This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2457417, IEEE Transactions on Affective Computing

The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing

Florian Eyben, Klaus Scherer, Björn Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Devillers, Julien Epps, Petri Laukka, Shrikanth Narayanan, Khiet Truong

Abstract—Work on voice sciences over recent decades has led to a proliferation of acoustic parameters that are used quite selectively and are not always extracted in a similar fashion. With many independent teams working in different research areas, shared standards become an essential safeguard to ensure compliance with state-of-the-art methods allowing appropriate comparison of results across studies and potential integration and combination of extraction and recognition systems. In this paper we propose a basic standard acoustic parameter set for various areas of automatic voice analysis, such as paralinguistic or clinical speech analysis. In contrast to a large brute-force parameter set, we present a minimalistic set of voice parameters here. These were selected based on a) their potential to index affective physiological changes in voice production, b) their proven value in former studies as well as their automatic extractability, and c) their theoretical significance. The set is intended to provide a common baseline for evaluation of future research and eliminate differences caused by varying parameter sets or even different implementations of the same parameters. Our implementation is publicly available with the openSMILE toolkit. Comparative evaluations of the proposed feature set and large baseline feature sets of INTERSPEECH challenges show a high performance of the proposed set in relation to its size.

Index Terms—Affective Computing, Acoustic Features, Standard, Emotion Recognition, Speech Analysis, Geneva Minimalistic Parameter Set

1 Introduction

Interest in the vocal expression of different affect states has a long history with researchers working in various fields of research ranging from psychiatry to engineering. Psychiatrists have been attempting to diagnose affective states. Psychologists and communication researchers have been exploring the capacity of the voice to carry signals of emotion. Linguists and phoneticians have been discovering the role of affective pragmatic information in language production and perception. More recently, computer scientists and engineers have been attempting to automatically recognize and manipulate speaker attitudes and emotions to render information technology more accessible and credible for human users. Much of this research and development uses the extraction of acoustic parameters from the speech signal as a method to understand the patterning of the vocal expression of different emotions and other affective dispositions and processes. The underlying theoretical assumption is that affective processes differentially change autonomic arousal and the tension of the striate musculature and thereby affect voice and speech production on the phonatory and articulatory level and that these changes can be estimated by different parameters of the acoustic waveform [1].

Emotional cues conveyed in the voicce have been empirically documented recently by the measurement of emotion-differentiating parameters related to subglottal pressure, transglottal airflow, and vocal fold vibration ([2], [3], [4], [5], [6], [7], [8]). Mostly based on established procedures in phonetics and speech sciences to measure different aspects of phonation and articulation in speech, researchers have used a large number of acoustic parameters (see [9]; [10], for overviews), including parameters in the Time domain (e. g., speech rate), the Frequency domain (e. g., fundamental frequency (F_0) or formant frequencies), the Amplitude domain (e. g., intensity or energy), and the spectral distribution domain (e. g., relative energy in different frequency bands). Not all of these parameters have been standardized in terms of their exact

- *F. Eyben is with Technische Universität, München, Germany, Université de Genève, Geneva, Switzerland, and audEERING, Gilching, Germany.*
- *K. Scherer is with Université de Genève, Geneva, Switzerland.*
- *B. Schuller is with University of Passau, Passau, Germany, Imperial College, London, UK, and Université de Genève, Geneva, Switzerland.*
- *J. Sundberg is with KTH Royal Institute of Technology, Stockholm, Sweden.*
- *E. André is with Universität Augsburg, Germany.*
- *C. Busso is with University of Texas, Dallas, TX, USA.*
- *L. Devillers is with University of Paris-Sorbonne IV and CNRS/LIMSI, Paris, France.*
- *J. Epps is with University of New South Wales, Sydney, Australia and NICTA ATP Laboratory, Eveleigh, Australia.*
- *P. Laukka is with Stockholm University, Stockholm, Sweden.*
- *S. Narayanan is with SAIL at the University of Southern California, Los Angeles, CA, USA.*
- *K. Truong is with University of Twente, Enschede, The Netherlands.*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2457417, IEEE Transactions on Affective Computing

computation and thus results reported in the literature cannot always be easily compared. Even where parameters have been extracted using widely used tools like Praat [11], the exact settings used are not usually easily and publicly accessible. Furthermore, different studies often use sets of acoustic features that overlap only partially, again rendering comparison of results across studies exceedingly difficult and thus endangering the cumulation of empirical evidence. The recent use of machine learning algorithms for the recognition of affective states in speech has led to a proliferation in the variety and quantity of acoustic features employed, amounting often to several thousand basic (low-level) and derived (functionals) parameters (e. g., [12]). While this profusion of parameters allows to capture many acoustic characteristics in a comprehensive and reliable manner, this comes at the cost of serious difficulties in the interpretation of the underlying mechanisms.

However, applications such as the fine grained control of emotionality in speech synthesis (cf. [13], [14]), or dimensional approaches to emotion and mental state recognition that seek to quantify arousal, valence or depression severity, for example, along a single axis, all require a deeper understanding of the mechanism of production and perception of emotion in humans. To reach this understanding, finding and interpreting relevant acoustic parameters is crucial. Thus, based on many previous findings in the area of speech and voice analysis (e. g., , [2], [9], [15], [16], [17], [18], [19]), in this article the authors present a recommendation for a minimalistic standard parameter set for the acoustic analysis of speech and other vocal sounds. This standard set is intended to encourage researchers in this area to adopt it as a baseline and use it alongside any specific parameters of particular interest to individual researchers or groups, to allow replication of findings, comparison between studies from different laboratories, and greater cumulative insight from the efforts of different laboratories on vocal concomitants of affective processes.

Moreover, large brute-forced feature sets are well known to foster over-adaptation of classifiers to the training data in machine learning problems, reducing their generalisation capabilities to unseen (test) data (cf. [20]). Minimalistic parameter sets might reduce this danger and lead to better generalisation in cross-corpus experiments and ultimately in real-world test scenarios. Further, as mentioned above, the interpretation of the meaning of the parameters in a minimalistic set is much easier than in large brute-forced sets, where this is nearly impossible.

The remainder of this article is structured as follows: First, Section 2 provides a brief overview of acoustic analyses in the fields of psychology, phonetics, acoustics, and engineering which are the basis of the recommendation proposed in this article; next, in Section 3 we give a detailed description of the acoustic parameters contained in the recommended parameter set and the implementation thereof. The parameter set is extensively evaluated on six well-known affective speech databases and the classification performance is compared to all high-dimensional brute-forced sets of the INTERSPEECH Challenges on Emotion and Paralinguistics from 2009 to 2013 in Section 4. Final remarks on the parameters recommended in this article and the classification performance relative to other established sets as well as a discussion on the direction of future research in this field is given in 5.

2 Related Work

The minimalistic feature set proposed in this article is not the first joint attempt to standardise acoustic parameter sets. The CEICES initiative [21], for example, brought researchers together who were working on identification of emotional states from the voice. They combined the acoustic parameters they had used in their individual work in a systematic way in order to create large, brute-forced parameter sets, and thereby identify individual parameters by a unique naming (code) scheme. However, the exact implementation of the individual parameters was not well standardised. CEICES was a more engineering-driven "collector" approach where parameters which were successful in classification experiments were all included, while GeMAPS is a more interdisciplinary attempt to agree on a minimalistic parameter set based on multiple source, interdisciplinary evidence and theoretical significance or a few parameters.

Related programs for computation of acoustic parameters, which are used by both linguists and computer science researchers, include the popular Praat toolkit [11] or Wavesurfer[1].

This section gives a literature overview on studies where parameters that form the basis of our recommendation have been proposed and used for voice analysis and related fields.

An early survey [15] and a recent overview [17] nicely summarise a few decades of psychological literature on affective speech research and concludes from the empirical data presented that intensity (loudness), F0 (fundamental frequency) mean, variability, and range, as well as the high frequency content/energy of a speech signal show correlations with prototypical vocal affective expressions such as stress (Intensity, F0 mean), anger and sadness (all parameters), and boredom (F0 variability and range), for example. Further, speech and articulation rate was found to be important for all emotional expressions. For the case of automatic arousal recognition, [22] successfully builds an unsupervised recognition framework with these descriptors.

[16] perform acoustic analysis of various fundamental frequency and harmonics related parameters

1. http://www.speech.kth.se/wavesurfer/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2457417, IEEE Transactions on Affective Computing

on a small set of emotional speech utterances. The findings confirm that parameters related to F0 and spectral distribution are important cues to affective speech content. [16] introduce a ratio of the peak frequency to the fundamental frequency, and use spectral roll-off points (called distribution of frequency – DFB – there). More recently, [18], also validate the discriminatory power of amplitude, pitch, and spectral profile (tilt, balance, distribution) parameters for a larger set of vocal emotional expressions.

