-LIMSI, Spoken Language Processing Group, Orsay, France
schuller@IEEE.org, telephone: (33) 1 69 85 80 08, fax: (33) 1 69 85 80 88

Abstract

Paralinguistic analysis is increasingly turning into a mainstream topic in speech and language processing. This article aims to provide a broad overview of the constantly growing field by defining the field, introducing typical applications, presenting exemplary resources, and sharing a unified view of the chain of processing. It then presents the first broader Paralinguistic Challenge organised at INTERSPEECH 2010 by the authors including a historical overview of the Challenge tasks of recognising age, gender, and affect, a summary of methods used by the participants, and their results. In addition, we present the new benchmark obtained by fusion of participants' predictions and conclude by discussing ten recent and emerging trends in the analysis of paralinguistics in speech and language.

Keywords: Paralinguistics, Age, Gender, Affect, Survey, Trends, Challenge

1. Introduction

This special issue will address new approaches towards the analysis of paralinguistics in naturalistic speech and language; the aim of this overview article is to attempt a definition of this field with its manifold tasks, to provide insight into the history and state-of-the-art by discussing the results of the first challenge in this field, and to show promising recent and future trends in the recognition of realistic paralinguistic phenomena.

To start with, we define and illustrate paralinguistics in section 2. Then, in section 3, we present relevant applications. In section 4 we discuss the current status and challenges of speech and language resources for the analysis of paralinguistic phenomena, and in section 5 we sketch the common practice for their computational analysis. Section 6 describes the first Challenge on Paralinguistics held at INTERSPEECH 2010. In section 6.1, we give a historical overview of research in the fields of age, gender, and interest analysis in speech as featured in this Challenge. In section 6.2 the conditions of the challenge are discussed, and in sections 6.3 and 6.4 the tasks dealing with speaker traits and speaker states are, respectively, described, together with results obtained in the Challenge. Section 7 summarises ten recent and future trends in the field.

In a way, the different parts of this article are only loosely connected. The basics of paralinguistics, rooted in phonetics, linguistics, and all the other scholarships, are not yet tied to the methodologies of automatic speech processing. The topics of the challenge – age, gender, and interest computing – are not necessarily completely reflective of all aspects of paralinguistics. Thus, the sometimes weak integration of the different parts of this article is not a bug but a feature, mirroring the nascent state-of-the-art. Yet, we do hope that,

eventually, these efforts would contribute to a more complete integration of the multiple facets of the field. We attempt to join the loose ends of this article in section 8.

2. Paralinguistic Analysis: An Overview

Paralinguistics means 'alongside linguistics' (from the Greek preposition $\pi\alpha\rho\alpha$). Since it first came into use, in the middle of the last century, it was confined to the realm of human-human communication, but with a broad and a narrow meaning. We will follow (Crystal, 1974) who excludes visual communication and the like from the subject area and restricts the scope of the term to "vocal factors involved in paralanguage"; cf. (Abercrombie, 1968) for a definition along similar lines. "Vocal factor", however, in itself is not well-defined. Again, there can be a narrow meaning excluding linguistic/verbal factors, or a broad meaning including them. We will use the last one, defining paralinguistics as 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 restrict the term to everything that can be found in the speech signal, e. g., in telephone speech or in audio recordings, which cannot be described only in strictly phonetic and/or linguistic terms. Note that, in practice, information obtained from speech will often be combined with information obtained from vision, extra-linguistic context, and the like. There is a suitable term for that, namely multi-modal processing.

To give examples for acoustic phenomena: Everybody would agree that coughs are not linguistic events, but they are somehow embedded in the linguistic message. The same holds for laughter and filled pauses which display some of the characteristics of language, though, e. g., as far as grammatical position or phonotactics is concerned. All these phenomena are embedded in the word chain and are often modelled the same way as words in automatic speech processing; they can denote (health) state, emotion/mood, speaker idiosyncrasies, and the like. In contrast, high pitch as an indication of anxiety and breathy voice indicating attractivity, for example, are modulated onto the verbal message. As for the linguistic level, paralinguistics also deals with everything beyond pure phonology/morphology/syntax/semantics. Let us give an example from semantics: The 'normal' word for a being that can be denoted with these classic semantic features [+human, +female, +adult] is 'woman'. In contrast, 'slut' has the same denotation but a very different connotation, indicating a strong negative valence and, at the same time, the social class and/or the character of the speaker. Bunches of units, for instance the use of many and/or specific adjectives or particles, can indicate personality traits or emotional states.

The (formal) description of all these paralinguistic phenomena, be they purely acoustic or linguistic, quite often employs phonetic or linguistic terminology. *Phonetics* deals with the acoustic, perceptual, and production aspects of spoken language (speech) and *linguistics* with all aspects of written language; this is the traditional point of view. From an engineering point of view, there is a slightly different partition: normally, the focus is on recognising and subsequent understanding of the content of spoken or written language; for speech, acoustic modelling is combined with linguistic modelling whereas, naturally enough, (written) language can only be modelled by linguistic means. Another important partition is the one telling apart *form* and *function*: for instance, a phonetic form is constituted by some higher-level, structural shape or type which can be described holistically and analysed/computed using 1-n Low Level Descriptors (LLD) and functionals over time. A simple example is a high rising final tone which very often denotes, i. e., functions as indicating, a question. This is a genuine linguistic function. In addition, there are non-linguistic (*paralinguistic*) functions encoded in speech or in other vocal activities. Examples include a slurred voice if the speaker is inebriated, or a loud and high-pitched voice when a person is angry; further examples follow below.

Ever since the advent of structuralism (de Saussure, 1916), the study of (speech and) language has been more or less confined to the skeleton of language: phonology, morphology, syntax, and grammar in general; there were only rather anecdotal remarks on functions of language which go beyond pure linguistics, e. g., (Bloomfield, 1933): *"[...] pitch is the acoustic feature where gesture-like variations, non-distinctive but socially effective, border most closely upon genuine linguistic distinctions. The investigation of socially effective but non-distinctive patterns in speech, an investigation scarcely begun, concerns itself, accordingly, to a large extent with pitch."* In the same vein, cf. (Pike, 1945): *"Other intonation characteristics may be affected or*

caused by individual's physiological state - anger, happiness, excitement, age, sex, and so on. These help one to identify people and to ascertain how they are feeling [...]". Thus, the central focus of linguistics in the last century was on structural, on genuine linguistic and, as far as speech is concerned, on formal aspects within phonetics and phonology. Language was conceived of as part of semiotics which deals with *denotation*. Non-linguistic aspects were conceived of as fringe phenomena, often taken care of by neighbouring disciplines such as ethnology or psychology. However, in the middle of the last century, linguists and phoneticians began to be interested in all these phenomena mentioned by Bloomfield (1933); Pike (1945), i. e., in a broader conceptualisation of semiotics, dealing with *connotation* (e. g., affective/emotive aspects) as well. Terms such as 'extralinguistic', 'paralanguage', and 'paralinguistics' were coined, maybe for the first time, by Trager (1958), and later elaborated on by Crystal (1963, 1966, 1974).

Whereas the 'garden-fencing' within linguistics, i. e., the concentration on structural aspects, was mainly caused by theoretical considerations, a similar development can be observed within automatic speech (and language) processing which, however, was mainly caused by practical constraints. It began with concentrating on single words; then very constrained, read/acted speech, representing only one variety (i. e., one rather canonical speech register) was addressed. Nowadays, different speech registers, dialects, and spontaneous speech in general are being processed as well.

Within phonetics, the basic interest in formal aspects (i. e., phonetic, not semantic, content) meant that research concentrated mostly on purely phonetic phenomena, and rather later on the functions that can be attributed to them. In contrast, automatic speech and language processing is interested in the functional aspects, i. e., foremost in the denotational meaning. However, especially in the last decade, a new focus on connotational meaning, i. e., on paralinguistic phenomena, came into view. This goes along with the differences between phonetic and speech processing approaches, the former mainly using a few, carefully selected parameters, whereas the latter nowadays often relies on brute forcing, i. e., on using a general purpose, large feature vector with subsequent feature selection and/or reduction. Advantages and disadvantages are obvious: Phoneticians gain deep insights into a few features but do not know whether there are other ones equally important. Brute force approaches hopefully provide all the relevant features but chances are high that one may not be able to see the wood for the trees; moreover, performance and not interpretation is the main measure of quality.

Another founding principle of phonetics and linguistics is *segmentation*. Speech is composed of components that can be segmented on the time axis, from sub-phonic units to phones/phonemes to syllables/morphemes to words to constituents/phrases and, eventually, to whole narratives. In written language—the genuine topic of Natural Language Processing—segmentation is trivially given by blanks, punctuation marks, and paragraphs. All this is often obfuscated within the automatic processing of paralinguistics because, on the one hand, equally spaced analysis frames are used and, on the other hand, whole recordings, be this single utterances or whole narratives, are taken as unit of analysis. However, for a thorough and sound analysis, it is not irrelevant whether paralinguistic phenomena are really modulated onto (sort of dispersed over) speech, whether they can be clearly localised, or whether it is something in between.

Let us now introduce one specific sub-type of *phonation* as an example for paralinguistic forms and functions; "Phonation concerns the generation of acoustic energy [...] at the larynx, by the action of the vocal folds." (Laver, 1994), p. 132. One major phonation type is voicing with several sub-types: normal 'modal' voicing, laryngealisation, and falsetto. Laryngealisation (other terms are, e. g., irregular phonation, creak, vocal fry, creaky voice, or pulse register) shows up as irregular voiced stretches of speech. Mostly, it does not disturb pitch perception but is perceived as suprasegmental, differently shaped irritation modulated onto the pitch curve which can be found both in pathological and normal speech. As for the formal aspect, one can try and establish different types of laryngealisations (Batliner et al., 1993) and classify them automatically (Kießling et al., 1995; Ishi et al., 2005). As for the functional aspect, one can try and find different functions for this phenomenon. In spite of the fact that it is largely unnoticed by speakers even if they employ it themselves, there is a plethora of such functions: signalling utterance finality (Böhm and Shattuck-Hufnagel, 2007), word boundaries (Huber, 1988; Kushan and Slifka, 2006; Ní Chasaide and Gobl, 2004), or specific segments (Gerfen and Baker, 2005; Fischer-Jørgensen, 1989) (phonological/linguistic functions), holding or yielding the floor (Local and Kelly, 1986) (pragmatic functions). Laver (1994, p. 194ff.) lists different uses and functions of 'creak' phonation, among them the paralinguistic function 'bored resignation' in English

Received Pronunciation (RP), 'commiseration and complaint' in Tzeltal, and 'apology or supplication' in an Otomanguean language of Central America. Extra- and paralinguistically, laryngealisations can be a marker of personal identity and/or social class; normally, they are a marker of higher class speech. Wilden et al. (1998) quote evidence that, not only for human voices but for mammals in general, 'non-linear phenomena' (i. e., irregular phonation) can denote individuality and status (pitch as an indicator of a large body size and/or social dominance; *"... subharmonic components might be used to mimic a low-sounding voice"*). Bad news is communicated with breathy and creaky voice (Freese and Maynard, 1998), boredom with lax creaky voice, and, to a smaller extent, sadness with creaky voice (Gobl and Ní Chasaide, 2003). However, if speakers habitually produce laryngealised speech, then it is at least very difficult to tell apart other functions: It is an idiosyncratic trait. Further references are given in (Batliner et al., 2007).

Thus, a formal phenomenon, i. e., laryngealisation, can be short term or long term/persisting, or can cover almost any time slice in between these extremes. It can be both a formal (short term) *state* and a formal (long term) *trait*. In the same vein, we can speak of traits vs. states if it comes to paralinguistic functional phenomena. We choose laryngealisations as an example because it is, on the one hand, pronounced enough, displaying different forms and functions; on the other hand, neither taxonomy nor functions are fully understood. Phenomena like that will be the basis of differences found in automatic processing of paralinguistic functions but, most likely, they will not be detected or discussed because they are well hidden behind a forest of features and procedures.

At least amongst linguists, language has always been seen as the principal mode of communication for human beings (Trager, 1958) which is accompanied by other communication systems such as body posture, movement, facial expression, cf. (Crystal, 1966) where the formal means of indicating communicative stances are listed: (1) vocalisations such as 'mhm', 'shhh', (2) hesitations, (3) 'non-segmental' prosodic features such as tension (slurred, lax, tense, precise), (4) voice qualifiers (whispery, breathy, ...), (5) voice qualification (laugh, giggle, sob, cry), and (6) non-linguistic personal noises (coughs, sneezes, snores, heavy breathing, ...). Examples for some recent approaches that deal with vocal outbursts include sighs and yawns (Russell et al., 2003), laughs (Campbell et al., 2005; Batliner et al., 2011b), cries (Pal et al., 2006), hesitations and consent (Schuller et al., 2009a), and coughs (Matos et al., 2006).

The extensional differentiation between terms such as verbal/non-verbal or vocal/non-vocal is sometimes not easy to maintain and different usages do exist; as often, it might be favourable to employ a prototype concept with typical and fringe phenomena (Rosch, 1975). A fringe phenomenon, for example, is filled pauses which often are conceived of as non-verbal, vocal phenomena; however, they normally follow the native phonotactics, cannot be placed everywhere, can be exchanged by filler words such as *'well'*, and are modelled in automatic speech recognition (ASR) the same way as words.

Several formal types have been mentioned so far. As for different functions, the basic, meaningful main taxonomy is along the time axis as well, from long term traits to short term states. Again, the following listing is neither complete nor do we mention all possible varieties. For instance, we assume only two genders even if there exist more varieties in between.

- **Long term traits**:
  - *biological trait primitives* such as height (Mporas and Ganchev, 2009; Schuller et al., 2011e), weight, age (Schuller et al., 2010b), gender (Schuller et al., 2010b);
  - *group/ethnicity membership*: race/culture/social class with a weak borderline towards other linguistic concepts, i. e., speech registers such as dialect or nativeness (Omar and Pelecanos, 2010);
  - *personality traits*: likability (Weiss and Burkhardt, 2010; Bruckert et al., 2006);
  - *personality in general*, cf. below (Gocsl, 2009; Mohammadi et al., 2010; Polzehl et al., 2010);
  - a *bundle of traits* constitutes speaker idiosyncrasies, i. e., speaker-ID; speaker traits can be used to mutually improve classification performance (van Dommelen and Moxness, 1995; Schuller et al., 2011e).
- **Medium term between traits and states**:

- *(partly) self-induced more or less temporary states*: sleepiness (Krajewski et al., 2009), intoxication (e. g., alcoholisation (Levit et al., 2001; Schiel and Heinrich, 2009; Schuller et al., 2011c)), health state (Maier et al., 2009), mood (e. g., depression (Ellgring and Scherer, 1996));
- *structural (behavioural, interactional, social) signals*: role in dyads, groups, and the like (Laskowski et al., 2008), friendship and identity (Ipgrave, 2009), positive/negative attitude (Fujie et al., 2006), (non-verbal) social signals (Vinciarelli et al., 2009), entrainment (Lee et al., 2011a);

- **Short term states**:
  - *mode*: speaking style (Nose et al., 2007) and voice quality (Zhang and Hansen, 2007);
  - *emotions* (full-blown, prototypical): (Scherer, 2003);
  - *emotion-related states or affects*: for example, general (Batliner et al., 2011a,d, 2008b), stress (Hansen and Bou-Ghazale, 1997), intimacy (Batliner et al., 2008a), interest (Schuller et al., 2009a, 2010b), confidence (Pon-Barry, 2008), uncertainty (Black et al., 2008; Litman et al., 2009), deception (Enos et al., 2007; Bnzech, 2007), politeness (Nadeu and Prieto, 2011; Yildirim et al., 2005, 2011), frustration (Ang et al., 2002; Arunachalam et al., 2001; Lee et al., 2001), sarcasm (Rankin et al., 2009; Tepperman et al., 2006), pain (Belin et al., 2008).

All these traits and states can have different intensity, of course, apart from the ones that are binary or can be measured on an interval scale such as age or height. The temporal aspects of paralinguistics are fundamental, and hence need to be reflected in the computational approaches as well. Other possible taxonomies, e. g., honest vs. dishonest communication, felt vs. perceived affective states, and the like are 'higher-order' constructs and belong to the realm of psychology. Moreover, the default processing of paralinguistics normally takes phonetic and linguistic cues at face values; this is conditioned both by the normal way of collecting data via annotation and by the inherent difficulties in telling apart these different stances. Needless to say all these other taxonomies are both interesting and important for specific applications.

As for the rather permanent traits, the authors in (van Dommelen and Moxness, 1995) examined the ability of listeners to determine the speaker's height and weight from speech samples and found that, especially for male speakers, listeners are able to estimate a speaker's height and weight up to a certain degree. A similar study is documented in (Krauss et al., 2002); it deals with the assignment of photographs to voices as well as the estimation of a speaker's age, height, and weight via speech samples. The relationship between formant frequencies and body size was examined in (Gonzalez, 2004). Especially for female participants, a significant correlation between formant parameters and height could be found. Another study revealed significant negative correlations between fundamental frequency/formant dispersion and body shape and weight of male speakers (Evans et al., 2006).

One of the most important traits is personality, where research has a long tradition, leading to the now established Five-Factor Model of personality (Digman, 1990) modelling Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism ('OCEAN'). Mostly linguistic information has been used because self-assessment and peer-assessment of personality normally is conducted with the help of lists of verbal descriptors which subsequently are combined and condensed into descriptions of higher-level dimensions. *Meta language*, i. e., verbal descriptors, prevailed, and *object language*, i. e., the use of linguistic, phonetic, verbal and non-verbal markers in the speech of subjects, was less exploited. Scherer (1979) gives an overview of personality markers in speech and pertinent literature; in (Mokhtari and Campbell, 2008), listener tests were conducted, revealing that people consistently associate different tones of voice with certain speaker personality characteristics. A more recent account of the state-of-the-art, especially on the automatic recognition of personality with the help of speech and linguistic information, and experimental results can be found in (Mairesse et al., 2007). To give some examples of other topics: Gawda (2007) tests the associations among neuroticism, extraversion, and paralinguistic expression. Rosenberg and Hirschberg (2005, 2009) deal with acoustic/prosodic and lexical correlates of charismatic speech. Gregory and Gallagher (2002) demonstrate that US president election outcomes can be predicted on the basis of spectral information beneath .5 kHz, and Oberlander and Nowson (2006) employ textual features for personality classification of weblogs.

Obviously, correlations among speaker states and traits exist; for example, in (Byrd, 1994) it was shown that both the speaker's sex and the dialect region affect pronunciation. Sex and dialect related variation was studied using the TIMIT database, displaying effects on phonetic characteristics such as of central vowels, speech rate, flapping. Further, relationships between word usage and demographics were analysed in (Gillick, 2010). Demographic traits included gender, age, education level, ethnicity, and geographic region. Beyond speaker classification generally for paralinguistics, gender-dependencies have been reported consistently, e. g., by Provine (1993). Several studies indicate that considering gender information in an automatic emotion recognition system leads to higher recognition accuracies (Ververidis and Kotropoulos, 2004; Vogt and Andre, 2006); however, it is not settled yet whether this is simply due to the different pitch registers of male and female voices, or to gender-specific differences in the expression of emotion. Also, for forensic speaker classification, a wide range of different speaker characteristics such as dialect, foreign accent, sociolect, age, gender, and medical conditions has been employed (Jessen, 2007).

The search for formal paralinguistic parameters might have been dominated by acoustic parameters, probably, because of the long time prevailing experimental paradigm to use segmentally identical or at least tightly controlled experimental stimuli. However, specific functions such as the indication of evaluation/valence (positive vs. negative) are not good candidates for pure acoustic modelling (Scherer, 1981). As for the use of linguistic means, evaluation can also be influenced by subtle (seemingly 'innocent') linguistic structural means such as the use of transitive or non-transitive verbs (Fausey and Boroditsky, 2010) or of anaphoric pronouns (Batliner, 1984). However, frequently employed for indicating valence is, of course, semantics, i. e., denotations and especially connotations, via the use of specific words and word classes.

Both form and function of paralinguistic phenomena can be straightforward, easy to obtain, and frequent, or complex, difficult to extract/annotate, and sparse. The computation of pitch register is relatively straightforward; it is normally always obtainable for speech, and thus it is a frequent parameter, and one of its functions, namely telling apart males from females, is often encoded in databases as well. Moreover, telling gender apart automatically is attractive for many applications. So is age, such as for selecting acoustic and language models of speech recognisers according to children or senior speakers (Wöllmer et al., 2011). The ground truth can be provided fully objectively, and it is known that age influences, for example, voice quality such as the jitter and shimmer parameters. In contrast, for instance, laryngealisation as marker of social status cannot always be observed, and it is often speaker-specific; it is not easy to tell apart its different functions, or to annotate it. So far, we do not know of any large database annotated with this phenomenon and such functions. Therefore, age and gender were natural trait candidates for a first paralinguistic challenge; in addition, we employed a database with the emotion-related state 'interest', in order to cover both long term traits and short term states.

