Ten Recent Trends in Computational Paralinguistics

Björn Schuller and Felix Weninger

Institute for Human-Machine Communication, Technische Universität München,
Arcisstr. 21, 80333 München, Germany
{schuller,weninger}@tum.de

Abstract. The field of computational paralinguistics is currently emerging from loosely connected research on speaker states, traits, and vocal behaviour. Starting from a broad perspective on the state-of-the-art in this field, we combine these facts with a bit of 'tea leaf reading' to identify ten currently dominant trends that might also characterise the next decade of research: taking into account more tasks and task interdependencies, modelling paralinguistic information in the continuous domain, agglomerating and evaluating on large amounts of heterogeneous data, exploiting more and more types of features, fusing linguistic and non-linguistic phenomena, devoting more effort to optimisation of the machine learning aspects, standardising the whole processing chain, addressing robustness and security of systems, proceeding to evaluation in real-life conditions, and finally overcoming cross-language and cross-cultural barriers. We expect that following these trends we will see an increase in the 'social competence' of tomorrow's speech and language processing systems.

1 Introduction

Social competence, i. e., the ability to permanently analyse and re-assess dialogue partners with respect to their traits (e. g., personality or age) and states (e. g., emotion or sleepiness), and to react accordingly (by adjusting the discourse strategy, or aligning to the dialogue partner) remains one key feature of human communication that is not found in most of today's technical systems. By simulating such capabilities through signal processing and machine learning techniques, the emerging field of computational paralinguistics aims to increase the perceived social competence of technical systems for human-machine communication. One main application is to increase efficiency and hence, user satisfaction in task oriented dialogue systems by enabling naturalistic interaction. Furthermore, recognition of paralinguistic information in human signals can be used for multimedia retrieval (enabling queries by certain speaker traits), in surveillance applications (e. g., to monitor customer satisfaction or potential attackers), for efficient audio or video coding and speech-to-speech translation (e. g., resolving semantic ambiguities by recognising

intention, or synthesising translated speech with the original speaker's affect) and finally entertainment (e. g., to render states and traits in the voice of an avatar in accordance to the player).

As can be seen from these applications, computational paralinguistics comprise a variety of tasks [84]. A taxonomy can be established along the time axis, distinguishing long term *traits* from medium-term phenomena and short-term *states*. Long term traits include biological primitives such as height, weight, age, gender or race [55, 77, 81]. Interestingly, humans seem to exploit acoustic correlates of these primitives in their reasoning. For instance, age, height, and weight of speakers could be assigned by listeners to voices in repeated studies [22, 42]; acoustic correlates of body shape, size and weight include fundamental frequencies and other formant parameters [28, 35]. Other trait concepts such as group membership, ethnicity and culture overlap with linguistic phenomena such as dialect or nativeness [57]. For instance, the output of a speech recognition system can be used for classification of demographic traits including education level, ethnicity, and geographic region [33]. Besides, analysis of personality is an increasingly popular area of research [34, 53, 59] comprising acoustic and linguistic phenomena [65]. Medium term speaker attributes refer to temporary conditions, including sleepiness [41], (alcohol) intoxication [45, 68, 78], health [50] or depression [25], but also group roles [43], friendship and identity [38]. Finally, important short term states from an application point of view include voice quality, speaking style, and affect. In typical applications, one will rarely encounter full-blown, prototypical emotions such as sadness or disgust, but rather affect-related states including interest [98], uncertainty [47], frustration [2], stress level [37] or pain [5].

In summary, we hope that this unified view on the aspects of computational paralinguistics in speech may help to bridge the gap between some of the loosely connected fields in speech processing, including speech and speaker recognition, and the emerging domain of speaker classification. Further, this view enables us to outline *ten trends* that might characterise the field of computational paralinguistics in the following years. These trends are partially motivated by technological development—first and foremost, drastic decreases in the cost of computing power and storage space, the latter enabling access to virtually infinite amounts of speech data—but also conceptual advances in machine learning and signal processing. Altogether, we believe, these will allow technologies for computational paralinguistics to penetrate into daily life, which poses, in turn, several 'grand challenges' connected to real-life applications as opposed to 'in-the-lab' usage. While we put a strong focus on speech *analysis* in this chapter, many of the trends might be relevant for speech *synthesis* as well.

2 Ten Recent and Future Trends

2.1 More Tasks and Coupling of Tasks

Relevant tasks in computational paralinguistics are manifold, and we have mentioned a non-exhaustive list of relevant tasks above. Still, the lion's share of research is devoted to emotion and emotion-related states, followed by physical

traits (age, height) and personality[1]. It can be conjectured that addressing additional tasks will largely depend on the availability of annotated data. However, it could turn out that taking into account more and more seemingly novel tasks would be reinventing the wheel: A number of inderdependencies is already visible in the above list of paralinguistic states and traits. Following the taxonomy along the time axis, many dependencies on long term traits can be found. Long term traits themselves are coupled to some degree, e. g., height with age, gender and race. Medium term phenomena can depend on long term traits as well, e. g., health state can deteriorate with age, and group roles arguably depend on personality traits such as leadership emergence. Finally, also short term states are dependent on long term traits: The manifestation of emotion is dependent on personality [62, 63]; in [54], it was revealed that human listeners consistently associate different tones of voice with certain speaker personalities. Furthermore, gender-dependencies of non-linguistic vocalisations have been repeatedly reported, e. g., in [60] for laughter.