Most studies, such as the two previously mentioned, deal with the analysis of acoustic arousal and report fairly consistent parameters which are cues to vocal arousal (nicely summarised by [17]). The original findings that prosodic parameters (F0 and intensity) are relevant for arousal have been confirmed in many similar studies, such as [4], and more automatic, machine learning based parameter evaluation studies such as [23]. Regarding energy/intensity, [24] shows that a loudness measure, in which the signal energy in various frequency bands is weighted according to the human-hearing's frequency sensitivity, is better correlated to vocal affect dimensions than the simple signal energy alone. Further, it is shown there, that spectral flux has the overall best correlation for a single feature.

Recent work, such as [17] and [25], has dealt with other dimensions besides arousal – in particular valence (both) and the level of interest (LOI) [25]. For valence both of these studies conclude that spectral shape parameters could be important cues for vocal valence. Also, rhythm related parameters, such as speaking rate are correlated with valence. [26] confirms the importance of various spectral band energies, spectral slope, overall intensity, and the variance of the fundamental frequency, for the detection of angry speech. These parameters were also reported to be important for cognitive load [27] and psychomotor retardation [28].

[25] also shows a large importance of cepstral parameters (Mel-Frequency-Cepstral-Coefficients – MFCC), especially for LOI. These are closely related to spectral shape parameters. Especially the lower order MFCC, resemble spectral tilt (slope) measures to some extent over the full range of the spectrum (first coefficient), or in various smaller sub-bands (second and higher coefficient). The relevance of spectral slope and shape is also investigated and confirmed by [29], for example, and by [30] and [31].

In contrast to the findings in [15], for example, [25] suggests that the relative importance of prosodic parameters as well as voice quality parameters decreases in the case of degraded audio conditions (background noise, reverberation), while the relative importance of spectral shape parameters increases. This is likely due to degraded accuracy in the estimation of the prosodic parameters such as due to interfering harmonics or energy contributed by the noise components. Overall, we believe that the lower order MFCC are important to consider for various tasks and thus we include MFCC 1–4 in the parameter set proposed in this article.

For automatic classification, large-scale brute-force acoustic parameter sets are used (cf. e. g., [32], [33], [34], [12]). These contain parameters which are easily and reliably computable from acoustic signals. The general tendency in most studies is, that larger parameter sets perform better [34]. This might be due to the fact that in larger feature sets the 'right' features are more likely present, or due to the fact that the combination of all features is necessary. Another reason might be that with this many parameters (over 6000 in some cases), the machine learning methods simply over-adapt to the (rather) small training datasets. This is evident especially in cross-corpus classification experiments, where the large feature sets show poorer performance despite their higher performance in intra-corpus evaluations [20]. As said, it is thus our aim in this article to select relevant parameters, guided by the findings of previous, related studies.

Besides vocal emotional expressions, there are numerous other studies which deal with other vocal phenomena and find similar and very related features to be important. [27], for example, shows the importance of vowel-based formant frequency statistics, and [5], for example, shows the usefulness of glottal features when combined with prosodic features for identification of depression in speech. Voice source features, in particular the harmonic difference H1–H2, showed a consistent decrease with increasing cognitive load, based on a study employing manually corrected pitch estimates [35]. Recently, researchers have attempted to analyse further paralinguistic characteristics of speech, ranging from age and gender [36], to cognitive and physical load [37], for example.

Many automatically extracted brute-force parameter sets neglect formant parameters due to difficulties in extracting them reliably. For voice research and automatic classification, they are very important though. Formants have been shown sensitive to many forms of emotion and mental state and formants give approximately state of the art cognitive load classification results [27] and depression recognition and assessment results [31], [38], and can provide competitive emotion recognition performance [39] with a fraction of the feature dimension of other systems. A basic set of formant related features is thus included in our proposed set.

Due to the proven high importance of the fundamental frequency (cf. [6]) and amplitude/intensity, a robust fundamental frequency measure and a pseudo-auditory loudness measure are included in our proposed set. A wide variety of statistics are applied to both parameters over time, in order to capture distributional changes. To robustly represent the high frequency content and the spectral balance, the de-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2457417, IEEE Transactions on Affective Computing

scriptors alpha ratio, Hammarberg index, and spectral slope are considered in this article. The vocal timbre is encoded by Mel-Frequency Cepstral Coefficients (MFCC), and the quality of the vocal excitation signal by the period-to-period jitter and shimmer of F0. To allow for vowel-based voice research, and due to their proven relevance for certain tasks, formant parameters are also included in the set.

3 Acoustic Parameter Recommendation

The recommendation presented here has been conceived at an interdisciplinary meeting of voice and speech scientists in Geneva [2] and further developed at Technische Universität München (TUM). The choice of parameters has been guided (and is justified) by three criteria: 1) the potential of an acoustic parameter to index physiological changes in voice production during affective processes, 2) the frequency and success with which the parameter has been used in the past literature (see Section 2), and 3) its theoretical significance (see [2]; [1]).

Two versions of the acoustic parameter set recommendation are proposed here: a minimalistic set of parameters, which implements prosodic, excitation, vocal tract, and spectral descriptors found to be most important in previous work of the authors, and an extension to the minimalistic set, which contains a small set of cepstral descriptors, which – from the literature (e. g., [40]) – are consistently known to increase the accuracy of automatic affect recognition over a pure prosodic and spectral parameter set. Several studies on automatic parameter selection, such as [23], [24], suggest that the lower order MFCCs are more important for affect and paralinguistic voice analysis tasks. When looking at the underlying Discrete Cosine Transformation (DCT-II) base functions used when computing MFCCs, it is evident that the lower order MFCC are related to spectral tilt and the overall distribution of spectral energy. Higher order MFCCs would reflect more fine grained energy distributions, which are presumably more important to identify phonetic content than non-verbal voice attributes.

To encourage rapid community discussion on the parameter sets, as well as updates and additions from the community, a wiki-page[3] was set up where researchers can quickly connect and discuss issues with the parameter set. New ideas, if they are favoured by multiple contributors, will then be implemented and after a certain number of improvements or after a certain time frame, new versions of the parameter sets will be released publicly.

In the following sub-sections, we first give an overview over the minimalistic parameter recommendation (Section 3.1), and the extended parameter set (Section 3.2), before describing details of the algorithms used to compute the parameters in Section 6.1.

3.1 Minimalistic Parameter Set

The minimalistic acoustic parameter set contains the following compact set of 18 Low-level descriptors (LLD), sorted by parameter groups:

Frequency related parameters:

- **Pitch**, logarithmic F_0 on a semitone frequency scale, starting at 27.5 Hz (semitone 0).
- **Jitter**, deviations in individual consecutive F_0 period lengths.
- **Formant 1, 2, and 3 frequency**, centre frequency of first, second, and third formant
- **Formant 1**, bandwidth of first formant.

Energy/Amplitude related parameters:

- **Shimmer**, difference of the peak amplitudes of consecutive F_0 periods.
- **Loudness**, estimate of perceived signal intensity from an auditory spectrum.
- **Harmonics-to-Noise Ratio (HNR)**, relation of energy in harmonic components to energy in noise-like components.

Spectral (balance) parameters:

- **Alpha Ratio**, ratio of the summed energy from 50–1000 Hz and 1–5 kHz
- **Hammarberg Index**, ratio of the strongest energy peak in the 0–2 kHz region to the strongest peak in the 2–5 kHz region.
- **Spectral Slope 0–500 Hz and 500–1500 Hz**, linear regression slope of the logarithmic power spectrum within the two given bands.
- **Formant 1, 2, and 3 relative energy**, as well as the ratio of the energy of the spectral harmonic peak at the first, second, third formant's centre frequency to the energy of the spectral peak at F_0.
- **Harmonic difference H1–H2**, ratio of energy of the first F_0 harmonic (H1) to the energy of the second F_0 harmonic (H2).
- **Harmonic difference H1–A3**, ratio of energy of the first F_0 harmonic (H1) to the energy of the highest harmonic in the third formant range (A3).

All LLD are smoothed over time with a symmetric moving average filter 3 frames long (for pitch, jitter, and shimmer, the smoothing is only performed within voiced regions, i. e., not smoothing the transitions from 0 (unvoiced) to non 0). ***Arithmetic mean*** and ***coefficient of variation*** (standard deviation normalised by the arithmetic mean) are applied as ***functionals***

2. Conference organised by K. Scherer, B. Schuller, and J. Sundberg on September 1–2, 2013 at the Swiss Center of Affective Sciences in Geneva on *Measuring affect and emotion in vocal communication via acoustic feature extraction: State of the art, current research, and benchmarking* with the explicit aim of commonly working towards a recommendation for a reference set of acoustic parameters to be broadly used in the field.