3. Applications

In human-human communication, the determination of a speaker's states and traits takes place at all times; this we want to bundle under the heading *speaker classification*. It is valuable for the communication process, because people constantly adapt their manner of speaking, based on the assessment of their communication partner's age, gender, mood, or mother tongue; in addition, they exploit this information to interpret their communication partners' intentions. Thus it pays off to incorporate such strategies into automated voice services. Moreover, new applications will be possible. Various scenarios are envisaged in the following where speaker classification could be applied. Some of them have been repeatedly mentioned in the literature or in the public media; a few are even deployed as real-world applications. Many scenarios based in particular on emotion recognition are further discussed in (Cowie et al., 2001; Picard, 2003; Batliner et al., 2006).

Before we describe the types of applications, we have to put forth two caveats: performance and ethics. These are mutually interdependent: If the performance of a classification is low, it might either be not worthwhile or unethical to use it—or both. Lie detectors are a good example. Despite high expectations and promises, their performance is rather poor. Even if the—assumed—performance might please some juries, it is definitely unethical to base verdicts of guilty on such poor evidence. On the other hand, their performance could be sufficiently high to warrant their use for computer surveillance at airports—we do not know yet. Thus we have to tell apart single instance performance and cumulative and/or statistical

evidence. Employing the detection of costumers anger in call centre application might be risky because some false alarms are inevitable. Using such information in a cumulative way for checking call centre agents' quality might be promising; however, the questions about the ethics of such automation are still open. As the present article is rather a technical one, normally we will not detail ethical considerations when describing the applications.

Speech recognition and interpretation of speakers' intention: It seems obvious that 'what' has been said has to be interpreted in the light of 'how' it has been said; natural language understanding can indeed profit from paralinguistic information (Chen, 2009), e. g., when trying to recognise equivocation (Bello, 2006). Information about speaker state or traits can be exploited even in the acoustic layer to improve recognition of 'what' has been said, e. g., by acoustic model adaptation (Fernandez and Picard, 2003; Athanaselis et al., 2005; Steidl et al., 2010).

Conversation analysis, mediation, and transmission: Use-cases for computer-aided analysis of human-human conversations include the investigation of synchrony in the prosody of married couples (Lee et al., 2010), specific types of discourse (Romanyshyn, 2009) in psychology, or the analysis and summarisation of meetings (Kennedy and Ellis, 2003; Laskowski, 2009). For severely hearing-impaired persons with cochlear-implants (CI), this may be of interest as well because so far CI processors typically alter the spectral cues which are crucial for the perception of paralinguistic information (Massida et al., 2011). Children with autism may profit from the analysis of emotional cues as they may have difficulties understanding or displaying them (Demouy et al., 2011; Mower et al., 2011). Finally, transmitting paralinguistic information along with other message elements can be used to animate avatars (de Sevin et al., 2010), to enrich dictated text messages, or to label calls in voice mailboxes by symbols such as emoticons (Biever, 2005).

Adapting to callers in a voice portal: This describes in a generic way the idea to classify the customers of a call centre and pass them on to an agent whose profile matches the caller's class. One obvious example would be the language of the caller in a multilingual call-centre, e. g., a credit-card hotline. Detecting anger in the caller's voice (think of a complaint hotline) and subsequent handling of very angry customers by specially trained agents (Burkhardt et al., 2005b) might be very beneficial. In the case of a static speaker classification such as age or gender classification, one possibility would be to implement several designs and activate the one that fits best to the current user profile. This might consist of very subtle changes: For example, elderly customers might prefer a slower speech rate in system prompts. Here, a misclassification would not lead to a perceptible difficulty for callers. On the other hand, a dynamic speaker classification such as classifying emotion of speakers, could be used to adapt the dialogue dynamically. One of the most famous examples for such an application is the emotion-aware voice portal that detects anger in the user's voice and tries to soothe her by comforting dialogue strategies, as described in (Burkhardt et al., 2005b). However, a misclassification might lead to serious problems because callers that were not angry will probably get angry if 'accused unjustly'. The technology has already reached the market: Companies offer voice-portal services that include emotion, age, gender, or language detection modules such as the technology offered by NICE®[1].

Call centre quality management: Call centre managers have a strong interest in monitoring and optimising the quality of the services provided by their agents (Mishne et al., 2005). Speaker classification can be employed for this purpose in a variety of ways. For example, a classifier for the emotional state of the agents and/or the callers can be used to calculate scores that act as indicators of the average quality of service provided over a certain period of time. Based on a large number of calls, a classifier with state-of-the-art recognition rates on the utterance level might be suitable to detect relevant changes, such as an increasing number of angry callers or an increasing average stress level of the call centre agents. Speaker classification can also be employed to identify individual calls in a corpus of recorded call centre conversations. For example, calls of angry users can be selected for the purpose of training agents to cope with this situation. Further, one can assume the success of a call centre to be higher (this could be measured in customer satisfaction or sometimes even in revenue) when the characteristics of the caller and the agent match, i. e., they are in the same age range and social level. If it is possible to create a profile of the caller groups over time (in the morning, young male professionals call and in the afternoon, elderly housewives from middle class families),

[1] http://www.nice.com/

it makes sense trying to match the structure of the call centre agent groups to the caller structure, depending on the time of the day.

Target-group specific advertising: In analogy to the development of internet services, a rising number of telephone services will be financed by advertising. Knowledge of the user group can help in choosing products and ways of marketing. This could be used as an online application if the users get classified at the beginning of the conversation, and tailored advertisement is presented to them in the course of further interaction, e. g., while they wait for a connection. In an offline application, the main user groups of a specific voice-portal or branch of a voice portal could be identified during a data collecting phase, and later, advertisement targeted for the major user group can be chosen. This is non-critical application (Batliner et al., 2006) because the main target of the interaction (to obtain more detailed user informations) would not be put at risk by the choice of an inappropriate advertisement.

Gaming, fun: A related field of applications can be found in the gaming or entertainment sector. For example, the love detector by Nemesysco Ltd.[2] attempts to classify speech samples based on how much 'love' they convey. Numerous companies offer lie or stress detectors which are supposed to detect the truth of the spoken words by brain-wave analysis manifested in the acoustic signal. For example, THQ®Entertainment[3] recently introduced the game "Truth or Lies – Someone Will Get Caught" for video consoles that comes with a microphone and claims to detect lies. Note that here a poor performance of the lie detector might not be critical because the main motivation is not to detect the truth but to keep the game going. In automatic prize competitions over the telephone, the set of questions can be matched against the callers' characteristics. Usually, the idea behind such quiz games is that callers should have a quite good chance to get the right answers in order to provide a chance to establish a relation to the caller. If the system is aware of teens calling in, perhaps questions on currently successful pop bands is the right choice; for the older generation, questions on classical music might be more preferable. Last but not least, the credibility of virtual characters in role games can be enhanced by the detection of emotions or of other personality traits.

Therapeutic serious games: Today, the technologies of video games, together with the technologies of speech analysis (e. g., stress detection) and of vocal interaction, allow the design of immersive serious games with therapeutic purpose, based on the verbal interaction and the techniques of role play. These tools have a dual purpose: First, they want to assist the therapists in automating certain therapeutic processes such as to estimate, to follow, and to treat a higher number of patients. Second, they want to allow the patients to increase the hours of exercises of cognitive remediation with a tool that gives biofeedback, e. g., on the level of stress.

Health related: On the one hand, speech based classification can be used to help elderly people to live longer in their homes by using an acoustic pain surveillance for detecting and classifying distress calls automatically without using automated speech recognition (Belin et al., 2008). On the other hand, voice classification can be used to diagnose, monitor, and screen diseases and speech disorders (Schoentgen, 2006) such as Parkinson's disease (Rektorova et al., 2007), patients who had their larynx removed due to cancer, children with cleft lip and palate (Maier et al., 2009) or dysphonia (Malyska et al., 2005); or further pathological effects (Dibazar and Narayanan, 2002). At the University of Memphis, the analysis of soundtracks from a recording system worn by young children is used to find differences between typically developing kids and those previously diagnosed with autism or with language delays (Oller et al., 2010). At the University of Haifa, an acoustic analysis method has been to detect early stages of Parkinson's disease (Sapir et al., 2009). In Germany, the pharmaceutical company Boehringer Ingelheim, together with the Fraunhofer Institute, offers a telephone help line that classifies the cough of callers. They can determine whether the caller suffers more from a 'dry' or a 'congested' cough. In (Harrison and Horne, 2000), several differences are shown in the quality of the vocal articulation after a night of sleep deprivation (reduced intonation and a slowing down of the vocal flow); in (Bard et al., 1996), a reduction of the spontaneous dialogues and performance degradation of the subjects is observed under similar conditions. Generally speaking, these results suggest effects of sleep deprivation on communication, especially with a reduction of the spontaneous verbalisations, trouble finding words, and a degradation of the articulation. Subjects under

[2] http://www.nemesysco.com/
[3] http://www.thq.com/

sleep deprivation produce less details and show less empathy toward a team-mate (Caraty and Montacie, 2010). Some stressors such as alcohol are likely to influence articulators, which helps to explain intra-speaker and inter-speaker variability (Schiel et al., 2011).

Tutoring systems: In edutainment software and tutoring systems, user states such as uncertainty (Litman et al., 2009), interest, stress, cognitive load (Boril et al., 2010), or even deception can be employed to adapt the system and the teaching pace (Litman and Forbes, 2003; Ai et al., 2006); generally, paralinguistic cues are essential for tutors and students to make learning successful (Price et al., 2007). In addition, automatic voice coaching, e. g., to give better public speeches or simply to intonate appropriately when learning foreign languages, becomes possible (Pfister and Robinson, 2010).

Assistive and communicative robotics: Another field of applications is Robotics. The analysis of affective states (emotion, feeling) and personality is still very rudimentary in robotics and often limits itself to tactile interactions. With a better modelling of these states and traits, we will be able to add social competence to humanoid or other highly interactive and communicative robots (Martinez and Cruz, 2005; Batliner et al., 2011c), assistive robots (Delaborde and Devillers, 2010), or to (virtual) agents (Schröder et al., 2008).

Surveillance: There are many security related situations (surgical operation (Schuller et al., 2010a), crisis management, and all the tasks connected to piloting) where monitoring of stress level, sleepiness, intoxication, and such, may play a vital role (Ronzhin, 2005). Speech can be used as modality of analysis for these states. In addition, counter terrorism or counter vandalism surveillance may be aided by analysing paralinguistic cues such as aggressiveness of potential aggressors (Schuller et al., 2008b; Kwon et al., 2008), or fear of potential victims (Clavel et al., 2008).

Media retrieval: In the field of multimedia retrieval, paralinguistic information is of interest for manifold types of media searches, such as highlights in sports games by measuring the level of excitement in the reporter's speech (Boril et al., 2011) or simply looking for speakers belonging to specific classes (such as age, gender, or charisma (Schuller et al., 2011d)).

Encoding and compression: Paralinguistic information can additionally be used to encode or compress speech. For example, the MPEG4 standard comprises Ekman's 'Big Six' emotion categories. In future applications, further speaker classification can be used to exploit speaker state and trait information for high compression rates.

Controlling: In fact, even (cross-modal) control is possible by paralinguistic rather than linguistic means, e. g., in the aid of artists with upper limb disabilities to use the volume of their voice for controlling cursor movements to create drawings on the screen (Perera et al., 2009) or in programming non-verbal voice control of inanimate objects (Al Hashimi, 2009).

All these applications, modelling paralinguistic speech and language, illustrate the great potential in human-machine and human-robot interaction as well as in machine mediated human-human communication and media retrieval, apart from being a 'hot topic' in research.

4. Speech and Language Resources

Speech databases used for training and adaptation comprise the stored audio of exemplary speech for model learning (training) and testing; a transcription of the spoken content may be given, together with labels for phenomena such as emotion, age, or personality. It is common wisdom in automatic speech processing that training data should be as close as possible to the data used for testing. For some applications, read, i. e., non-spontaneous, non-realistic speech data will do because the data used for testing will be read as well; examples are the screening of pathological or non-native speech or speaker verification for access control. A majority of applications, however, calls for non-prompted, non-acted speech data. There is an obvious clash between this demand and the majority of studies that still deal with prompted, acted data. It is sometimes argued that such data can be tightly controlled and are better suited to investigate felt, not only displayed, emotions. Notwithstanding such theoretical debates, the proof of the pudding is speech obtained from realistic environments. Inevitably, algorithms trained with data basically different from the data used for testing will yield significantly lower recognition performance. We know of a few studies that proved this

statement to be true, for instance (Batliner et al., 2000); we do not know of a single study that proved the opposite.

There are other requirements that fall under the heading 'nice and very useful to have': large number of speakers and tokens; data with both close talk microphone and room microphone, displaying noise and reverberation (as is the case for real-life data); meaningful categorisations (cf. emotion categories vs. dimensions) (Mori et al., 2011); reliable annotations either by the speaker herself or by a higher number of annotators to avoid skewness (at least three but more is better); additional perception tests by independent labellers to provide a comparison of human performance for the task; balanced distribution of instances across classes or within the dimensional continuum; knowledge of the targeted distribution; high diversity of speaker ages, gender, ethnicity, language, etc.; and normally, high spoken content variation. Finally, one wishes for well-defined test, develop, and training partitions without prototypical selection of 'friendly' cases for classification; free availability of the database; and well-documented meta-data. At the same time, the privacy of the speakers has to be preserved, which can be in contradiction to the other requirements (Wyatt et al., 2007).

As shown above, a broad range of topics is related to paralinguistic speech but very few corpora with realistic data are public. Furthermore, the annotation protocol (annotation scheme, number of annotators) and the validation protocols of the annotation (alpha and kappa measures, perception tests) are rarely detailed. Most of the corpora also contain only a few speakers and small amounts of data which often are segmented based on coarse and suboptimal criteria. Often, partitions are not well-defined, which makes exact repetition of experiments impossible. Most of the time, there is no development partition defined; and meta information such as speaker age or height is often missing. This is particularly unfortunate because such meta-information could be used for parallel assessment of several states and traits at a time and can exploit mutual dependencies.

Looking at data collection, we can regroup several types of speaker states and traits into two big classes of databases. First, the class of databases with 'common' traits (e. g., age, gender), which always are present despite the fact that pertaining information is not always annotated. Obviously, there is more of this type of corpora than of the second class of corpora, which are especially tailored for the study of 'more specific' traits. Second, the class of databases with 'less usual' states and traits that are sparser in data collected in real-life scenarios, and more difficult to record in large quantity; often, their labelling is more ambiguous (Mower et al., 2009). Due to this sparseness in real-life, specific databases are built for specific analyses such as alcoholised speech, sleepy speech (Schuller et al., 2011a), or emotions. In the last ten years, large efforts in the community have been undertaken towards collecting such corpora. Regrettably, most of them used by different teams have high levels of privacy and cannot be shared (Devillers et al., 2005). For a few databases, several types of paralinguistic factors can be analysed: In the aGender database, age, gender, and likability (Burkhardt et al., 2011); in the TUM AVIC database, interest and diverse types of non-linguistic vocalisations. These corpora are introduced in sections 6.3 and 6.4. Some corpora have initially not been recorded aiming at modelling speaker states and traits; however, their rich meta-data makes this possible: The TIMIT corpus (Fisher et al., 1986), originally recorded for automatic speech recognition, can be used for speaker trait analysis (Mporas and Ganchev, 2009; Schuller et al., 2011e); the "Vera am Mittag" (VAM) corpus (Grimm et al., 2008), recorded for three-dimensional affect recognition, can be used for age and gender recognition (Schuller et al., 2011d).

Note that speaker *trait* databases require a basically different approach compared to speaker *state* databases. For a reasonable modelling, a very high number of different speakers is needed (typically around 1 000). For speaker states, one can collect data from only a few speakers (typically around ten to 100, cf., e. g., (Stuhlsatz et al., 2011)), but with varying state. Speaker independence is a must when dealing with traits – unless longitudinal studies can be conducted, and speaker adaptation may not be an option, while one can choose to deal differently with these topics when looking at states.

5. Computational Analysis

In Figure 1, a unified overview of typical paralinguistic speech and language analysis systems is given. Each component in the chain of processing is described in the following.

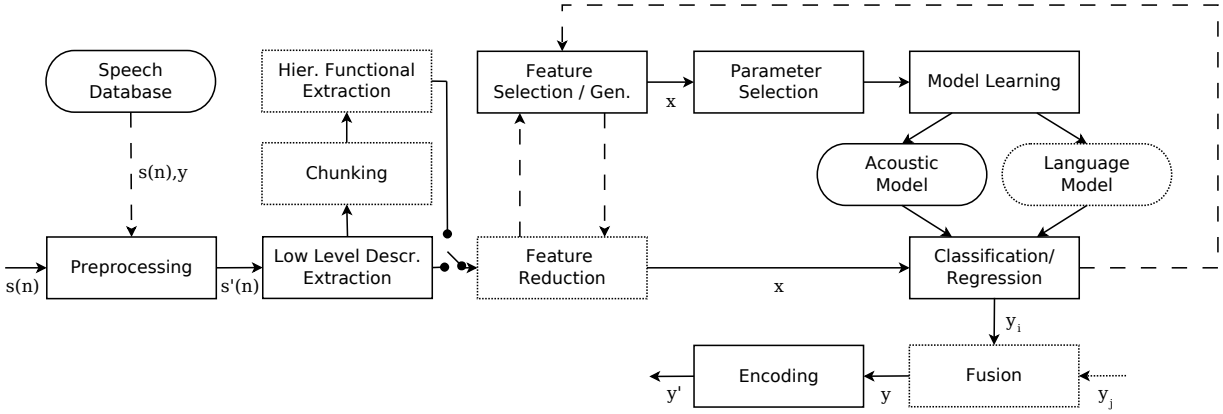


Figure 1: Unified overview of typical paralinguistic speech analysis systems. Dotted boxes indicate optional components. Dashed lines indicate steps carried out only during system training or adaptation phases, where $s(n)$, x, y are the speech signal, feature vector, and target vector, respectively high comma indicates altered versions, and subscripts diverse vectors.

Speech databases (training/adaptation phase): They comprise the stored audio of exemplary speech for model learning and testing. In addition, a transcription of the spoken content may be given together with labels for emotion, age, personality, or other phenomena (section 4).

Preprocessing: Subsequent to capturing the speech sound—a complex sequence of changes in air pressure—by a single microphone or an array of microphones, and its sampling and quantisation, preprocessing follows. This step usually aims at the enhancement of the speech signal or at the separation of multiple speakers' signals. Usually, de-noising is dealt with in the literature more frequently than de-reverberation that aims at reducing the influence of varying room impulse responses. Popular speech enhancement algorithms comprise Independent Component Analysis in the case of multiple microphones/arrays, and Non-Negative Matrix Factorisation (NMF) (Schmidt and Olsson, 2006) in the case of single microphones for separation of signals.

Low Level Descriptor extraction: At this stage, feature vectors are extracted—at approximately 100 frames per second with typical window sizes of 10–30 ms for acoustics. Windowing functions are usually rectangular for extraction of LLDs in the time domain and smooth (e. g., Hamming or Hann) for extraction in the frequency or time-frequency (TF, e. g., Gaussian or general wavelets) domains. Many systems process features on this level directly, either to provide a frame-by-frame estimate, by sliding windows of feature vectors of fixed length, or by dynamic approaches that provide some sort of temporal alignment and warping such as through Hidden Markov Models or general Dynamic Bayesian Networks. Typical acoustic LLDs in the field cover: intonation (pitch, etc.), intensity (energy, Teager functions, etc.), linear prediction cepstral coefficients (LPCC), perceptual linear prediction (PLP) parameters, cepstral coefficients (Mel frequency cepstral coefficients (MFCCs), etc.), formants (amplitude, position, width, etc.), spectrum (Mel frequency bands (MFB), NMF-based components, MPEG-7 audio spectrum projection, roll-off, etc.), TF transformation (wavelets, Gabor, etc.), harmonicity (harmonics-to-noise ratio (HNR), noise-to-harmonics Ratio (NHR), etc.), and perturbation (jitter, shimmer, etc.). These are often added by deriving further LLDs based on the raw LLDs (deltas, regression coefficients, correlation coefficients, etc.). Further, diverse filtering (smoothing, normalising, etc.) may be applied. In addition, typical linguistic LLDs comprise linguistic entities (phoneme sequences, word sequences, etc.), non-linguistic strings (laughter, sighs, etc.), and disfluencies (false starts, filled/unfilled pauses, etc.). Again, deriving further LLDs may be considered (stemmed, Part-Of-Speech tagged, semantically tagged, etc.). Finally, these may be tokenised in different ways, such as word (back-off) N-Grams, or character N-Grams; a more detailed explanation is given in (Schuller et al., 2011b). Note that their extraction usually requires automatic speech recognition. This further allows providing speech recognition confidences as LLD.

Chunking (optional): This stage is needed prior to the application of 'functionals' to determine the temporal unit of analysis. Different types of such units have been investigated requiring more or less 'intelligence' in the segmentation effort. These comprise a fixed number of frames, acoustic chunking (e. g., by Bayesian Information Criterion), voiced/unvoiced parts, phonemes, syllables, words, sub-turns in the sense of syntactically or semantically motivated chunks below the turn level, or complete turns (Batliner et al., 2010). Note that only an additional effort towards chunking is optional – and very likely beneficial; implicitly, chunking always takes place, even if just the recorded speech signal from first to last sample point is chosen as unit of analysis.

