Acoustic Emotion Recognition: A Benchmark Comparison of Performances

Björn Schuller [#,1], Bogdan Vlasenko [*,2], Florian Eyben [#], Gerhard Rigoll [#], Andreas Wendemuth [*]

[#] *Institute for Human-Machine Communication, Technische Universität München*
D-80333 München, Germany
[1] schuller@tum.de

[*] *Cognitive Systems, IESK, Otto-von-Guericke University*
Magdeburg, Germany
[2] bogdan.vlasenko@ovgu.de

***Abstract*—In the light of the first challenge on emotion recognition from speech we provide the largest-to-date benchmark comparison under equal conditions on nine standard corpora in the field using the two pre-dominant paradigms: modeling on a frame-level by means of Hidden Markov Models and supra-segmental modeling by systematic feature brute-forcing. Investigated corpora are the ABC, AVIC, DES, EMO-DB, eNTERFACE, SAL, SmartKom, SUSAS, and VAM databases. To provide better comparability among sets, we additionally cluster each database's emotions into binary valence and arousal discrimination tasks. In the result large differences are found among corpora that mostly stem from naturalistic emotions and spontaneous speech vs. more prototypical events. Further, supra-segmental modeling proves significantly beneficial on average when several classes are addressed at a time.**

I. Introduction

Few works in the field of emotion recognition from speech consider several corpora at a time for comparative test-runs and evaluations. Further, these usually consider mostly two [1] to three [2] or maximal four [3] databases at a time. However, in the light of the first open comparative challenge on emotion recognition [4] we want to provide baseline benchmark results on a multiplicity of popular databases and decided for nine typical such. We therefore choose two standard popular and freely available toolkits for the recognition of emotion: a very basic approach employing the Hidden Markov Toolkit that is used for a very broad selection of speech and general audio recognition tasks, and our more specifically tailored openEAR emotion recogniser.

The remainder of the paper is structured as follows: we first describe the nine chosen data sets and the clustering of emotions to binary arousal and valence tasks in Sec. II. Next, frame-level (Sec. III) and supra segmental modeling (Sec. IV) are introduced prior to the presentation of result and drawn conclusions (Sec. V).

II. Nine Popular Databases

One of the major needs of the community ever since - maybe even more than in many related pattern recognition tasks - is the constant need for data sets [5], [6]. In the early days of the late 1990s these have not only been few, but also small (≈ 500 turns) with few subjects (≈ 10), uni-modal, recorded in studio noise conditions, and acted. Further, the spoken content was mostly predefined (DES [7], Berlin Emotional Speech-Database [8], SUSAS [9]). These were seldom made public and few annotators - if any at all - usually labeled exclusively the perceived emotion. Additionally, these were partly not intended for analysis, but for quality measurement of synthesis (e. g. DES, Berlin Emotional Speech-Database). However, any data is better than none. Today we are happy to see more diverse emotions covered, more elicited or even spontaneous sets of many speakers, larger amounts of instances (5k -10k) of more subjects (up to more than 100), multimodal data that is annotated by more labelers (4 (AVIC [10]) - 17 (VAM [11])), and that is made publicly available. Thereby it lies in the nature of collecting acted data that equal distribution among classes is easily obtainable. In more spontaneous sets this is not given, which forces one to either balance in the training or shift from reporting of simple recognition rates to F-measures or unweighted recall values, best per class (e. g. FAU AIBO [4], and the AVIC databases). However, some acted and elicited data sets with pre-defined content are still seen (e. g. eNTERACE [12]), yet these also follow the trend of more instances and speakers. Positively, also transcription is becoming more and more rich: additional annotation of spoken content and non-linguistic interjections (e. g. FAU AIBO, AVIC databases), multiple annotator tracks (e. g. VAM corpus), or even manually corrected pitch contours (FAU AIBO database) and additional audio tracks in different recordings (e. g. close-talk and room-microphone), phoneme boundaries and manual phoneme labelling (e. g. EMO-DB database), different chunkings (e. g. FAU AIBO database) and prototypicality levels. At the same time these are partly also recorded under more realistic conditions (or taken from the media). However, in future sets multilinguality and subjects of diverse cultural backgrounds will be needed in addition to all named positive trends.