3. http://audeering.com/research-and-open-source/gemaps: will be launched when the article is published

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2457417, IEEE Transactions on Affective Computing

to all 18 LLD, yielding 36 parameters. To *loudness and pitch* the following 8 functionals are additionally applied: ***20-th, 50-th, and 80-th percentile***, the ***range*** of 20-th to 80-th percentile, and the ***mean and standard deviation of the slope of rising/falling signal parts***. All functionals are applied to voiced regions only (non-zero F0), with the exception of all the functionals which are applied to loudness. This gives a total of 52 parameters. Also, the arithmetic mean of the Alpha Ratio, the Hammarberg Index, and the spectral slopes from 0–500 Hz and 500–1500 Hz over all unvoiced segments are included, totalling 56 parameters. In addition, 6 **temporal features** are included:

- the **rate of loudness peaks**, i. e., the number of loudness peaks per second,
- the **mean length** and the **standard deviation** of continuously **voiced regions** ($F_0 > 0$),
- the **mean length** and the **standard deviation** of **unvoiced regions** ($F_0 = 0$; approximating pauses),
- the **number of continuous voiced regions per second** (pseudo syllable rate).

No minimal length is imposed on voiced or unvoiced regions, i. e., in the extreme case they could be only one frame long. The Viterbi-based smoothing of the F_0 contour, however, prevents single voiced frames which are missing by error, for example, effectively. In total, 62 parameters are contained in the *Geneva Minimalistic Standard Parameter Set*.

3.2 Extended Parameter Set

The minimalistic set does not contain any cepstral parameters and only very few dynamic parameters (i. e., it contains no delta regression coefficients and no difference features; only the slopes of rising and falling F_0 and loudness segments encapsulate some dynamic information). Further, especially cepstral parameters have proven highly successful in modelling of affective states, e. g., by [23], [40], [41]. Thus, an *extension* set to the minimalistic set is proposed which contains the following 7 LLD in addition to the 18 LLD in the minimalistic set:

Spectral (balance/shape/dynamics) parameters:

- **MFCC 1–4** Mel-Frequency Cepstral Coefficients 1–4.
- **Spectral flux** difference of the spectra of two consecutive frames.

Frequency related parameters:

- **Formant 2–3 bandwidth** added for completeness of Formant 1–3 parameters.

As ***functionals***, the ***arithmetic mean*** and the ***coefficient of variation*** are applied to all of these 7 additional LLD to all segments (voiced and unvoiced together), except for the formant bandwidths to which the functionals are applied only in voiced regions. This adds 14 extra descriptors. Additionally, the arithmetic mean of the spectral flux in unvoiced regions only, the arithmetic mean and coefficient of variation of the spectral flux and MFCC 1–4 in voiced regions only is included. This results in another 11 descriptors. Additionally the **equivalent sound level** is included. This results in 26 extra parameters. In total, when combined with the Minimalistic Set, the *extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS)* contains 88 parameters.

4 Baseline Evaluation

The proposed minimalistic parameter set and the extended set are both evaluated for the task of automatic recognition in binary arousal and binary valence dimensions. The labels of six standard databases of affective speech were mapped to binary dimensional labels, as described in Section 4.2: Levels of Interest (TUM AVIC database), acted speech emotions in the Geneva Multimodal Emotion Portrayals (GEMEP) corpus and the German Berlin Emotional Speech database (EMO-DB), emotions portrayed in the singing voice of professional opera singers (SING), valence in childrens' speech from the FAU AIBO corpus [42] as used for the INTERSPEECH 2009 Emotion Challenge [43], as well as real-life emotions from German talk-show recordings (Vera-am-Mittag corpus (VAM)). The proposed minimal sets are compared to five large-scale, brute-forced baseline acoustic feature sets of the INTERSPEECH 2009 Emotion Challenge [43] (384 parameters), the INTERSPEECH 2010 Paralinguistic Challenge [36] (1,582 parameters), the INTERSPEECH 2011 Speaker State Challenge [44] (4,368 parameters), the INTERSPEECH 2012 Speaker Trait Challenge [45] (6,125 parameters), and the INTERSPEECH 2013 Computational Paralingusitics ChallengE (ComParE) [12] set (6,373 parameters), which is also used for the INTERSPEECH 2014 Computational Paralinguistics ChallengE [37].

4.1 Data-sets

4.1.1 FAU AIBO

FAU AIBO served as the official corpus for the world's first international Emotion Challenge [43]. It contains recordings of children who are interacting with the Sony pet robot *Aibo*. It thus contains spontaneous, German speech which is emotionally coloured. The children were told that the Aibo robot was responding to their voice commands regarding directions. However, the robot was in fact controlled by a human operator, who caused the robot to behaved disobediently sometimes, to provoke strong emotional reactions from the children. The recordings were performed at two different schools, referred to as MONT and OHM, from 51 children in total (age 10–13, 21 males, 30 females; approx. 9.2 hours of speech without pauses). The recorded audio was segmented automatically into speech turns with a speech-pause threshold of 1 s. The

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2457417, IEEE Transactions on Affective Computing

data are labelled for emotional expression on the word level. As given in [43] 5 emotion class labels are used: Anger, Emphatic, Neutral, Positive, and Rest. For a two-class valence task, all negative emotions (Anger and Emphatic – NEG) and all non-negative emotions (Neutral, Positive, and Rest – IDL) are combined.

4.1.2 TUM Audiovisual Interest Corpus

The TUM Audiovisual Interest Corpus (TUM-AVIC) contains audiovisual recordings of spontaneous affective interactions with non-restricted spoken content [46]. It was used as data-set for the INTERSPEECH 2010 Paralinguistics Challenge [36]. In the set-up, a product presenter walks a subject through a commercial presentation. The language used is English, although most of the product presenters were German native speakers. The subjects were mainly from European and Asian nationalities. 21 subjects (10 female) were recorded in the corpus.

The Level of Interest (LOI) is labelled for every sub-turn (which are found by a manual pause based sub-division of speaker turns) in three labels ranging from *boredom* (subject is bored with the conversation or the topic or both, she/he is very passive and does not follow the conversation; also referred to as *loi1*), over *neutral* (she/he follows and participates in the conversation but it can not be judged, whether she/he is interested in or indifferent towards the topic; also referred to as *loi2*) to *joyful* interaction (showing a strong desire of the subject to talk and to learn more about the topic, i. e., he/she shows a high interest in the discussion; also referred to as *loi3*). For the evaluations here, all 3 002 phrases (sub-turns) as in [47] are used – in contrast to the only 996 phrases with high inter-labeller agreement as, e. g., employed in [46].

4.1.3 Berlin Emotional Speech Database

A very well known and widely used set to test the effectiveness of automatic emotion classification is the Berlin Emotion Speech Database, also commonly known as EMO-DB. It was introduced by [48]. It contains sentences spoken in the emotion categories anger, boredom, disgust, fear, joy, neutrality, and sadness. The linguistic content is pre-defined by ten German short sentences, which are emotionally neutral, such as *"Der Lappen liegt auf dem Eisschrank" (The cloth is lying on the fridge.)*. Ten (five of them female) professional actors speak ten sentences in each of the seven emotional states. While the whole set contains over 700 utterances, in a listening test only 494 phrases are labelled as a minimum 60 % naturally sounding and a minimum 80 % identifiable (with respect to the emotion) by 20 people. A mean accuracy of 84.3 % is achieved for identification of the emotions by the subjects in the listening experiment on this reduced set of 494 utterances. This set is used in most other studies related to this database (cf. [47]), therefore, it is also adopted here.

4.1.4 The Geneva Multimodal Emotion Portrayals

The Geneva Multimodal Emotion Portrayals (GEMEP) corpus is a collection of 1,260 multimodal emotion expressions enacted by ten French-speaking actors [49]. The list of emotions includes those most frequently encountered in the literature (e. g., anger, fear, joy, and sadness) as well as more subtle variations of these categories (e. g., anger vs. irritation, and fear vs. anxiety). Specifically, the 12 following emotions are considered, which are distributed across all four quadrants of the activation-valence space: *amusement*, *pride*, *joy*, *relief*, *interest*, *pleasure*, *hot anger*, *panic fear*, *despair*, *irritation* (cold anger), *anxiety* (worry), and *sadness* (depression). 1,075 instances (approx. 90 per emotion) are in this set.

The actors portrayed each emotion through three different verbal contents (one sustained vowel and two pseudo-sentences) and several expression regulation strategies. During this process the subjects were recorded with three cameras and one microphone. All devices were synchronised. In order to increase realism and spontaneity in the recordings, a professional director helped the respective actor to choose a personal scenario for each emotion – e. g., by recall or mental imagery – which was personally relevant for the actor. The actors did not receive any instructions on how the emotions were to be expressed and they were free to use any movement and speech techniques they felt were appropriate.