Hierarchical functional extraction (optional): In this stage, functionals are applied per LLD or spanning over LLDs (Pachet and Roy, 2009; Eyben et al., 2010b). The intention is a further information reduction and projection of the time series of potentially unknown length to a scalar value per applied functional. By that, analysis is shifted to the 'supra-segmental' level, which is known to be sufficient or even beneficial, in particular for prosodic information. For acoustic LLDs, these functionals comprise: extremes (minimum, maximum, range, etc.), mean (arithmetic, absolute, etc.), percentiles (quartiles, ranges, etc.), standard deviation, higher moments (skewness, kurtosis, etc.), peaks (number, distances, etc.), segments (number, duration, etc.), regression (coefficients, error, etc.), spectral (Discrete Cosine Transformation coefficients, etc.) and temporal (durations, positions, etc.) parameters, as provided, e. g., by the openSMILE feature extractor (Eyben et al., 2010b). For linguistic LLDs, the following functionals can be computed per chunk: vector space modelling (bag-of-words, etc.), look-up (word lists, concepts, etc.), statistical and information theoretic measures (salience, information gain, etc.). Also at this level, further and altered features can be obtained from the raw functionals (hierarchical, cross-LLD, cross-chunking, contextual, etc.). Finally, another stage of filtering (smoothing, normalising, etc.) is frequently adopted.

Feature reduction: This step usually first transforms the feature space—typically by a translation into the origin of the feature space, and a rotation to reduce covariances outside the main diagonal of the covariance matrix, in order to reduce covariance between features in the transformed space. This is typically obtained by Principal Component Analysis (PCA) (Jolliffe, 2002). Linear Discriminant Analysis (LDA) additionally employs target information (usually discrete, i. e., class-labels) in order to maximise the distance between class centres and to minimise dispersion of classes. Next, a dimension reduction by selecting a limited number of features in the new space takes place—in the case of PCA and LDA, done by choosing components with highest eigenvalues. Note that these features usually still require extraction of all features in the original space as they are mostly linear combinations of these.

Feature selection/generation (training/adaptation phase): This procedure decides which features actually to keep in the feature space. This may be of interest if a new task—e. g., estimation of a speaker's weight, body surface, race, or heart rate from acoustic properties—is not well known. In such a case, a multiplicity of features can be 'brute-forced'. From these, the ones well suited for the task can be kept. Typically, a target function is defined first. In the case of 'open loop' selection, typical target functions are based on information theoretic criteria such as information gain, or statistical ones such as correlation among features or correlation of features with the target of the task. In the case of 'closed loop', these are often the learning algorithm's error to be reduced. Usually a search function is needed in addition because an exhaustive search in the feature space is computationally hardly feasible. Such search may start with an empty set, adding features in 'forward' direction or start with the full set deleting features in 'backward' direction, or start 'somewhere in the middle' and perform bidirectional processing. Often randomness is injected or the search is based entirely on random selection guided by principles such as evolutionary (genetic) algorithms. As the search is usually based on accepting a sub-optimal solution by reducing computation effort, 'floating' is often added to overcome nesting effects (Pudil et al., 1994; Ververidis and Kotropoulos, 2006). In the case of forward search, (limited) backward steps are added to avoid too 'greedy' a search. This 'Sequential Forward Floating Search' is among the most popular in the field, as one typically selects a small number of final features out of a large set. In addition, generation of further feature variants can be considered within the selection of features, e. g., by applying single feature or multiple feature mathematical operations such as logarithm, or division which can lead to better representation in the feature space.

Parameter selection (training/adaptation phase): Parameter selection 'fine tunes' the learning algorithm. In the example of neuronal networks, this can comprise optimisation of their topology such as the

number of neurons in the hidden layer, network initialisation, the type of function in the neurons, or step size in the gradient-based back-propagation in the learning phase. Indeed, the performance of a machine learning algorithm can be significantly influenced by optimal or suboptimal parametrisation. While this step is seldom carried out systematically apart from varying expert-picked 'typical' values, the most popular approach is often grid search. As for feature selection, it is crucial not to 'tune' on speech instances used for evaluation because obviously, this would lead to overestimation of performance.

Model learning (training/adaptation phase): This is the actual training phase in which the classifier or regressor model is built, based on labelled data. There are classifiers or regressors that do not need this phase (so called lazy learners) as they only decide at run-time which class to choose, e. g., by the training instance with shortest distance in the feature space to the testing ones. However, these are seldom used, as they typically do not lead to sufficient accuracy or robustness in the rather complex task of speech analysis.

Classification/regression: This step assigns the actual target to an unknown test instance. In the case of classification, these are discrete labels such as Ekman's 'big six' emotion classes anger, disgust, fear, happiness, sadness, and surprise. In the case of regression, the output is a continuous value like a speaker's height in centimetres or age in years; in the case of emotion dimensions like potency, arousal, and valence, or the 'big five' personality dimensions of openness, conscientiousness, extraversion, agreeableness, and neuroticism, this is a real value typically ranging from -1 to +1. In general, a great diversity exists in the field, partly owing to the diverse requirements arising from the varieties of task. As discussed earlier, an increasing number of different speaker states and traits such as intoxication and sleepiness, or personality, likability, or height have recently been considered for classification or regression tasks (Mporas and Ganchev, 2009). With a growing amount of such target tasks, the question arises how knowledge of non-target speaker state and trait information may help the task at hand throughout classification or regression. If such information is not available, the follow-up question will be how combined assessment may help individual tasks, as in multi-task learning. Since learning multiple classification and regression tasks simultaneously allows to model mutual information between the tasks—which in turn can result in enhanced recognition performance for the individual tasks, multi-task learning has recently attracted a lot of attention in the machine learning community. Applying Support Vector Machines (SVM) with kernel functions that use a task-coupling parameter, or multi-task learning based on minimisation of regularisation functionals, outperformed single-task SVMs (Evgeniou and Pontil, 2004). In (Micchelli and Pontil, 2005), the authors use matrix-valued functions as kernels for multi-task learning. Further, (Ni et al., 2007) propose a hierarchical Bayesian model for multi-task learning with sequential data, and (Roy and Kaelbling, 2007) presents a hierarchical extension of the classic Naive Bayes classifier, coupling multiple Naive Bayes classifiers for multi-task learning. In (Obozinski and Taskar, 2006), joint feature selection across a group of related classification and regression tasks is examined. For the task of combined stress and speech recognition, a multi-task approach representing an efficient alternative to the strategy of applying a front-end stress classification system as preprocessing step for a stress dependent recognition system has been introduced in (Womack and Hansen, 1999). The authors focus on generalising a standard one-dimensional Hidden Markov Model (HMM) to an N-channel HMM to model different stress conditions. By that is it possible to model stress at the sub-phoneme level, resulting in enhanced recognition rates. A general framework for 'lifelong learning' by applying knowledge transfer between different machine learning tasks has been proposed in (Thrun and Mitchell, 1995).

Fusion (optional): This stage exists if information is fused on the 'late semantic' level rather than on early feature level (cf., e. g., (Bocklet et al., 2010)).

Encoding (optional): Once the final decision is made, the information needs to be represented in an optimal way for system integration, e. g., in a dialogue system (De Melo and Paiva, 2007). Here, standards may be employed to ensure utmost re-usability such as VoiceXML, Extensible MultiModal Annotation markup language (EMMA) (Baggia et al., 2007), Emotion Markup Language (EmotionML) (Schröder et al., 2007), Multimodal Interaction Markup Language (MIML) (Mao et al., 2008), etc. Additional information such as confidences can reasonably be added to allow for disambiguation strategies and such.

Acoustic models: These consist of the learnt dependencies between acoustic observations and classes, or continuous values in the case of regression.

Language models: They resemble acoustic models; they store the learnt dependencies of linguistic

observations and according assignments.

6. The First Paralinguistic Challenge: Age, Gender, and Affect

We now move from general paralinguistics to concrete examples of speaker states and traits. In this section, we first sketch the history of the specific speaker classification tasks—age, gender, and affect—addressed in the first Paralinguistic Challenge held at INTERSPEECH 2010, and then introduce the Challenge conditions and results.

6.1. Historical Overview

Automatic age and gender classification from speech has been a topic of interest from as early as the 1950's (Mysak, 1959). Two main classes of features have been predominantly used for this task: Long term (mostly prosodic) features, and short term features based on MFCCs. Extensive work has also been done on both refining and measuring the significance of the long term features to the age classification task (Müller, 2006), as well as on how to optimally combine the two feature classes (Müller and Burkhardt, 2007).

In a binary classification task of perceived speaker age on Japanese read speech, (Minematsu et al., 2002a,b) used MFCCs and delta regression coefficients modelled with a Gaussian Mixture Model (GMM). Forty-three speakers previously judged as elderly, and equally as many speakers judged as non-elderly, were chosen for the study. 90.0 % of the test sentences (with a length of five seconds each) were correctly classified. By including speech rate (morae per time unit) and local perturbation of power (power peaks per time unit), the accuracy was increased to 95.3 %. Shafran et al. (2003) used HMM based classifiers on MFCCs and F0 features to recognise age in a five class task: < 25, $'25$, $26 - 50$, $'50$ and >50, all in years. A database of spontaneous English phone calls to a customer care system (65 % women, 35 % men) was used for training. Results for age recognition showed: 68.4 % correct classifications using only MFCCs, and 70.2 % correct using a combination of cepstral and F0 features. Minematsu et al. (2003) conducted a study with male speakers on Japanese read speech (age groups $6 - 12$, $20 - 60$, and $60 - 90$ years). The direct age was estimated by students in perception experiments from single sentences. Then each speaker was modelled with GMMs using MFCCs, Delta-MFCCs and Delta-Power as features yielding a correlation of 0.89 between human judgements and automatic classification.

Schötz (2006, 2007) used classification and regression trees (CART) in two studies on Swedish single-words with the aim to learn more about the relative importance of acoustic-phonetic features for automatic speaker age recognition. 50 features (e. g., measures of F0, duration and formant frequencies) were used from the phoneme segments of 2K versions of one Swedish word (*rasa*, engl. 'collapse'), produced by 214 females and 214 males. The CART obtained an accuracy of 72 %. The best correlation between direct chronological and recognised age was 0.45. Although humans and CARTs used similar cues, in a perception experiment, human listeners (mean error $\pm$ 8.9 years) were better judges of age than the CART estimators ($\pm$ 14.5 years). Schötz (2006) used 748 speakers and 78 features to construct separate estimators of age for female, male, and all speakers. Results showed that F0 and duration were the most important single features. The best recogniser of Schötz (2006) led to similar results.

Müller et al. (2003) compared six classification methods: Decision trees (DT), multilayer perceptron (MLPs), k-Nearest Neighbour (kNN), Naive Bayes (NB), and SVM in a study of automatic classification of age group and gender. Microvariations of fundamental frequency (jitter and shimmer) were extracted automatically and used as acoustic features. Two speech corpora comprising 393 speakers (about 10 000 samples from 347 speakers $>$ 60 years, about 5 000 samples from 46 speakers $<$ 60 years) were used in the study. Results showed that all six methods performed significantly better than the baselines always predicting the more frequent class (elderly: 88 %, male: 59 %). The MLP performed best with 96.6 % correct age group estimations. Müller (2005) extended the automatically extracted acoustic features to include not only jitter and shimmer but also F0, HNR, and speech rate (syllables per second) as well as pause duration and frequency. The number of speakers encompassed a total of 507 female and 657 male speakers divided into four age classes. The majority of the speakers were children and seniors. Models were trained using the same five classification methods as in (Müller et al., 2003). The best accuracy for the seven class task was again obtained using MLPs, with an overall accuracy of 63.5 %.

A first non-public comparative study of speaker age recognition is reported in (Metze et al., 2007). The evaluation data were taken from the German SpeechDat II corpus, which is annotated with age and gender labels as given by callers at the time of recording. The database consists of 4 000 native German speakers who called a recording system over the telephone and read a set of numbers, words and sentences. For each class, 80 speakers were selected for training and 20 speakers for testing (weighted age and gender structure). Training data consisted of the whole utterance set of each person, up to 44 utterances. For further analysis, a sub-set of short utterances and another set of longer sentences was created. In order to evaluate the performance on data that originate from a different domain, the systems were also tested on VoiceClass data. These data were collected at the Deutsche Telekom and consists of 660 native speakers of German who called a voice recorder and freely talked for about five to 30 seconds on the topic of their favourite dish. Here, age structure was not controlled, and the data consist of many children and youth but almost no seniors.

Four systems A, B, C, and D were compared which are described in more detail in (Metze et al., 2007). The accuracy for the entire evaluation set ranged between 27 % and 54 % for precision while recall ranged between 46 % and 55 %. System *A* (based on class-specific phone recognisers) yielded the best performance. However, performance dropped for the short utterances which was attributed to the temporal structure realised in the phone bi-grams. System *B* (based on multiple prosodic features computed for the entire signal) and its accuracy showed very little dependency on the length of the utterance. Results of the out-of-domain task were similar which was interpreted as good robustness of the approaches against data from different domains and channels.

Müller and Burkhardt (2007) repeated the experiment comparing various combinations of short term cepstral and long term prosodic features. The best performing system for the entire evaluation set was a set of Gaussian Mixture Models using frame-based MFCCs plus a set of SVMs using utterance-level pitch statistics combined on the score level. It achieved a precision of 50 % and a recall of 49 %. Significant improvements were reported later by Bocklet et al. (2008) using a GMM-SVM supervector approach. With the best parameter set, this system achieved a performance of 77 % precision and 74 % recall on the entire test set. Note that the test conditions in (Bocklet et al., 2008) were different from the original evaluation conditions.

Concerning 'affect', a broad overview of its analysis in speech and language is given in (Schuller et al., 2011b). Here, we only give a very short historical overview of the broader field before focusing on the actual Challenge task, i. e., interest analysis. The time-line of automatic affect analysis can roughly be broken into three phases: Some spurious papers on recognition of affect in speech during the second half of the nineties (less than ten per year), a growing interest until 2004 (maybe some 30 per year), and then, a steep rise until today ($>$ 100 per year) (Schuller et al., 2011b). What still can be observed nowadays is on the one hand, a more and more sophisticated employment of statistical procedures while, on the other hand, that these procedures often do not keep up with the requirements of processing application-oriented, realistic speech data, due to over-fitting and lack of generalisation ability. Too often, the data used are still un-realistic, i. e., prompted and acted, and thus not representative of real-life. Moreover, few standardised corpora and test conditions exist to compare performances under exactly the same conditions. Apart from the first Emotion Challenge (Schuller et al., 2009b), a multiplicity of evaluation strategies is employed—such as cross-validation or percentage splits without proper instance definition—which prevents exact reproducibility.

Considering in particular the analysis of speaker's interest using speech and language cues—the task in the Challenge—, not much literature existed prior to the Challenge introduced below. In several studies it is believed that information on interest or disinterest of users has great potential for general Human-Computer Interaction (Pentland and Madan, 2005; Shriberg, 2005) and many commercial applications, such as sales and advertisement systems or virtual guides. Within the sparse existing literature that deals with human interest sensing, the following topics have been addressed so far: Contextual (Suzuki et al., 2005), vision-based approaches (Qvarfordt et al., 2005; Koshizen et al., 2007) for curiosity detection, e. g., for topic switching in infotainment systems or in customer service systems; multimodal (Stiefelhagen et al., 2002) and audivisual interest detection in meetings (Gatica-Perez et al., 20005); or (children's) tutoring systems (Mota and Picard, 2003). Moreover, there has been some research on interest detection on the Challenge corpus (TUM AVIC, cf. section 6.4) prior to the Challenge. This work was mainly based on acoustic cues (Vlasenko et al., 2008; Schuller and Rigoll, 2009), additional linguistic cues including non-verbal vocal outbursts (Schuller et al.,

2006a), and even audiovisual cues (Schuller et al., 2007a, 2009a). Note that in these studies, non-verbal vocal outbursts were recognised automatically, cf. also for the same dataset (Schuller et al., 2008a; Schuller and Weninger, 2010; Eyben et al., 2011a).

Looking at the 'opposite side' of interest, i. e., *boredom*, there exist several studies recognising it in line with other prototypical emotion categories; however, they usually employ acted data. Probably the most frequently used (acted) data set is the Berlin Emotional Speech database (Burkhardt et al., 2005a)—for results on this set cf., e. g., (Schuller et al., 2010c; Gaurav, 2008). Besides this set, others exist, such as the one used in (Pao et al., 2010).

6.2. Challenge Conditions

We now flesh out the details of the tasks and conditions put forth in the Paralinguistic Challenge. In the *Age Sub-Challenge*, the four groups—children, youth, adults, and seniors—had to be discriminated. In the *Gender Sub-Challenge*, a three-class classification task had to be solved separating female, male, and children. Finally, the *Affect Sub-Challenge* featured the state 'interest' in ordinal representation. Thus, regression was used for this task; participants could include linguistic features but only by incorporating automatic speech recognition. To this end, transcription of the training and development sets, including those of non-linguistic vocalisations, were known. Contextual knowledge could be used, as the sequence of 'sub-speaker-turns' was known.

All Sub-Challenges allowed contributors to find their own features with their own classification algorithm. However, a standard acoustic feature set (cf. Table 1) was given for each corpus that could be used. Participants, however, had to stick to the definition of train, develop, and test sets. They could report on results obtained for the development set, but had only two trials to upload their results on the test set whose labels were unknown to them. The use of well-known and obtainable further language resources, e. g., for speech recognition, was permitted. Each participation had to be accompanied by a paper presenting the results that underwent peer-review. The organisers preserved the right to re-evaluate the findings, but did not participate themselves in the Challenge.

In this Challenge, an extended set of features compared to the INTERSPEECH 2009 Emotion Challenge (Schuller et al., 2009b) was given to the participants, again choosing the open-source toolkit openSMILE (Eyben et al., 2009). This extension intended to better reflect paralinguistic information (Ishi et al., 2006; Müller, 2007). 1 582 acoustic features were obtained by systematic 'brute-force' feature (over)generation in three steps: First, the 38 low-level descriptors shown in Table 1 were extracted at 100 frames per second with varying window type and size (Hamming, 25 ms for all LLDs apart from pitch which was computed with a Gaussian window and 60 ms length), and smoothed by simple moving average low-pass filtering with a window length of three frames. Next, their first order regression coefficients were added in full compliance with the HTK toolkit (Young et al., 2006). Then, 21 functionals were applied (cf. Table 1) per audio file. However, 16 zero-information features (e. g., minimum F0, which is always zero) were discarded. Finally, the two single features 'F0 number of onsets' and 'turn duration' were added.

Due to the size of the aGender corpus, a limited set was provided for this corpus consisting of 450 features (missing descriptors and functionals are marked by '−' in Table 1). We reduced the number of descriptors from 38 to 29, and that of functionals from 21 to 8. However, the configuration file to extract the same features as for TUM AVIC with openSMILE was provided.

6.3. The Traits: Age and Gender

In the *Age* and *Gender Sub-Challenge*, the 'aGender' corpus was used for analyses and comparison (Burkhardt et al., 2010). An external company was employed by the corpus owner (Deutsche Telekom) to identify possible speakers of the targeted age and gender groups. The subjects received written instructions on the procedure and a financial reward. They were asked to call the recording system six times free of charge. Each time they were prompted by an automated Interactive Voice Response system to repeat given utterances or to produce free form spoken content. The speakers obtained individual prompt sheets containing the utterances and additional instructions. Between each session, a break of one day was scheduled to ensure more variations of the voices. Each subject's six calls had to be done with a mobile phone alternating indoor

Table 1: *Provided feature sets: 38 low-level descriptors with regression coefficients, 21 functionals. Details in the text. A '$^-$' indicates those only used for the TUM AVIC baseline. Abbreviations: DDP: difference of difference of periods, LSP: line spectral pairs, Q/A: quadratic, absolute.*

Descriptors	**Functionals**
PCM loudness$^-$	position$^-$ max./min.
MFCC [0-14]	arithmetic mean, standard deviation
log Mel Freq. Band [0–7]$^-$	skewness, kurtosis
LSP Frequency [0-7]	linear regression coefficients$^-$ 1/2
F0 by Sub-Harmonic Summation	linear regression error Q/A$^-$
F0 Envelope	quartile$^-$ 1/2/3
Voicing Probability	quartile range$^-$ 2–1/3–2/3–1
Jitter local	percentile 1/99
Jitter DDP	percentile range 99–1
Shimmer local	up-level time$^-$ 75/90

Table 2: *Age and gender classes of the aGender corpus, where f and m denote female and male, and x represents children w/o gender discrimination. The last two columns display the number of speakers/instances per set (Train and Develop).*

class	**group**	**age**	**gender**	**# Train**	**# Develop**
1	CHILD	7—14	x	68 / 4 406	38 / 2 396
2	YOUTH	15–24	f	63 / 4 638	36 / 2 722
3	YOUTH	15–24	m	55 / 4 019	33 / 2 170
4	ADULT	25–54	f	69 / 4 573	44 / 3 361
5	ADULT	25–54	m	66 / 4 417	41 / 2 512
6	SENIOR	55–80	f	72 / 4 924	51 / 3 561
7	SENIOR	55–80	m	78 / 5 549	56 / 3 826

and outdoor, in order to obtain different recording environments. The caller was connected by mobile network or ISDN and PBX to the recording system, which consisted of an application server hosting the recording application and a VoiceXML telephony server (Genesys Voice Platform). During the call, the utterances were recorded by the VoiceXML interpreter using the recording feature provided by the voice platform, and sent to the application server. The utterances were stored on the application server as 8 bit, 8 kHz, A-law data. To validate the data, the associated age cluster was compared with a manual transcription of the self stated date of birth.