For the following investigations, we chose nine among the most popular. Only such available to the community were considered. These should cover a broad variety reaching from acted speech (the Danish (DES, [7]) and the Berlin Emotional

978-1-4244-5480-8/09/$26.00 © 2009 IEEE

Speech ($EMO-DB$, [8]) databases), over story guided as the eNTERFACE corpus [12] with fixed spoken content and the Airplane Behaviour Corpus (ABC, [13]), to spontaneous with fixed spoken content represented by the Speech Under Simulated and Actual Stress (SUSAS, [9]) database, to more modern corpora with respect to the number of subjects involved, spontaneity, and free language covered by the Audiovisual Interest Corpus ($AVIC$, [10]), the Sensitive Artificial Listener (SAL, [14]), the SmartKom [15], and the Vera-Am-Mittag (VAM, [11]) databases.

An overview on properties of the chosen sets is found in table I. Next, we will shortly introduce the sets.

A. Danish Emotional Speech

The Danish Emotional Speech (DES) [7] database has been chosen as first set as one of the 'traditional representatives' for our study, because it is easily accessible and well annotated. The data used in the experiments are 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 contains 419 speech utterances (i. e., speech segments between two silence pauses) which are expressed by four professional actors, two males and two females. All utterances are equally separated for each gender. Speech is expressed in five emotional states: *anger, happiness, neutral, sadness, and surprise*. Twenty judges (native speakers from 18 to 58 years old) verified the emotions with a score rate of 67 %.

B. Berlin Emotional Speech Database

A further well known set chosen to test the effectiveness of emotion classification is the popular studio recorded Berlin Emotional Speech Database (EMO-DB) [8], which covers *anger, boredom, disgust, fear, joy, neutral, and sadness* speaker emotions. The spoken content is again pre-defined by ten German emotionally neutral sentences as *"Der Lappen liegt auf dem Eisschrank" (The cloth is lying on the fridge.)*. As DES, it thus provides a high number of repeated words in diverse emotions. Ten (five female) professional actors speak ten German emotionally undefined sentences. While the whole set comprises around 900 utterances, only 494 phrases are marked as minimum 60 % natural and minimum 80 % assignable by 20 subjects in a listening experiment. 84.3 % mean accuracy is the result of this perception study for this limited 'more prototypical' set. As this set is usually used in the manifold works reporting results on the corpus we restrict ourselves to this selection, as well.

C. eNTERFACE

The eNTERFACE [12] corpus is a further public, yet audiovisual emotion database. It consists of induced *anger*, *disgust*, *fear*, *joy*, *sadness*, *surprise* speaker emotions. 42 subjects (eight female) from 14 nations are included. It consists of office environment recordings of pre-defined spoken content in English. Each subject was instructed to listen to six successive short stories, each of them eliciting a particular emotion.

They then had to react to each of the situations by uttering previously read phrases that fit the short story. Five phrases are available per emotion as *"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 consists of 1 277 samples.

D. Airplane Behaviour Corpus

Another audiovisual emotion database is the Airplane Behaviour Corpus (ABC) [13] crafted for the special target application of public transport surveillance. In order to induce a certain mood, a script was used, which lead the subjects through a guided storyline: prerecorded announcements by five different speakers were automatically played back controlled by a hidden test-conductor. As a general framework a vacation flight with return flight was chosen, consisting of 13 and 10 scenes as start, serving of wrong food, turbulences, falling asleep, conversation with a neighbor, or touch-down. The general setup consisted of an airplane seat for the subject, positioned in front of a blue screen. 8 subjects in gender balance from 25–48 years (mean 32 years) took part in the recording. The language throughout recording is German. A total of 11.5 h video was recorded and annotated independently after pre-segmentation by three experienced male labelers within a closed set. The average length of the 431 clips in total is 8.4 s.

E. Speech Under Simulated and Actual Stress

The Speech Under Simulated and Actual Stress (SUSAS) database [9] serves as a first reference for spontaneous recordings. As additional challenge speech is partly masked by field noise. We decided for the 3 593 actual stress speech samples recorded in subject motion fear and stress tasks. Seven speakers, three of them female, in roller coaster and free fall actual stress situations are contained in this set. Next to *neutral* speech and *fear* two different stress conditions have been collected: *medium stress*, and *high stress*, and *screaming*. SUSAS is also restricted to a pre-defined spoken text of 35 English air-commands, such as *"brake"*, *"help"* or *"no"*. Likewise, only single words are contained similar to DES where this is also mostly the case.

