## INTRODUCTION

There are no standard agreed-upon acoustic features for the automatic recognition of human emotions. This is because: (1) few data exist; (2) there is a large range and set of emotions, requiring different sets of acoustic features for discrimination; (3) many of the features considered are multifunctional, serving not only as markers for emotions but also for linguistic and speaker characteristics.

Moreover, there are not enough cross-corpora and cross-linguistic studies that support a universally-agreed-upon standard set of acoustic features for automatic emotion recognition. Most studies are carried out on single data sets, where the types of emotions, the speakers, and the acoustic conditions are the same throughout. The few existing studies which deal with feature-set evaluation on multiple databases usually consider mostly two [1] to three [2] or maximally four [3, 4] databases at a time. A universal acoustic feature set, however, needs to be suitable for emotion recognition from speech independently of the speaker, the domain, the type of emotions, and the recording conditions.

A number of acoustic features have been employed for automatically classifying emotions. These include: prosodic (suprasegmental) features (pitch, intensity, duration, rhythm) and spectral features (Mel-Frequency Cepstral Coefficients (MFCCs) and alike, formants, spectral statistics, etc.). Of these, MFCCs proved to be among the most valuable features for identifying emotions, as was found in recent comparisons (INTERSPEECH Challenges); but this can, at least partly, be due to challenge participants being well aquainted with them because they often use them in their own Automatic Speech Recognition (ASR) systems. Prosodic features (such as excursion of pitch or intensity) may be more relevant for identifying the level of arousal, while spectral features may be more relevant for identifying valence. The level of arousal indicates the "activeness" of a person. It is therefore also referred to as "activation". Examples of emotions with high arousal are happiness and hot anger. Boredom and cold anger have a low level of arousal. Valence, also known as "evaluation", refers to the pleasantness of an emotion, i. e., how positive or negative an emotion is. Anger and fear, for example, carry negative valence, while happiness and pride have positive valence.

In the present article, we use eight widely-used, publicly available databases, and apply six different, large-scale acoustic feature sets in the analysis of the emotion dimensions arousal and valence. Different emotion labels taken from the eight databases are mapped onto the two binary labels arousal (high/low) and valence (positive/negative). These two dimensions span the most basic space in which most emotions can be adequately described. We anticipate this to be a first step towards identifying which feature sets prove to be most valuable and accurate in classifying emotions along those two basic dimensions. In order to cover a broad range of available acoustic features and to ensure reproducibility and comparability, we employ large acoustic feature sets that have been used in public challenges during the last few years.

The paper is structured as follows: (a) description of the eight databases and the mapping of the emotion categories to binary arousal and valence values; (b) presentation of the six feature sets; (d) presentation and discussion of the results; and (e) conslusions.

## EIGHT EMOTIONAL SPEECH DATABASES USED

Before describing the eight corpora used in this study (cf. Table 1), we will give a brief overview on the history of emotional speech databases. In the late 1990s, the first emotional speech databases were collected for the purpose of automatic emotion analysis and synthesis. These sets were small ($\approx$ 500 sentences/phrases) and contained data from only few subjects ($\approx$ 10 speakers). The content was mostly speech read from predefined prompts (DES [5], Berlin Emotional Speech-Database [6], SUSAS [7]), the emotions were acted, and the recording was made with high quality equipment in a noise free environment. Very few annotators – if any at all – labeled the perceived emotion in few discrete categories.

Today we are happy to see more diverse emotions covered, which are elicited or even spontaneous (not acted and the text is not read from scripts), larger amounts of instances (5 k – 10 k) of more subjects (up to more than 100), multimodal data that are annotated by more labelers (4 (AVIC [8]) - 17 (VAM [9])), and databases that are made publicly available.

In this study, we chose eight among the most popular, publicly (or upon request) available databases, mixing old and newer databases. They cover a broad variety of emotional speech reaching from acted speech (the Danish (*DES*, [5]) and the Berlin Emotional Speech ($EMO-DB$, [6]) databases), over story guided such as the eNTERFACE corpus [10] and the Airplane Behaviour Corpus (ABC, [11]), to spontaneous emotions within a linguistically restricted domain (cf. Speech Under Simulated and Actual Stress (SUSAS, [7]) database), to more naturalistic corpora such as the Audiovisual Interest Corpus ($AVIC$, [8]), the Sensitive Artificial Listener (SAL, [12]), and the Vera-Am-Mittag (VAM,

**TABLE 1.** 8 speech emotion databases used in this study, given in alphabetic order: Airplance Behaviour Corpus (ABC), Audiovisual Interest Corpus (AVIC), Danish Emotional Speech (DES) database, Berlin Emotional Speech Database (EMO-DB), eNTERFACE'05 database (eNTERFACE), Sensitive Artificial Listener (SAL) database, Speech Under Simulated and Actual Stress (SUSAS) database, and the Vera-Am-Mittag (VAM) database. Emotion categories, mapped to high(+)/low(-)) arousal and positive/negative valence, and number of instances in each category. ***Abbreviations****: `agre`: aggressive, `angr`: angry, bore: boredom, chee: cheerful, disg: disgust, happ: happy, hist: high stress, into: intoxicated, loi1–3: level of interest 1–3, meds: medium stress, nerv: nervous, neut: neutral, q1–q4: quadrants in the arousal-valance plane, cf.* ***SAL*** *and* ***VAM*** *and explanation in text, sadn: sadness, surp - surprise, tire: tired*