Indeed, it has been repeatedly confirmed that modelling 'contextual' knowledge from different paralinguistic tasks benefits the performance in practice. Such knowledge can be integrated by building age, gender or height dependent models for any of the other tasks. For example, several studies indicate that considering gender information enables higher accuracy of automatic speech emotion recognition [93, 95]; however, it is an open question whether this can be attributed to low-level acoustic differences in the pitch registers of male and female voices, or to higher-level differences in the expression of emotion. To exploit mutual information from the speaker identity, speaker adaptation or normalisation can be performed [9]. Finally, related state and trait information can be added as a feature: In [81] first beneficial effects are shown by providing knowledge on speaker dialect region, education level and race as ground truth along with acoustic features in the assessment of speaker traits including age, gender and height. Such addition of speaker traits as ground truth features can be relevant in practical situations, where for example height can be determined from camera recordings.

An alternative to such explicit modelling of dependencies using prior knowledge is to automatically learn them from training data. For example, the rather simple strategy of using pairs of age and gender classes as learning target instead of each attribute individually can already be advantageous, as has been proven in the first Paralinguistic Challenge [77]. In the future, enhanced modelling of multiple correlated target variables could be performed through multitask learning [15]. Here, a representation of the input features is shared among tasks, such as the internal activations in the hidden layer of a neural network. Recurrent neural networks, in particular, allow accessing past predictions for any of the variables for analysing the current time frame [89]—in some sense, this is similar to replacing the ground truth speaker features in the above setup by (time-varying) classifier predictions. In this context, one of the peculiarities of computational paralinguistics is found in the representation of task variables

[1] According to a Scopus search for the title ('speech' or 'speaker') AND ... in February 2011.

by various data types (continuous, ordinal, nominal), which additionally often differ by their time scale (e. g., gender is constant in a speech turn while emotion or speaking style may vary). Considering methods for multi-scale fusion, one could also exploit multi-task learning for integrating phoneme recognition with analysis of paralinguistic information, in order to increase robustness of conversational speech recognition. Coupling the speech recognition task with, for example, gender or dialect recognition could be beneficial since in [10] it was shown that both these traits affect speech rate, flapping and central vowels.

2.2 More Continuous Modelling

The classic approach to computational paralinguistics is classification into 2-n classes, e. g., the big 6 emotions, gender, or age groups [77]. However, this often corresponds to an artificial discretisation, implying loss of information. For instance, the ground truth is continuous in case of intoxication (blood or breath alcohol concentration) or physical speaker traits (age, weight and height). Concepts to measure emotion and personality are often based on continuous valued dimensions, of which the most common are the arousal-valence model [66], or the five-factor 'OCEAN' model of personality [20]. For some states, annotation is performed using ordinal scales, e. g., using the Karolinska Sleepiness Scale (KSS), resulting in a quasi-continuum when ratings from multiple annotators are fused, e. g., by averaging; emotion annotation is sometimes performed directly in continuous dimensions, e. g., by the Feeltrace toolkit [19]. Conversely, machine learning research provides a rich set of tools for predicting continuous quantities including (extensions of) logistic regression, support vector regression, (recurrent) neural networks or random forests (ensembles of regression trees) which can be applied to paralinguistic analysis [77, 98]. Evaluation procedures for regression are readily available as well, and include correlation (Pearson), rank-correlation (Spearman) and determination coefficients (R^2), mean absolute or (root) mean squared error. In addition to continuous valued annotation, short-term variations of speaker states can be captured by a representing them as a function of time. For example, the Feeltrace toolkit [19] allows annotating emotion with a 'sampling frequency' of 10 ms. On the recognition side, this allows for dynamic classification or regression techniques, and investigation of diverse units of analysis including syllables, words or turns [97].

2.3 More, Synthesised, Agglomerated, and Cross Data

While it is a common belief in pattern recognition that there is 'no data like more data', publicly available speech data with rich annotation of paralinguistic information are still sparse. In fact, there are increasingly more databases ready for experimentation; the crux is that these often come with different labelling schemes (discrete, continuous, dimensional, categorical) and, in the context of speaker states, different strategies for elicitation (acted, induced, natural). This makes data agglomeration and evaluation across multiple corpora less straightforward than for other tasks, such as automatic speech recognition. On the other

hand, multi-corpus and cross-corpus evaluation, such as done in [56] for age and gender and recently in [79,91] for emotion, is crucial to assess generalisation of the models. In fact, experiments in cross-corpus emotion recognition suggest some overfitting to single corpora [79] which can only partly be alleviated by corpus or speaker normalisation. To make things worse, common techniques to reduce overfitting such as feature selection may exhibit low cross-data generalisation themselves [29]. Hence, acquiring more data for building robust and generalising emotion models can be seen as one of the great challenges for the future. Recent results show that combining different databases in a unified labelling scheme through data agglomeration or voting significantly improves performance [85]. Still, such unification of the labelling schemes introduces 'information loss'; late fusion techniques for multiple classifiers trained on single corpora using distinct labelling schemes could be an interesting direction for the future. In addition, the efficacy of semi-supervised learning[2] to leverage unlabelled speech data for emotion recognition has been repeatedly demonstrated [39, 48, 100, 103]; yet, large-scale studies across multiple speaker states and traits, and using large amounts of data acquired from the web, are still to follow. Finally, a promising technique is synthesis of training data: In fact, it has been shown that generalisation properties of emotion models in a cross-corpus setting can be improved through joint training with both human and synthetic speech [72]. This result is very promising since synthetic speech can be easily produced in large quantities, and a variety of combinations of speaker states and traits can be simulated. It is hoped that this will yield good generalisation of models and facilitate learning of multiple tasks and their interdependencies (cf. above).