F. Audiovisual Interest Corpus

To add spontaneous emotion samples of non-restricted spoken content, we further decided for the Audiovisual Interest Corpus (AVIC) [10], another audiovisual emotion corpus. In its scenario setup, a product presenter leads one of 21 subjects (10 female) through an English commercial presentation. The level of interest is annotated for every sub-speaker turn reaching from *boredom* (subject is bored with listening and talking about the topic, very passive, 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 recognised, if she/he is interested or indifferent in

the topic; 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 are labeled in the AVIC set. For our evaluation we use all 3 002 phrases, in contrast to only 996 phrases with high inter-labeler agreement as e. g. employed in [10].

G. Sensitive Artificial Listener

The Belfast Sensitive Artificial Listener (SAL) data is part of the final HUMAINE database [16]. We consider the subset used e. g. in [14] which 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 natural human-computer conversations that were recorded through a SAL interface designed to let users work through a range of emotional states. The data has been labeled continuously in real time by four annotators with respect to valence and activation using a system based on FEELtrace [17]: the annotators used a sliding controller to annotate both emotional dimensions separately whereas the adjusted values for valence and activation were sampled every 10 ms to obtain a temporal quasi-continuum. To compensate linear offsets that are present among the annotators, the annotations were normalised to zero mean globally. Further, to ensure common scaling among all annotators, each annotator's labels were scaled so that 98 % of all values are in the range from -1 to +1. The 25 recordings have been split into turns using an energy based Voice Activity Detection. A total of 1 692 turns is accordingly contained in the database. Labels for each turn are computed by averaging the frame level valence and activation labels over the complete turn. Apart from the necessity to deal with continuous values for time and emotion, the great challenge of the SAL database is the fact that one must deal with all data - as recorded - and not only manually pre-selected 'emotional prototypes' as in practically any other database apart from [4].

H. SmartKom

We further include a second audiovisual corpus of spontaneous speech and natural emotion in our tests: the SmartKom[15] multi-modal corpus consists of Wizard-Of-Oz dialogs in German and English. For our evaluations we use German dialogs recorded during a public environment technical scenario. As with SUSAS, noise is overlaid (street noise). The database contains multiple audio channels and two video channels (face, body from side). The primary aim of the corpus was the empirical study of human - computer interaction in a number of different tasks and technical setups. It is structured into sessions which contain one recording of approximately 4.5 min length with one person. Utterances are labeled in seven broader emotional states: *neutral*, *joy*, *anger*, *helplessness*, *pondering*, *surprise* are contained together with unidentifiable episodes.

I. Vera-Am-Mittag

The Vera-Am-Mittag (VAM) corpus [11] consists of audio-visual recordings taken from a German TV talk show. The set used contains 946 spontaneous and emotionally coloured utterances from 47 guests of the talk show which were recorded from unscripted, 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 dialogs contains a large amount of colloquial expressions as well as non-linguistic vocalisations and partly covers different German dialects. For annotation of the speech data, the audio recordings were manually segmented to the utterance level, whereas each utterance contained at least one phrase. A large number of human labelers was used for annotation (17 labelers for one half of the data, six for the other). The labeling bases on a discrete five point scale for three dimensions mapped onto the interval of [-1,1]: the average results for the standard deviation are 0.29, 0.34, and 0.31 for valence, activation, and dominance. The averages for the correlation between the evaluators are 0.49, 0.72, and 0.61, respectively. The correlation coefficients for activation and dominance show suitable values, whereas the moderate value for valence indicates that this emotion primitive was more difficult to evaluate, but may partly also be a result of the smaller variance of valence.