| **Corpus** | **Emotion, Arousal/Valence Mapping, #** | | | | | | | **Language, Text, Emotion** | **time [h:mm]** | **#m / #f spk.** | **Recording conditions** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **ABC** | agre<br>+-<br>95 | chee<br>++<br>105 | into<br>+-<br>33 | nerv<br>+-<br>93 | neut<br>-+<br>79 | tire<br>--<br>25 | | German, fixed, induced | 1:15 | 4 / 4 | studio, 16 kHz |
| **AVIC** | loi1<br>--<br>553 | loi2<br>++<br>2279 | loi3<br>++<br>170 | | | | | English, free, natural | 1:47 | 11 / 10 | studio, 44 kHz |
| **DES** | angr<br>+-<br>85 | happ<br>++<br>86 | neut<br>-+<br>85 | sad<br>--<br>84 | surp<br>++<br>84 | | | Danish, fixed, acted | 0:28 | 2 / 2 | studio, 20 kHz |
| **EMO-DB** | angr<br>+-<br>127 | bore<br>--<br>79 | disg<br>--<br>38 | fear<br>+-<br>55 | happ<br>++<br>64 | neut<br>-+<br>78 | sadn<br>--<br>53 | German, fixed, acted | 0:22 | 5 / 5 | studio, 16 kHz |
| **eNTERFACE** | angr<br>+-<br>215 | disg<br>--<br>215 | fear<br>+-<br>215 | happ<br>++<br>207 | sadn<br>--<br>210 | surp<br>++<br>215 | | English, fixed, induced | 1:00 | 34 / 8 | studio, 16kHz |
| **SAL** | q1<br>++<br>459 | q2<br>-+<br>320 | q3<br>--<br>564 | q4<br>+-<br>349 | | | | English, free, natural | 1:41 | 2 / 2 | studio, 16 kHz |
| **SUSAS** | hist<br>+-<br>1202 | meds<br>+-<br>1276 | neut<br>-+<br>701 | scre<br>+-<br>414 | | | | English, fixed, natural | 1:01 | 4 / 3 | noisy, 8 kHz |
| **VAM** | q1<br>++<br>21 | q2<br>-+<br>50 | q3<br>--<br>451 | q4<br>+-<br>424 | | | | German, free, natural | 0:47 | 15 / 32 | noisy, 16 kHz |

[9]) databases. Three languages are represented by these databases: English, German, and Danish; see the Appendix for a more in-depth description.

The mapping from emotion categories to binary arousal (passive and active) and valence (positive and negative) labels can be found in Table 1. As mentioned in the introduction, this mapping is required to ensure that all databases share the same emotion labels. Not all mappings, however, are straightforward. The emotional state "neutral", for example, could be mapped to either one of the binary arousal/valence labels because it actually is a third state (neither passive nor active, neither positive nor negative) in the center of the two dimensional arousal space. The databases SAL and VAM already have emotion annotations in terms of quadrants in the arousal/valence space. Thereby the quadrants q1–q4 correspond to the emotion groups happy/exciting, relaxed/serene, sad/bored, and angry/anxious.

## FEATURE SETS AND EVALUATION

We evaluate the suitability of six feature sets for emotion recognition on all eight corpora. An acoustic feature set is a collection of audio descriptors, which potentially carry information about affective cues in the voice. Very basic examples of such descriptors are the signal energy and the pitch of the voice. A common approach in the field of emotion recognition is to use a large set of virtually all available features and let the machine learning algorithm automatically figure out the relevant features and correlations between these and the emotion label(s). This is a so

**TABLE 2.** openSMILE standard features sets by LLDs. EC: INTERSPEECH 2009 Emotion Challenge, PC: INTERSPEECH 2010 Paralinguistic Challenge, SSC: INTERSPEECH 2011 Speaker State Challenge, STC: INTERSPEECH 2011 Speaker Trait Challenge, AVEC: Audio/Visual Emotion Challenges 2011/2012. [1]Only used for the TUM AVIC baseline (PC). For the PC and SSC feature sets, as additional feature the number of voiced segments ($F_0$ onsets) were added.

| | **EC** | **PC** | **SSC** | **STC** | **AVEC11** | **AVEC12** |
|---|---|---|---|---|---|---|
| Number of descriptors | | | | | | |
| # LLDs | 16 | 38 | 59 | 64 | 31 | 31 |
| # Functionals | 12 | 21 | 39 | 40 | 42 | 38 |
| # Features | 384 | 1 582 | 4 368 | 6 125 | 1 941 | 1 841 |
| LLDs | | | | | | |
| RMS Energy | ✓ | | ✓ | ✓ | ✓ | ✓ |
| sum of auditory spectrum (loudness) | | ✓[1] | ✓ | ✓ | ✓ | ✓ |
| sum of RASTA-sytle filtered auditory spectrum | | | ✓ | ✓ | | |
| Zero-crossing rate | ✓ | | ✓ | ✓ | ✓ | ✓ |
| energy in bands from 250 – 650 Hz, 1 kHz – 4 kHz | | | ✓ | ✓ | ✓ | ✓ |
| spectral roll-off points 25 %, 50 %, 75 %, 90 % | | | ✓ | ✓ | ✓ | ✓ |
| spectral flux | | | ✓ | ✓ | ✓ | ✓ |
| spectral entropy | | | ✓ | ✓ | ✓ | ✓ |
| spectral variance | | | ✓ | ✓ | ✓ | ✓ |
| spectral skewness | | | ✓ | ✓ | ✓ | ✓ |
| spectral kurtosis | | | ✓ | ✓ | ✓ | ✓ |
| spectral slope | | | ✓ | ✓ | | |
| psychoacoustic sharpness | | | | ✓ | ✓ | ✓ |
| harmonicity | | | | ✓ | ✓ | ✓ |
| MFCC 0 | | ✓ | | | | |
| MFCC 1 – 10 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| MFCC 11 – 12 | ✓ | ✓ | ✓ | ✓ | | |
| MFCC 13 – 14 | | ✓ | | ✓ | | |
| log Mel frequency band 0 – 7 | | ✓[1] | | | | |
| LSP frequency 0 – 7 | | ✓ | | | | |
| RASTA-style auditory spectrum bands 1 – 26 (0 – 8 kHz) | | | ✓ | ✓ | | |
| $F_0$ (ACF based) | ✓ | | | | | |
| $F_0$ (SHS based) | | ✓ | | | | |
| $F_0$ (SHS based followed by Viterbi smoothing) | | | ✓ | ✓ | ✓ | ✓ |
| $F_0$ envelope | | ✓ | | | | |
| probability of voicing | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| jitter | | ✓ | ✓ | ✓ | ✓ | ✓ |
| jitter (delta: 'jitter of jitter') | | ✓ | ✓ | ✓ | ✓ | ✓ |
| shimmer | | ✓ | ✓ | ✓ | ✓ | ✓ |
| logarithmic HNR | | | | ✓ | ✓ | ✓ |