Four age groups—CHILD, YOUTH, ADULT, and SENIOR—were defined. The choice was not motivated by any physiological aspects arising from the development of the human voice with increasing age but solely on market aspects stemming from the application such as dialogue control in call centres. Since children are not subdivided into female and male, this results in seven classes as shown in Table 2. Note that the given age in years might differ by one year due to birthdays close to the date of speech collection. Also there are six cases where youth stated an incorrect age. Nonetheless, the (external) speaker recruiter assured that these n speakers indeed are youth.

The following requirements were communicated to the company assigned with the speaker recruitment: At least 100 German speakers for each class acquired from all German Federal States without perfect balance of German dialects needed to be included. Multiple speakers from one household were allowed. The ability to read the given phrases was a precondition for the participation of children. As further minimum requirement, we defined age sub-clusters of equal size. To account for the different age intervals of the groups, CHILDren and YOUTH should be uniformly distributed into two year clusters, and ADULTs and SENIORs into five year clusters. This means, for example, that 25 children from seven to eight years and 20 young-aged females

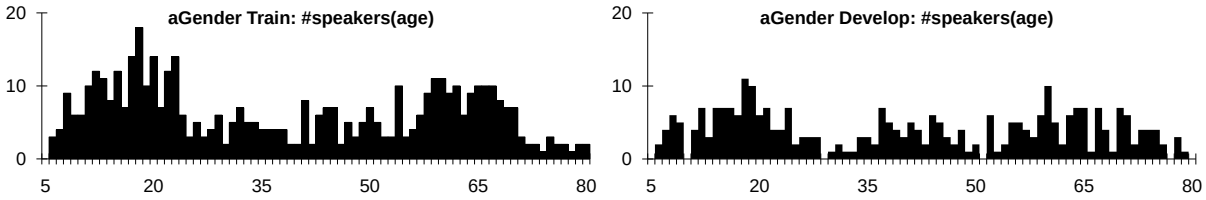


Figure 2: Age (in years) histograms for the Train and Develop sets of aGender.

between 17 to 18 years should participate. All age groups, including CHILDren, should have equal gender distribution.

The content of the database was designed in the style of the SpeechDat corpora. Each of the six recording sessions contained 18 utterances taken from a set of utterances listed in detail in (Burkhardt et al., 2010). The topics of these included *command words*, *embedded commands*, *month*, *week day*, *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 appropriate 'eliciting' questions such as *"Please tell us any date, for example the birthday of a family member."*

On an accompanying instruction sheet, all items relevant for the specific recording session were listed. There was no control that the personal dates, such as names or birthday, correspond with the information given at the screening of the callers. There was no need to repeat the same names on every recording session. Within the set of the pre-set words, it was ensured that the content for each speaker did not recur. In total, 47 hours of speech in 65 364 single utterances from 945 speakers were collected. Note that not all volunteers completed all six calls and there were cases where some of them called more often than six times, resulting in different numbers of utterances per speaker. The mean utterance length was 2.58 sec.

For each of the seven classes, we selected randomly 25 speakers as a fixed Test set (17 332 utterances, 12.45 hours), and the other 770 speakers as a Train set (53 076 utterances, 38.16 hours). The latter was further subdivided into Train (32 527 utterances in 23.43 hours of speech of 471 speakers) and Develop (20 549 utterances in 14.73 hours of speech of 299 speakers) sets. Overall, this random speaker-based partitioning resulted roughly in a 40 %/30 %/30 % Train/Develop/Test data distribution. Table 2 lists the number of speakers and the number of utterances per class in the Train and Develop sets and Figure 2 depicts the number of speakers as a histogram over their age.

Decisive for the *Age Sub-Challenge* was the age group $\{\mathrm{C}, \mathrm{Y}, \mathrm{A}, \mathrm{S}\}$ by classification as indicated in Table 2, and not the age in years. The age group could be handled either as combined age/gender task by classes $\{1, \ldots, 7\}$, or as age group task independent of gender by classes $\{\mathrm{C}, \mathrm{Y}, \mathrm{A}, \mathrm{S}\}$. For the official comparison of results, though, only the age group information was used for the competition in the *Age Sub-Challenge*, by mapping $\{1, \ldots, 7\} \rightarrow \{\mathrm{C}, \mathrm{Y}, \mathrm{A}, \mathrm{S}\}$ as denoted in Table 2. For the *Gender Sub-Challenge*, the classes $\{f, m, x\}$ had to be classified, as gender discrimination of children is considerably difficult; yet we decided to keep all instances for both tasks.

For the baselines, we exclusively exploited acoustic feature information. For transparency and easy reproducibility, we used the WEKA data mining toolkit for classification and regression (Witten and Frank, 2005; Hall et al., 2009). An overview of the baseline calculation steps in analogy to a general system as shown in Figure 1 is given in Figure 3. Note that we did not carry out optimisations because we wanted to keep the process easily reproducible. Thus, for the baseline, the Train and Develop sets are only used in union.

Table 3 shows unweighted and weighted accuracy per class for the *Age* and *Gender Sub-Challenge*. As the distribution among classes is not balanced, the competition measure is UA. The 'blind' Test set shows better results likely due to the now larger training set. In almost all cases, a 7-group sub-model, separating age groups for gender recognition and vice versa, performs slightly better than direct modelling.

Out of 32 sites registrating for the Challenge, nine research groups succeeded in taking part in the INTERSPEECH 2010 Paralinguistic Challenge with valid result submissions and accepted papers for presentation at INTERSPEECH 2010; five from Europe (Germany (2), Portugal, Czech Republic, Slovenia), one from Israel, two from the United States (one in cooperation with Korea), and one from Australia. As

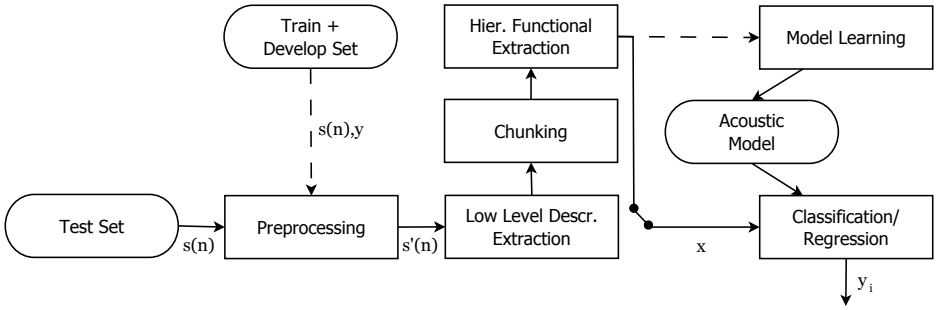


Figure 3: Baseline calculation. Dashed lines indicate steps carried out only during system training, where $s(n)$, x, y are the speech signal, feature vector, and target vector; high comma indicates altered versions, and subscripts diverse vectors.

Table 3: *Age and Gender Sub-Challenge baseline results by Sequential Minimum Optimisation learned pairwise Support Vector Machines with linear Kernel; UA: unweighted accuracy, WA: weighted accuracy.*

Sub-Ch.	**Task**	**% UA**	% WA
Train vs. Develop			
–	$\{1, \ldots, 7\}$	44.24	44.40
Age	$\{1, \ldots, 7\} \rightarrow \{\mathrm{C}, \mathrm{Y}, \mathrm{A}, \mathrm{S}\}$	47.11	46.17
	$\{\mathrm{C}, \mathrm{Y}, \mathrm{A}, \mathrm{S}\}$	46.22	45.85
Gender	$\{1, \ldots, 7\} \rightarrow \{x, f, m\}$	77.28	84.60
	$\{x, f, m\}$	76.99	86.76
Train + Develop vs. Test			
–	$\{1, \ldots, 7\}$	44.94	45.60
Age	$\{1, \ldots, 7\} \rightarrow \{\mathrm{C}, \mathrm{Y}, \mathrm{A}, \mathrm{S}\}$	48.83	46.71
	$\{\mathrm{C}, \mathrm{Y}, \mathrm{A}, \mathrm{S}\}$	**48.91**	46.24
Gender	$\{1, \ldots, 7\} \rightarrow \{x, f, m\}$	**81.21**	84.81
	$\{x, f, m\}$	80.42	86.26

Table 4: Participants of the three Sub-Challenges of the INTERSPEECH 2011 Paralinguistic Challenge

Sub-Challenges			Number of
Age	Gender	Affect	participants
✓	–	–	1
✓	✓	–	6
–	✓	✓	1
–	–	✓	1

shown in Table 4, most of the participants took part in both the Age and the Gender Sub-Challenge. One research group focused on the Age Sub-Challenge, one took part in the Affect Sub-Challenge only, and one took part in both the Gender and the Affect Sub-Challenge.

As described in detail above, we as the organisers provided a set of 450 acoustic features as part of our baseline system. Two groups focused on the improvement of the classification technique using this feature set. Nguyen et al. (2010) used fuzzy Support Vector Machines instead of the standard SVM of the baseline system and achieved small but not significant improvements on the Test set, compared to the baseline, for both the Age and the Gender Sub-Challenge. Lingenfelser et al. (2010) explored ensemble classification techniques with a set of different Naive Bayes classifiers, each of them trained on a different feature subset, with the goal to recognise one particular class especially well. Different fusion techniques were evaluated with significant improvements over their baseline, which was obtained with a single Naive Bayes classifier. However, the results remained clearly below the official baseline.

Other research groups focused on MFCC features only. Porat et al. (2010) used twelve MFCCs and Δ coefficients with cepstral mean subtraction and variance normalisation, and trained a GMM universal background model (UBM) on the non-silent parts of the training set. Then, they computed supervectors of weights: Each element of this vector represents one component of the Gaussian mixture model and measures how 'relevant' this component is for the production of the speech of the given speaker; i. e., it is the relative frequency how often this component is one of the top n components producing the sequence of feature vectors of one speaker. Thus, these supervectors characterise the given speaker and are then classified in a second step with a Support Vector Machine to obtain the age of the speaker. The authors took part only in the Age Sub-Challenge; although this approach is interesting and novel, the results on the Test set unfortunately clearly stayed behind the official baseline. The gender classification system of Gajšek et al. (2010) is a rather 'simple' one but astonishingly quite competitive. It is based on 'standard HTK' MFCCs 1–12 and the short time energy and their Δ coefficients with cepstral mean and variance normalisation. Three full-covariance GMMs with 512 components each were trained on non-silent speech regions for the three gender classes. This system outperformed their alternative gender detection system based on the GMM-UBM model adapted to the three gender classes using maximum a-posteriori (MAP) adaptation to the speech of the specific speaker and subsequent classification of the GMM supervector with a Support Vector Machine. Their 'simple' system reached the third place in the challenge being only slightly (and not significantly) behind the system on the second place.

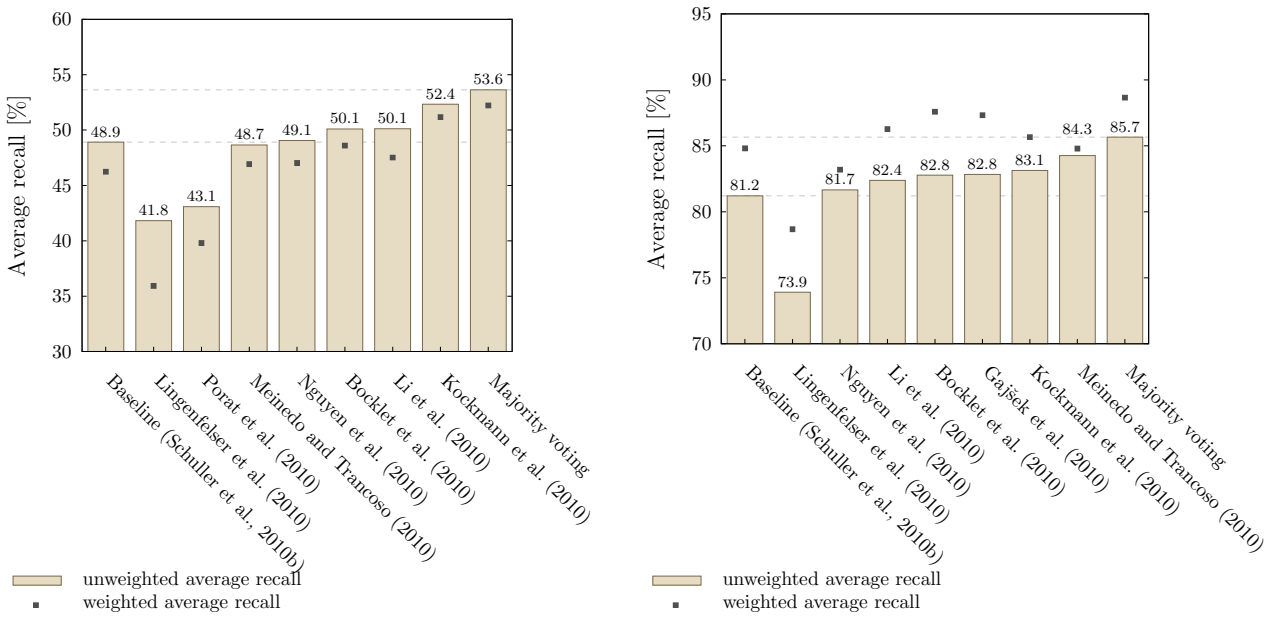
The other four participants in the Age Sub-Challenge all built various classification systems and used late fusion in order to combine them. Li et al. (2010) trained four different GMM systems, each of them based on a GMM universal background model. As features, they used 39 MFCC features (coefficients 0–12 and Δ and $\Delta\Delta$ features) with cepstral mean and variance normalisation and voice activity detection (VAD) to eliminate non-speech frames. In their first system, the GMM-UBM is adapted to the age and gender classes using MAP adaptation; the adapted models are used directly for age and gender classification. In the second system, the GMM-UBM is adapted to the speech of one speaker; GMM supervectors are created by concatenation of the mean vectors of all Gaussian components. In a second step, these supervectors were classified using 21 one-vs.-one classifiers and seven one-vs.-rest classifiers to obtain "discriminative aGender characterisation score vectors", which were subsequently classified by a back end SVM classifier. The third system used maximum likelihood linear regression (MLLR) adaptation to adapt the GMM-UBM to the speech of one speaker. Then, the MLLR matrix supervector was directly classified by a multi-class SVM classifier. In the

fourth system, the GMM-UBM is used to calculate a Tandem posterior probability supervector, which is subsequently classified with a linear kernel multi-class SVM. Additionally to these four systems, an SVM system based on the openSMILE feature set (450 features) provided by the organisers of the challenge was used. These five systems are finally fused using a linear combination of the posterior scores of the individual systems. For the final system submitted to the challenge, the authors determined the weights automatically, based on the inverse entropy. In both the Age and the Gender Sub-Challenge, the results on the Test set were significantly above the baseline results.

In many aspects, the system of the winners of the Age Sub-Challenge, Kockmann et al. (2010), is very similar to the one presented by Li et al. This system is also based on late fusion of various sub-systems. In this case, six sub-systems are fused using multi-class logistic regression. Again, one of these sub-systems is based on the official openSMILE feature set of the challenge and SVM classification. Furthermore, four sub-systems are based on MFCCs 0–12 and Δ and $\Delta\Delta$ features. A Hungarian phone recogniser is used to discard non-speech segments and a standard RelAtive SpecTrAl (RASTA) filter is used to remove slow and very fast spectral changes, which seem not to be characteristic for natural speech. The first of the four MFCC based systems consists of seven class-specific GMMs with task-specific back ends for both the Age and the Gender task. Each class-specific GMM was obtained by MAP adaptation of the UBM. For the second system, class-specific GMMs were trained directly using conventional Maximum Likelihood (ML) estimation, which serve as starting point for a further discriminative re-estimation of the mean and variance parameters using the Maximum Mutual Information (MMI) criterion. The third system is based on *speaker factors* that control the position of the speaker in the eigenvoice space from a Joint Factor Analysis (JFA) based speaker recognition system. 50 eigenvoices are trained from 309 speakers of different age and gender from the *2004 NIST Speaker Recognition Evaluation* corpus. Then, a 50-dimensional *speaker factor* vector is estimated from an aGender utterance and classified with a multi-class SVM. The fourth system is another system based on a JFA based speaker recognition system. Data of 235 speakers of the aGender Train set are used to build speaker dependent models. New data are scored against these models, resulting in a 235-dimensional feature vector, which is then classified with a multi-class SVM. The sixth system is a GMM-SVM system based on 26 perceptual linear predictive (PLP) features. Kockmann et al. won the Age Sub-Challenge and reached second place in the Gender Sub-Challenge, significantly behind the system of Meinedo et al., who presented two similar, but separate systems for the Age and the Gender Sub-Challenge.

The gender classification system of Meinedo and Trancoso (2010) is also a system based on late fusion of six sub-systems using multi-class linear logistic regression. The first two sub-systems are based on the official openSMILE feature set provided by the organisers. The first system uses a linear kernel SVM, the second one a multi-layer perceptron (MLP) with two hidden layers. Systems 3 and 4 are both MLP systems based on twelfth order PLP features, short time energy, and Δ features. Like Kockmann et al., Meinedo et al. used additional speech corpora in order to train their systems: the CMU Kids corpus, the PF-STAR Children corpus, and the European Portuguese Broadcast News BN ALERT corpus. The former two corpora contain children's speech, the latter one adults' speech with gender annotation representing a higher speaker variability and a more diverse audio background. System 3 is trained on the aGender corpus and the additional children's speech data, system 4 is trained on all listed corpora including the broadcast news corpus. For the fifth system, 28 static modulation spectrogram features are extracted which are then classified with a GMM classifier whose class-dependent models are obtained by MAP adaptation of a universal background model. This system was trained on all listed corpora, too. In addition to these five systems, the output of the age classification system was included. This system is similar to the gender-dependent system and consists of four sub-systems. The broadcast news corpus was not used since it does not contain any age information. Consequently, the corresponding system to system 4 of the gender classification system is missing. The results of the age classification are close, but slightly worse than the official baseline results. However, the authors' system for gender classification outperformed the systems of all other participants.

Bocklet et al. did not take part in the official challenge as they are with the same research group as two of the organisers. In their work, the authors compare early feature level fusion with late score level fusion. The latter fusion technique resulted in better results for the combined 7-class problem on the Development set. Five different systems/feature types are fused: The first system is based on MFCCs 1–12 and the short time logarithmic energy and their Δ features which are modelled with a GMM-UBM. MAP adaptation was used



(a) Age Sub-Challenge

(b) Gender Sub-Challenge

Figure 4: Results of the participants of the INTERSPEECH 2010 Paralinguistic Challenge for the Age and Gender Sub-Challenge.

to adapt the UBM to the utterance of a specific speaker. The GMM supervector obtained by concatenation of the mean vectors of the Gaussian components is then used as a feature vector. For the second system, the first 13 cepstral coefficients of the PLP model spectrum and their Δ features are extracted. Again, a GMM supervector is calculated. For the third system, TempoRAL PatternS (TRAPS) of a 310 ms context in 18 Mel-frequency bands are computed resulting in a 558-dimensional feature vector which is finally reduced to a dimension of 24 using LDA prior to the computation of the GMM supervector. The fourth system uses 73 prosodic features modelling F_0, energy, duration, pauses; jitter, and shimmer are extracted for voiced speech segments. On the utterance level, the three functionals *minimum*, *maximum*, and *arithmetic mean* are applied, resulting in a 219-dimensional feature vector. The fifth system is based on nine glottis parameters of a physical mass-spring vocal fold model introduced by Stevens (Stevens, 1998). Again, the three functionals *minimum*, *maximum*, and *mean* are applied resulting in a 27-dimensional feature vector. For early fusion, the three GMM supervectors, the 219-dimensional prosodic feature vector, and the 27-dimensional feature vector of the glottis parameters are concatenated and classified with a SVM. For late fusion, the feature vectors of the five systems are classified with separate SVMs and the output scores are finally combined with multi-class logistic regression. The results on the Test set are significantly better than the ones of the baseline system for both the Age and the Gender Sub-Challenge.

This overview of the techniques used by the participants of the Age and Gender Sub-Challenges showed that a diversity of features and classification techniques has been used. The individual results of the participants are shown in Figure 4. Figure 4a shows the results of the seven participants of the Age Sub-Challenge, Figure 4b those of the seven participants of the Gender Sub-Challenge. Figure 5 displays absolute improvements needed to be significantly better than a given result for four levels of significance based on a one-sided z-test (Dietterich, 1998). To be significantly better than the age baseline on a level of $\alpha = 0.01$, for example, an absolute improvement of 1.25 % is needed. Hence, results of 50.2 % unweighted average recall or better are significantly better than the baseline of 48.9 % UAR. For the Gender Sub-Challenge, results $\geq$ 82.2 % UAR are better than the baseline of 81.2 % UAR for $\alpha = 0.01$.