2.4 More and Novel Features

The features used in early research on speaker states and traits were motivated by the adjacent fields of automatic speech and speaker recognition. Thus, usage of spectral or cepstral features (Mel frequency cepstral coefficients, MFCCs) prevailed. In the meantime, a plethora of novel, mostly expert-crafted acoustic features, including perceptually motivated ones [49, 78, 99] or such that base on pre-classification [64] have been proposed and evaluated for paralinguistic analysis, along with the addition of more or less brute forced linguistic features (e. g., Bag of Words or Bag of N-grams). Furthermore, it has repeatedly been shown that enlarging the feature space can help boost accuracy [78, 80]. An alternative direction is supervised generation of features through evolutionary algorithms [74] or unsupervised learning of features, e. g., through deep belief networks or sparse coding. Still, the challenge is less the efficient computation, or combination of features in more or less brute force approaches, but to systematically investigate the relations between different types of features, especially in

[2] In the context of automatic speech recognition, this is often referred to as *unsupervised* learning—we prefer the more common term *semi-supervised* to highlight the difference to purely unsupervised techniques such as clustering or latent semantic analysis.

terms of generalisation to cross-corpus or cross-task analyses: After all, it is not clear whether novel features indeed add new information, or observed increases in performance stem from (over-)fitting to specific data sets, acoustic conditions, speakers or content (such as in fixed language speaker state corpora).

2.5 More (Coupling of) Linguistics and Non-linguistics

Transmitting information through non-verbal channels is a crucial part of human-human communication. Besides the low-level acoustic manifestations of speaker states and traits, such non-verbal channels also include the use of non-linguistic vocalisations. Recently, there is renewed interest in the use of such vocalisations in computer-mediated human-human and human-machine communication [12, 83]. Just as human communication uses both non-verbal and verbal expression, the ultimate solution will, of course, not be to define new research domains dealing only with non-verbal phenomena (Social Signal Processing [94]), or to differentiate between non-linguistic vocalisations alone (such as in [13]), but to attain joint access to the linguistic / non-linguistic channels by machines. On the analysis side, there are already a couple of studies on fusion of linguistic with non-linguistic information. The simplest, yet effective and efficient strategy is to integrate non-linguistic vocalisations as word-like entities into the linguistic string [83, 98]; in contrast, a late fusion approach has been investigated in [32].

2.6 More Optimisation

With the increased maturity of computational paralinguistics, and an established basic methodology, more and more efforts are devoted to optimisation of the whole processing chain. First, the systematic optimisation of machine learning algorithms including feature selection, reduction and classification is facilitated through the increasing availability of public corpora with well-defined partitioning into training and test sets, such as the ones used for the first paralinguistic challenges [76–78]. More precisely, such optimisation steps can involve 'global' as well as 'local' feature selection for sub-sets of classes in hierarchical classification [44] or for different sub-units of speech [7]. Additionally, more and more optimisations are applied in classifier training, including balancing of skewed class distributions (e. g., by synthetic minority oversampling [16, 76] or similar techniques), or instance selection, i. e., pruning of noisy training data or outliers [26, 82]. The importance of selecting appropriate classifier parameters is well known in machine learning and consequently also for paralinguistic information retrieval, as reported, e. g., in [78]. Besides, there is an increasing trend towards fusion of multiple systems, as has been evident in the sequence of paralinguistic challenges [76–78]. Fusion can be applied to classifier decisions in hierarchical [44, 101], hybrid [75] or ensemble architectures [73, 86]; at an even higher level, fusing the output of entire recognition systems can successfully exploit their complementarity; for instance, majority voting among the systems from the best participants in the Interspeech 2009 Emotion Challenge yields the best result reported so far on the challenge corpus [71].

Apart from such general machine learning techniques, speech analysis provides specific starting points for optimisation, including speech clustering by emotional state for speaker identification [23, 46] and speaker adaptation / normalisation, which is nowadays observed particularly for speaker state analysis [9]. Finally, also the process of capturing speech signals itself can be optimised, e. g., by using silent speech interfaces for stress detection and speaker verification [58].

2.7 More Standardisation

Arguably, the more mature and closer to real-life application the field of computational paralinguistic gets, the greater is the need for standardisation. Similarly as in the argument made in the previous section, standardisation efforts can be categorised along the signal processing chain. They include documentation and well-motivated grouping of features such as the CEICES Feature Coding Scheme [4], standardised feature sets as provided by the openSMILE [31] and openEAR [30] toolkits in this field, and machine learning frameworks such as the Weka environment [36]. Such Standardised feature extraction / classification allows to evaluate the feature extraction and classification components of a recognition system separately. To further increase the reproducibility and comparability of results, well-defined evaluation settings are needed, such as the ones provided by recurring 'challenge' events [76–78]. Finally, communication between system components in real-life applications requires standardisation of recognition results for dialogue management or speech synthesis, etc. This is currently achieved by markup languages for description of emotional states (EMMA [3], EmotionML [69], MIML [51]) or extensions of VoiceXML to model speaker states in dialogue systems.