*Description of feature name abbreviations:*
RMS . . . : Root Mean-Square . . . ; MFCC: Mel-Frequency Cepstral Coefficients.
RASTA: Refers to a technique of bandpass filtering in the log-spectral domain as used in PLP feature extraction.
PLP: Perceptual Linear Predictive coding.
LSP: Line Spectral Pair (frequencies), derived from linear predictive coding coefficients.
$F_0$: Fundamental frequency of the speech signal; ACF: Auto-Correlation Function; SHS: Sub-Harmonic Summation.
HNR: Harmonics-to-Noise Ratio.

called data-driven approch. This is opposed to the alternative way where one uses hand-crafted features and rules – a so called rule-based approach. Such features and rules are usually derived from psychological observations. They are, however, often limited to very specific emotions, and cannot be implemented robustly enough – e. g., when we implement rules based on the pitch of the voice, we face the problem of errors made by the automatic pitch detection algorithm. In some cases such errors might completely invert or randomize a result. As we are dealing with diverse emotions from eight different data collections, we chose the data-driven approach and attempt to automatically find relevant features.

The six feature sets we evaluate here are well known sets of acoustic features, which have been provided as baseline feature sets for several affective evaluation challenges – co-organized by the authors – over the last few

years: the INTERSPEECH 2009 Emotion Challenge [13], the INTERSPEECH 2010 Paralinguistics Challenge [14], the INTERSPEECH 2011 Speaker State Challenge [15], the INTERSPEECH 2012 Speaker Trait Challenge [16], and the first and second Audiovisual Emotion Challenge (AVEC 2011 [17] and AVEC 2012 [18]). The sets vary considerably in number and types of features contained.

**TABLE 3.** openSMILE standard features sets by functionals. EC: INTERSPEECH 2009 Emotion Challenge, PC: INTERSPEECH 2010 Paralinguistic Challenge, SSC: INTERSPEECH 2011 Speaker State Challenge, STC: INTERSPEECH 2011 Speaker Trait Challenge, AVEC: Audio/Visual Emotion Challenges 2011/2012. [1]Only used for the TUM AVIC baseline (PC). [2]Only applied to $F_0$. [3]Not applied to delta coefficient contours. [4]For delta coefficients the mean of only positive values is applied, otherwise the arithmetic mean is applied. [5]Not applied to voicing related LLD. [6]Only applied to voicing related LLD. For the PC and SSC feature sets, as additional feature the number of voiced segments ($F_0$ onsets) was added.