4.1.5 Singing voice emotion database

This database of singing emotional speech was first introduced by [50]. Here, an extended set of the database is used (abbreviated as SING). Compared to the original set which contains three singers, additional recordings of five professional opera singers have been added following the same protocol. In total the recordings present are from five male and three female singers. The singers sung three different phrases and tone scales in ten emotion categories: neutral (no expression), panic/fear, passionate love, tense arousal, animated joy, triumphant pride, anger, sadness, tenderness, calm/serenity, condescension. Every recording session was recorded in one continuous stream without pause. The recordings were afterwards manually split into the phrase and scale parts. In this way, a set of 300 single instances of sung speech was obtained. The distribution of the instances across all emotion classes is almost balanced .

4.1.6 Vera-Am-Mittag

The Vera-Am-Mittag (VAM) corpus [51] consists of videos extracted from the German TV show "Vera am Mittag". In this show, the host (Vera) moderates

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2457417, IEEE Transactions on Affective Computing

discussions between the guests, e. g., by using questions to guide the discussion. The database contains 947 emotionally rich, spontaneous speech utterances sampled from 47 talk show guests. The discussions were authentic and not scripted and due to the nature of the show and the selection of guests these discussions rather quite affective and contain a large variety of highly emotional states. The topics discussed in the show were mostly personal issues, like friendship crises, fatherhood questions, or love affairs. At the time of the recording of the TV show, the subjects were not aware that the recordings were ever going to be analysed in scientific studies. The emotion within the VAM corpus is described in terms of three dimensions: activation, valence, and dominance/power[4].

During annotation, raters used an icon-based method which let them choose an image from an array of five images for each emotion dimension. Each annotator had to listen to each utterance (manually segmented prior to the rating) and then choose an icon for each emotion dimension that best described the emotion in that utterance. The choice of these icons was afterwards mapped onto a five category scale for each dimension evenly distributed across the range $[-1; 1]$ and averaged over annotators under consideration of a weighting function that accounts for annotator certainty as described by [52]. To enable comparative evaluations here, the continuous valence and activation labels were discretised to four classes which represent the four quadrants of the activation-valence space (q1, q2, q3, and q4, corresponding to positive-active, positive-passive, negative-passive, and negative-active, respectively).

4.2 Common mapping of emotions

In order to be able to compare results and feature set performance across all the data-sets (cf. [20]), the corpus specific affect labels were mapped to a common binary arousal and valence representation (cf. [53]) as suggested by [43], [47] and [49] (for GEMEP). THe mapping for SING was performed in analogy to the procedure used for GEMEP. Table 1 gives the mapping of emotion categories to binary activation and valence labels.

4.3 Experimental Protocol

All experiments, except those on AIBO, are performed using the Leave-One-Speaker(group)-Out (LOSO) cross-validation. Thereby, if the number of speakers in the corpus is smaller or equal to eight (only for SING), data from each speaker is seen as one cross-validation fold. For more than eight speakers, the speaker IDs are arranged randomly into 8 speaker groups and the data is partitioned into eight folds according to this grouping. The cross-validation is then performed by training eight different models, each on data from 7 folds, leaving out the first fold for testing for the first model, the second fold for testing for the second model, and so on. In this way predictions for the whole data-set are produced without an overlap in training and testing data. For FAU AIBO, a two fold cross-validation is used, i. e., training on OHM and evaluating on MONT and the inverse, i. e., training on MONT and evaluating on OHM.

As classifier, the most widely used static classifier in the field of paralinguistics is chosen: Support-Vector Machines (SVMs). The SVMs are trained with the Sequential Minimal Optimisation algorithm as implemented in WEKA [54]. A range of values for the model complexity C are evaluated, and results are averaged over the full range in order to obtain more stable results wrt. to the performance of the parameter set. The range spans 17 C values according to the following scheme: $C_1 = 0.000025$, $C_2 = 0.00005$, $C_3 = 0.000075$, $C_4 = 0.0001$, ..., $C_{15} = 0.075$, $C_{16} = 0.1$, $C_{17} = 0.25$.

Each training partition is balanced in order to have the same number of instances for each class. This is required for the implementation of SVMs [54] used here to avoid learning an a priori bias for the majority classes in the model. Up-sampling is employed for this purpose, i. e., randomly selected instances in the minority classes are duplicated until the same number of instances as in the majority class is reached.

For SVMs to be numerically efficient, all acoustic parameters must be normalised to a common value range. To this end, z-normalisation, i. e., a normalisation to 0 mean and unit variance is performed. Three different methods for computing (and applying) the normalisation parameters are investigated in this article: a) computing the means and variances from the whole training partition (*std*), b) computing the means and variances individually for each speaker (*spkstd*) similarly to [55], and c) computing the means and variances individually for the training and test partitions (*stdI*).

4.4 Results

We compare the results obtained with the proposed minimalistic parameter sets with large state-of-the-art brute-forced parameter sets from the series of Interspeech Challenges on Emotion in 2009 [43] (InterSp09), Age and Gender as well as Level of Interest in 2010 [36] (InterSp10), Speaker States in 2011 [44] (InterSp11), Speaker Traits in 2012 [45] (InterSp12), and Computational Paralinguistics in 2013 and 2014 [12], [37] (ComParE).

Table 2 shows the summarised results obtained for binary arousal and binary valence classification. In order to eliminate all variables except the parameter set, the results are averaged over five databases (all,

4. the dominance dimension is not used in this study, as it was found to be highly correlated with activation.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2457417, IEEE Transactions on Affective Computing

Corpus	Activation		Valence	
	low	high	negative	positive
FAU AIBO	-		NEG	IDL
TUM AVIC	loi1	loi2, loi3	loi1	loi2, loi3
EMO-DB	boredom, disgust, neutral, sadness	anger, fear, happiness	angry, sad	happy, neutral, surprise
GEMEP	pleasure, relief, interest, irritation, anxiety, sadness	joy, amusement, pride, hot anger, panic fear, despair	hot anger, panic fear, despair, irritation, anxiety, sadness	joy, amusement, pride, pleasure, relief, interest
SING	neutral, tenseness, sadness, tenderness, calm/serenity, condescension	fear, passionate love, animated joy, triumphant pride, anger	fear, tense arousal, anger, sadness, condescension	neutral, passionate love, animated joy, triumphant pride, tenderness, calm/serenity
VAM	q2, q3	q1, q4	q3, q4	q1, q2

TABLE 1
Mapping of data-set specific emotion categories to binary activation labels (low/high) and binary valence labels (negative/positive). Note, that for FAU AIBO, due to the nature of the original 5 class labels, only a mapping to binary valence is feasible.

except FAU AIBO) and the highest 9 SVM complexity settings, starting at $C = 0.0025$. The decision to average only over the higher complexity settings was taken because at complexities lower than this threshold, performance drops significantly for the smaller feature sets, which biases the averaging.

Parameter Set	average UAR	
	Arousal	Valence
GeMAPS	79.59	65.32
eGeMAPS	**79.71**	66.44
InterSp09	76.08	64.88
InterSp10	76.50	64.44
InterSp11	76.43	65.96
InterSp12	77.26	66.71
ComParE	78.00	**67.17**

TABLE 2
Leave-one-speaker out classification of binary arousal/valence. UAR averaged over all databases (except FAU AIBO) and 9 highest SVM complexities $C \geq 0.0025$ – set text (both unweighted averages). Per speaker standardisation, instance up-sampling for balancing of training set.

Database	Best parameter set	Best UAR [%] with:		
		best set	GeMAPS	eGeMAPS
FAU AIBO	ComParE	**43.1**[5]	40.4	41.5
TUM-AVIC	InterSp12	**69.4**	68.8	68.5
EMO-DB	ComParE	**86.0**	80.0	81.1
GEMEP	InterSp12	**43.6**	36.9	38.5
SING	ComParE	**38.8**	29.4	34.0
VAM	InterSp12	**43.9**	38.5	38.9
EMO-DB (A)	InterSp09	**97.8**	95.1	95.3
GEMEP (A)	**eGeMAPS**	**84.6**	84.5	**84.6**
SING (A)	ComParE	**77.2**	75.5	75.1
VAM (A)	InterSp11	**77.4**	74.7	75.3
FAU AIBO (V)	InterSp10	**76.2**[5]	73.1	73.4
TUM-AVIC (V)	InterSp11	**75.9**	73.1	73.4
EMO-DB (V)	ComParE	**86.7**	77.1	78.1
GEMEP (V)	InterSp10	**71.4**	64.3	65.6
SING (V)	**eGeMAPS**	**67.8**	66.5	**67.8**
VAM (V)	**eGeMAPS**	**54.1**	53.2	**54.1**

TABLE 3
Leave-one-speaker out classification of affective categories of each database (see each database for description) and binary arousal (A) and valence (V). UAR obtained with best SVM complexity C. Per speaker standardisation, instance up-sampling for balancing of training set.

A high efficiency of the GeMAPS sets is shown by the average results. The eGeMAPS set performs best for arousal, reaching almost 80% UAR, while it is third best for arousal (close behind the two largest sets - ComParE and the Interspeech 2012 speaker trait set).