2.8 More Robustness

Robustness issues in the context of paralinguistic analysis can be categorised into technical robustness on the one hand and security on the other hand. Technical robustness refers to robustness against signal distortions including additive noise, e. g., environmental noise or interfering speakers (cocktail party problem) and reverberation, but also artefacts of transmission due to package loss and coding. Many of these issues have been extensively studied in the context of automatic speech recognition, and a wealth of methods is available, including speech enhancement, robust feature extraction, model-based techniques (i. e., learning the distortions) and novel recognition architectures such as graphical models. On another level, the *security* of paralinguistic analysis systems pertains to recognising malicious mis-use, i. e., attempted fraud. Examples for fraud include feigning of age (e. g., in an audio-based system for parental control), degree of intoxication, or emotion (e. g., by faking anger in an automated voice portal system in order to be redirected to a human operator).

Still, the majority of research in computational paralinguistics assumes laboratory conditions, i. e., a direct connection to the recogniser via high-quality audio interfaces, and data is recorded from (often paid) volunteers instead of real users

with potentially malicious intentions. There do exist a few studies on technical robustness of affect analysis, e. g., [70, 92, 96]—other speaker classification tasks are yet to follow. Yet, studies on the security of computational paralinguistics are currently sparse; these include detection of fake emotions from facial expressions [102] and recognition of feigned depression and sleepiness [14, 61]. This is in stark contrast to the efforts devoted to speaker verification, i. e., robustness of speaker recognition systems against feigning speaker identity [11]. Besides, little attention has been paid to the 'goats' of paralinguistic analysis: This is how non-malicious system users that systematically cause false alarms have been termed in the 'zoo' of speaker verification [21]. For instance, it is known that speaker identification is hindered by emotion [87], and personality analysis is influenced by the use of second language [18]. Future research should broaden this analysis to other influence factors such as tiredness or intoxication; multi-task learning of paralinguistic information could help systems to model these influences.

2.9 More Realism

Basically, there is agreement that in order to evaluate systems for paralinguistic analysis in conditions close to real-life application, realistic data are needed: That is, natural occurrences of states and traits such as sleepiness or personality, recorded in real-life acoustic environments and interaction scenarios, are required. Still, progress is slow; one of the reasons might be the high effort of collecting and annotating such data. Of course, the required type of data depends on the particular application; in many cases, realistic data corresponds to spontaneous and conversational, i. e., verbally unrestricted speech. Besides, realism concerns the choice of testing instances. In order to obtain a realistic estimate of system performance, these should not be restricted to prototypical, straightforward cases, such as ones with high human agreement [90]. If pre-selection is applied, e. g., to gain performance bounds, this should follow transparent, objective criteria instead of an 'intuitive' selection by experts. Realism further relates to pre-processing of data such as chunking according to acoustic, phonetic or linguistic criteria. Such chunking should either be oriented on low-level acoustic features (i. e., a voice activity based chunking, which can already be challenging in reverberant or noisy acoustic conditions). Alternatively, if linguistic or phonetic criteria are employed, these should be evaluated on speech recognition output, such as in [52], not forced alignment based on manual transliteration, such as in many of today's emotional corpora, e. g., [76]. If additional meta-information or common knowledge is exploited in the analysis process, this information should be obtained from publicly available sources, e. g., by web-based queries, rather than by including expert knowledge. Finally, real-life applications imply the requirement of speaker independence in most cases. This can be established by partitioning into train, development and test sets [76]; however, often cross-validation is employed especially in case of small data sets, in order to ensure significance of results and to avoid overfitting in case of small data sets. Using a three-fold speaker independent and stratified subdivision according to

simple criteria (e. g., splitting according to subject IDs) seems to be a reasonable compromise between transparency and statistical significance in that case.

2.10 More Cross-Cultural and Cross-Lingual Evaluation

One of the barriers to overcome if paralinguistic information retrieval systems are to be widely employed is to enable their use across cultural borders. Yet, cross-cultural effects make this task even more challenging. Concerning speech, it is still an open question which speaker states and traits manifest consistently across cultures and languages [8]. It seems intuitive that, for example, linguistic features used to express certain emotional states differ; yet, often one-to-one mappings between languages can be found. However, generally little attention is paid to the more subtle effects of the cultural background. Among others, the relative robustness of speaker identification to the language being spoken has been confirmed [6], resulting in performance differences that are small in magnitude, although they may be statistically significant [40]. Emotion recognition, on the other hand, has been shown to depend strongly on the language being spoken [17,27,88]; multimodal fusion might be a promising approach since some non-verbal behavioural signals, including laughter [67] or facial expressions [24] have been found to be largely independent of cultural background. Indeed, there is evidence that multimodality helps humans in cross-cultural emotion recognition [1]. In general, it might turn out that cross-cultural recognition of paralinguistic information is just another instance of learning correlated tasks: Recognising the race, ethnicity, dialect region, etc. of a person could help in determining his or her emotion state, but possibly even biological primitives such as age or gender. Thus, while the most obvious strategy to perform cross-cultural recognition is to build specifically adapted models, any of the other strategies discussed above in Section 2.1 could be promising as well.