| Functional: | **EC** | **PC** | **SSC** | **STC** | **AVEC11** | **AVEC12** |
|---|---|---|---|---|---|---|
| positive arithmetic mean | | | | ✓[4] | ✓[4] | ✓[4] |
| arithmetic mean | ✓ | ✓ | ✓ | ✓[4] | ✓[4] | ✓[4] |
| root quadratic mean | | | | ✓ | ✓ | ✓ |
| contour centroid | | | ✓ | ✓ | | |
| standard deviation | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| flatness | | | | ✓ | ✓ | ✓ |
| skewness | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| kurtosis | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| quartiles 1, 2, 3 | | ✓[1] | ✓ | ✓ | ✓ | ✓ |
| inter-quartile ranges 2-1, 3-2, 3-1 | | ✓[1] | ✓ | ✓ | ✓ | ✓ |
| percentile 1 %, 99 % | | ✓ | ✓ | ✓ | ✓ | ✓ |
| percentile range 1 %–99 % | | ✓ | ✓ | ✓ | ✓ | ✓ |
| % frames signal is above minimum + 25%, 50% of range | | | | ✓ | ✓ | ✓ |
| % frames signal is above minimum + 75 % of range | | ✓[1] | | ✓ | | |
| % frames signal is above minimum + 90 % of range | | ✓[1] | ✓ | ✓ | ✓ | ✓ |
| % frames signal is below minimum + 25% of range | | | ✓ | ✓ | | |
| % frames signal is below minimum + 50%, 75%, 90% of range | | | | ✓ | | |
| % frames signal is rising | | | ✓ | ✓ | ✓ | ✓ |
| % frames signal is falling | | | ✓ | ✓ | | |
| % frames signal has left/right curvature | | | ✓[6] | ✓[6] | | |
| % frames that are non-zero | | | ✓[2] | ✓[2] | | |
| linear regression offset | ✓ | ✓[1] | | ✓[3] | | |
| linear regression slope | ✓ | ✓[1] | ✓ | ✓[3] | ✓[3] | ✓[3] |
| linear regression approximation error (linear) | | ✓[1] | | ✓[3] | ✓[3] | ✓[3] |
| linear regression approximation error (quadratic) | ✓ | ✓[1] | ✓ | | | |
| quadratic regression coefficient *a* | | | ✓ | ✓[3] | ✓[3] | ✓[3] |
| quadratic regression coefficient *b* | | | ✓ | ✓[3] | | |
| quadratic regression approximation error (linear) | | | | | ✓[3] | ✓[3] |
| quadratic regression approximation error (quadratic) | | | ✓ | ✓[3] | | |
| maximum, minimum | ✓ | | | | | |
| maximum - minimum (range) | ✓ | | | | | |
| rising, falling slopes (min to max) mean, standard deviation | | | | ✓[3] | ✓[3] | ✓[3] |
| inter maxima distances mean, standard deviation | | | | ✓[3] | ✓[3] | ✓[3] |
| amplitude mean of maxima | | | | ✓[3] | ✓[3] | ✓[3] |
| amplitude mean of minima | | | | ✓[3] | ✓[3] | ✓[3] |
| amplitude range of maxima | | | | ✓[3] | ✓[3] | ✓[3] |
| relative position of maximum, minimum | ✓ | ✓[1] | | ✓ | | |
| Linear Predictive coding gain | | | ✓ | ✓ | ✓[3,5] | ✓[3,5] |
| Linear Predictive coding coefficients 1 – 5 | | | ✓ | ✓ | ✓[3,5] | ✓[3,5] |
| peak distances mean | | | ✓ | ✓[3] | | |
| peak distances standard deviation | | | ✓ | ✓[3] | | |
| peak value mean | | | ✓ | ✓[3] | | |
| peak value mean – arithmetic mean | | | ✓ | ✓[3] | | |
| segment length mean, max, min, standard deviation | | | ✓[2] | ✓ | ✓[5] | |
| input duration in seconds | | ✓[2] | ✓[2] | | | |

All six feature sets contain standard supra-segmental features. This means that the acoustic descriptor signals such

as energy and pitch (which are sampled at a fixed rate – typically 5 or 10 ms), are summarized over a given segment (of variable length) into a single feature vector of constant length. This is achieved by applying statistical functionals to the acoustic low-level descriptors (LLD). Thereby each functional maps each LLD signal to a single value for the given segment. Examples for functionals are mean, standard deviation, higher order statistical moments, quartiles, etc. Our open-source feature extraction toolkit openSMILE [19] is used to extract all features. Table 2 and 3 show the LLDs and functionals used in each of these six sets. All LLDs are computed from short, overlapping windows of the original audio signal. The windows are typically 20 – 60 ms long and are sampled every 5 or 10 ms. To remove artifacts introduced by this windowing, LLDs are filtered over time by a simple moving average (SMA) low-pass filter. First order delta coefficients (equivalent to the first derivative) are computed for each LLD. Now, the total number of features is – in principle – obtained by multiplying the number of LLD times two (because of delta coefficients) times the number of functionals. For the INTERSPEECH 2009 Emotion Challenge feature set (EC), for example, 16 LLDs and 16 delta LLDs times 12 functionals yields 384 features. However, for the other feature sets exceptions hold from this strict brute-forcing rule. These are indicated as footnotes in Tables 2 and 3 and explained in the captions. The exceptions eliminate features which do not contribute any meaningful information, e. g., because they are always constant.

After feature extraction, the values in the feature vectors are shifted and scaled to mean zero and variance one (this is called standardization) separately for each speaker. We refer to this as speaker standardization. For classification of emotions from these features, we use Support-Vector Machines (SVM) with a linear kernel function. The SVMs are trained by the Sequential Minimal Optimisation (SMO) algorithm [20]. For the SVMs we use a complexity parameter $C = 0.2$ for all experiments. All evaluations are carried out in a Leave-One-Speaker-Out (LOSO) cross validation manner in order to obtain speaker independence of the results. Thereby, for a given database, $N$ evaluation runs are carried out – $N$ being the number of speakers. In each evaluation run data from one of the speakers is left out for testing, while the remaining data are used for training the SVM model. The final result is the weighted average – by number of instances in each test partition wrt. the total number of instances in the database – of the $N$ runs' results.

As evaluation measure, we employ the standard accuracy (Acc.) measure, defined as the total number of correctly classified samples divided by the total number of samples in the test-set. It can also be referred to as weighted average of class-wise recall rates (WAR). The recall rate for a class $c$ thereby is the number of samples classified as belonging to $c$ divided by the total number of samples belonging to $c$ (from the ground-truth labels) in the test-set. Another measure, often used in related work to normalize for an imbalance of samples across classes, is the unweighted average of class-wise recall rates (UAR). However, for this study, UAR would not allow us to compare resuls averaged over all eight databases. This is because the number of instances in the databases is not constant throughout, and a weighted (by number of samples in the database) averaging of UAR per class is not legitimate.