J. Clustering of Emotions

The chosen sets cover a broad variety reaching from acted (DES, EMO-DB) over induced (ABC, eNTERFACE) to natural emotion (AVIC, SmartKom, SUSAS, VAM) with strictly limited textual content (DES, EMO-DB, SUSAS) over more variation (eNTERFACE) to full variance (ABC, AVIC, SAL, SmartKom, VAM). Further Human-Human (AVIC, VAM) as well as Human-Computer (SAL, SmartKom) interaction are contained. Three languages (English, German, and Danish) are comprised. However, these three all belong to the same family of Germanic languages. The speaker ages and backgrounds vary strongly, and so do of course microphones used, room acoustics, and coding (e. g. sampling rate reaching from 8 kHz to 44.1 kHz) as well as the annotators.

For better comparability of obtained performances among corpora we thus decided to map the diverse emotion groups onto the two most popular axes in the dimensional emotion model: arousal (i.e. passive vs. active) and valence (i.e. positive vs. negative). The chosen mappings are depicted in Tables II and III, accordingly. Note that these mappings are not straight forward. This is especially true for the neutral emotion, which could have been chosen as a third state. Sadly, however, not all databases provide such a state. Thus, the mapping can be seen as compromise in favour of better balance among the target classes. We further discretised in the arousal-valence plane for the databases SAL and VAM to provide numbers exclusively on classification rather than mixed with regression tasks. We consider only four quadrants obtained by discretising into binary tasks as described above, but now handling the problem as a four-class problem. The according quadrant's q1–q4 (counterclockwise, starting in positive quadrant, assuming valence as ordinate and arousal as abscissa) can also be assigned

TABLE I
OVERVIEW OF THE SELECTED EMOTION CORPORA.

Corpus	Content	# / Emotion							# Arousal		# Valence		# All	hh:mm	# Sub	Rec	kHz
									low	high	-	+					
ABC	German fixed	agre 95	chee 105	into 33	nerv 93	neut 79	tire 25	-	104	326	213	217	431	01:15	8 4 f	acted stud	16
AVIC	English variable	loi1 553	loi2 2279	loi3 170	-	-	-	-	553	2449	553	2449	3002	01:47	21 10 f	spon norm	44.1
DES	Danish fixed	angr 85	happ 86	neut 85	sad 84	surp 84	-	-	169	250	169	250	419	00:28	4 2 f	acted norm	20
EMO-DB	German fixed	angr 127	bore 79	disg 38	fear 55	happ 64	neut 78	sadn 53	248	246	352	142	494	00:22	10 5 f	acted stud	16
eNTERF.	English fixed	angr 215	disg 215	fear 215	happ 207	sadn 210	surp 215	-	425	852	855	422	1277	01:00	42 8 f	acted norm	16
SAL	English variable	q1 459	q2 320	q3 564	q4 349	-	-	-	884	808	917	779	1692	01:41	4 2 f	spon norm	16
SmartKom	German variable	angr 220	help 161	joy 284	neut 2179	pond 643	surp 70	unid 266	3088	735	381	3442	3823	07:08	79 47 f	spont noisy	16
SUSAS	English fixed	hist 1202	meds 1276	neut 701	scre 414	-	-	-	701	2892	1616	1977	3593	01:01	7 3 f	mixed noisy	8
VAM	German variable	q1 21	q2 50	q3 451	q4 424	-	-	-	501	445	875	71	946	00:47	47 32 f	spon norm	16

Abbreviations: Sub: subjects (f stands for the number of female subjects), Rec: recording characteristics. agre: aggressive, angr: angry, bore: boredom, chee: cheerful, disg: disgust, happ: happy, help: helplessness, hist: high stress, into: intoxicated, loi1–3: level of interest 1–3, meds: medium stress, nerv: nervous, neut: neutral, pond: pondering, q1–q4: quadrants in the arousal-valance plane, sadn: sadness, surp - surprise, tire: tired, unid: unidentifiable

emotion tags: "happy / exciting" (q1), "angry / anxious" (q2), "sad / bored" (q3), and "relaxed / serene" (q4).

TABLE II
MAPPING OF EMOTIONS FOR THE CLUSTERING TO A BINARY AROUSAL DISCRIMINATION TASK.