3 Conclusions

Starting from a broad and unified overview of the field of computational paralinguistics, we outlined ten dominant trends that can be summarised as: extending the field to new and combined tasks and more variety in data, taking into account recent paradigms in machine learning, and moving from 'out-of-the-lab' to real-life application contexts. Despite these recent developments, there remain some 'black spots' in literature. These include the generalisation of features and models across paralinguistic information retrieval tasks; determination of meaningful confidence measures for paralinguistics in general, for instance, for use in dialogue systems; and finally, bridging the gap between analysis and synthesis of speaker states and traits, by transferring methodologies and the broader view on computational paralinguistics to enable multi-faceted speech synthesis, voice transformation, and benefit from it for (ad-hoc) training of analysis systems. Following these trends, we expect higher generalisation abilities of future systems for computational paralinguistics, and we look forward to experiencing their increasing application in real world contexts.

References

1. Abelin, A.: Cross-Cultural Multimodal Interpretation of Emotional Expressions - An Experimental Study of Spanish and Swedish. In: Proc. of Speech Prosody, ISCA (2004); no pagination
2. Ang, J., Dhillon, R., Shriberg, E., Stolcke, A.: Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In: Proc. Interspeech, pp. 2037–2040. Denver (2002)
3. Baggia, P., Burnett, D.C., Carter, J., Dahl, D.A., McCobb, G., Raggett, D.: EMMA: Extensible MultiModal Annotation markup language (2007), `http://www.w3.org/TR/emma/`
4. Batliner, A., Steidl, S., Schuller, B., Seppi, D., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Aharonson, V., Amir, N.: Whodunnit – Searching for the Most Important Feature Types Signalling Emotional User States in Speech. Computer Speech and Language 25, 4–28 (2011)
5. Belin, P., Fillion-Bilodeau, S., Gosselin, F.: The montreal affective voices: A validated set of nonverbal affect bursts for research on auditory affective processing. Behavior Research Methods 40(2), 531–539 (2008)
6. Bellegarda, J.R.: Language-independent speaker classification over a far-field microphone. In: Mueller, C. (ed.) Speaker Classification II: Selected Projects, pp. 104–115. Springer, Berlin (2007)
7. Bitouk, D., Verma, R., Nenkova, A.: Class-level spectral features for emotion recognition. Speech Communication 52(7-8), 613–625 (2011)
8. Boden, M.: Mind as Machine: A History of Cognitive Science, ch. 9. Oxford Univ. Press, New York (2008)
9. Bone, D., Black, M.P., Li, M., Metallinou, A., Lee, S., Narayanan, S.: Intoxicated Speech Detection by Fusion of Speaker Normalized Hierarchical Features and GMM Supervectors. In: Proc. of Interspeech, Florence, Italy, pp. 3217–3220 (2011)
10. Byrd, D.: Relations of sex and dialect to reduction. Speech Communication 15(1-2), 39–54 (1994)
11. Campbell, J.: Speaker recognition: a tutorial. Proceedings of the IEEE 85(9), 1437–1462 (1997)
12. Campbell, N.: On the use of nonverbal speech sounds in human communication. In: Proc. of COST 2102 Workshop, Vietri sul Mare, Italy, pp. 117–128 (2007)
13. Campbell, N., Kane, J., Moniz, H.: Processing 'yup!' and other short utterances in interactive speech. In: Proc. of ICASSP, Prague, Czech Republic, pp. 5832–5835 (2011)
14. Cannizzaro, M., Reilly, N., Snyder, P.J.: Speech content analysis in feigned depression. Journal of Psycholinguistic Research 33(4), 289–301 (2004)
15. Caruana, R.: Multitask learning: A knowledge-based source of inductive bias. Machine Learning 28, 41–75 (1997)
16. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16, 321–357 (2002)
17. Chen, A.: Perception of paralinguistic intonational meaning in a second language. Language Learning 59(2), 367–409 (2009)
18. Chen, S.X., Bond, M.H.: 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 (2010)

19. Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., Schröder, M.: 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 (2000)
20. Digman, J.M.: Personality Structure: emergence of the Five-Factor Model. Ann. Rev. Psychol. 41, 417–440 (1990)
21. Doddington, G., Liggett, W., Martin, A., Przybocki, M., Reynolds, D.: Sheep, Goats, Lambs and Wolves: A Statistical Analysis of Speaker Performance in the NIST 1998 Speaker Recognition Evaluation. In: Proc. of ICSLP (1998); no pagination
22. van Dommelen, W.A., Moxness, B.H.: Acoustic parameters in speaker height and weight identification: Sex-specific behaviour. Language and Speech 38(3), 267–287 (1995)
23. Dongdong, L., Yingchun, Y.: Emotional speech clustering based robust speaker recognition system. In: Proceedings of the 2009 2nd International Congress on Image and Signal Processing, CISP 2009, Tianjin, China, pp. 1–5 (2009)
24. Elfenbein, H., Mandal, M.K., Ambady, N., Harizuka, S.: Cross-Cultural Patterns in Emotion Recognition: Highlighting Design and Analytical Techniques. Emotion 2(1), 75–84 (2002)
25. Ellgring, H., Scherer, K.R.: Vocal Indicators of Mood change in Depression. Journal of Nonverbal Behavior 20, 83–110 (1996)
26. Erdem, C.E., Bozkurt, E., Erzin, E., Erdem, A.T.: RANSAC-based training data selection for emotion recognition from spontaneous speech. In: AFFINE 2010 - Proceedings of the 3rd ACM Workshop on Affective Interaction in Natural Environments, Co-located with ACM Multimedia 2010, Florence, Italy, pp. 9–14 (2010)
27. Esposito, A., Riviello, M.T.: 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 (2011)
28. Evans, S., Neave, N., Wakelin, D.: 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 (2006)
29. Eyben, F., Batliner, A., Schuller, B., Seppi, D., Steidl, S.: 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 (2010)
30. Eyben, F., Wöllmer, M., Schuller, B.: openEAR - Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit. In: Proc. ACII, Amsterdam, pp. 576–581 (2009)
31. Eyben, F., Wöllmer, M., Schuller, B.: openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor. In: Proc. ACM Multimedia, Florence, Italy, pp. 1459–1462 (2010)
32. Eyben, F., Wöllmer, M., Valstar, M., Gunes, H., Schuller, B., Pantic, M.: 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 (2011)
33. Gillick, D.: Can conversational word usage be used to predict speaker demographics? In: Proc. of Interspeech, Makuhari, Japan, pp. 1381–1384 (2010)
34. Gocsál: Female listeners' personality attributions to male speakers: The role of acoustic parameters of speech. Pollack Periodica 4(3), 155–165 (2009)

35. Gonzalez, J.: Formant frequencies and body size of speaker: a weak relationship in adult humans. Journal of Phonetics 32(2), 277–287 (2004)
36. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.: The WEKA Data Mining Software: An Update. SIGKDD Explorations 11 (2009)
37. Hansen, J., Bou-Ghazale, S.: Getting started with susas: A speech under simulated and actual stress database. In: Proc. EUROSPEECH 1997, Rhodes, Greece, vol. 4, pp. 1743–1746 (1997)
38. Ipgrave, J.: The language of friendship and identity: Children's communication choices in an interfaith exchange. British Journal of Religious Education 31(3), 213–225 (2009)
39. Jia, L., Chun, C., Jiajun, B., Mingyu, Y., Jianhua, T.: 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 (2007)
40. Kleynhans, N.T., Barnard, E.: 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 (November 2005)
41. Krajewski, J., Batliner, A., Golz, M.: Acoustic sleepiness detection: Framework and validation of a speech-adapted pattern recognition approach. Behavior Research Methods 41, 795–804 (2009)
42. Krauss, R.M., Freyberg, R., Morsella, E.: Inferring speakers physical attributes from their voices. Journal of Experimental Social Psychology 38(6), 618–625 (2002)
43. Laskowski, K., Ostendorf, M., Schultz, T.: 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 (2008)
44. Lee, C., Mower, E., Busso, C., Lee, S., Narayanan, S.: Emotion recognition using a hierarchical binary decision tree approach. In: Proc. Interspeech, Brighton, pp. 320–323 (2009)
45. Levit, M., Huber, R., Batliner, A., Nöth, E.: 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 (2001)
46. Li, D., Wu, Z., Yang, Y.: Speaker recognition based on pitch-dependent affective speech clustering. Moshi Shibie yu Rengong Zhineng/Pattern Recognition and Artificial Intelligence 22(1), 136–141 (2009)
47. Litman, D., Rotaru, M., Nicholas, G.: Classifying Turn-Level Uncertainty Using Word-Level Prosody. In: Proc. Interspeech, Brighton, UK, pp. 2003–2006 (2009)
48. Mahdhaoui, A., Chetouani, M.: 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, pp. 1–6. IEEE, Grenoble (2009)
49. Mahdhaoui, A., Chetouani, M., Kessous, L.: Time-Frequency Features Extraction for Infant Directed Speech Discrimination. In: Solé-Casals, J., Zaiats, V. (eds.) NOLISP 2009. LNCS (LNAI), vol. 5933, pp. 120–127. Springer, Heidelberg (2010)
50. Maier, A., Haderlein, T., Eysholdt, U., Rosanowski, F., Batliner, A., Schuster, M., Nöth, E.: PEAKS - A system for the automatic evaluation of voice and speech disorders. Speech Communication 51, 425–437 (2009)
51. Mao, X., Li, Z., Bao, H.: An Extension of MPML with Emotion Recognition Functions Attached. In: Prendinger, H., Lester, J.C., Ishizuka, M. (eds.) IVA 2008. LNCS (LNAI), vol. 5208, pp. 289–295. Springer, Heidelberg (2008)