## RESULTS

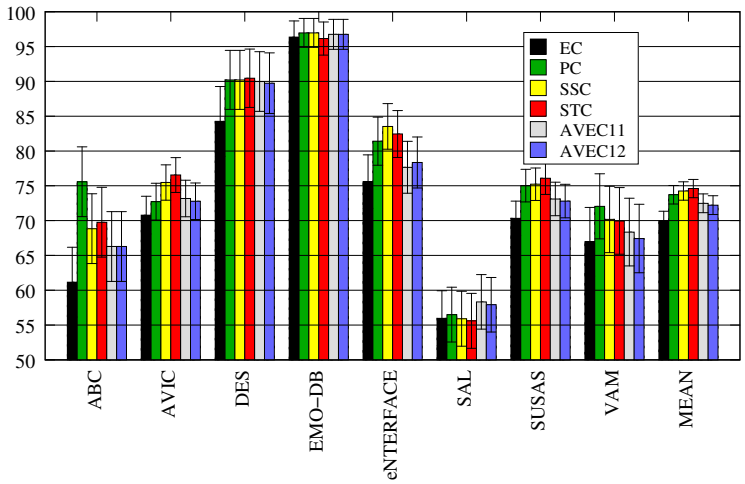


**FIGURE 1.** Results (Accuracy, [%]) for the classification of high/low arousal for all corpora (and mean of all corpora) and feature sets. The error bars indicate the significance interval of a two-tailed t-test for a significance level of 0.01. See text for more details.

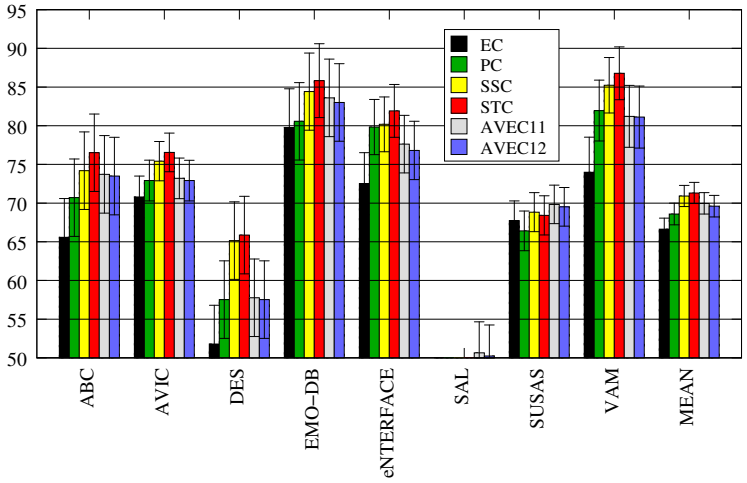


**FIGURE 2.** Results (Accuracy, [%]) for classification of positive/negative valence for all corpora (and mean of all corpora) and feature sets. The error bars indicate the significance interval of a two-tailed t-test for a significance level of 0.01. See text for more details.

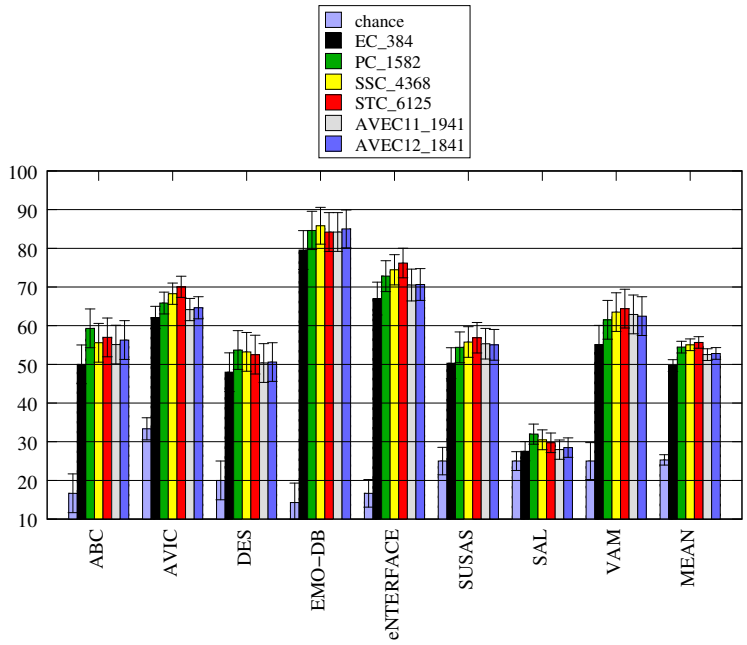


**FIGURE 3.** Results (Accuracy, [%]) for categories for all corpora (and mean of all corpora) and feature sets. First category is chance level performance ($1/N$) when an even instance distribution among classes is assumed. The error bars indicate the significance interval of a two-tailed t-test for a significance level of 0.01. See text for more details.

When interpreting our results we consider the relationship of both the database type and the feature set with respect to automatic emotion recognition performance in terms of accuracy. An average accuracy for all databases is computed by weighting the result obtained for each database by the number of instances in each database. It is displayed in the last column of Figures 1 – 3, labeled with "MEAN". Significance of all results is computed via a two-tailed t-test for a significance level of 0.01. The test only takes into account the number of instances in the database and the difference in accuracy between two results. Thus, with the number of instances given, we can compute the interval within which the accuracy can vary at the given significance level.

Figure 1 shows the accuracies obtained with all six feature sets on all eight databases for classification of high vs. low arousal. We see a clear separation in terms of database type: The more spontaneous/induced databases with free textual content (ABC, AVIC, SAL, VAM) yield lower performance. They are closely followed by SUSAS and

eNTERFACE, which also contain induced emotions. The best performance is obtained on EMO-DB and DES, which contain acted emotional speech read from pre-defined text chunks.