When looking at individual results (Table 3), i. e., when selecting the best C value for each feature set and database, the GeMAPS sets are outperformed for the classification of categories always by the large ComParE or Interspeech 2012 sets, and are outperformed in many cases by the Interspeech 2009–2011 sets for binary arousal and valence classification. More detailed results are given in plots in the Appendix (Section 6.2). The eGeMAPS set gives the best result for binary arousal classification on the GEMEP database and for binary valence classification on the SING database. However, it can be concluded that the eGeMAPS set is always superior or equal to the GeMAPS set, which is an indication that the additional parameters (MFCC and spectral flux in particular) are important. This is in particular the case for valence where the average difference between GeMAPS and eGeMAPS is larger, suggesting the importance of those parameters for acoustic valence. Yet, also for valence, the difference between the GeMAPS sets and the large Interspeech Challenge sets (esp. ComParE with its 6 373 parameters) is large compared with arousal (except for the databases SING and VAM – again, the latter not being representative for valence; SING contains sung speech, which is different in nature). Again, this suggests that for valence further

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2457417, IEEE Transactions on Affective Computing

important parameters must be identified in future work, starting with a deep parameter analysis of the ComParE set, for example.

Although slightly behind the large-scale parameter sets on average, overall, the GeMAPS sets show remarkably comparable performance given their minimalistic size of less than 2% of the largest (ComParE) set. In future studies it should be investigated, whether the proposed minimalistic sets are able to obtain better generalisation in cross-database classification experiments.

5 Discussion and Conclusion

One of the essential preconditions for accumulation of knowledge in science is the agreement on fundamental methodological procedures, specifically the nature of the central variables and their measurement. This condition is hard to achieve, even in a single discipline, let alone in interdisciplinary endeavors. In consequence, the initiative described in this article, carried out by leading researchers in different disciplines interested in the objective measurement of acoustic parameters in affective vocalizations is an important step in the right direction. It will make the replication of results across different studies far more convincing, given the direct comparability of parameters that have often been labeled differently and often measured in non-standardized fashion. As the instrument that embodies the minimal acoustic parameter set is open-source and thus readily available it could also lead to a higher degree of sophistication in a complex research domain. It is important to underline that the GeMAPS has been conceived as an open, constantly evolving system, encouraging contributions by the research community both with respect to the number and definition of specific parameters as well as the algorithms used to extract them from the speech wave. From the start, great emphasis has been placed on the stringent evaluation of the contribution of the parameters to explain variance in empirical corpora and thus it is hoped that GeMAPS will become a standard measurement rod in new work on affective speech corpora and voice analysis.

GeMAPS is based on an automatic extraction system which extracts an acoustic paramter set from an audio waveform without manual interaction or correction. Not all parameters which have been found to be relevant or correlated to certain phenomena can be reliably extracted automatically though. For example a vowel-based formant analysis requires a reliable automatic vowel detection and classification system. Thus, with GeMAPS, only those parameters which can be extracted reliably and without supervision in clean acoustic conditions have been included.

Another potential danger of automatic extraction of a standard parameter set is that the link to production phenomena may be neglected. In choosing the parameter set we have taken care to highlight these links and use the underlying vocal mechanisms as one of the criteria for collection. It is expected that further research will strengthen these underpinnings and provide new insights. For instance, it seems reasonable to expect that arousal is associated with quick phonatory and/or articulatory gestures, and that a peaceful character results from slow gestures [56]. In the future, therefore, it would be worthwhile to expand our understanding of the acoustic output of affective phonation beyond sound level, pitch and other basic parameters with to the underlying, physiologically relevant parameters. In this context glottal adduction is a particularly relevant parameter. Increasing adduction has the effect of lengthening the closed phase and decreasing the amplitude of the transglottal airflow pulses. Acoustically, this should result in attenuation of the voice source fundamental, or, more specifically, in reducing the level difference between the two lowest voice source partials. In the radiated sound this level difference is affected also by the frequency of the first formant mainly, which may be of secondary importance to the affective coloring of phonation. The future development of the GeMAPS could include the addition of techniques for inverse filtering the acoustic output signal to directly measure voice source parameter (see e.g., [57]). Such analysis of affective vocalization can allow determination of physiological correlates of various characteristics of the acoustic output [7], [58] and thus strengthen our knowledge about the mechanisms whereby emotional arousal affects voice production.

Acknowledgments

We would like to thank Elisabeth André, Tanja Bänziger, Pascal Belin, Carlos Busso, Laurence Devillers, Julien Epps, Olivier Lartillot, Petri Laukka, Shrikanth Narayanan, and Khiet Truong for their helpful contributions and inspiring discussions at the Geneva Bridge Meeting, which started our effort to create this joint parameter set recommendation.

This research was supported by an ERC Advanced Grant in the European Community's 7th Framework Programme under grant agreement 230331-PROPEREMO (Production and perception of emotion: an affective sciences approach) to Klaus Scherer and by the National Center of Competence in Research (NCCR) Affective Sciences financed by the Swiss National Science Foundation (51NF40-104897) and hosted by the University of Geneva.

References

[1] K. R. Scherer, "Vocal affect expression: A review and a model for future research," *Psychological Bulletin*, vol. 99, pp. 143–165, 1986.

[2] R. Banse and K. R. Scherer, "Acoustic Profiles in Vocal Emotion Expression," *Journal of Personality and Social Psychology*, vol. 70, no. 3, pp. 614–636, 1996.

[3] P. N. Juslin and P. Laukka, "Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion," *Emotion*, vol. 1, pp. 381–412, 2001.

[4] S. Yildirim, M. Bulut, C. Lee, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, "An acoustic study of emotions expressed in speech," in *Proc. of the 8th International Conference on Spoken Language Processing (ICSLP 04)*, Jeju Island, Korea, Oct. 2004, pp. 2193–2196.

[5] E. Moore, M. Clements, J. Peifer, and L. Weisser, "Critical analysis of the impact of glottal features in the classification of clinical depression in speech," *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 1, pp. 96–107, Jan. 2008.

[6] C. Busso, S. Lee, and S. Narayanan, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 4, pp. 582–596, May 2009.

[7] J. Sundberg, S. Patel, E. Bjorkner, and K. R. Scherer, "Interdependencies among voice source parameters in emotional speech," *IEEE Transactions on Affective Computing*, vol. 2, no. 3, pp. 162–174, Jul. 2011.

[8] T. F. Yap, "Production under cognitive load: Effects and classification," Dissertation, The University of New South Wales, Sydney, Australia, 2012.

[9] P. N. Juslin and P. Laukka, "Communication of emotions in vocal expression and music performance: Different channels, same code?" *Psychological Bulletin*, vol. 129, no. 5, pp. 770–814, Sep. 2003.

[10] S. Patel and K. R. Scherer, *Vocal behaviour*. Berlin: Mouton-DeGruyter, 2013, pp. 167–204.

[11] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[12] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani *et al.*, "The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism," in *Proc. of INTERSPEECH*. Lyon, France: ISCA, 2013, pp. 148–152.

[13] M. Schröder, *Speech and Emotion Research. An Overview of Research Frameworks and a Dimensional Approach to Emotional Speech Synthesis*, ser. Reports in Phonetics, University of the Saarland. Institute for Phonetics, University of Saarbrücken, 2004, vol. 7.

[14] M. Schröder, F. Burkhardt, and S. Krstulovic, "Synthesis of emotional speech," in *Blueprint for Affective Computing*, K. R. Scherer, T. Bänziger, and E. Roesch, Eds. Oxford: Oxford University Press, 2010, pp. 222–231.

[15] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Communication*, vol. 40, pp. 227–256, 2003.

[16] K. Hammerschmidt and U. Jürgens, "Acoustical correlates of affective prosody," *Journal of Voice*, vol. 21, pp. 531–540, 2007.

[17] M. Goudbeek and K. R. Scherer, "Beyond arousal: Valence and potency/control cues in the vocal expression of emotion," *Journal of the Acoustical Society of America (JASA)*, vol. 128, pp. 1322–1336, 2010.

[18] D. A. Sauter, F. Eisner, A. J. Calder, and S. K. Scott, "Perceptual cues in nonverbal vocal expressions of emotion," *Quarterly Journal of Experimental Psychology*, vol. 63, pp. 2251–2272, 2010.

[19] P. Laukka and H. A. Elfenbein, "Emotion appraisal dimensions can be inferred from vocal expressions," *Social Psychological and Personality Science*, vol. 3, pp. 529–536, 2012.

[20] B. Schuller, B. Vlasenko, F. Eyben, M. Wöllmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-corpus acoustic emotion recognition: Variances and strategies," *IEEE Transactions on Affective Computing (TAC)*, vol. 1, no. 2, 2010.