52. Metze, F., Batliner, A., Eyben, F., Polzehl, T., Schuller, B., Steidl, S.: Emotion recognition using imperfect speech recognition. In: Proc. Interspeech 2010, Makuhari, Japan, pp. 478–481 (2011)
53. Mohammadi, G., Vinciarelli, A., Mortillaro, M.: The Voice of Personality: Mapping Nonverbal Vocal Behavior into Trait Attributions. In: Proc. SSPW 2010, Firenze, Italy, pp. 17–20 (2010)
54. Mokhtari, A., Campbell, N.: Speaking style variation and speaker personality. In: Proc. of Speech Prosody, Campinas, Brazil, pp. 601–604 (2008)
55. Mporas, I., Ganchev, T.: Estimation of unknown speakers' height from speech. International Journal of Speech Technology 12(4), 149–160 (2009)
56. Müller, C., Wittig, F., Baus, J.: 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 (2003)
57. Omar, M.K., Pelecanos, J.: A novel approach to detecting non-native speakers and their native language. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, Dallas, Texas, pp. 4398–4401 (2010)
58. Patil, S.A., Hansen, J.H.L.: The physiological microphone (pmic): A competitive alternative for speaker assessment in stress detection and speaker verification. Speech Communication 52(4), 327–340 (2010)
59. Polzehl, T., Möller, S., Metze, F.: Automatically assessing personality from speech. In: Proceedings - 2010 IEEE 4th International Conference on Semantic Computing, ICSC 2010, Pittsburgh, PA, pp. 134–140 (2010)
60. Provine, R.: Laughter punctuates speech: linguistic, social and gender contexts of laughter. Ethology 15, 291–298 (1993)
61. Reilly, N., Cannizzaro, M.S., Harel, B.T., Snyder, P.J.: Feigned depression and feigned sleepiness: A voice acoustical analysis. Brain and Cognition 55(2), 383–386 (2004)
62. Reisenzein, R., Weber, H.: Personality and Emotion. In: Corr, P.J., Matthews, G. (eds.) The Cambridge Handbook of Personality Psychology, pp. 54–71. Cambridge University Press, Cambridge (2009)
63. Revelle, W., Scherer, K.: Personality and Emotion. In: Oxford Companion to the Affective Sciences, pp. 1–4. Oxford University Press, Oxford (2009)
64. Ringeval, F., Chetouani, M.: 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 (2008)
65. Rosenberg, A., Hirschberg, J.: Acoustic/Prosodic and Lexical Correlates of Charismatic Speech. In: Proc. of Interspeech, Lisbon, pp. 513–516 (2005)
66. Russel, J.A.: A circumplex model of affect. Journal of Personality and Social Psychology 39, 1161–1178 (1980)
67. Sauter, D.A., Eisner, F., Ekman, P., Scott, S.K.: Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. Proc. of the National Academy of Sciences of the U.S.A. 107(6), 2408–2412 (2010)
68. Schiel, F., Heinrich, C.: Laying the foundation for in-car alcohol detection by speech. In: Proc. INTERSPEECH 2009, Brighton, UK, pp. 983–986 (2009)
69. Schröder, M., Devillers, L., Karpouzis, K., Martin, J.-C., Pelachaud, C., Peter, C., Pirker, H., Schuller, B., Tao, J., Wilson, I.: What Should a Generic Emotion Markup Language Be Able to Represent? In: Paiva, A.C.R., Prada, R., Picard, R.W. (eds.) ACII 2007. LNCS, vol. 4738, pp. 440–451. Springer, Heidelberg (2007)
70. Schuller, B.: Affective speaker state analysis in the presence of reverberation. International Journal of Speech Technology 14(2), 77–87 (2011)

71. Schuller, B., Batliner, A., Steidl, S., Seppi, D.: 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 (2011)
72. Schuller, B., Burkhardt, F.: 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 (2010)
73. Schuller, B., Jiménez Villar, R., Rigoll, G., Lang, M.: Meta-classifiers in acoustic and linguistic feature fusion-based affect recognition. In: Proc. ICASSP, Philadelphia, pp. I:325–I:328 (2005)
74. Schuller, B., Reiter, S., Rigoll, G.: Evolutionary feature generation in speech emotion recognition. In: Proc. Int. Conf. on Multimedia and Expo, ICME 2006, Toronto, Canada, pp. 5–8 (2006)
75. Schuller, B., Rigoll, G., Lang, M.: 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 (2004)
76. Schuller, B., Steidl, S., Batliner, A.: The INTERSPEECH 2009 Emotion Challenge. In: Proceedings of 11th European Conference on Speech Communication and Technology, Interspeech 2009 – Eurospeech, Brighton, UK, September 6-10, pp. 312–315 (2009)
77. Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., Narayanan, S.: The INTERSPEECH 2010 Paralinguistic Challenge – Age, Gender, and Affect. In: Proceedings of 11th International Conference on Spoken Language Processing, Interspeech 2010 – ICSLP, Makuhari, Japan, September 26-30, pp. 2794–2797 (2010)
78. Schuller, B., Steidl, S., Batliner, A., Schiel, F., Krajewski, J.: The Interspeech 2011 Speaker State Challenge. In: Proc. Interspeech, Florence, Italy, pp. 3201–3204 (2011)
79. Schuller, B., Vlasenko, B., Eyben, F., Wöllmer, M., Stuhlsatz, A., Wendemuth, A., Rigoll, G.: Cross-corpus acoustic emotion recognition: Variances and strategies. IEEE Transactions on Affective Computing 1(2), 119–131 (2010)
80. Schuller, B., Wimmer, M., Mösenlechner, L., Kern, C., Arsic, D., Rigoll, G.: Brute-Forcing Hierarchical Functionals for Paralinguistics: a Waste of Feature Space? In: Proc. ICASSP, Las Vegas, pp. 4501–4504 (2008)
81. Schuller, B., Wöllmer, M., Eyben, F., Rigoll, G., Arsic, D.: Semantic Speech Tagging: Towards Combined Analysis of Speaker Traits. In: Proc. AES 42nd International Conference, Ilmenau, Germany, pp. 89–97 (2011)
82. Schuller, B., Zhang, Z., Weninger, F., Rigoll, G.: Selecting training data for cross-corpus speech emotion recognition: Prototypicality vs. generalization. In: Proc. 2011 Afeka-AVIOS Speech Processing Conference, Tel Aviv, Israel (2011)
83. Schuller, B., Müller, R., Eyben, F., Gast, J., Hörnler, B., Wöllmer, M., Rigoll, G., Höthker, A., Konosu, H.: 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 (2009)
84. Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., Narayanan, S.: Paralinguistics in Speech and Language—State-of-the-Art and the Challenge. Computer Speech and Language, Special Issue on Paralinguistics in Naturalistic Speech and Language (2011) (to appear)
85. Schuller, B., Zhang, Z., Weninger, F., Rigoll, G.: Using Multiple Databases for Training in Emotion Recognition: To Unite or to Vote? In: Proc. of INTERSPEECH, pp. 1553–1556. ISCA, Florence (2011)