Figure 2 shows the accuracies obtained for all six feature sets on all eight databases for the classification of positive vs. negative valence. VAM, EMO-DB, and eNTERFACe give the best results, while SAL and DES give the worst results. For VAM, however, there is only a very little number of instances with positive valence; thus, this result is probably not too reliable and must be interpreted with care. Moreover, in VAM the valence labels are correlated to the arousal labels and mostly negative valence occurs due to the nature of the conversations. Therefore, the valence task is more of an arousal task; thus, the better performance. For all other databases, the results show that the task of recognising valence is more challenging than recognising arousal. For the SAL database, all results are not significantly better than random guess (50%); some are even below chance level (not significant) and thus do not appear in figure 2, where the y-axis starts at 50%. Note, that the results for AVIC are the same for arousal and valence, due to identical mapping.

In terms of feature set we can get the most reliable result from the mean accuracy over all databases. A common trend is observed over all tasks: The SSC and STC sets perform best, significantly better than EC and the AVEC sets. Between SSC and STC, no significant difference can be found, although STC always tends to outperform SSC slightly. EC, being the smallest set, gives the lowest performance on average.

From our results we conclude that the size of the feature set is an important factor for classification accuracy. Even for the small databases (EMO-DB, DES), a larger feature set consistently improves performance.

For reference, we provide the results that are obtained with the emotion categories of each data set in figure 3. The chance level performance is also included in this plot for comparison. In terms of performance per data set, we see similar trends as for the binary arousal/valence tasks. EMO-DB and eNTERFACE are among the best performing, followed by AVIC. The performance on DES in the categorical task is quite low, which we can attribute to the poor separation performance on the valence axis. The performance on SAL is almost at chance level in all three cases. This is in line with previous findings (e. g., [21]). Possible reasons for this are: data from only 4 speakers are contained in the set, thus only 3 speakers can be used for training, which limits the generalisation ability of the classifier; SAL also contains dicussions on a higher intellectual level than VAM. This might mean, that for SAL – compared to other databases – the cues for valence are even more exclusively found in the textual content than in the acoustic features.

## CONCLUSION AND OUTLOOK

From our results we can now draw two major conclusions concerning the type of emotional speech data: (a) the type of data has large impact on the classification performance, while (b) the choice of an optimal feature set does not depend on the type of database. If a feature-set in the binary arousal/valence classification task yields significantly (significance level 0.01) better performance than the other feature sets, it is always SSC and/or STC.

Further, we can draw the conclusion that there is (a) a very minor effect of the feature set on classification performance on average, but the performance varies significantly (up to 10% absolute) in individual cases. The trend, however, is always the same: larger feature sets give better performance, even on small databases, where we would suppose that data-sparseness becomes a problem, especially in a high dimensional feature space.

As larger feature sets also contain more descriptors, they are more likely to contain the correct descriptors. However, the kind of feature brute-forcing used (applying all functionals to all LLDs), yields very little information about which specific features are relevant for which task; the large amount of features obsurces these details. Therefore, to advance over the current state-of-the-art, we must identify and isolate important features from these sets and combine them, while removing unimportant features.

In follow up work we will evaluate the relevance of each low-level feature type (prosodic, spectral, cepstral, etc.) and statistical functional individually to gain more insight into what features exactly cause the performance differences reported in this article. In order to do this, we must first overcome some problems which we will be facing in such a study. No feature by itself will yield high performance, so combinations of features must be investigated. However, when combining features and comparing combined sets of features, the results are hard to compare, if the number of features in the sets differs. A possible solution to this is the use of balanced sets of selected acoustic descriptors as demonstrated in [22, 23]. If we manage to solve these problems, we can get more information on the relevance of all acoustic features known to date in a systematic and automated way. This will allow emotion recognition technology to move to the next level and quickly improve systems in new domains of paralinguistics and speakers states and traits.

## ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 289021(ASC-Inclusion).

## APPENDIX

An extended description of the eight emotional speech databases used in this study is given here.

### Danish Emotional Speech

The Danish Emotional Speech (DES) [5] database contains recordings of nine Danish sentences, two words and chunks that are located between two silent segments of two passages of fluent text. For example: *"Nej" (No)*, *"Ja" (Yes)*, *"Hvor skal du hen?" (Where are you going?)*. The set used here contains 419 speech utterances (i. e., speech segments between two pauses). These are produced in equal amounts by four professional actors, two males and two females. The words are read in five emotional states: *anger, happiness, neutral, sadness,* and *surprise*. Twenty judges (native speakers from 18 to 58 years old) verified the emotions by listening and identifying them. They achieved an average accuracy of 67 % on this task.

### Berlin Emotional Speech Database

The Berlin Emotional Speech Database (EMO-DB) [6] covers the emotions *anger, boredom, disgust, fear, joy, neutral, and sadness*. The spoken content is pre-defined as ten German, emotionally neutral sentences such as *"Der Lappen liegt auf dem Eisschrank" (The cloth is lying on the fridge.)*. All sentences are read/acted in all seven emotional states. The data are produced by professional actors (five males, five females). In total, approximately 900 utterances are recorded. These are verified by 20 subjects in a listening experiment. Those uttereances with less than 40% agreement among the subjects are considered as ambiguous and are discarded. Only 494 utterances remain. The average accuracy of the 20 listeners in identifying the emotions in these 494 utterances is 84.3 %. This reduced set is typically used in other studies which report results on the Berlin Emotional Speech Database. Therefore, we restrict ourselves to this selection as well.