[21] A. Batliner, S. Steidl, B. Schuller, D. Seppi, K. Laskowski, T. Vogt, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson, "Combining Efforts for Improving Automatic Classification of Emotional User States," in *Proceedings 5th Slovenian and 1st International Language Technologies Conference, ISLTC 2006*. Ljubljana, Slovenia: Slovenian Language Technologies Society, October 2006, pp. 240–245.

[22] D. Bone, C.-C. Lee, and S. Narayanan, "Robust unsupervised arousal rating:a rule-based framework with knowledge-inspired vocal features," *IEEE Transactions on Affective Computing*, vol. 5, no. 2, pp. 201–213, April 2014.

[23] B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson, "The Relevance of Feature Type for the Automatic Classification of Emotional User States: Low Level Descriptors and Functionals," in *Proceedings INTERSPEECH 2007, 8th Annual Conference of the International Speech Communication Association*, ISCA. Antwerp, Belgium: ISCA, August 2007, pp. 2253–2256.

[24] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer, "On the Acoustics of Emotion in Audio: What Speech, Music and Sound have in Common," *Frontiers in Emotion Science*, vol. 4, no. Article ID 292, pp. 1–12, May 2013.

[25] F. Eyben, F. Weninger, and B. Schuller, "Affect recognition in real-life acoustic conditions – A new perspective on feature selection," in *Proc. of INTERSPEECH 2013*. Lyon, France: ISCA, Aug. 2013, pp. 2044–2048.

[26] M. Tahon and L. Devillers, "Acoustic measures characterizing anger across corpora collected in artificial or natural context," in *Proc. of Speech Prosody*, Chicago, IL, USA, 2010.

[27] T. F. Yap, J. Epps, E. Ambikairajah, and E. H. C. Choi, "Formant frequencies under cognitive load: Effects and classification," *EURASIP Journal on Advances in Signal Processing*, vol. 2011, pp. 1:1–1:11, Jan. 2011. [Online]. Available: http://dx.doi.org/10.1155/2011/219253

[28] A. C. Trevino, T. F. Quatieri, and N. Malyska, "Phonologically-based biomarkers for major depressive disorder," *EURASIP Journal on Advances in Signal Processing*, pp. 1–18, 2011.

[29] L. Tamarit, M. Goudbeek, and K. R. Scherer, "Spectral slope measurements in emotionally expressive speech," in *Proc. of SPKD-2008*. ISCA, 2008, paper 007.

[30] P. Le, E. Ambikairajah, J. Epps, V. Sethu, and E. H. C. Choi, "Investigation of spectral centroid features for cognitive load classification," *Speech Communication*, vol. 54, no. 4, pp. 540–551, 2011.

[31] N. Cummins, J. Epps, M. Breakspear, and R. Goecke, "An investigation of depressed speech detection: Features and normalizatio," in *Proc. of Interspeech 2011, Florence, Italy*, 2011, pp. 2997–3000.

[32] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Communication*, vol. 48, no. 9, pp. 1162–1181, Sep. 2006.

[33] B. Schuller, M. Wimmer, L. Mösenlechner, C. Kern, D. Arsić, and G. Rigoll, "Brute-Forcing Hierarchical Functionals for Paralinguistics: a Waste of Feature Space?" in *Proceedings 33rd IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2008*, IEEE. Las Vegas, NV: IEEE, April 2008, pp. 4501–4504.

[34] F. Eyben, A. Batliner, and B. Schuller, "Towards a standard set of acoustic features for the processing of emotion in speech," *Proceedings of Meetings on Acoustics (POMA)*, vol. 9, no. 1, pp. 1–12, Jul. 2012.

[35] T. F. Yap, J. Epps, E. Ambikairajah, , and E. H. C. Choi, "Voice source features for cognitive load classification," in *Proc. of ICASSP 2011, Prague, Czech Republic*. IEEE, 2011, pp. 5700–5703.

[36] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "The INTERSPEECH 2010 Paralinguistic Challenge," in *Proceedings INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association*, ISCA. Makuhari, Japan: ISCA, September 2010, pp. 2794–2797.

[37] B. Schuller, S. Steidl, A. Batliner, J. Epps, F. Eyben, F. Ringeval, E. Marchi, and Y. Zhang, "The interspeech 2014 computational paralinguistics challenge: Cognitive and physical load," in *Proc. of INTERSPEECH 2014, Singapore*. ISCA, 2014, to appear.

[38] J. R. Williamson, T. F. Quatieri, B. S. Helfer, R. Horwitz, B. Yu, and D. D. Mehta, "Vocal biomarkers of depression based on motor incoordination," in *Proc. 3rd ACM Int. Workshop on Audio/Visual Emotion Challenge*. ACM, 2013.

[39] V. Sethu, E. Ambikairajah, and J. Epps, "On the use of speech parameter contours for emotion recognition," *EURASIP Journal on Audio, Speech and Music Processing (JASMP)*, 2013.

[40] B. Schuller and G. Rigoll, "Recognising Interest in Conversational Speech – Comparing Bag of Frames and Supra-segmental Features," in *Proceedings INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication*

Association, ISCA. Brighton, UK: ISCA, September 2009, pp. 1999–2002.
[41] E. Marchi, A. Batliner, B. Schuller, S. Fridenzon, S. Tal, and O. Golan, "Speech, Emotion, Age, Language, Task, and Typicality: Trying to Disentangle Performance and Feature Relevance," in *Proceedings First International Workshop on Wide Spectrum Social Signal Processing (WS^{3}P 2012), held in conjunction with the ASE/IEEE International Conference on Social Computing (SocialCom 2012)*, ASE/IEEE. Amsterdam, The Netherlands: IEEE, September 2012.
[42] S. Steidl, *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech*. Berlin: Logos Verlag, 2009.
[43] B. Schuller, S. Steidl, A. Batliner, and F. Jurcicek, "The INTERSPEECH 2009 Emotion Challenge," in *Proc. of INTERSPEECH*, Brighton, UK, Sep. 2009, pp. 312–315.
[44] B. Schuller, A. Batliner, S. Steidl, F. Schiel, and J. Krajewski, "The INTERSPEECH 2011 Speaker State Challenge," in *Proc. of INTERSPEECH*, Florence, Italy, Aug. 2011, pp. 3201–3204.
[45] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, "The INTERSPEECH 2012 Speaker Trait Challenge," in *Proceedings INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association*, ISCA. Portland, OR: ISCA, September 2012.
[46] B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker, and H. Konosu, "Being Bored? Recognising Natural Interest by Extensive Audiovisual Integration for Real-Life Application," *Image and Vision Computing, Special Issue on Visual and Multimodal Analysis of Human Spontaneous Behavior*, vol. 27, no. 12, pp. 1760–1774, November 2009.
[47] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth, "Acoustic emotion recognition: A benchmark comparison of performances," in *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU) 2009, Merano, Italy*. IEEE, Nov. 2009, pp. 552–557.
[48] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Proceedings Interspeech 2005, Lissabon, Portugal*, 2005, pp. 1517–1520.
[49] T. Bänziger, M. Mortillaro, and K. R. Scherer, "Introducing the Geneva Multimodal Expression corpus for experimental research on emotion perception," *Emotion*, vol. 12, no. 5, pp. 1161–1179, 2012.
[50] K. R. Scherer, J. Sundberg, L. Tamarit, and G. L. Salom ao, "Comparing the acoustic expression of emotion in the speaking and the singing voice," *Computer Speech and Language*, vol. 01, 2013.
[51] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German Audio-Visual Emotional Speech Database," in *Proc. of the IEEE International Conference on Multimedia and Expo (ICME)*, Hannover, Germany, 2008, pp. 865–868.
[52] M. Grimm, E. Mower, K. Kroschel, and S. Narayanan, "Primitives based estimation and evaluation of emotions in speech," *Speech Communication*, vol. 49, pp. 787–800, 2007.
[53] H. Schlosberg, "Three dimensions of emotion," *Psychology Review*, vol. 61, pp. 81–88, 1954.
[54] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.
[55] V. Sethu, E. Ambikairajah, , and J. Epps, "Speaker normalization of speech-based emotion recognition," in *Proc. of the IEEE International Conference on Digital Signal Processing, Cardiff, UK*, 2007, pp. 611–614.
[56] I. Fonagy, "La vive voix," *Payo*, 1983.
[57] M. Airas, H. Pulakka, T. Bäckström, and P. Alku, "A toolkit for voice inverse filtering and parametrisation," in *Proc. of Interspeech 2005*. ISCA, 2005, pp. 2145–2148.
[58] S. Patel, K. R. Scherer, J. Sundberg, and E. Björkner, "Mapping emotions into acoustic space: The role of voice production," *Biological Psychology*, vol. 87, pp. 93–98, 2011.
[59] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," in *Proc. of ACM MM 2013, Barcelona, Spain*. New York, NY, USA: ACM, 2013, pp. 835–838.
[60] D. J. Hermes, "Measurement of pitch by subharmonic summation," *Journal of the Acoustical Society of America*, vol. 83, no. 1, pp. 257–264, 1988.
[61] B. Schuller, *Intelligent Audio Analysis*, ser. Signals and Communication Technology. Springer, 2013.
[62] E. Zwicker and H. Fastl, *Psychoacoustics – Facts and Models*. Springer, 1999.
[63] H. Hermansky, "Perceptual linear predictive (plp) analysis for speech," *Journal of the Acoustical Society of America*, vol. 87, pp. 1738–1752, 1990.
[64] J. Makhoul and L. Cosell, "Lpcw: An lpc vocoder with linear predictive spectral warping," in *Proc. of ICASSP'76, Philadelphia, USA*. IEEE, 1976, pp. 466–469.
[65] B. Hammarberg, B. Fritzell, J. Gauffin, J. Sundberg, and L. Wedin, "Perceptual and acoustic correlates of abnormal voice qualities," *Acta Otolaryngologica*, vol. 90, pp. 441–451, 1980.
[66] S. Patel, K. R. Scherer, J. Sundberg, and E. Björkner, "Acoustic markers of emotions based on voice physiology," in *Proc. of Speech Prosody 2010, Chicago, IL, USA*, 2010.
[67] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 5, pp. 561–580, Apr. 1975.
[68] S. Young, G. Evermann, M. Gales, T. Hain, D. , X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book for HTK Version 3.4*. Cambridge University Engineering Department, 2006.