Corpus	Negative	Positive
ABC	neutral, tiered	aggressive, cheerful, intoxicated, nervous,
AVIC	loi1	loi2, loi3
DES	neutral, sad	angry, happy, surprise
EMO-DB	boredom, disgust, neutral, sadness	anger, fear, happiness
eNTERFACE	disgust, sadness	anger, fear, happiness, surprise
SAL	q2, q3	q1, q4
SmartKom	neutral, pondering, unidentifiable	anger, helplessness, joy, surprise
SUSAS	neutral	high stress, medium stress, screaming
VAM	q2, q3	q1, q4

TABLE III
MAPPING OF EMOTIONS FOR THE CLUSTERING TO A BINARY VALENCE DISCRIMINATION TASK.

Corpus	Negative	Positive
ABC	aggressive, nervous, tired	cheerful, intoxicated, neutral
AVIC	loi1	loi2, loi3
DES	angry, sad	happy, neutral, surprise
EMO-DB	anger, boredom, disgust, fear, sadness	happiness, neutral
eNTERFACE	anger, disgust, fear, sadness	happiness, surprise
SAL	q3, q4	q1, q2
SmartKom	anger, helplessness,	joy, neutral, pondering, surprise, unidentifiable
SUSAS	high stress, screaming	medium stress, neutral
VAM	q3, q4	q1, q2

III. FRAME-LEVEL MODELING

Speech input is processed using a 25 ms Hamming window, with a frame rate of 10 ms. As in typical speech recognition we employ a 39 dimensional feature vector per each frame consisting of 12 MFCC and log frame energy plus speed and acceleration coefficients. Cepstral Mean Substraction (CMS) and variance normalisation are applied to better cope with channel characteristics.

We consider using a speaker recognition system to recognise emotion from speech in the first place. Likewise, instead of the usual task to deduce the most likely speaker (from a known speaker set) Ω_k from a given sequence X of M acoustic observations x, we will recognise the current emotion. This is solved by a stochastic approach using the following equation:

$$\Omega_k = \underset{\Omega}{\operatorname{argmax}} P(\Omega|X) = \underset{\Omega}{\operatorname{argmax}} \frac{P(X|\Omega)P(\Omega)}{P(X)} \qquad (1)$$

where $P(X|\Omega)$ is called the emotion acoustic model, $P(\Omega)$ is the prior user behaviour information and Ω is one of all system known emotions. In case of turn level analysis the emotion acoustic model is designed by s single state HMMs. This state is associated with an emission-probability $P(X|s)$ which for continuous variables x is replaced with its probability density function (PDF). These PDFs are realised using weighted sums of elementary Gaussian PDFs (Gaussian Mixtures Models,

GMM). Each emotion is modeled by its own GMM. One emotion is assigned for a full dialog turn.

The HTK toolkit [18] was used to build these models, using standard techniques such as forward-backward and Baum-Welch re-estimation algorithms.

IV. Supra-segmental Modeling

Using the openEAR toolkit [19], 6 552 features are extracted as 39 functionals of 56 acoustic low-level descriptors (LLD) and corresponding first and second order delta regression coefficients.

TABLE IV
33 Low-Level Descriptors (LLD) used in acoustic analysis with openEAR.

Feature Group	Features in Group
Raw Signal	Zero-crossing-rate
Signal energy	logarithmic
Pitch	Fundamental frequency F_0 in Hz via Cepstrum and Autocorrelation (ACF). Exponentially smoothed F_0 envelope.
Voice Quality	Probability of voicing ($\frac{ACF(T_0)}{ACF(0)}$)
Spectral	Energy in bands 0 - 250 Hz, 0 - 650 Hz, 250 - 650 Hz, 1 - 4 kHz 25 %, 50 %, 75 %, 90 % roll-off point, centroid, flux, and rel. pos. of spectrum max. and min.
Mel-spectrum	Band 1-26
Cepstral	MFCC 0-12

Table V lists the statistical functionals, which were applied to the LLD as shown in Table IV to map a time series of variable length onto a static feature vector.

TABLE V
39 functionals applied to LLD contours and regression coefficients of LLD contours.