Figures 4a and 4b also show the result of a majority voting of the contributions of the three and the five

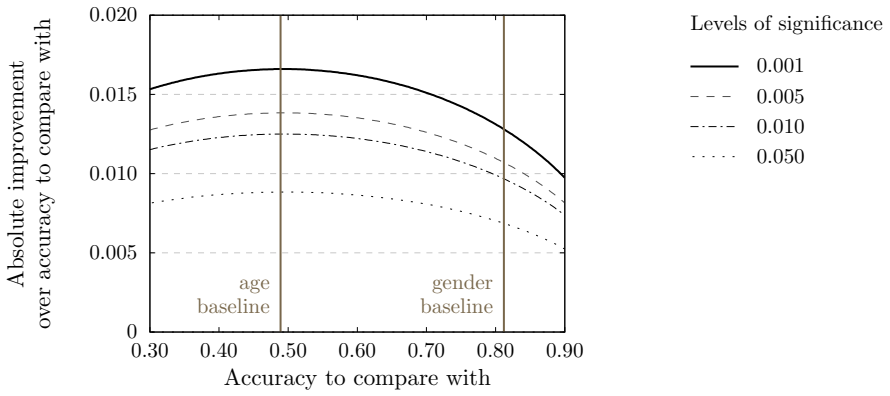
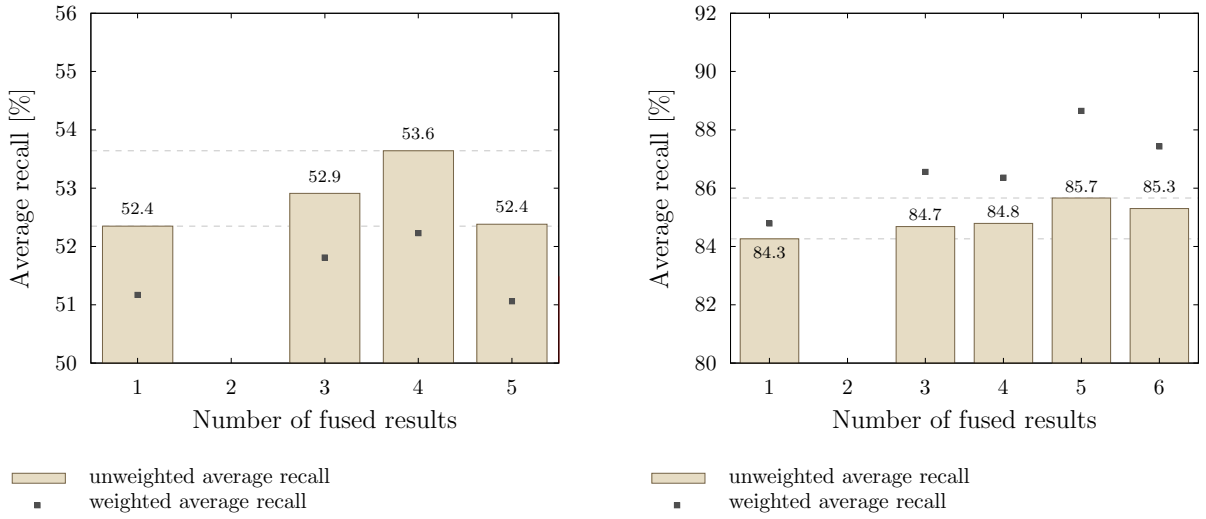


Figure 5: Lines of significant absolute improvements for different levels of significance based on one-sided z-test.



(a) Age Sub-Challenge

(b) Gender Sub-Challenge

Figure 6: Combination of the results of the n best participants of the INTERSPEECH 2010 Paralinguistic Challenge for the Age and Gender Sub-Challenge by (unweighted) majority voting. Note that this type of voting is not defined for the fusion of two systems.

best participants in the Age and the Gender Sub-Challenge, respectively. In the rare event of a draw, the first class is preferred, each, in the order of C, Y, A, S (Age Sub-Challenge) and *x*, *f*, *m* (Gender Sub-Challenge). In both cases, the improvement by fusing the best individual contributions is significant at $\alpha = 0.01$. It has to be noted that the number of contributions that are used in the majority vote is optimised on the Test set—the result is thus to be considered as an upper benchmark. Figures 6a and 6b show the results for different numbers of fused contributions. For the Age Sub-Challenge, better results than the best single contribution are only obtained if the best three or four contributions are fused. The best single contribution in the Gender Sub-Challenge can be improved by fusing the best n contributions with n ranging from three to six. However, the best result is obtain if the best five systems are used.

6.4. The States: Affect

For the *Affect Sub-Challenge*, we selected the Audiovisual Interest Corpus recorded at the Technische Universität München ('TUM AVIC') as described in (Schuller et al., 2009a). In the scenario setup, an experimenter and a subject are sitting on opposite sides of a desk. The experimenter plays the role of a product presenter and leads the subject through a commercial presentation. The subject's role is to listen to explanations and topic presentations of the experimenter, ask several questions of her/his interest, and actively interact with the experimenter, considering his/her interest in the addressed topics. The subject was explicitly asked not to worry about being polite to the experimenter, e. g., by always showing a certain level of 'polite' attention, in order to increase data variability. Visual and voice data were recorded by a camera and two microphones, one headset and one far-field microphone. For the Challenge, the lapel microphone recordings at 44.1 kHz, 16 bit were used. 21 subjects took part in the recordings, three of them Asian, the remaining European. The language throughout experiments is English, and all subjects are non-native, but experienced English speakers. More details on the subjects are summarised in Table 5.

Table 5: *Details on subjects contained in the TUM AVIC database. Further details in the text.*

Group	**# subjects**	**mean age**	**rec. time [h]**
All	21	29.9	10:22:30
Male	11	29.7	5:14:30
Female	10	30.1	5:08:00
Age <30	11	23.4	5:13:10
Age 30–40	7	32.0	3:37:50
Age >40	3	47.7	1:31:30

In the Challenge, the speech data of the subjects were used exclusively for analysis. To acquire reliable labels of a subject's 'Level of Interest' (LoI), the entire video material was segmented into speaker- and sub-speaker-turns and subsequently labelled by four male annotators, independently from each other. The annotators were undergraduate students of psychology. The intention was to annotate observed interest in the common sense. A speaker-turn is defined as continuous speech segment produced solely by one speaker—back channel interjections ("*mhm*", etc.) are ignored, i. e., every time there is a speaker change, a new speaker turn begins. This is in accordance with the common understanding of the term 'turn-taking'. Speaker-turns thus can contain multiple and especially long sentences. In order to provide Level of Interest analysis on a finer time scale, the speaker turns were further segmented at grammatical phrase boundaries: A turn lasting longer than two seconds is split by punctuation and syntactical and grammatical rules, until each segment is shorter than two seconds. These segments are referred to as sub-speaker-turns.

The LoI is annotated for each sub-speaker turn. In order to get an impression of a subject's character and behaviour prior to the actual annotation, the annotators had to watch approximately five minutes of a subject's video. This helps to find the range of intensity within which the subject expresses her/his curiosity. As the focus of interest based annotation lies on the sub-speaker turn, each of those had to be viewed at least once to find out the LoI displayed by the subject. Five Levels of Interest were distinguished:

LoI−2—*Disinterest* (subject is tired of listening and talking about the topic, is totally passive, and does not follow)

LoI−1—*Indifference* (subject is passive, does not give much feedback to the experimenter's explanations, and asks unmotivated questions, if any)

LoI0—*Neutrality* (subject follows and participates in the discourse; it cannot be recognised, if she/he is interested or indifferent about the topic)

LoI+1—*Interest* (subject wants to discuss the topic, closely follows the explanations, and asks questions)

LoI+2—*Curiosity* (strong wish of the subject to talk and learn more about the topic).

Additionally, the spoken content has been transcribed, and *long pause*, *short pause*, and non-linguistic vocalisations have been labelled. These vocalisations are *breathing* (452), *consent* (325), *hesitation* (1 147), *laughter* (261), and *coughing, other human noise* (716). There is a total of 18 581 spoken words, and 23 084

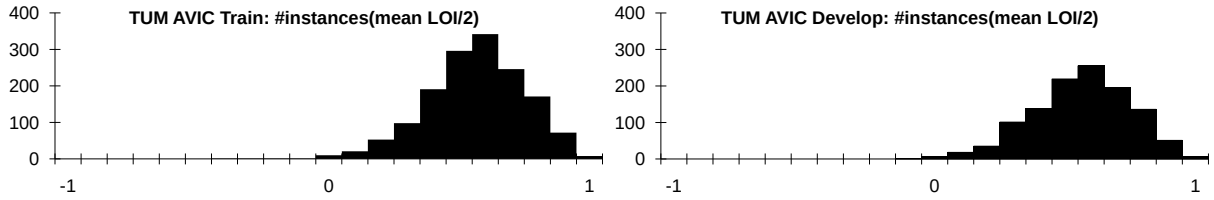


Figure 7: Mean Level of Interest (LoI, divided by 2) histograms for the Train and Develop sets of TUM AVIC.

Table 6: *Affect Sub-Challenge baseline results. Unpruned REP-Trees (25 cycles) in Random-Sub-Space meta-learning (500 Iterations, sub-space size 5 %).*

Sub-Challenge	**CC**	MLE
Train vs. Develop		
Baseline	0.604	0.118
Gajšek et al. (2010)	0.630	0.123
Train + Develop vs. Test		
Baseline	**0.421**	0.146
Gajšek et al. (2010)	0.390	0.143
Jeon et al. (2010)	0.428	0.146

word-like units including fragments and 2 901 non-linguistic vocalisations. In summary, the overall annotation contains per sub-speaker-turn segment: spoken content, non-linguistic vocalisations, individual annotator tracks, and mean LoI.

For the Challenge, ground truth is established by shifting to a continuous scale obtained by averaging the single annotator LoI. The histogram for this mean LoI is depicted in Figure 7. As can be seen, the subjects still seemed to have been somewhat polite, as almost no negative average LoI is found. Note that here the original LoI scale reaching from LoI-2 to LoI$+2$ is mapped onto $[-1, 1]$ by division by two in accordance with the scaling adopted in other corpora, e. g., (Grimm et al., 2008). Apart from higher precision, this representation form allows for subtraction of a subject's long term interest profile. Note that the Level of Interest introduced herein is highly correlated to arousal. However, at the same time there is an obvious strong correlation to valence, as, e. g., boredom has a negative valence, while strong interest is characterised by positive valence. The annotators, however, labelled interest in the common sense, thus comprising both aspects.

As before, we partitioned the 21 speakers (and 3 880 sub-speaker-turns) speaker-independently in the best achievable balance with priority on gender, followed by age and then ethnicity into three sets for Train (1 512 sub-speaker-turns in 51:44 minutes of speech, four female, four male speakers), Develop (1 161 sub-speaker-turns in 43:07 minutes of speech, three female, three male speakers), and Test (1 207 sub-speaker-turns in 42:44 minutes of speech, three female, four male speakers).

For the baselines, we again exclusively exploit acoustic feature information. Linguistic information—well known to be useful (Lee and Narayanan, 2005)—could be used for the *Affect Sub-Challenge*, but the word level transcription is given exclusively for the Train and Develop sets of TUM AVIC, as the Challenge aims at 'real-life' conditions as if for a running system 'in the wild' (Devillers and Vidrascu, 2006). Spoken content of the Test set thus needs to be recognised by automatic speech recognition rather than using perfect transcription—in the end, even recognition of affective speech may be a challenge (Steidl et al., 2010). However, to evaluate best suited textual analysis methods for interest determination, the Develop set providing perfect transcription could be used.

Table 6 depicts the results for the *Affect Sub-Challenge* baseline. The measures for this task are cross correlation (CC) and mean linear error (MLE), as found in other studies (e. g., (Grimm et al., 2008)), whereby

CC is the primary measure. Here, a clear downgrade is observed for the apparently more 'challenging' Test condition.

Only two research groups succeeded in taking part in the Affect Sub-Challenge whose papers got accepted for presentation at INTERSPEECH 2010. Gajšek et al. (2010) used the same set of low-level descriptors as in the official feature set of the baseline provided by the organisers. They formed four groups of LLDs: prosodic features, MFCC features, linear spectral frequency (LSF) pairs, and logarithmic Mel-frequency bands (LMFB). Each set of features was modelled by its own GMM-UBM, which was then adapted by MAP adaptation to a given speech sample. The GMM supervector was formed by concatenation of the mean vectors of the adapted GMM; the Level of Interest was subsequently predicted using two regression techniques: WEKA's REP-Trees with Random-Sub-Space learning and support vector regression (SVR). For each set of features, the number of Gaussian mixtures, the type of the covariance matrix (diagonal or full covariance) and the type of the regression technique were optimised on the Develop set. The four systems were finally combined by sum rule fusion; the weights were optimised using a four-fold cross-validation on the Develop set. Although the authors obtained better results on the Develop set (cf. Table 6), the results on the Test set in terms of the cross-correlation were slightly worse than the official baseline results.

The second research group taking part in the Affect Sub-Challenge was Jeon et al. (2010), who also won this Sub-Challenge. They could improve the baseline results on the Test set slightly (cf. Table 6) with a decision level fusion of three sub-systems. The first two systems used acoustic information. In both systems the official openSMILE feature set provided by the organisers was used. In the first system, the Level of Interest was predicted with Random-Sub-Space with REP-Tree and in the second one, a Gaussian Process model with normalised polynomial kernel was used. The third sub-system was based on lexical information obtained from the output of an ASR system. Bag-of-words features, the number of words in a turn, the number of strong subjective words in a turn based on the subjective dictionary, and the maximum LoI value among all words in a turn were used. The LoI value of a word, defined as the average LoI value of all utterances that contain this word, was estimated on the Train set. A super-classifier (SVM) was used to combine the outputs of these three sub-systems. The most improvement was due to the integration of lexical information, although the word error rate of the ASR system seems to be very high (estimated to be around 70 %; the human transcription was not available for the Test set).

Overall, this Challenge has seen several highly 'tuned' systems using additional data and fusing several engines; however, the best result was reached only by fusion of all participants' estimates, i. e., by employing even more systems.

7. Ten Recent and Future Trends

The variety of approaches that have been used in the Challenge mirrors the general trends described in section 5. In this final section, we will in addition present a list of *ten trends* that possibly—and hopefully—will characterise the field of automatic processing of paralinguistics in the years to come; some of the trends have been exemplified above, some others will only be mentioned.

More tasks and coupling of tasks: A variety of tasks has been mentioned passim so far; this list is, however, not exhaustive. It will simply be a matter of availability of annotated data which additional tasks will be addressed. The taxonomy along the time axis, displayed in section 2, is a convenient point of departure for (inter-)dependencies and coupling of tasks: Long term traits are interdependent to some degree (e. g., age with height and weight); medium-term phenomena can be dependent on long term traits, e. g., health state can deteriorate with age, and short term states such as emotion are of course dependent on personality traits (Reisenzein and Weber, 2009; Revelle and Scherer, 2009). Simultaneously learning multiple classification targets or exploiting contextual knowledge (automatic estimation of certain speaker or speech characteristics for a related classification task) is especially promising whenever there exist correlations between the characteristics and prediction targets. For example, a benefit of modelling age and gender in parallel over dealing with each attribute individually could be proven in our first Paralinguistic Challenge. For paralinguistic information extraction, a number of different methods exists in theory and partly in practice to exploit such additional 'contextual' knowledge such as age, emotion, gender, height, race, etc.: These methods comprise separate models per class, normalisation, or more complex adaptation by type

or addition of speaker state and trait information as additional feature. In (Schuller et al., 2011e), first beneficial effects are shown by providing knowledge on speaker traits as ground truth feature information, when assessing other speaker traits. This may occur in practical situations where, for example, height can be assessed by analysing camera recordings. The authors choose the TIMIT corpus mentioned above and the three main speaker traits age, gender, and height with the additional speaker information of speaker dialect region, education level, and race. While this work focused on knowledge inclusion by feature addition, there obviously are many other approaches to exploit such knowledge, e. g., by building age, gender, or height dependent models for any of the other tasks. This will require further experience in the case of age and height dependent models because reasonable quantisation is required. A further step will be to find methods to automatically estimate any of these parameters at the same time by mutual exploitation of each other—which is interesting, given the different task representations (continuous, ordinal, or binary). Methods such as multi-task learning of neural nets generally allow for this (Stadermann et al., 2005); however, they have to be tailored accordingly.

More continuous modelling: The classic approach is classification with two to N classes, e. g., the big six emotions, gender, or age groups. However, in most cases, this goes along with some reduction of information, especially where continuous or ordinal measures are the basic units, as is the case for intoxication (blood alcohol percentage), age, weight, height, and the like. Further, human annotation is often performed using ordinal scales, be this degree of intoxication or sleepiness, of emotion-related states such as level of interest, or based on dimensional approaches towards emotion (activation, valence, power, etc.) (Gunes et al., 2011), or personality (the 'big five OCEAN' dimensions). If continuous or ordinal measures are available, regression procedures such as (Recurrent) Neural Networks, Support Vector Regression, or REP-Trees can be applied, as well as evaluation procedures such as correlation coefficient, mean linear error, or mean square error. In addition to a continuous representation, annotation over time is becoming 'more continuous' for short term speaker states by real-time labelling per dimension (sliders can be shifted with a sampling frequency of 10 ms, for example with the Feeltrace toolkit (Cowie et al., 2000)). This continuous annotation allows for mapping to larger chunks such as syllables, words, turns, etc. (Wöllmer et al., 2010), e. g., by averaging.

More, synthesised and agglomerated data, and cross-corpus modelling: Paralinguistic databases are still sparse—in particular publicly available ones. But, this seems to be changing: There are more databases increasingly available, and hopefully, this trend will continue in the future. Then, multi-corpus and cross-corpus evaluations such as the ones described above for age and gender (cf. section 6.1) and as recently done for emotion (Stuhlsatz et al., 2011; Schuller et al., 2010c; Eyben et al., 2010a) could be part of future research on combined speaker trait analysis. In addition, this allows for data agglomeration (Schuller et al., 2011f), i. e., for a combination of multiple corpora for model training—a standard procedure in automatic speech recognition. It has further been shown that synthesised speech can be used for improved cross-corpus model training in speaker state analysis (Schuller and Burkhardt, 2010)—a promising result, as synthesised speech can be produced in high variety, altering all states and traits to produce very general learning models. Finally, semi-supervised learning or even unsupervised learning can be employed for adding large amounts of speech and language data (Jia et al., 2007; Mahdhaoui and Chetouani, 2009; Yamada et al., 2010).

More and novel features: A plethora of different (types of) new features and varieties of established features has been evaluated and described in numerous papers; it has been repeatedly shown that enlarging the feature space can help boost accuracy (Schuller et al., 2008c, 2011a). All these features can be combined in different ways—the problem being less the possibility to compute and combine them in brute force approaches, but finding out whether they really add new information and by that, help boosting performance in more general or cross-corpus and cross-task analyses, and help doing feature selection and reduction. An interesting alternative is to investigate expert-crafted higher-level features (Mubarak et al., 2006) such as perceptually motivated features (Wu et al., 2011; Mahdhaoui et al., 2010), or features based on pre-classification (Ringeval and Chetouani, 2008).

More (coupling of) linguistics and non-linguistics: Recently, there is renewed interest in non-verbal/non-linguistic phenomena and their use in human-human/human-machine communication. The ultimate solution will of course not be to define new research domains as dealing (only) with non-verbal phenomena (Social Signal Processing (Vinciarelli et al., 2009)), but to combine acoustic/vocal with linguistic/verbal—and

multimodal—analysis. There are, already, a couple of papers that integrate such non-linguistic information either directly into the linguistic string (Schuller et al., 2009a) or in a late fusion (Eyben et al., 2011b).

More optimisation: Based on publicly available corpora with well-defined partitioning, and based on the experience gained with first challenges (Schuller et al., 2009b, 2010b), more systematic optimisation is encouraged and reported. These optimisation steps usually involve feature space optimisation by reduction and selection; this can be done 'globally', i. e., for all classes at a time, or 'locally', e. g., for sub-sets of classes in hierarchical classification (Lee et al., 2009, 2011b), or for different phonemes, etc. (Bitouk et al., 2011)). Also, feature generation (Schuller et al., 2006b) can be employed. Recently, more optimisation for data representation is utilised such as instance balancing (e. g., by SMOTE (Schuller et al., 2009b) or similar techniques) or instance selection (Erdem et al., 2010; Schuller et al., 2011f), and more systematic classifier and regressor optimisation and tailored architectures (hierarchical (Yoon and Park, 2011) and hybrid (Schuller et al., 2004) or ensembles (Schuller et al., 2005; Schwenker et al., 2010)). The Challenge reported above has also shown an increasing trend towards fusion of multiple systems. More specific to speech analysis than to machine learning in general, speaker clustering for state analysis or speech clustering by state for trait analysis (Dongdong and Yingchun, 2009; Li et al., 2009), and adaptation/normalisation is observed in particular for speaker state analysis. Finally, also in (speech) signal capturing optimisation efforts can be observed increasingly in the literature, such as by use of silent speech interfaces for stress detection and speaker verification (Patil and Hansen, 2010).