6 APPENDIX

6.1 Implementation Details

All the parameters are extracted with the open-source toolkit openSMILE [59]. In the latest version (2.1[5]) of this toolkit, configuration files and a graphical user interface are provided which can be used to extract both the minimalistic and the extended set "out-of-the-box". Further, it is also possible to only extract the LLD without the summarisation over segments by the functionals. This ensures that teams across the world, who are working with these standard parameter sets are able to use a common implementation of these descriptors as a starting point for further analysis, such as statistical inspection of corpora, or machine learning experiments for various affective computing and paralinguistics tasks.

The remainder of this section describes details of the LLD extraction process. Full details and descriptions of the algorithms are found in the supplementary material provided with this article .

All input audio samples are scaled to the range $[-1; +1]$ and stored as 32-bit floating point numbers, in order to work with normalised values regardless of the actual bit-depth of the inputs. F_0, harmonic differences, HNR, jitter, and shimmer are computed from overlapping windows which are 60 ms long and 10 ms apart. The frames are multiplied with a Gaussian window with $\sigma = 0.4$ in the time domain prior to the transformation to the frequency domain (with an FFT) – for jitter and shimmer, which are computed in the time domain, no window function is applied. Loudness, spectral slope, spectral energy proportions, Formants, Harmonics, Hammarberg Index, and Alpha Ratio are computed from 20 ms windows which are 10 ms apart; a Hamming function is applied to these windows. Zero-padding is applied to all windows to

5. to be released to the public together with this article

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2457417, IEEE Transactions on Affective Computing

the next power-of-2 (samples) frame size in order to be able to efficiently perform the FFT.

The **fundamental frequency** (F_0) is computed via sub-harmonic summation (SHS) in the spectral domain as described by [60]. Spectral smoothing, spectral peak enhancement, and auditory weighting are applied as in [60]. 15 harmonics are considered, i. e., the spectrum is octave shift-added 15 times, and a compression factor of 0.85 is used at each shifting ([60]). $F_0 = 0$ is defined for unvoiced regions. The voicing probability is determined by the ratio of the harmonic summation spectrum peak belonging to an F_0 candidate and the average amplitude of all harmonic summation spectrum bins, scaled to a range $[0, 1]$. A maximum of 6 F_0 candidates in the range of 55–1000 Hz are selected. On-line Viterbi post-smoothing is applied to select the most likely F_0 path through all possible candidates. A voicing probability threshold of 0.7 is then applied to discern voiced from unvoiced frames. After Viterbi smoothing the F_0 range of 55–1000 Hz is enforced by setting all voiced frames outside the range to unvoiced frames ($F_0 = 0$). The final F_0 value is converted from its linear Hz scale to a logarithmic scale – a semitone frequency scale starting at 27.5 Hz (semitone 0). However, as 0 is reserved for unvoiced frames, every value below semitone 1 (29.136 Hz) is clipped to 1.

For computing **jitter** and **shimmer** it is required to know the exact locations and lengths of individual pitch periods. The SHS algorithm described above delivers only an average F_0 value for a 60 ms window, which can contain between 4–40 (depending on the actual F_0 frequency) pitch periods in the defined range. In order to determine the exact lengths of the individual pitch periods, a correlation based waveform matching algorithm is implemented. The matching algorithm uses the frame average estimate of $T_0 = 1/F_0$ found via the SHS algorithm, to limit the range of the period cross-correlation to improve both the robustness against noise and computational efficiency. The waveform matching algorithm operates directly on unwindowed 60 ms audio frames.

Jitter, is computed as the average (over one 60 ms frame) of the absolute local (period to period) jitter $J_{pp}(n')$ scaled by the average fundamental period length. For two consecutive pitch periods, with the length of the first period $n'-1$ being $T_0(n'-1)$ and the length of the second period n' being $T_0(n')$, the absolute period to period jitter, also referred to as absolute local jitter, is given as follows [61]:

$$J_{pp}(n') = |T_0(n') - T_0(n'-1)| \text{ for } n' > 1. \quad (1)$$

This definition yields one value for J_{pp} for every pitch period, starting with the second one. To obtain a single jitter value per frame for N' local pitch periods $n' = 1 \ldots N'$ within one analysis frame, the average local jitter $\overline{J_{pp}}$ is given by:

$$\overline{J_{pp}} = \frac{1}{N'-1} \sum_{n'=2}^{N'} |T_0(n') - T_0(n'-1)|. \quad (2)$$

In order to make the jitter value independent of the underlying pitch period length, it is scaled by the average pitch period length. This yields the average relative jitter, used as the jitter measure in our parameter set:

$$\overline{J_{pp,rel}} = \frac{\frac{1}{N'-1} \sum_{n'=2}^{N'} |T_0(n') - T_0(n'-1)|}{\frac{1}{N'} \sum_{n'=1}^{N'} T_0(n')}. \quad (3)$$

Shimmer is computed as average (over on frame) of the relative peak amplitude differences expressed in dB. Because the phase of the pitch period segments found by the waveform matching algorithm is random, the maximum and minimum amplitude ($x_{max,n'}$ and $x_{min,n'}$) within each pitch period are identified. By analogy with jitter, the local period to period shimmer is expressed as:

$$S_{pp}(n') = |A(n') - A(n'-1)|, \quad (4)$$

with the peak to peak amplitude difference $A(n') = x_{max,n'} - x_{min,n'}$.

As for jitter, the period to period shimmer values are averaged over each 60 ms frame in order to synchronise the rate of this descriptor with the constant rate of all other short-time descriptors. The averaged, relative shimmer is referred to as $\overline{S_{pp,rel}}$. It is expressed as amplitude ratios, i. e., the per period amplitude values are normalised to the per frame average peak amplitude:

$$\overline{S_{pp,rel}} = \frac{\frac{1}{N'-1} \sum_{n'=2}^{N'} S_{pp}(n')}{\frac{1}{N} \sum_{n'=1}^{N'} A(n')} \quad (5)$$

$$(6)$$

Loudness is used here as a more perceptually relevant [62] alternative to the signal energy. In order to approximate humans' non-linear perception of sound, an auditory spectrum as is applied in the Perceptual Linear Prediction (PLP) technique [63] is adopted. A non-linear Mel-band spectrum is constructed by applying 26 triangular filters distributed equidistant on the Mel-frequency scale from 20–8000 Hz to a power spectrum computed from a 25 ms frame. An auditory weighting with an equal loudness curve as used by [63] and originally adopted from [64] is performed. Next, a cubic root amplitude compression is performed for each band b of the equal loudness weighted Mel-band power spectrum [63]. resulting in a spectrum which is referred to as *auditory spectrum*. Loudness is then computed as the sum over all bands of the auditory spectrum.

The **equivalent sound level** (LEq) is computed by converting the average of the per-frame RMS energies to a logarithmic (dB) scale.

The **Harmonics-to-Noise Ratio (HNR)** gives the energy ratio of the harmonic signal parts to the noise signal parts in dB. It is estimated from the short-time autocorrelation function (ACF) (60 ms window) as the logarithmic ratio of the ACF amplitude at F_0 and the total frame energy, expressed in dB, as given by [61]:

$$HNR_{acf,log} = 10\log_{10}\left(\frac{ACF_{T_0}}{ACF_0 - ACF_{T_0}}\right)\text{dB}. \quad (7)$$

where ACF_{T_0} is the amplitude of the autocorrelation peak at the fundamental period (derived from the SHS-based F_0 extraction algorithm described above) and ACF_0 is the 0-th ACF coefficient (equivalent to the quadratic frame energy). The logarithmic HNR value is floored to -100 dB to avoid highly negative and varying values for low-energy noise.