Functionals, etc.	#
Respective rel. position of max./min. value	2
Range (max.-min.)	1
Max. and min. value - arithmetic mean	2
Arithmetic mean, Quadratic mean	2
Number of non-zero values	1
Geometric, and quadratic mean of non-zero values	2
Mean of absolute values, Mean of non-zero abs. values	2
Quartiles and inter-quartile ranges	6
95 % and 98 % percentile	2
Std. deviation, variance, kurtosis, skewness	4
Centroid	1
Zero-crossing rate	1
# of peaks, mean dist. btwn. peaks, arth. mean of peaks, arth. mean of peaks - overall arth. mean	4
Linear regression coefficients and corresp. approximation error	4
Quadratic regression coefficients and corresp. approximation error	5

As pre-processing steps speaker (group) standardisation (cf. next section) and balancing of the training partition were carried out. The classifier of choice is Support Vector Machines with polynomial Kernel and pairwise multi-class discrimination based on Sequential Minimal Optimisation. Moreover, standardisation of each training fold in contrast to speaker standardisation was evaluated.

V. Results and Conclusion

For all databases test-runs are carried out in Leave-One-Speaker-Out (LOSO) or Leave-One-Speakers-Group-Out (LOSGO) manner to face speaker independence, as required by most applications. In the case of 10 or less speakers in one corpus we apply the LOSO strategy; otherwise, namely for AVIC, eNTERFACE, SmartKom, and VAM, we select 5 speaker groups with utmost equal amount of male and female speakers and samples per group for LOSGO evaluation.

As evaluation measures we employ the weighted (WA, i. e. accuracy) and unweighted (UA, thus better reflecting unbalance among classes) average of class-wise recall rates as demanded in [4].

The results for frame-level (Table VI) and supra-segmental modeling (Table VII) are found for all emotion classes contained per database and for the clustered two-class tasks of binary arousal and valence discrimination as described. Note that for supra-segmental modeling SVM with speaker standardisation in constant parametrisation are used for the given results. The delta of the mean in Table VII to the mean of the best performing individual configurations is 1.7 % (UA) and 0.7 % (WA) for class-wise results, 0.2 % (UA) and 1.8 % (WA) for arousal and 9.4 % (UA) and 9.5 % (WA) for valence (mostly due to variations on SAL).

TABLE VI
Results of the HMM/GMM based emotion recognition engine

Corpus	All [%]		Arousal [%]		Valence [%]	
	UA	WA	UA	WA	UA	WA
ABC	48.8	57.7	71.5	74.7	81.1	81.2
AVIC	65.5	66.0	74.5	77.5	74.5	77.5
DES	45.3	45.3	82.0	84.2	55.6	58.0
EMO-DB	73.2	77.1	91.5	91.5	78.0	80.4
eNTERFACE	67.1	67.0	74.9	76.8	78.7	80.5
SAL	34.0	32.7	61.2	61.6	57.2	57.0
SmartKom	28.6	47.9	58.2	64.6	57.1	68.4
SUSAS	55.0	47.9	56.0	68.0	67.3	67.8
VAM	38.4	70.2	76.5	76.5	49.2	89.9
Mean	50.7	56.9	71.8	75.0	66.5	73.4

TABLE VII
Results of the SVM based emotion recognition engine.

Corpus	All [%]		Arousal [%]		Valence [%]	
	UA	WA	UA	WA	UA	WA
ABC	55.5	61.4	61.1	70.2	70.0	70.0
AVIC	56.5	68.6	66.4	76.2	66.4	76.2
DES	59.9	60.1	87.0	87.4	70.6	72.6
EMO-DB	84.6	85.6	96.8	96.8	87.0	88.1
eNTERFACE	72.5	72.4	78.1	79.3	78.6	80.2
SAL	29.9	30.6	55.0	55.0	50.0	49.9
SmartKom	23.5	39.0	59.1	64.1	53.1	75.6
SUSAS	61.4	56.5	63.7	77.3	67.7	68.3
VAM	37.6	65.0	72.4	72.4	48.1	85.4
Mean	53.5	59.9	71.1	75.4	64.5	68.3

Among the two result tables very similar trends can be observed: the best performance is achieved on the databases containing acted, prototypical emotions, where only emotions with high inter-labeler agreement were selected (EMO-DB,

eNTERFACE). A little exception here is the DES corpus, where performance is well behind EMO-DB, even though DES also contains acted, prototypical emotions. This difference is not so obvious for the arousal task as it is for the full classification task. One reason for this might be that no selection wrt. high inter-labeler agreement was done on DES and labelers may agree more upon arousal than on the emotion categories. The remaining six corpora are more challenging since they contain non-acted or induced emotions. On the lower end of recognition performance the SAL, SmartKom, and VAM corpora can be found, which contain the most spontaneous and naturalistic emotions, which in turn are also the most challenging to label. Moreover, SmartKom contains long pauses with a high noise level and annotations are multi-modal, i.e. mimic and audio based, thus the target emotion might not always be detectable from speech. The results for the SAL corpus are only marginally above chance level, which is due to speaker independent evaluation on highly naturalistic data with only four speakers in total. In previous work on this database only speaker dependent evaluations were presented, e.g. [14].