### eNTERFACE

The eNTERFACE'05 Emotion [10] database is an audiovisual emotional speech database. It was recorded during the 2005 eNTERFACE Summer Workshop on Multimodal Interfaces. Speech from 42 subjects (eight female) from 14 nations is contained. The recordings were performed in an office environment. Six emotion categories were considered: *anger*, *disgust*, *fear*, *joy*, *sadness*, *surprise*. Each subject was instructed to listen to six successive short stories, each of them eliciting a particular emotion by describing a certain situation. After each story the subject had to react to the situation by speaking predefined phrases that fit the short story. Five phrases are available per emotion, e. g., *"I have nothing to give you! Please don't hurt me!"* in the case of fear. Two experts judged whether the reaction expressed the emotion in an unambiguous way. Only if this was the case, the sample was added to database. Overall, the database contains 1 277 samples.

### Airplane Behaviour Corpus

The Airplane Behaviour Corpus (ABC) [11] was built for the special application of public transportation surveillance. In order to induce a certain mood a script was used, which lead the subjects through a guided storyline: pre-recorded announcements were automatically played via loudspeakers, controlled by a hidden supervisor. As scenarios a vacation flight and return flight were chosen, consisting of 13 and 10 scenes each, such as takeoff, serving of wrong

food, turbulences, falling asleep, conversation with a neighbor, and landing. The recording setup consisted of an airplane seat for the subject, positioned in front of a blue screen. 8 subjects (4 male, 4 female) from 25–48 years of age (mean age: 32 years) participated in the recording. All conversations are in German. A total of 11.5 hours of audio and video was recorded. After manual pause removal and pre-segmentation on an utterance level, 431 utterances remain. They were annotated independently by three experienced male annotators using the following set of affective labels: *agressive*, *cheerful*, *intoxicated*, *nervous*, *neutral*, and *tired*. The total length of these 431 utterances is one hour and 20 seconds. The average utterance length is 8.4 seconds.

## Speech Under Simulated and Actual Stress

The Speech Under Simulated and Actual Stress (SUSAS) database [7] is a database which contains recordings of induced emotions. Speech is partly masked by loud background noise, which makes recognition of speech and emotions within this set even more challenging. The database contains recordings of speech under actual stress (during a roller coaster ride and a free fall) and under simulated stress. For this study we decided to use only 3 593 actual stress samples. Seven speakers, three of them female, are contained in this set. Next to *neutral* speech and *fear*, different stress conditions have been collected: *medium stress*, *high stress*, and *screaming*. The linguistic content in SUSAS is restricted to a pre-defined vocabulary of 35 English air-commands, such as *"brake"*, *"help"*, or *"no"*.

## Audiovisual Interest Corpus

A database of emotional speech samples with non-restricted spoken content is the Audiovisual Interest Corpus (AVIC) [8]. The recording setup consisted of a scenario in which a product presenter lead the subject through an interactive English commercial presentation. 21 subjects (11 male, 10 female) participated in the recording. The level of interest is annotated for every speech turn reaching from *boredom* (subject is bored with listening and talking about the topic, very passive, she/he does not follow the discourse; this state is also referred to as level of interest (loi) 1, i.e. loi1), over *neutral* (subject follows and participates in the discourse, it can not be recognized if she/he is interested or indifferent; loi2) to *joyful* interaction (strong wish of the subject to talk and learn more about the topic; loi3). Additionally, the spoken content and non-linguistic vocalisations (laughter, breathing, consent "mh hm", coughing) are labeled. The database contains 3 002 segments of speech in total from all subjects. In contrast to the only 996 phrases, which have a high inter-labeler agreement and were used in [8], we use all segments in this study.

## Sensitive Artificial Listener

The Belfast Sensitive Artificial Listener (SAL) data collection is part of the HUMAINE database [24]. The subset used contains 25 recordings in total from 4 speakers (2 male, 2 female) with an average length of 20 minutes per speaker. The data contains audio-visual recordings from non-acted human-computer conversations. The conversations were recorded through a SAL interface designed to elicit a range of emotional states in the subjects. The recordings were labeled continuously in real time by four judges using a system based on FEELtrace [25]. Each judge used a sliding controller to annotate the dimensions valence and arousal separately while listening to the recordings twice. The values from the sliders were sampled every 10 ms. To compensate for different perceptual scales of the judges, the annotations were shifted to zero mean globally and scaled by a factor so that 98 % of all the values are in the range from -1 to +1. The 25 recordings have been split into turns using an energy based Voice Activity Detection, resulting in a total of 1 692 utterances. Labels for each turn are computed by averaging the frame level valence and arousal labels over the complete turn.

## Vera-Am-Mittag

The Vera-Am-Mittag (VAM) corpus [9] consists of audio-visual recordings taken from a German TV talk show. It contains 946 spontaneous and emotionally coloured utterances from 47 guests of the talk show which were recorded from unscripted, semi-authentic discussions. The topics were mainly personal issues such as friendship

crises, fatherhood questions, or romantic affairs. To obtain non-acted data, a talk show in which the guests were not being paid to perform as actors was chosen. The speech extracted from the dialouges contains a large amount of colloquial expressions as well as non-linguistic vocalisations and partly covers different German dialects. Before annotation, the audio recordings were manually segmented into utterances where each utterance contains at least one phrase. 17 judges labeled one half of the data, 6 the other half. The labeling is based on a discrete five point scale for each of the three dimensions arousal, potency, and valence. Potency indicates the dominance of a person in a conversation with at least one other person. The standard deviation between the judges is 0.34, 0.31, and 0.29 for activation, potency, and valence. The average correlation coefficients between the evaluators are 0.72, 0.61, and 0.49 respectively. The correlation coefficients for activation and potency show reasonably high agreement of the judges, whereas the moderate value for valence indicates that this affective dimension is more difficult to label.