More robustness: Normally, speech is preferred if recorded under acoustically favourable conditions, i. e., with a close-talk microphone. Robustness refers to 'technical' robustness against noise (Tabatabaei and Krishnan, 2010) (technical noise, cocktail party effect), reverberation (Weninger et al., 2011; Schuller, 2011), packet loss, and coding. These phenomena have, however, almost exclusively been addressed for affect analysis so far—other speaker classification tasks are yet to follow. Non-technical robustness refers to phenomena such as: attempted fraud (feigning a speaker trait such as identity or age, or a speaker state such as degree of intoxication, or emotion which is not one's own (Cannizzaro et al., 2004; Reilly et al., 2004)), correct identification even if influenced or distorted by intervening factors such as tiredness or emotion (Shahin, 2009), or the use of non-idiosyncratic/non-native language (Chen and Bond, 2010).

More standardisation: This is sorely needed as measures of performance, for instance, vary widely between studies. Examples for standardisations comprise markup languages such as the ones named above (EMMA (Baggia et al., 2007), EmotionML (Schröder et al., 2007), MIML (Mao et al., 2008), VoiceXML), documentation and well-motivated grouping of features such as the CEICES Feature Coding Scheme (Batliner et al., 2011d), and well-defined evaluation settings (Schuller et al., 2009b, 2010b, 2011c), and feature sets and classification frameworks as provided by the openSMILE (Eyben et al., 2010b) and openEAR (Eyben et al., 2009) toolkits in this field.

More realism: This is an ever-lasting topic: Basically there is agreement that less acted, more realistic data are needed but progress is slow, due to the high effort of collecting and subsequently processing realistic data. Realism concerns type of data (more spontaneous, conversational, verbally non-restricted speech) and transparent choice of instances, i. e., no preselection of nice and straightforward cases, especially not on an intuitive basis (Steidl et al., 2009). It further concerns the types of processing such as fully automatic chunking based on acoustic, phonetic, or linguistic criteria, dealing with ASR output (Metze et al., 2011) and not with forced alignment (Schuller et al., 2007b), or manual segmentation and transliteration of the spoken word chain. If additional meta-information or common knowledge is exploited in the analysis process, this should be accessed by actual web-based queries rather than providing exactly the correct and flawless data, etc. Evaluation should follow well established criteria such as partitioning into train, develop, and test set, and subject independence, i. e., not (only) randomised and by that, subject dependent many-fold cross-validation.

Cross-cultural aspects: The task is getting even more challenging if we look at cross-cultural effects. Speech has both universal aspects, on the one hand, and language- or culture-specific aspects, on the other hand. Despite centuries of theoretical debate (Boden, 2008), the balance between these two classes of properties is still highly controversial. At a superficial level, it is clear that spoken language differs across cultures and languages along a multiplicity of dimensions, ranging from phonetics through grammar, vocabulary and metaphor, to pragmatics and discourse strategies. (Some of these differences, such as those

involving metaphor or phonetics, may be pronounced even for cultural groups that share the same language, whereas other factors such as grammar tend to be more widely shared by speakers of the same language.) However, so far the effects of culture on the various facets of speaker classification have received comparatively little attention. Various authors have reported that modern approaches to speaker identification are reasonably insensitive to the language being spoken (see, e. g., (Bellegarda, 2007)). Although statistically significant differences in the performance of speaker-verification algorithms on different languages from the same corpus have been reported (Kleynhans and Barnard, 2005), these differences are relatively small in magnitude. Emotion recognition, on the other hand, has been shown to depend strongly on the language being spoken (Shami and Verhelst, 2007; Chen, 2009; Esposito and Riviello, 2011).

With the gain in accuracies and higher generalisation abilities of future automatic paralinguistic analysis systems to be expected following these trends, we look forward to seeing more applications in a real-world-context.

8. Concluding Remarks

In conclusion, we herewith summarise the multiple threads of ideas and results presented in this article. We started with a short historical overview and a description of the most important aspects of defining the realm of paralinguistics. Then, we sketched promising applications. In order to end up with successful applications, we have to first establish successful computational approaches toward detecting, classifying, and recognizing paralinguistic phenomena, be these distinctive classes or continua. Of course, a full foundation of this field is far beyond what this article can achieve. We thus concentrated on a specific use case, namely the first Paralinguistic Challenge which has to be taken as exemplar for state-of-the-art computational approaches. Obviously, there is a gap between these approaches and basic—phonetic or linguistic—research: Successful computational approaches use more or less sophisticated (combinations/fusions of) classifiers together with either some standard set of features—mostly MFCCs—or with brute force feature sets, in combination with some feature reduction or selection. Naturally enough, in such a challenge where performance is the most important measure, interpretation of phenomena and pertinent features comes second—or not at all. However, this nicely mirrors the state-of-the-art in general: Results or interpretations from basic research are simply taken as suggestion for putting into operation the machinery of data mining and pattern recognition. Most likely, for the next time to come, the two approaches will simply run in parallel, without too much interdependence. It is evident, though, that optimal performance can be reached only with a deeper understanding and harnessing of the underlying processes. We have mentioned some of these aspects in section 7, and the basic conceptual inventory has been sketched in section 2. Needless to say, there is much more to take into account for implementing a fully functional paralinguistic module within applications including generation of speech and facial gestures and the use of multimodality in general, the incorporation of non-linguistic/social context, philosophical considerations and psychological understanding and last but definitely not least, ethical considerations. Indeed, the balance between what is computationally possible and what is societally meaningful are critical but largely open questions of ongoing inquiry.

9. Acknowledgement

The authors would like to thank the sponsors of the challenge, the HUMAINE Association and Deutsche Telekom Laboratories. The responsibility lies with the authors.

This work was supported by a fellowship within the postdoc program of the German Academic Exchange Service (DAAD).

References

Abercrombie, D., 1968. Paralanguage. International Journal of Language & Communication Disorders 3, 55–59.

Ai, H., Litman, D., Forbes-Riley, K., Rotaru, M., Tetreault, J., Purandare, A., 2006. Using System and User Performance Features to Improve Emotion Detection in Spoken Tutoring Dialogs. In: Proc. Interspeech. Pittsburgh, pp. 797–800.

Al Hashimi, S., 2009. Vocal telekinesis: Towards the development of voice-physical installations. Universal Access in the Information Society 8 (2), 65–75.
Ang, J., Dhillon, R., Shriberg, E., Stolcke, A., 2002. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In: Proc. Interspeech. Denver, pp. 2037–2040.
Arunachalam, S., Gould, D., Anderson, E., Byrd, D., Narayanan, S., 2001. Politeness and frustration language in child-machine interactions. In: Proc. Eurospeech. Aalborg, pp. 2675–2678.
Athanaselis, T., Bakamidis, S., Dologlu, I., Cowie, R., Douglas-Cowie, E., Cox, C., 2005. ASR for emotional speech: Clarifying the issues and enhancing performance. Neural Networks 18, 437–444.
Baggia, P., Burnett, D. C., Carter, J., Dahl, D. A., McCobb, G., Raggett, D., 2007. EMMA: Extensible MultiModal Annotation markup language. `http://www.w3.org/TR/emma/`.
Bard, E. G., Sotillo, C., Anderson, A. H., Thompson, H. S., Taylor, M. M., 1996. The DCIEM map task corpus: Spontaneous dialogue under SD and drug treatment. Speech Communication 20, 71–84.
Batliner, A., 1984. The comprehension of grammatical and natural gender: a cross-linguistic experiment. Linguistics 22, 831–856.
Batliner, A., Burger, S., Johne, B., Kiessling, A., 1993. MÜSLI: A Classification Scheme For Laryngealizations. In: Proc. of ESCA Workshop on Prosody. Lund University, Department of Linguistics, Lund, pp. 176–179.
Batliner, A., Burkhardt, F., van Ballegooy, M., Nöth, E., 2006. A Taxonomy of Applications that Utilize Emotional Awareness. In: Proceedings of IS-LTC 2006. Ljubliana, pp. 246–250.
Batliner, A., Fischer, K., Huber, R., Spilker, J., Nöth, E., 2000. Desperately Seeking Emotions: Actors, Wizards, and Human Beings. In: Proceedings of the ISCA Workshop on Speech and Emotion. Newcastle, Northern Ireland, pp. 195–200.
Batliner, A., Schuller, B., Schaeffler, S., Steidl, S., 2008a. Mothers, Adults, Children, Pets — Towards the Acoustics of Intimacy. In: Proc. ICASSP 2008. Las Vegas, pp. 4497–4500.
Batliner, A., Schuller, B., Seppi, D., Steidl, S., Devillers, L., Vidrascu, L., Vogt, T., Aharonson, V., Amir, N., 2011a. The Automatic Recognition of Emotions in Speech, 1st Edition. Cognitive Technologies. Springer, Berlin Heidelberg, pp. 71–99.
Batliner, A., Seppi, D., Steidl, S., Schuller, B., 2010. Segmenting into adequate units for automatic recognition of emotion-related episodes: a speech-based approach. Advances in Human-Computer InteractionVol. 2010, Article ID 782802, 15 pages.
Batliner, A., Steidl, S., Eyben, F., Schuller, B., 2011b. On Laughter and Speech Laugh, Based on Observations of Child-Robot Interaction. In: Trouvain, J., Campbell, N. (Eds.), The Phonetics of Laughing. Mouton de Gruyter, Berlin, to appear.
Batliner, A., Steidl, S., Hacker, C., Nöth, E., 2008b. Private emotions vs. social interaction — a data-driven approach towards analysing emotions in speech. User Modeling and User-Adapted Interaction 18, 175–206.
Batliner, A., Steidl, S., Nöth, E., 2007. Laryngealizations and Emotions: How Many Babushkas? In: Proceedings of the International Workshop on Paralinguistic Speech – between Models and Data (ParaLing'07). Saarbrücken, pp. 17–22.
Batliner, A., Steidl, S., Nöth, E., 2011c. Associating Children's Non-Verbal and Verbal Behaviour: Body Movements, Emotions, and Laughter in a Human-Robot Interaction. In: Proceedings of ICASSP. Prague, pp. 5828–5831.
Batliner, A., Steidl, S., Schuller, B., Seppi, D., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Aharonson, V., Amir, N., 2011d. Whodunnit – Searching for the Most Important Feature Types Signalling Emotional User States in Speech. Computer Speech and Language 25, 4–28.
Belin, P., Fillion-Bilodeau, S., Gosselin, F., 2008. The montreal affective voices: A validated set of nonverbal affect bursts for research on auditory affective processing. Behavior Research Methods 40 (2), 531–539.
Bellegarda, J. R., 2007. Language-independent speaker classification over a far-field microphone. In: Mueller, C. (Ed.), Speaker Classification II: Selected Projects. Springer-Verlag, Berlin, pp. 104–115.
Bello, R., 2006. Causes and paralinguistic correlates of interpersonal equivocation. Journal of Pragmatics 38 (9), 1430–1441.
Biever, C., 2005. You have three happy messages. New Scientist 185 (2481), 21.
Bitouk, D., Verma, R., Nenkova, A., 2011. Class-level spectral features for emotion recognition. Speech Communication 52 (7–8), 613–625.
Black, M., Chang, J., Narayanan, S., 2008. An Empirical Analysis of User Uncertainty in Problem-Solving Child-Machine Interactions. In: Proceedings of the 1st Workshop on Child, Computer and Interaction. Chania, Greece, p. no pagination.
Bloomfield, L., 1933. Language. Holt, Rinhart and Winston, New York, british edition 1935, London, Allen and Unwin.
Bnzech, M., 2007. Vrit et mensonge : l'valuation de la crdibilit en psychiatrie lgale et en pratique judiciaire. Annales Medico-Psychologiques 165 (5), 351–364.
Bocklet, T., Maier, A., Bauer, J. G., Burkhardt, F., Nöth, E., 2008. Age and Gender Recognition for Telephone Applications Based on GMM Supervectors and Support Vector Machines. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). Las Vegas, NV, pp. 1605–1608.
Bocklet, T., Stemmer, G., Zeissler, V., Nöth, E., 2010. Age and gender recognition based on multiple systems – early vs. late fusion. In: ICSLP (2010), pp. 2830–2833.
Boden, M., 2008. Mind as Machine: A History of Cognitive Science. Oxford Univ. Press, New York, NY, Ch. 9.
Böhm, T., Shattuck-Hufnagel, S., 2007. Listeners recognize speakers' habitual utterancefinal voice quality. In: Proceedings of the International Workshop on Paralinguistic Speech – between Models and Data (ParaLing'07). Saarbrücken, pp. 29–34.
Boril, H., Sadjadi, S., Kleinschmidt, T., Hansen, J., 2010. Analysis and detection of cognitive load and frustration in drivers' speech. In: Proc. Interspeech 2010. Makuhari, Japan, pp. 502–505.
Boril, H., Sangwan, A., Hasan, T., Hansen, J., 2011. Automatic excitement-level detection for sports highlights generation. In: Proc. Interspeech 2010. Makuhari, Japan, pp. 2202–2205.
Bruckert, L., Lienard, J., Lacroix, A., Kreutzer, M., Leboucher, G., 2006. Women use voice parameter to assess men's characteristics. Proc. R. Soc. B. 237 (1582), 83–89.
Burkhardt, F., Eckert, M., Johannsen, W., Stegmann, J., 2010. A Database of Age and Gender Annotated Telephone Speech. In: Proc. 7th International Conference on Language Resources and Evaluation (LREC 2010). Valletta, Malta, pp. 1562–1565.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B., 2005a. A Database of German Emotional Speech. In: Proc. Interspeech. Lisbon, pp. 1517–1520.
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.
Burkhardt, F., van Ballegooy, M., Englert, R., Huber, R., 2005b. An emotion-aware voice portal. In: Proc. Electronic Speech Signal Processing ESSP. pp. 123–131.
Byrd, D., 1994. Relations of sex and dialect to reduction. Speech Communication 15 (1-2), 39–54.
Campbell, N., Kashioka, H., Ohara, R., 2005. No laughing matter. In: Proc. Interspeech. Lisbon, pp. 465–468.
Cannizzaro, M., Reilly, N., Snyder, P. J., 2004. Speech content analysis in feigned depression. Journal of psycholinguistic research 33 (4), 289–301.
Caraty, M., Montacie, C., 2010. Multivariate analysis of vocal fatigue in continuous reading. In: Proc. Interspeech 2010. Makuhari, Japan, pp. 470–473.
Chen, A., 2009. Perception of paralinguistic intonational meaning in a second language. Language Learning 59 (2), 367–409.
Chen, S. X., Bond, M. H., 2010. Two languages, two personalities? examining language effects on the expression of personality in a bilingual context. Personality and Social Psychology Bulletin 36 (11), 1514–1528.
Clavel, C., Vasilescu, I., Devillers, L., Richard, G., Ehrette, T., 2008. Fear-type emotion recognition for future audio-based surveillance systems. Speech Communication 50 (6), 487–503.
Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., Schröder, M., 2000. Feeltrace: An instrument for recording perceived emotion in real time. In: Proceedings of the ISCA Workshop on Speech and Emotion. Newcastle, Northern Ireland, pp. 19–24.
Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J., 2001. Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine 18 (1), 32–80.
Crystal, D., 1963. A perspective for paralanguage. Le Maître Phonétique 120, 25–29.
Crystal, D., 1966. The linguistic status of prosodic and paralinguistic features. Proceedings of the University of Newcastle-upon Tyne Philosophical Society 1, 93–108.
Crystal, D., 1974. Paralinguistics. In: Sebeok, T. (Ed.), Current trends in linguistics 12. Mouton, The Hague, pp. 265–295.
De Melo, C., Paiva, A., 2007. Expression of emotions in virtual humans using lights, shadows, composition and filters. Vol. 4738 LNCS of Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer, pp. 549–560.
de Saussure, F., 1916. Cours de linguistique générale. Payot, Paris.
de Sevin, E., Bevacqua, E., Pammi, S., Pelachaud, C., Schröder, M., Schuller, B., 2010. A multimodal listener behaviour driven by audio input. In: Proc. 9th Intenrational Conference on Autonomous Agents and Multiagent Systems (AAMAS) ACM International Workshop on Interacting with ECAs as Virtual Characters. Toronto, Canada, pp. 1–4.
Delaborde, A., Devillers, L., 2010. Use of non-verbal speech cues in social interaction between human and robot: Emotional and interactional markers. In: AFFINE'10 - Proceedings of the 3rd ACM Workshop on Affective Interaction in Natural Environments, Co-located with ACM Multimedia 2010. Florence, Italy, pp. 75–80.
Demouy, J., Plaza, M., Xavier, J., Ringeval, F., Chetouani, M., Prisse, D., Chauvin, D., Viaux, S., Golse, B., Cohen, D., Robel, L., 2011. Differential language markers of pathology in autism, pervasive developmental disorder not otherwise specified and specific language impairment. Research in Autism Spectrum Disorders 5 (4), 1402–1412.
Devillers, L., Vidrascu, L., 2006. Real-life emotions detection with lexical and paralinguistic cues on Human-Human call center dialogs. In: Proceedings of the International Conference on Spoken Language Processing (Interspeech 2006 – ICSLP). Pittsburgh, pp. 801–804.
Devillers, L., Vidrascu, L., Lamel, L., 2005. Challenges in real-life emotion annotation and machine learning based detection. Neural Networks 18, 407–422.
Dibazar, A., Narayanan, S., 2002. A system for automatic detection of pathological speech. In: Proc. Conference Signals, Systems, and Computers. Asilomar, CA, no pagination.
Dietterich, T. G., 1998. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Computation 10, 1895–1923.
Digman, J. M., 1990. Personality Structure: emergence of the Five-Factor Model. Ann. Rev. Psychol. 41, 417–440.
Dongdong, L., Yingchun, Y., 2009. Emotional speech clustering based robust speaker recognition system. In: Proceedings of the 2009 2nd International Congress on Image and Signal Processing, CISP'09. Tianjin, China, pp. 1–5.
Ellgring, H., Scherer, K. R., 1996. Vocal Indicators of Mood change in Depression. Journal of Nonverbal Behavior 20, 83–110.
Enos, F., Shriberg, E., Graciarena, M., Hirschberg, J., Stolcke, A., 2007. Detecting deception using critical segments. In: Proc. Interspeech. Antwerp, pp. 2281–2284.
Erdem, C. E., Bozkurt, E., Erzin, E., Erdem, A. T., 2010. RANSAC-based training data selection for emotion recognition from spontaneous speech. In: AFFINE'10 - Proceedings of the 3rd ACM Workshop on Affective Interaction in Natural Environments, Co-located with ACM Multimedia 2010. Florence, Italy, pp. 9–14.
Esposito, A., Riviello, M. T., 2011. The cross-modal and cross-cultural processing of affective information. In: Proceeding of the 2011 conference on Neural Nets WIRN10: Proceedings of the 20th Italian Workshop on Neural Nets. Vol. 226. pp. 301–310.
Evans, S., Neave, N., Wakelin, D., 2006. Relationships between vocal characteristics and body size and shape in human males: An evolutionary explanation for a deep male voice. Biological Psychology 72 (2), 160–163.
Evgeniou, T., Pontil, M., 2004. Regularized multi-task learning. In: Proc. of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. Seattle, WA, USA, pp. 109–117.
Eyben, F., Batliner, A., Schuller, B., Seppi, D., Steidl, S., 2010a. Cross-Corpus Classification of Realistic Emotions Some Pilot Experiments. In: Proc. 3rd International Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion

and Affect. Valetta, pp. 77–82.
Eyben, F., Petridis, S., Schuller, B., Tzimiropoulos, G., Zafeiriou, S., Pantic, M., 2011a. Audiovisual classification of vocal outbursts in human conversation using long-short-term memory networks. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP). Prague, pp. 5844–5847.
Eyben, F., Wöllmer, M., Schuller, B., 2009. openEAR - Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit. In: Proc. ACII. Amsterdam, pp. 576–581.
Eyben, F., Wöllmer, M., Schuller, B., 2010b. openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor. In: Proc. ACM Multimedia. Florence, Italy, pp. 1459–1462.
Eyben, F., Wöllmer, M., Valstar, M., Gunes, H., Schuller, B., Pantic, M., 2011b. String-based audiovisual fusion of behavioural events for the assessment of dimensional affect. In: Proc. 9th International IEEE Conference on Face and Gesture Recognition 2011 (FG 2011). Santa Barbara, CA, pp. 322–329.
Fausey, C. M., Boroditsky, L., 2010. Subtle linguistic cues influence perceived blame and financial liability. Psychonomic Bulletin & Review 17, 644–650.
Fernandez, R., Picard, R. W., 2003. Modeling drivers' speech under stress. Speech Communication 40, 145–159.
Fischer-Jørgensen, E., 1989. Phonetic analysis of the stød in standard Danish. Phonetica 46, 1–59.
Fisher, W., Doddington, G., Goudie-Marshall, K., 1986. The darpa speech recognition research database: Specifications and status. In: Proc. of DARPA Workshop on Speech Recognition. pp. 93–99.
Freese, J., Maynard, D. W., 1998. Prosodic features of bad news and good news in conversation. Language in Society 27, 195–219.
Fujie, S., Ejiri, Y., Kikuchi, H., Kobayashi, T., 2006. Recognition of positive/negative attitude and its application to a spoken dialogue system. Systems and Computers in Japan 37 (12), 45–55.
Gajšek, R., Žibert, J., Justin, T., Štruc, V., Vesnicer, B., Mihelič, F., 2010. Gender and Affect Recognition Based on GMM and GMM-UBM modeling with relevance MAP estimation. In: ICSLP (2010), pp. 2810–2813.
Gatica-Perez, D., McCowan, I., Zhang, D., Bengio, S., March 20005. Detecting group interest-level in meetings. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP). Philadelphia, pp. 489–492.
Gaurav, M., 2008. Performance analysis of spectral and prosodic features and their for emotion recognition in speech. In: 2008 IEEE Workshop on Spoken Language Technology, SLT 2008 - Proceedings. Goa, India, pp. 313–316.
Gawda, B., 2007. Neuroticism, extraversion, and paralinguistic expression. Psychological reports 100 (3 I), 721–726.
Gerfen, C., Baker, K., 2005. The production and perception of laryngealized vowels in Coatzospan Mixtec. Journal of Phonetics, 311–334.
Gillick, D., 2010. Can conversational word usage be used to predict speaker demographics? In: Proc. of Interspeech. Makuhari, Japan, pp. 1381–1384.
Gobl, C., Ní Chasaide, A., 2003. The role of voice quality in communicating emotion, mood and attitude. Speech Communication 40, 189–212.
Gocsl, ., 2009. Female listeners' personality attributions to male speakers: The role of acoustic parameters of speech. Pollack Periodica 4 (3), 155–165.
Gonzalez, J., 2004. Formant frequencies and body size of speaker: a weak relationship in adult humans. Journal of Phonetics 32 (2), 277–287.
Gregory, S., Gallagher, T., 2002. Spectral Analysis of Candidates Nonverbal Vocal Communication: Predicting U.S. Presidential Election Outcomes. Social Psychology Quarterly 65, 298–308.
Grimm, M., Kroschel, K., Narayanan, S., 2008. The Vera am Mittag German Audio-Visual Emotional Speech Database. In: Proc. of the IEEE International Conference on Multimedia and Expo (ICME). Hannover, Germany, pp. 865–868.
Gunes, H., Schuller, B., Pantic, M., Cowie, R., 2011. Emotion representation, analysis and synthesis in continuous space: A survey. In: Proc. International Workshop on Emotion Synthesis, rePresentation, and Analysis in Continuous spacE (EmoSPACE 2011) held in conjunction with the 9th International IEEE Conference on Face and Gesture Recognition 2011 (FG 2011). Santa Barbara, CA, pp. 827–834.
Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I., 2009. The WEKA Data Mining Software: An Update. SIGKDD Explorations 11.
Hansen, J., Bou-Ghazale, S., 1997. Getting started with susas: A speech under simulated and actual stress database. In: Proc. EUROSPEECH-97. Vol. 4. Rhodes, Greece, pp. 1743–1746.
Harrison, Y., Horne, J., 2000. The impact of sleep deprivation on decision making: A review. Journal of Experimental Psychology: Applied 6, 236–249.
Huber, D., 1988. Aspects of the Communicative Function of Voice in Text Intonation. Ph.D. thesis, Chalmers University, Göteborg/Lund.
ICSLP, 2010. Interspeech 2010 – ICSLP, 11th International Conference on Spoken Language Processing, September 26-30, 2010, Makuhari, Japan, Proceedings.
Ipgrave, J., 2009. The language of friendship and identity: Children's communication choices in an interfaith exchange. British Journal of Religious Education 31 (3), 213–225.
Ishi, C., Ishiguro, H., Hagita, N., 2005. Proposal of Acoustic Measures for Automatic Detection of Vocal Fry. In: Proc. 9th Eurospeech - Interspeech 2005. Lisbon, pp. 481–484.
Ishi, C., Ishiguro, H., Hagita, N., 2006. Using Prosodic and Voice Quality Features for Paralinguistic Information Extraction. In: Proc. of Speech Prosody 2006. Dresden, pp. 883–886.
Jeon, J. H., Xia, R., Liu, Y., 2010. Level of interest sensing in spoken dialog using multi-level fusion of acoustic and lexical evidence. In: ICSLP (2010), pp. 2802–2805.
Jessen, M., 2007. Speaker classification in forensic phonetics and acoustics. In: Müller, C. (Ed.), Speaker Classification I. Vol.

4343. Springer Berlin / Heidelberg, pp. 180–204.
Jia, L., Chun, C., Jiajun, B., Mingyu, Y., Jianhua, T., 2007. Speech emotion recognition using an enhanced co-training algorithm. In: Proceedings of the 2007 IEEE International Conference on Multimedia and Expo, ICME 2007. Beijing, China, pp. 999–1002.
Jolliffe, I. T., 2002. Principal component analysis. Springer, Berlin.
Kennedy, L., Ellis, D., December 2003. Pitch-based emphasis detection for characterization of meeting recordings. In: Proc. ASRU. Virgin Islands, pp. 243–248.
Kießling, A., Kompe, R., Niemann, H., Nöth, E., Batliner, A., 1995. Voice Source State as a Source of Information in Speech Recognition: Detection of Laryngealizations. In: Rubio Ayuso, A., López Soler, J. (Eds.), Speech Recognition and Coding. New Advances and Trends. Vol. 147 of NATO ASI Series F. Springer, Berlin, pp. 329–332.
Kleynhans, N. T., Barnard, E., Nov. 2005. Language dependence in multilingual speaker verification. In: Proceedings of the 16th Annual Symposium of the Pattern Recognition Association of South Africa. Langebaan, South Africa, pp. 117–122.
Kockmann, M., Burget, L., Černocký, J., 2010. Brno University of Technology System for Interspeech 2010 Paralinguistic Challenge. In: ICSLP (2010), pp. 2822–2825.
Koshizen, T., Kon, M., Prendinger, H., Hasegawa, Y., Aihara, K., 2007. User interest estimation using cross-modal computation. International Journal of Computational Intelligence Research 3 (3), 177–191.
Krajewski, J., Batliner, A., Golz, M., 2009. Acoustic sleepiness detection: Framework and validation of a speech-adapted pattern recognition approach. Behavior Research Methods 41, 795–804.
Krauss, R. M., Freyberg, R., Morsella, E., 2002. Inferring speakers physical attributes from their voices. Journal of Experimental Social Psychology 38 (6), 618–625.
Kushan, S., Slifka, J., 2006. Is irregular phonation a reliable cue towards the segmentation of continuous speech in American English? In: Proc. of Speech Prosody 2006. Dresden, pp. 795–798.
Kwon, H., Berisha, V., Spanias, A., 2008. Real-time sensing and acoustic scene characterization for security applications. In: 3rd International Symposium on Wireless Pervasive Computing, ISWPC 2008, Proceedings. pp. 755–758.
Laskowski, K., 2009. Contrasting Emotion-Bearing Laughter Types in Multiparticipant Vocal Activity Detection for Meetings. In: Proc. ICASSP. IEEE, Taipei, Taiwan, pp. 4765–4768.
Laskowski, K., Ostendorf, M., Schultz, T., 2008. Modeling Vocal Interaction for Text-Independent Participant Characterization in Multi-Party Conversation. In: Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue. Columbus, pp. 148–155.
Laver, J., 1994. Principles of Phonetics. Cambridge University Press, Cambridge.
Lee, C., Mower, E., Busso, C., Lee, S., Narayanan, S., 2009. Emotion recognition using a hierarchical binary decision tree approach. In: Proc. Interspeech. Brighton, pp. 320–323.
Lee, C., Narayanan, S., Pieraccini, R., 2001. Recognition of Negative Emotions from the Speech Signal. In: Proc. ASRU. Madonna di Campiglio, Italy, pp. 240–243.
Lee, C.-C., Black, M., Katsamanis, A., Lammert, A., Baucom, B., Christensen, A., Georgiou, P., Narayanan, S., 2010. Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples. In: Proc. Interspeech 2010. Makuhari, Japan, pp. 793–796.
Lee, C.-C., Katsamanis, A., Black, M., Baucom, B., Georgiou, P., Narayanan, S., 2011a. An analysis of pca-based vocal entrainment measures in married couples' affective spoken interactions. In: Proc. of Interspeech. Florence, Italy, pp. 3101–3104.
Lee, C.-C., Mower, E., Busso, C., Lee, S., Narayanan, S., 2011b. Emotion recognition using a hierarchical binary decision tree approach. Speech Communication 53 (9–10), 1162–1171.
Lee, C. M., Narayanan, S. S., 2005. Toward detecting emotions in spoken dialogs. IEEE Transactions on Speech and Audio Processing 13 (2), 293–303.
Levit, M., Huber, R., Batliner, A., Nöth, E., 2001. Use of prosodic speech characteristics for automated detection of alcohol intoxination. In: Bacchiani, M., Hirschberg, J., Litman, D., Ostendorf, M. (Eds.), Proc. of the Workshop on Prosody and Speech Recognition 2001. Red Bank, NJ, pp. 103–106.
Li, D. ., Wu, Z. ., Yang, Y. ., 2009. Speaker recognition based on pitch-dependent affective speech clustering. Moshi Shibie yu Rengong Zhineng/Pattern Recognition and Artificial Intelligence 22 (1), 136–141.
Li, M., Jung, C.-S., Han, K. J., 2010. Combining five acoustic level modeling methods for automatic speaker age and gender recognition. In: ICSLP (2010), pp. 2826–2829.
Lingenfelser, F., Wagner, J., Vogt, T., Kim, J., André, E., 2010. Age and gender classification from speech using decision level fusion and ensemble based techniques. In: ICSLP (2010), pp. 2798–2801.
Litman, D., Forbes, K., 2003. Recognizing emotions from student speech in tutoring dialogues. In: Proc. ASRU. Virgin Island, pp. 25–30.
Litman, D., Rotaru, M., Nicholas, G., 2009. Classifying Turn-Level Uncertainty Using Word-Level Prosody. In: Proc. Interspeech. Brighton, UK, pp. 2003–2006.
Local, J., Kelly, J., 1986. Projection and 'silences': notes on phonetic and conversational structure. Human Studies 9, 185–204.
Mahdhaoui, A., Chetouani, M., 2009. A new approach for motherese detection using a semi-supervised algorithm. In: Machine Learning for Signal Processing XIX - Proceedings of the 2009 IEEE Signal Processing Society Workshop, MLSP 2009. IEEE, Grenoble, France, pp. 1–6.
Mahdhaoui, A., Chetouani, M., Kessous, L., 2010. Time-frequency features extraction for infant directed speech discrimination. Vol. 5933 LNAI of Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer, pp. 120–127.
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.
Malyska, N., Quatieri, T., Sturim, D., 2005. Automatic dysphonia recognition using biologically inspired amplitude-modulation features. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP). Vol. I. Prague, pp. 873–876.
Mao, X., Li, Z., Bao, H., 2008. An extension of MPML with emotion recognition functions attached. Vol. 5208 LNAI of Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer, pp. 289–295.
Martinez, C. A., Cruz, A., 2005. Emotion recognition in non-structured utterances for human-robot interaction. In: IEEE International Workshop on Robot and Human Interactive Communication. Nashville, pp. 19–23.
Massida, Z., Belin, P., James, C., Rouger, J., Fraysse, B., Barone, P., Deguine, O., 2011. Voice discrimination in cochlear-implanted deaf subjects. Hearing research 275 (1–2), 120–129.
Matos, S., Birring, S., Pavord, I., Evans, D., 2006. Detection of cough signals in continuous audio recordings using hidden markov models. IEEE Trans. Biomedical Engineering, 1078–108.
Meinedo, H., Trancoso, I., 2010. Age and gender classification using fusion of acoustic and prosodic features. In: ICSLP (2010), pp. 2818–2821.
Metze, F., Ajmera, J., Englert, R., Bub, U., Burkhardt, F., Stegmann, J., Müller, C., Huber, R., Andrassy, B., Bauer, J. G., Littel, B., April 2007. Comparison of four approaches to age and gender recognition for telephone applications. In: ICASSP. Honolulu, Hawaii, pp. 1089–1092.
Metze, F., Batliner, A., Eyben, F., Polzehl, T., Schuller, B., Steidl, S., 2011. Emotion recognition using imperfect speech recognition. In: Proc. INTERSPEECH 2010. Makuhari, Japan, pp. 478–481.
Micchelli, C. A., Pontil, M., 2005. Kernels for multi-task learning. In: Proc. of the 18th Conference on Neural Information Processing Systems. Vancouver, BC, pp. 1–8.
Minematsu, N., Sekiguchi, M., Hirose, K., 2002a. Automatic estimation of one's age with his/her speech based upon acoustic modeling techniques of speakers. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 02). Orlando, Florida, pp. 137–140.
Minematsu, N., Sekiguchi, M., Hirose, K., 2002b. Performance improvement in estimating subjective agedness with prosodic features. In: Proceedings of Speech Prosody 2002. Aix-en-Provence, France, pp. 507–510.
Minematsu, N., Yamauchi, K., Hirose, K., 2003. Automatic Estimation of Perceptual Age Using Speaker Modeling Techniques. In: Proceedings of the Eighth European Conference on Speech Communication and Technology (Eurospeech 2003). Geneva, Switzerland, pp. 3005 – 3008.
Mishne, G., Carmel, D., Hoory, R., Roytman, A., Soffer, A., 2005. Automatic analysis of call-center conversations. In: Proc. CIKM'05. Bremen, Germany, pp. 453–459.
Mohammadi, G., Vinciarelli, A., Mortillaro, M., 2010. The Voice of Personality: Mapping Nonverbal Vocal Behavior into Trait Attributions. In: Proc. SSPW 2010. Firenze, Italy, pp. 17–20.
Mokhtari, A., Campbell, N., 2008. Speaking style variation and speaker personality. In: Proc. of Speech Prosody. Campinas, Brazil, pp. 601–604.
Mori, H., Satake, T., Nakamura, M., Kasuya, H., 2011. Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics. Speech Communication 53 (1), 36–50.
Mota, S., Picard, R. W., 2003. Automated Posture Analysis for Detecting Learner's Interest Level. In: Conference on Computer Vision and Pattern Recognition. Madison, Wisconsin, pp. 49–56.
Mower, E., Black, M., Flores, E., Williams, M., Narayanan, S., 2011. Design of an Emotionally Targeted Interactive Agent for Children with Autism. In: Proc. IEEE International Conference on Multimedia and Expo (ICME 2011). Barcelona, Spain, pp. 1–6.
Mower, E., Metallinou, A., Lee, C.-C., Kazemzadeh, A., Busso, C., Lee, S., Narayanan, S., 2009. Interpreting Ambiguous Emotional Expressions. In: Proc. ACII. Amsterdam, pp. 662–669.
Mporas, I., Ganchev, T., 2009. Estimation of unknown speakers' height from speech. International Journal of Speech Technology 12 (4), 149–160.
Mubarak, O. M., Ambikairajah, E., Epps, J., 2006. Novel features for effective speech and music discrimination. In: IEEE International Conference on Engineering of Intelligent Systems, ICEIS 2006. Islamabad, pp. 1–5.
Müller, C., 2005. Zweistufige kontextsensitive Sprecherklassifikation am Beispiel von Alter und Geschlecht [Two-layered Context-Sensitive Speaker Classification on the Example of Age and Gender]. Ph.D. thesis, Computer Science Institute, University of the Saarland, Germany.
Müller, C., September 2006. Automatic recognition of speakers' age and gender on the basis of empirical studies. In: Interspeech. Pittsburgh, Pennsylvania, pp. 1–4.
Müller, C., 2007. Classifying speakers according to age and gender. In: Müller, C. (Ed.), Speaker Classification II. Vol. 4343 of Lecture Notes in Computer Science / Artificial Intelligence. Springer, Heidelberg - New York - Berlin.
Müller, C., Burkhardt, F., August 2007. Combining short-term cepstral and long-term prosodic features for automatic recognition of speaker age. In: Interspeech. Antwerp, Belgium, pp. 1–4.
Müller, C., Wittig, F., Baus, J., 2003. Exploiting Speech for Recognizing Elderly Users to Respond to their Special Needs. In: Proceedings of the Eighth European Conference on Speech Communication and Technology (Eurospeech 2003). Geneva, Switzerland, pp. 1305 – 1308.
Mysak, E. D., 1959. Pitch duration characteristics of older males. Journal of Speech and Hearing Research 2, 46–54.
Nadeu, M., Prieto, P., 2011. Pitch range, gestural information, and perceived politeness in catalan. Journal of Pragmatics 43 (3),

841–854.
Nguyen, P., Le, T., Tran, D., Huang, X., Sharma, D., 2010. Fuzzy support vector machines for age and gender classification. In: ICSLP (2010), pp. 2806–2809.
Ni, K., Carin, L., Dunson, D., 2007. Multi-task learning for sequential data via ihmms and the nested dirichlet process. In: Proc. of the 24th International Conference on Machine Learning (ICML). Corvalis, Oregon, pp. 689–696.
Ní Chasaide, A., Gobl, C., 2004. Voice Quality and f_0 in Prosody: Towards a Holistic Account. In: Proc. of Speech Prosody 2004. Nara, Japan, pp. 4, no pagination.
Nose, T., Kato, Y., Kobayashi, T., 2007. Style estimation of speech based on multiple regression hidden semi-markov model. In: Proc. Interspeech. Antwerp, pp. 2285–2288.
Oberlander, J., Nowson, S., 2006. Whose thumb is it anyway? Classifying author personality from weblog text. In: Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions. Sydney, pp. 627–634.
Obozinski, G., Taskar, B., 2006. Multi-task feature selection. In: Workshop of structural Knowledge Transfer for Machine Learning in the 23rd International Conference on Machine Learning (ICML). Pittsburgh, Pennsylvania, pp. 1–15.
Oller, D. K., Niyogic, P., Grayd, S., Richards, J. A., Gilkerson, J., Xu, D., Yapanel, U., Warrene, S. F., 2010. Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development. Proceedings of the National Academy of Sciences of the United States of America (PNAS) 107.
Omar, M. K., Pelecanos, J., 2010. A novel approach to detecting non-native speakers and their native language. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings. Dallas, Texas, pp. 4398–4401.
Pachet, F., Roy, P., 2009. Analytical features: A knowledge-based approach to audio feature generation. EURASIP Journal on Audio, Speech, and Music Processing, 23 pages.
Pal, P., Iyer, A., Yantorno, R., 2006. Emotion detection from infant facial expressions and cries. In: Proc. ICASSP. Toulouse, pp. 809–812.
Pao, T., Yeh, J., Tsai, Y., 2010. Recognition and analysis of emotion transition in mandarin speech signal. In: Proc. IEEE International Conference on Systems, Man and Cybernetics. Istanbul, Turkey, pp. 3326–3332.
Patil, S. A., Hansen, J. H. L., 2010. The physiological microphone (pmic): A competitive alternative for speaker assessment in stress detection and speaker verification. Speech Communication 52 (4), 327–340.
Pentland, A., Madan, A., October 2005. Perception of social interest. In: Proc. IEEE Int. Conf. on Computer Vision, Workshop on Modeling People and Human Interaction (ICCV-PHI). Beijing, no pagination.
Perera, D., Eales, R. T. J., Blashki, K., 2009. Supporting the creative drive: Investigating paralinguistic voice as a mode of interaction for artists with upper limb disabilities. Universal Access in the Information Society 8 (2), 77–88.
Pfister, T., Robinson, P., 2010. Speech emotion classification and public speaking skill assessment. In: Proc. International Workshop on Human Behaviour Understanding. Istanbul, Turkey, pp. 151–162.
Picard, R., 2003. Affective Computing: Challenges. Journal of Human-Computer Studies 59, 55–64.
Pike, K. L., 1945. The Intonation of American English. University of Michigan Press, Ann Arbor.
Polzehl, T., Mller, S., Metze, F., 2010. Automatically assessing personality from speech. In: Proceedings - 2010 IEEE 4th International Conference on Semantic Computing, ICSC 2010. Pittsburgh, PA, pp. 134–140.
Pon-Barry, H., 2008. Prosodic manifestations of confidence and uncertainty in spoken language. In: INTERSPEECH 2008 - 9th Annual Conference of the International Speech Communication Association. Brisbane, Australia, pp. 74–77.
Porat, R., Lange, D., Zigel, Y., 2010. Age recognition based on speech signals using weights supervector. In: ICSLP (2010), pp. 2814–2817.
Price, L., Richardson, J. T. E., Jelfs, A., 2007. Face-to-face versus online tutoring support in distance education. Studies in Higher Education 32 (1), 1–20.
Provine, R., 1993. Laughter punctuates speech: linguistic, social and gender contexts of laughter. Ethology 15, 291–298.
Pudil, P., Novovicova, J., Kittler, J., 1994. Floating search methods in feature selection. Pattern Recognition Letters 15, 1119–1125.
Qvarfordt, P., Beymer, D., Zhai, S. X., 2005. Realtourist - a study of augmenting human-human and human-computer dialogue with eye-gaze overlay. In: INTERACT 2005. Vol. LNCS 3585. pp. 767–780.
Rankin, K. P., Salazar, A., Gorno-Tempini, M. L., Sollberger, M., Wilson, S. M., Pavlic, D., Stanley, C. M., Glenn, S., Weiner, M. W., Miller, B. L., 2009. Detecting sarcasm from paralinguistic cues: Anatomic and cognitive correlates in neurodegenerative disease. NeuroImage 47 (4), 2005–2015.
Reilly, N., Cannizzaro, M. S., Harel, B. T., Snyder, P. J., 2004. Feigned depression and feigned sleepiness: A voice acoustical analysis. Brain and cognition 55 (2), 383–386.
Reisenzein, R., Weber, H., 2009. Personality and Emotion. In: Corr, P. J., Matthews, G. (Eds.), The Cambridge Handbook of Personality Psychology. Cambridge University Press, Cambridge, pp. 54–71.
Rektorova, I., Barrett, J., Mikl, M., Rektor, I., Paus, T., 2007. Functional abnormalities in the primary orofacial sensorimotor cortex during speech in parkinson's disease. Movement Disorders 22 (14), 2043–2051.
Revelle, W., Scherer, K., 2009. Personality and Emotion. In: Oxford Companion to the Affective Sciences. Oxford University Press, Oxford, pp. 1–4.
Ringeval, F., Chetouani, M., 2008. A vowel based approach for acted emotion recognition. In: INTERSPEECH 2008 - 9th Annual Conference of the International Speech Communication Association. Brisbane, Australia, pp. 2763–2766.
Romanyshyn, N., 2009. Paralinguistic maintenance of verbal communicative interaction in literary discourse (on the material of w. s. maugham's novel "theatre"). In: Experience of Designing and Application of CAD Systems in Microelectronics - Proceedings of the 10th International Conference, CADSM 2009. Polyana-Svalyava, Ukraine, pp. 550–552.
Ronzhin, A. L., 2005. Estimating psycho-physiological state of a human by speech analysis. In: Proceedings of SPIE - The International Society for Optical Engineering. Vol. 5797. pp. 170–181.