86. Schwenker, F., Scherer, S., Schmidt, M., Schels, M., Glodek, M.: Multiple Classifier Systems for the Recogonition of Human Emotions. In: El Gayar, N., Kittler, J., Roli, F. (eds.) MCS 2010. LNCS, vol. 5997, pp. 315–324. Springer, Heidelberg (2010)
87. Shahin, I.: Verifying speakers in emotional environments. In: IEEE International Symposium on Signal Processing and Information Technology, ISSPIT 2009, Ajman, UAE, pp. 328–333 (2009)
88. Shami, M., Verhelst, W.: Automatic classification of expressiveness in speech: A multi-corpus study. In: Mueller, C. (ed.) Speaker Classification II: Selected Projects, pp. 43–56. Springer, Berlin (2007)
89. Stadermann, J., Koska, W., Rigoll, G.: Multi-task learning strategies for a recurrent neural net in a hybrid tied-posteriors acoustic mode. In: Proc. of Interspeech 2005, pp. 2993–2996. ISCA, Lisbon (2005)
90. Steidl, S., Schuller, B., Batliner, A., Seppi, D.: The Hinterland of Emotions: Facing the Open-Microphone Challenge. In: Proc. ACII, Amsterdam, pp. 690–697 (2009)
91. Stuhlsatz, A., Meyer, C., Eyben, F., Zielke, T., Meier, G., Schuller, B.: Deep Neural Networks for Acoustic Emotion Recognition: Raising the Benchmarks. In: Proc. ICASSP, Prague, Czech Republic, pp. 5688–5691 (2011)
92. Tabatabaei, T.S., Krishnan, S.: Towards robust speech-based emotion recognition. In: Proc. IEEE International Conference on Systems, Man and Cybernetics, Istanbul, Turkey, pp. 608–611 (2010)
93. Ververidis, D., Kotropoulos, C.: Automatic speech classification to five emotional states based on gender information. In: Proc. of 12th European Signal Processing Conference, Vienna, Austria, pp. 341–344 (2004)
94. Vinciarelli, A., Pantic, M., Bourlard, H.: Social signal processing: Survey of an emerging domain. Image and Vision Computing 27, 1743–1759 (2009)
95. Vogt, T., Andre, E.: Improving automatic emotion recognition from speech via gender differentiation. In: Proc. of Language Resources and Evaluation Conference (LREC 2006), Genoa, Italy, pp. 1–4 (2006)
96. Weninger, F., Schuller, B., Batliner, A., Steidl, S., Seppi, D.: Recognition of non-prototypical emotions in reverberated and noisy speech by nonnegative matrix factorization. Eurasip Journal on Advances in Signal Processing 2011(Article ID 838790), 16 pages (2011)
97. Wöllmer, M., Schuller, B., Eyben, F., Rigoll, G.: 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 (2010)
98. Wöllmer, M., Weninger, F., Eyben, F., Schuller, B.: Acoustic-Linguistic Recognition of Interest in Speech with Bottleneck-BLSTM Nets. In: Proc. of INTERSPEECH, Florence, Italy, pp. 77–80 (2011)
99. Wu, S., Falk, T.H., Chan, W.: Automatic speech emotion recognition using modulation spectral features. Speech Communication 53(5), 768–785 (2011)
100. Yamada, M., Sugiyama, M., Matsui, T.: Semi-supervised speaker identification under covariate shift. Signal Processing 90(8), 2353–2361 (2010)
101. Yoon, W., Park, K.: Building robust emotion recognition system on heterogeneous speech databases. In: Digest of Technical Papers - IEEE International Conference on Consumer Electronics, pp. 825–826 (2011)
102. Zhang, Z., Singh, V., Slowe, T., Tulyakov, S., Govindaraju, V.: Real-time Automatic Deceit Detection from Involuntary Facial Expressions. In: Proc. of CVPR, pp. 1–6 (2007)
103. Zhang, Z., Weninger, F., Wöllmer, M., Schuller, B.: Unsupervised Learning in Cross-Corpus Acoustic Emotion Recognition. In: Proc. Automatic Speech Recognition and Understanding Workshop (ASRU 2011). IEEE, Big Island (2011)