The **spectral slope** for the bands 0–500 Hz and 500–1500 Hz is computed from a logarithmic power spectrum by linear least squares approximation [29]. Next to the exact spectral slope, features closely related to the spectral slope can be used. [29] describes the **Hammarberg index** in this context: It was defined by [65] as the ratio of the strongest energy peak in the 0–2 kHz region to the strongest peak in the 2–5 kHz region. Hammarberg defined a fixed static pivot point of 2 kHz where the low and high frequency regions are separated. Formally the Hammarberg index η is defined as:

$$\eta = \frac{\max_{m=1}^{m_{2k}} X(m)}{\max_{m=m_{2k}+1}^{M} X(m)}, \quad (8)$$

with $X(m)$ being a magnitude spectrum with bins $m = 1..M$, and where m_{2k} is the highest spectral bin index where $f \leq 2$ kHz is still true. According to more recent findings, e. g., [29], it could be beneficial to pick the pivot point dynamically based upon the speaker's fundamental frequency. This is, however, on purpose not considered here because it would break the strictly static nature of all the extraction methods of all the parameters suggested for this set.

Similar to the Hammarberg index, the **Alpha Ratio** [66] is defined as the ratio between the energy in the low frequency region and the high frequency region. More specifically, it is the ratio between the summed energy from 50–1000 Hz and 1–5 kHz.

$$\rho_\alpha = \frac{\sum_{m=1}^{m_{1k}} X(m)}{\sum_{m=m_{1k}+1}^{M} X(m)} \quad (9)$$

where m_{1k} is the highest spectral bin index where $f \leq 1$ kHz is still true. In applications of emotion recognition from speech, this parameter most often – like other spectral slope related parameters – is computed from a logarithmic representation of a band-wise long-term average spectrum (LTAS, cf. [50], [66]). Here, however, in order be able to provide all parameters on a frame level, the alpha ratio is computed per frame (20 ms) and then, the functionals mean and variance are applied to summarise it over segments of interest.

Both **formant bandwidth** and **formant centre frequency** are computed from the roots of Linear Predictor (LP) [67] coefficient polynomial. The algorithm follows the implementation of [11].

The **formant amplitude** is estimated as the amplitude of the spectral envelope at F_i in relation to the amplitude of the spectral F_0 peak. More precisely, it is computed as the ratio of the amplitude of the highest F_0 harmonic peak in the range $[0.8{\cdot}F_i; 1.2{\cdot}F_i]$ (F_i is the centre frequency of the first formant) to the amplitude of the F_0 spectral peak.

Similarly, **harmonic differences** or *harmonic ratios*, are computed from the amplitudes of F_0 harmonic peaks in the spectrum normalised by the amplitude of the F_0 spectral peak. In the proposed parameter set, in particular the ratios H1–H2, i. e., the ratio of the first to the second harmonic, and H1–A3, which is the ratio of the first harmonic to the third formant's amplitude (as described in the previous paragraph).

Spectral Energy Proportions are computed from the linear frequency scale power spectrum by summing the energy of all bins in the bands 0–500 Hz and 0–1000 Hz and normalising by the total frame energy (sum of all power spectrum bins).

The first four **Mel-Frequency Cepstral Coefficients (MFCC)** (1–4) are computed as described by [68] from a 26-band power Mel-spectrum (20–8000 Hz). In contrast to all other descriptors, the audio samples are not normalised to $[-1; +1]$, but to the range of a signed 16-bit integer in order to maintain compatibility with [68]. Liftering of the cepstral coefficients with $L = 22$ is performed.

The **spectral flux** S_{flux} represents a quadratic, normalised version of the simple spectral difference, i. e., the bin-wise difference between the spectra of two consecutive speech frames. The definition of the unnormalised spectral flux for frame k and magnitude spectra $X(m)$ is as follows:

$$S_{flux}^{(k)} = \sum_{m=m_l}^{m_u} \left(X^{(k)}(m) - X^{(k-1)}(m)\right)^2, \quad (10)$$

where m_l and m_u are the lower and upper bin indices of the spectral range to be considered for spectral flux computation. Here, they are set such that the spectral range is set to 0 – 5000 Hz.

6.2 Detailed Results

This section shows detailed results in plots which compare all investigated acoustic parameter sets for each database over a range of SVM complexity constants. For details on the experimental set-up, please refer to Section 4.3.

Results for the TUM-AVIC database are shown in Figure 1, for EMO-DB in Figure 2, for GEMEP in

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2015.2457417, IEEE Transactions on Affective Computing

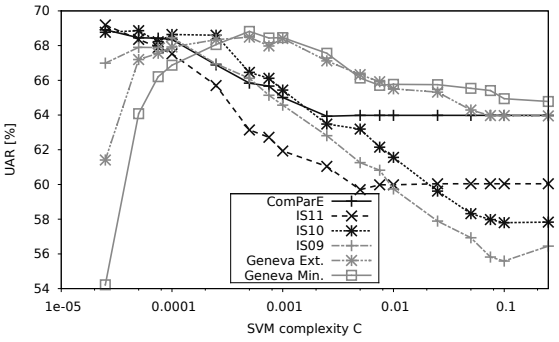

Fig. 1. Individual results (UAR [%] vs. SVM complexity – all 17 values, see Section 4.3) for the TUM-AVIC database (categories: 3 levels of interest).

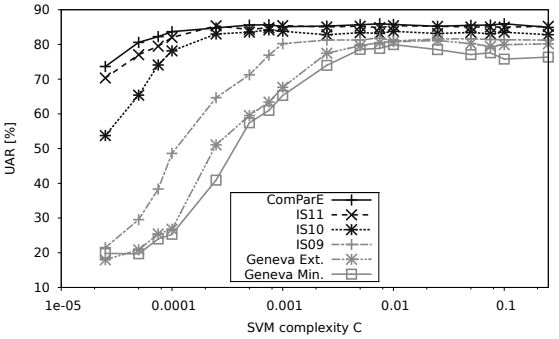

Fig. 2. Individual results (UAR [%] vs. SVM complexity – all 17 values, see Section 4.3) for the EMO-DB database (categories: 6 basic emotions and neutral).

Figure 3, for SING in Figure 4, and for the VAM database in Figure 5.

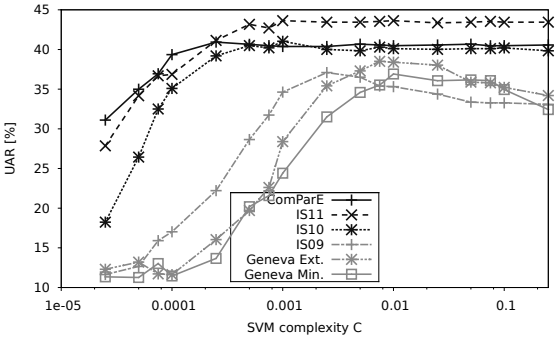

Fig. 3. Individual results (UAR [%] vs. SVM complexity – all 17 values, see Section 4.3) for the GEMEP database (categories: 12 emotions).

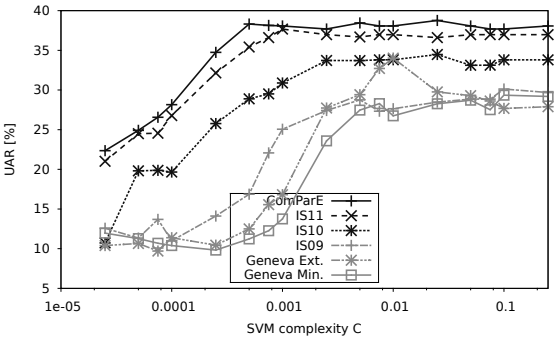

Fig. 4. Individual results (UAR [%] vs. SVM complexity – all 17 values, see Section 4.3) for the Geneva Singing Voice (SING) database (categories: 11 sung emotions).

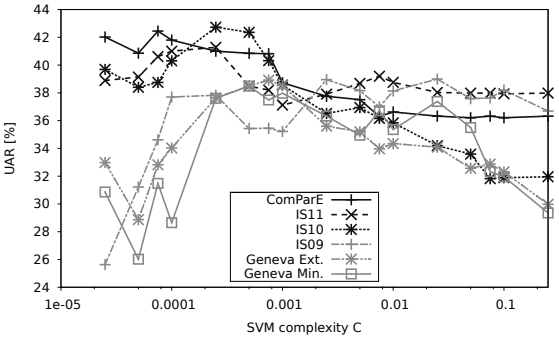

Fig. 5. Individual results (UAR [%] vs. SVM complexity – all 17 values, see Section 4.3) for the Vera-am-Mittag (VAM) database (categories: 4 quadrants of the arousal/valence space).