## REFERENCES


1. M. Shami, and W. Verhelst, "Automatic Classification of Expressiveness in Speech: A Multi-corpus Study," in *Speaker Classification II*, edited by C. Müller, Springer Berlin / Heidelberg, 2007, vol. 4441 of *LNCS / AI*, pp. 43–56.
2. B. Schuller, D. Seppi, A. Batliner, A. Meier, and S. Steidl, "Towards more Reality in the Recognition of Emotional Speech," in *Proc. ICASSP*, Honolulu, 2007, pp. 941–944.
3. B. Schuller, M. Wimmer, D. Arsić, T. Moosmayr, and G. Rigoll, "Detection of Security Related Affect and Behaviour in Passenger Transport," in *Proceedings INTERSPEECH 2008*, 2008, pp. 265–268.
4. F. Eyben, A. Batliner, B. Schuller, D. Seppi, and S. Steidl, "Cross-Corpus Classification of Realistic Emotions – Some Pilot Experiments," in *Proceedings 3rd International Workshop on EMOTION: Corpora for Research on Emotion and Affect, satellite of LREC 2010*, edited by L. Devillers, B. Schuller, R. Cowie, E. Douglas-Cowie, and A. Batliner, Valletta, Malta, 2010, pp. 77–82.
5. I. S. Engbert, and A. V. Hansen, Documentation of the danish emotional speech database des, Tech. rep., Center for Person Kommunikation, Aalborg University, Denmark (2007).
6. F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A Database of German Emotional Speech," in *Proc. Interspeech*, Lisbon, 2005, pp. 1517–1520.
7. J. Hansen, and S. Bou-Ghazale, "Getting Started with SUSAS: A Speech Under Simulated and Actual Stress Database," in *Proc. EUROSPEECH-97*, Rhodes, Greece, 1997, pp. 1743–1746.
8. 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," in *Image and Vision Computing Journal (IMAVIS), Special Issue on Visual and Multimodal Analysis of Human Spontaneous Behavior*, 2009, 17 pages.
9. 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.
10. O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The enterface'05 Audio-Visual Emotion Database," in *Proc. IEEE Workshop on Multimedia Database Management*, Atlanta, 2006.
11. B. Schuller, M. Wimmer, D. Arsic, G. Rigoll, and B. Radig, "Audiovisual Behavior Modeling by Combined Feature Spaces," in *Proc. ICASSP 2007, Honolulu*, 2007, vol. II, pp. 733–736.
12. M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie, "Abandoning Emotion Classes - Towards Continuous Emotion Recognition with Modelling of Long-Range Dependencies," in *Proc. 9th Interspeech 2008*, Brisbane, Australia, 2008, pp. 597–600.
13. B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 Emotion Challenge," in *Proc. Interspeech*, ISCA, Brighton, UK, 2009.
14. B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "The INTERSPEECH 2010 Paralinguistic Challenge," in *Proceedings INTERSPEECH 2010*, Makuhari, Japan, 2010, pp. 2794–2797.
15. B. Schuller, A. Batliner, S. Steidl, F. Schiel, and J. Krajewski, "The INTERSPEECH 2011 Speaker State Challenge," in *Proceedings INTERSPEECH 2011*, Florence, Italy, 2011, pp. 3201–3204.
16. 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*, Portland, OR, 2012, to appear.
17. B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, "AVEC 2011 – The First International Audio/Visual Emotion Challenge," in *Proceedings First International Audio/Visual Emotion Challenge and Workshop, AVEC 2011, in conjunction with ACII 2011, Memphis, TN*, edited by B. Schuller, M. Valstar, R. Cowie, and M. Pantic, Springer, 2011, pp. 415–424.
18. B. Schuller, M. Valstar, R. Cowie, and M. Pantic, editors, *Proceedings of the Second International Audio/Visual Emotion Challenge and Workshop – An Introduction*, Santa Monica, CA, 2012, grand Challenge and Satellite of ACM ICMI 2012, to appear.
19. F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proceedings of ACM Multimedia 2010*, Florence, Italy, 2010, pp. 1459–1462.

20. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The WEKA Data Mining Software: An Update," in *SIGKDD Explorations*, 2009, vol. 11.
21. B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth, "Acoustic Emotion Recognition: A Benchmark Comparison of Performances," in *Proceedings of ASRU 2009*, Merano, Italy, 2009, pp. 552–557.
22. A. Batliner, S. Steidl, B. Schuller, D. Seppi, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, V. Aharonson, L. Kessous, and N. Amir, "Whodunnit – Searching for the Most Important Feature Types Signalling Emotion-Related User States in Speech," in *Computer Speech and Language*, Elsevier, 2011, vol. 25, pp. 4–28.
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 *Proc. Interspeech*, Antwerp, 2007, pp. 2253–2256.
24. E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J.-C. Martin, L. Devillers, S. Abrilan, A. Batliner, N. Amir, and K. Karpousis, "The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data," in *Affective Computing and Intelligent Interaction*, edited by A. Paiva, R. Prada, and R. W. Picard, Springer, Berlin-Heidelberg, 2007, pp. 488–500.
25. R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder, "FEELTRACE: An instrument for recording perceived emotion in real time," in *Proc. ISCA Workshop on Speech and Emotion*, Newcastle, Northern Ireland, 2000, pp. 19–24.