Rosch, E., 1975. Cognitive Representations of Semantic Categories. Journal of Experimental Psychology: General 104 (3), 192–233.
Rosenberg, A., Hirschberg, J., 2005. Acoustic/Prosodic and Lexical Correlates of Charismatic Speech. In: Proc. of Interspeech. Lisbon, pp. 513–516.
Rosenberg, A., Hirschberg, J., 2009. Charisma perception from text and speech. Speech Communication 51 (7), 640–655.
Roy, D. M., Kaelbling, L. P., 2007. Efficient bayesian task-level transfer learning. In: Proc. of the 20th international joint conference on Artifical intelligence. Hyderabad, India, pp. 2599–2604.
Russell, J., Bachorowski, J., Fernandez-Dols, J., 2003. Facial and vocal expressions of emotion. Annual Review of Psychology, 329–349.
Sapir, S., Ramig, L. O., Spielman, J. L., Fox, C., 2009. Formant Centralization Ratio: A Proposal for a New Acoustic Measure of Dysarthric Speech. Journal of Speech, Language, and Hearing Research 53.
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.
Scherer, K. R., 1981. Speech and emotional states. In: Darby, J. (Ed.), Speech evaluation in psychiatry. Grune & Stratton, New York, pp. 115–135.
Scherer, K. R., 2003. Vocal communication of emotion: A review of research paradigms. Speech Communication 40, 227–256.
Schiel, F., Heinrich, C., 2009. Laying the foundation for in-car alcohol detection by speech. In: Proc. INTERSPEECH 2009. Brighton, UK, pp. 983–986.
Schiel, F., Heinrich, C., Barfuesser, S., 2011. Alcohol language corpus – the first public corpus of alcoholized german speech. Language Resources and EvaluationDOI: 10.1007/s10579-011-9139-y.
Schmidt, M. N., Olsson, R. K., 2006. Single-channel speech separation using sparse non-negative matrix factorization. In: Proc. of Interspeech. Pittsburgh, Pennsylvania, pp. 2–5.
Schoentgen, J., 2006. Vocal cues of disordered voices: An overview. Acta Acustica united with Acustica 92 (5), 667–680.
Schötz, S., 2006. Perception, Analysis and Synthesis of Speaker Age. Ph.D. thesis, University of Lund, Sweden.
Schötz, S., 2007. Acoustic Analysis of Adult Speaker Age. In: Müller, C. (Ed.), Speaker Classification. Vol. 4343 of Lecture Notes in Computer Science / Artificial Intelligence. Springer, Heidelberg - Berlin - New York, pp. 88–107, this issue.
Schröder, M., Cowie, R., Heylen, D., Pantic, M., Pelachaud, C., Schuller, B., 2008. Towards responsive sensitive artificial listeners. In: Proc. 4th Intern. Workshop on Human-Computer Conversation. Bellagio, p. no pagination.
Schröder, M., Devillers, L., Karpouzis, K., Martin, J.-C., Pelachaud, C., Peter, C., Pirker, H., Schuller, B., Tao, J., Wilson, I., 2007. What Should a Generic Emotion Markup Language Be Able to Represent? In: Paiva, A., Prada, R., Picard, R. W. (Eds.), Affective Computing and Intelligent Interaction. Springer, Berlin-Heidelberg, pp. 440–451.
Schuller, B., 2011. Affective speaker state analysis in the presence of reverberation. International Journal of Speech Technology 14 (2), 77–87.
Schuller, B., Batliner, A., Steidl, S., Schiel, F., Krajewski, J., 2011a. The INTERSPEECH 2011 Speaker State Challenge. In: Proc. of Interspeech. Florence, Italy, pp. 3201–3204.
Schuller, B., Batliner, A., Steidl, S., Seppi, D., 2011b. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech Communication 53, Special Issue on Sensing Emotion and Affect - Facing Realism in Speech Processing (9/10), 1062–1087.
Schuller, B., Burkhardt, F., 2010. Learning with Synthesized Speech for Automatic Emotion Recognition. In: Proc. 35th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Dallas, pp. 5150–5153.
Schuller, B., Eyben, F., Can, S., Feussner, H., 2010a. Speech in Minimal Invasive Surgery - Towards an Affective Language Resource of Real-life Medical Operations. In: Proc. 3rd ELRA International Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect. Valetta, pp. 5–9.
Schuller, B., Eyben, F., Rigoll, G., 2008a. Static and dynamic modelling for the recognition of non-verbal vocalisations in conversational speech. In: André, E. (Ed.), Proceedings of the 4th IEEE Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-based Systems (PIT 2008), Kloster Irsee, Germany. Vol. LNCS 5078. Springer, pp. 99–110.
Schuller, B., Jiménez Villar, R., Rigoll, G., Lang, M., 2005. Meta-classifiers in acoustic and linguistic feature fusion-based affect recognition. In: Proc. ICASSP. Philadelphia, pp. I:325–328.
Schuller, B., Köhler, N., Müller, R., Rigoll, G., 2006a. Recognition of Interest in Human Conversational Speech. In: Proc. Interspeech. Pittsburgh, pp. 793–796.
Schuller, B., Müller, R., Eyben, F., Gast, J., Hörnler, B., Wöllmer, M., Rigoll, G., Höthker, A., Konosu, H., 2009a. Being Bored? Recognising Natural Interest by Extensive Audiovisual Integration for Real-Life Application. Image and Vision Computing Journal, Special Issue on Visual and Multimodal Analysis of Human Spontaneous Behavior 27, 1760–1774.
Schuller, B., Müller, R., Hörnler, B., Höthker, A., Konosu, H., Rigoll, G., 2007a. Audiovisual recognition of spontaneous interest within conversations. In: Proc. 9th Int. Conf. on Multimodal Interfaces (ICMI), Special Session on Multimodal Analysis of Human Spontaneous Behaviour. ACM SIGCHI, Nagoya, Japan, pp. 30–37.
Schuller, B., Reiter, S., Rigoll, G., 2006b. Evolutionary feature generation in speech emotion recognition. In: Proc. Int. Conf. on Multimedia and Expo ICME 2006. Toronto, Canada, pp. 5–8.
Schuller, B., Rigoll, G., 2009. Recognising Interest in Conversational Speech – Comparing Bag of Frames and Supra-segmental Features. In: Proc. Interspeech. Brighton, pp. 1999–2002.
Schuller, B., Rigoll, G., Lang, M., 2004. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In: Proc. ICASSP. Montreal, pp. 577–580.
Schuller, B., Seppi, D., Batliner, A., Meier, A., Steidl, S., 2007b. Towards more Reality in the Recognition of Emotional Speech. In: Proc. ICASSP. Honolulu, pp. 941–944.
Schuller, B., Steidl, S., Batliner, A., 2009b. The INTERSPEECH 2009 Emotion Challenge. In: Interspeech 2009 – Eurospeech,

11th European Conference on Speech Communication and Technology, September 6-10, 2009, Brighton, UK, Proceedings. pp. 312–315.
Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., Narayanan, S., 2010b. The INTERSPEECH 2010 Paralinguistic Challenge – Age, Gender, and Affect. In: ICSLP (2010), pp. 2794–2797.
Schuller, B., Steidl, S., Batliner, A., Schiel, F., Krajewski, J., 2011c. The Interspeech 2011 Speaker State Challenge. In: Proc. Interspeech. Florence, Italy, pp. 3201–3204.
Schuller, B., Vlasenko, B., Eyben, F., Wöllmer, M., Stuhlsatz, A., Wendemuth, A., Rigoll, G., 2010c. Cross-corpus acoustic emotion recognition: Variances and strategies. IEEE Transactions on Affective Computing 1 (2), 119–131.
Schuller, B., Weninger, F., 2010. Discrimination of speech and non-linguistic vocalizations by non-negative matrix factorization. In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP). Dallas, TX, USA, pp. 5054–5057.
Schuller, B., Wimmer, M., Arsic, D., Moosmayr, T., Rigoll, G., 2008b. Detection of security related affect and behaviour in passenger transport. In: Proc. Interspeech. Brisbane, pp. 265–268.
Schuller, B., Wimmer, M., Mösenlechner, L., Kern, C., Arsic, D., Rigoll, G., 2008c. Brute-Forcing Hierarchical Functionals for Paralinguistics: a Waste of Feature Space? In: Proc. ICASSP. Las Vegas, pp. 4501–4504.
Schuller, B., Wöllmer, M., Eyben, F., Rigoll, G., 2011d. Retrieval of Paralinguistic Information in Broadcasts. In: Maybury, M. (Ed.), Multimedia Information Extraction: Advances in video, audio, and imagery extraction for search, data mining, surveillance, and authoring. IEEE Computer Society Press, Ch. 14, pp. 1–21.
Schuller, B., Wöllmer, M., Eyben, F., Rigoll, G., Arsic, D., 2011e. Semantic Speech Tagging: Towards Combined Analysis of Speaker Traits. In: Proc. AES 42nd International Conference. Ilmenau, Germany, pp. 89–97.
Schuller, B., Zhang, Z., Weninger, F., Rigoll, G., 2011f. Selecting training data for cross-corpus speech emotion recognition: Prototypicality vs. generalization. In: Proc. 2011 Afeka-AVIOS Speech Processing Conference. Tel Aviv, Israel, no pagination.
Schwenker, F., Scherer, S., Schmidt, M., Schels, M., Glodek, M., 2010. Multiple classifier systems for the recogonition of human emotions. Vol. 5997 LNCS of Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer.
Shafran, L., Riley, M., Mohri, M., 2003. Voice Signatures. In: Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2003). Virgin Islands, USA, pp. 31–36.
Shahin, I., 2009. Verifying speakers in emotional environments. In: IEEE International Symposium on Signal Processing and Information Technology, ISSPIT 2009. Ajman, UAE, pp. 328–333.
Shami, M., Verhelst, W., 2007. Automatic classification of expressiveness in speech: A multi-corpus study. In: Mueller, C. (Ed.), Speaker Classification II: Selected Projects. Springer-Verlag, Berlin, pp. 43–56.
Shriberg, E., 2005. Spontaneous speech: How peoply really talk and why engineers should care. In: Proc. EUROSPEECH 2005. Lisbon, Portugal, pp. 1781–1784.
Stadermann, J., Koska, W., Rigoll, G., 2005. Multi-task learning strategies for a recurrent neural net in a hybrid tied-posteriors acoustic mode. In: Proc. of Interspeech 2005. ISCA, Lisbon, Portugal, pp. 2993–2996.
Steidl, S., Batliner, A., Seppi, D., Schuller, B., 2010. On the Impact of Children's Emotional Speech on Acoustic and Language Models. EURASIP Journal on Audio, Speech, and Music Processing 2010, 1–14.
Steidl, S., Schuller, B., Batliner, A., Seppi, D., 2009. The Hinterland of Emotions: Facing the Open-Microphone Challenge. In: Proc. ACII. Amsterdam, pp. 690–697.
Stevens, K. N., 1998. Acoustic Phonetics. The MIT Press, Cambridge, MA.
Stiefelhagen, R., Yang, J., Waibel, A., July 2002. Modeling focus of attention for meeting indexing based on multiple cues. IEEE Transactions on Neural Networks 13 (4), 928 – 938.
Stuhlsatz, A., Meyer, C., Eyben, F., Zielke, T., Meier, G., Schuller, B., 2011. Deep Neural Networks for Acoustic Emotion Recognition: Raising the Benchmarks. In: Proc. ICASSP. Prague, Czech Republic, pp. 5688–5691.
Suzuki, T., Koshizen, T., Aihara, K., Tsujino, H., 2005. Learning to estimate user interest utilising the variational bayes' estimator. In: Proc. of 5th of International Conference on Intelligtent Systems Design and Applications. Wro-claw, Poland, pp. 94–99.
Tabatabaei, T. S., Krishnan, S., 2010. Towards robust speech-based emotion recognition. In: Proc. IEEE International Conference on Systems, Man and Cybernetics. Istanbul, Turkey, pp. 608–611.
Tepperman, J., Traum, D., Narayanan, S., 2006. "Yeah Right": Sarcasm Recognition for Spoken Dialogue Systems. In: Proc. of Interspeech. Pittsburgh, Pennsylvania, pp. 1838–1841.
Thrun, S., Mitchell, T. M., 1995. Learning one more thing. In: Proc. of the International Joint Conference on Artificial Intelligence (IJCAI). Montreal, Canada, pp. 1217–1223.
Trager, G. L., 1958. Paralanguage: A First Approximation. Studies in Linguistics 13, 1–12.
van Dommelen, W. A., Moxness, B. H., 1995. Acoustic parameters in speaker height and weight identification: Sex-specific behaviour. Language and Speech 38 (3), 267–287.
Ververidis, D., Kotropoulos, C., 2004. Automatic speech classification to five emotional states based on gender information. In: Proc. of 12th European Signal Processing Conference. Vienna, Austria, pp. 341–344.
Ververidis, D., Kotropoulos, C., 2006. Fast Sequential Floating Forward Selection applied to emotional speech features estimated on DES and SUSAS data collection. In: Proc. of European Signal Processing Conf. (EUSIPCO 2006). Florence, p. no pagination.
Vinciarelli, A., Pantic, M., Bourlard, H., 2009. Social signal processing: Survey of an emerging domain. Image and Vision Computing 27, 1743–1759.
Vlasenko, B., Schuller, B., Mengistu, T. K., Rigoll, G., A., W., 2008. Balancing spoken content adaptation and unit length in the recognition of emotion and interest. In: Proc. Interspeech. Brisbane, Australia, pp. 805–808.
Vogt, T., Andre, E., 2006. Improving automatic emotion recognition from speech via gender differentiation. In: Proc. of Language

Resources and Evaluation Conference (LREC 2006). Genoa, Italy, pp. 1–4.
Weiss, B., Burkhardt, F., 2010. Voice attributes affecting likability perception. In: Proc. INTERSPEECH. Makuhari, Japan, pp. 2014–2017.
Weninger, F., Schuller, B., Batliner, A., Steidl, S., Seppi, D., 2011. Recognition of nonprototypical emotions in reverberated and noisy speech by nonnegative matrix factorization. Eurasip Journal on Advances in Signal Processing 2011 (Article ID 838790), 16 pages.
Wilden, I., Herzel, H., Peters, G., Tembrock, G., 1998. Subharmonics, biphonation, and deterministic chaos in mammal vocalization. Bioacoustics 9, 171–196.
Witten, I. H., Frank, E., 2005. Data mining: Practical machine learning tools and techniques, 2nd Edition. Morgan Kaufmann, San Francisco.
Wöllmer, M., Schuller, B., Batliner, A., Steidl, S., Seppi, D., 2011. Tandem decoding of children's speech for keyword detection in a child-robot interaction scenario. ACM Transactions on Speech and Language Processing 7, Special Issue on Speech and Language Processing of Children's Speech for Child-machine Interaction Applications (4), article 12, 22 pages.
Wöllmer, M., Schuller, B., Eyben, F., Rigoll, G., 2010. Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening. IEEE Journal of Selected Topics in Signal Processing 4 (5), 867–881.
Womack, B. D., Hansen, J. H. L., 1999. N-channel hidden markov models for combined stressed speech classification and recognition. IEEE Transactions on Speech and Audio Processing 7 (6), 668–677.
Wu, S., Falk, T. H., Chan, W. ., 2011. Automatic speech emotion recognition using modulation spectral features. Speech Communication 53 (5), 768–785.
Wyatt, D., Choudhury, T., Kautz, H., 2007. Capturing spontaneous conversation and social dynamics: A privacy-sensitive data collection effort. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings. Vol. 4. pp. IV213–IV216.
Yamada, M., Sugiyama, M., Matsui, T., 2010. Semi-supervised speaker identification under covariate shift. Signal Processing 90 (8), 2353–2361.
Yildirim, S., Lee, C., Lee, S., Potamianos, A., Narayanan, S., 2005. Detecting Politeness and Frustration State of a Child in a Conversational Computer Game. In: Proc. of Interspeech 2005. ISCA, Lisbon, Portugal, pp. 2209–2212.
Yildirim, S., Narayanan, S., Potamianos, A., 2011. Detecting Emotional State of a Child in a Conversational Computer Game. Computer Speech and Language 25, 29–44.
Yoon, W. ., Park, K. ., 2011. Building robust emotion recognition system on heterogeneous speech databases. In: Digest of Technical Papers - IEEE International Conference on Consumer Electronics. pp. 825–826.
Young, S., Evermann, G., Gales, M., Hain, T., D.Kershaw, Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P., 2006. The HTK Book, for htk version 3.4 Edition. Cambridge University Engineering Department.
Zhang, C., Hansen, J. H. L., 2007. Analysis and classification of speech mode: Whispered through shouted. In: International Speech Communication Association - 8th Annual Conference of the International Speech Communication Association, Interspeech 2007. Vol. 4. pp. 2396–2399.

Vitae

Björn Schuller received his diploma in 1999 and his doctoral degree in 2006, both in electrical engineering and information technology from TUM in Munich/Germany where he is tenured as Senior Researcher and Lecturer in Pattern Recognition and Speech Processing. From 2009 to 2010 he was with the CNRS-LIMSI in Orsay/France and a visiting scientist in the Imperial College London's Department of Computing in London/UK. Dr. Schuller is a member of the ACM, HUMAINE Association, IEEE and ISCA and (co-)authored more than 250 peer reviewed publications leading to more than 2 300 citations – his current H-index equals 25.


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 Berkley/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 30 and more than 3 000 citations. His research interests are all aspects of prosody and paralinguistics in speech processing. Dr. Batliner repeatedly served as Workshop/Session (co)-organiser and is Associated Editor for the IEEE Transactions on Affective Computing.


Felix Burkhardt has a longstanding background in language technology. Originally an expert of Speech Synthesis at the Technical University of Berlin, he has been working for the Deutsche Telekom AG since 2000. He does tutoring, consulting, research, and development in the working fields VoiceXML based Voice-Portal architectures, Text-to-Speech synthesis, speaker classification, ontology based language modelling, and emotional human-machine interfaces. He has been a member of the European Network of Excellence HUMAINE on emotion-oriented computing and the W3C Emotion Markup Language Incubator Group.


Laurence Devillers is Associate Professor since 1995 at the Computer Science Division of the University Paris-XI/France, member of the LIMSI-CNRS. Her research interests include analysis of emotional

behaviours and corpora of Human-Human dialogues. She passed her habilitation in 2006. At present, she is the head of the 'Speech and Emotion' topic of research created in 2004 within the Spoken Language Processing group at LIMSI-CNRS. She (co-)authored more than 100 publications - her current H-index equals 18 -, co-organised international workshops, co-edited a special issue of Computer Speech and Language on affective speech, and is member of the HUMAINE association, IEEE, and ISCA.

Christian Müller is Senior Researcher at the German Research Center for Artificial Intelligence. His research interest is user-adaptive multimodal Human-Machine Interfaces with a special focus on speaker classification. Application areas of his research are primarily automotive interfaces and telecommunication applications. From 2006 to 2008, he has been visiting researcher at the International Computer Science Institute (ICSI) in Berkeley/USA. Dr. Müller earned a Ph.D. in Computer Science at Saarland University/Germany in January 2006 dealing with Speaker Classification. He edited a book on Speaker Classification that was published in 2007 by Springer in the LNCS/LNAI series and constitutes a state-of-the-art survey of the field.

Shrikanth Narayanan received his M.S., Engineer, and Ph.D. in electrical engineering from UCLA in 1990, 1992, 1995. 1995-2000 he was with AT&T Labs-Research, Florham Park. Currently, he is Professor at USC and holds joint appointments as Professor in Computer Science, Linguistics, and Psychology. He is Editor for Computer, Speech and Language, IEEE Transactions on Multimedia, Journal of Acoustical Society of America, and previously of the IEEE Transactions of Speech and Audio Processing, and IEEE Signal Processing Magazine. He is Fellow of the Acoustical Society of America, the IEEE, recipient of manifold awards, published over 350 papers, and has 7/10 granted/pending U.S. patents.