When comparing the frame-level modeling with the supra-segmental modeling an interesting conclusion can be drawn: frame-level modeling seems to be slightly superior for corpora containing variable content (AVIC, SAL, SmartKom, VAM), i.e. the subjects were not restricted to a predefined script, while supra-segmental modeling outperforms frame-level modeling by large on corpora where the topic/script is fixed (ABC, DES, EMO-DB, eNTERFACE, SUSAS), i.e. where there is an overlap in verbal content between test and training set. This can be explained by the nature of supra-segmental modeling: in corpora with non-scripted content, turn lengths may strongly vary. While frame-level modeling is mostly independent of highly varying turn length, in supra-segmental modeling each turn gets mapped onto one feature vector, which might not always be appropriate.

Still, supra-segmental modeling using openEAR on average outperforms frame-level modeling using HTK. In future work, supra-segmental modeling should be further improved by addressing the issues regarding varying structure and turn length by adding other types of functionals and using other segments than turns or considering word, syllable or phoneme dependent emotion models.

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. 211486 (SEMAINE).

The work has further been partly conducted in the framework of the NIMITEK project, within the framework of the "Excellence Program Neurowissenschaften" of the federal state of Sachsen-Anhalt, Germany (FKZ: XN3621A/1005M). This project is associated and supported by the Transregional Collaborative Research Centre SFB/TRR 62 "Companion-Technology for Cognitive Technical Systems" funded by the German Research Foundation (DFG).

REFERENCES

[1] M. Shami and W. Verhelst, "Automatic Classification of Expressiveness in Speech: A Multi-corpus Study," in *Speaker Classification II*, ser. LNCS / AI. Springer, 2007, vol. 4441, 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. Arsic, T. Moosmayr, and G. Rigoll, "Detection of security related affect and behaviour in passenger transport," in *Proc. 9th Interspeech 2008*. Brisbane, Australia: ISCA, 2008, pp. 265–268.

[4] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 Emotion Challenge," in *Proc. Interspeech*. Brighton, UK: ISCA, 2009.

[5] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, "Emotional speech: Towards a new generation of databases," *Speech Communication*, vol. 40, no. 1-2, pp. 33–60, 2003.

[6] D. Ververidis and C. Kotropoulos, "A review of emotional speech databases," in *PCI 2003, 9th Panhellenic Conference on Informatics, November 1-23, 2003, Thessaloniki, Greece*, 2003, pp. 560–574.

[7] I. S. Engbert and A. V. Hansen, "Documentation of the danish emotional speech database des," Center for PersonKommunikation, Aalborg University, Denmark, Tech. Rep., 2007.

[8] 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.

[9] J. Hansen and S. Bou-Ghazale, "Getting started with susas: A speech under simulated and actual stress database," in *Proc. EUROSPEECH-97*, vol. 4, Rhodes, Greece, 1997, pp. 1743–1746.

[10] 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 Journal (IMAVIS), Special Issue on Visual and Multimodal Analysis of Human Spontaneous Behavior*, 2009, 17 pages.

[11] 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.

[12] 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.

[13] B. Schuller, M. Wimmer, D. Arsic, G. Rigoll, and B. Radig, "Audiovisual behavior modeling by combined feature spaces," in *Proc. ICASSP 2007*, vol. II. Honolulu, Hawaii, USA: IEEE, 2007, pp. 733–736.

[14] 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: ISCA, 2008, pp. 597–600.

[15] S. Steininger, F. Schiel, O. Dioubina, and S. Raubold, "Development of user-state conventions for the multimodal corpus in smartkom," in *Proc. Workshop on Multimodal Resources and Multimodal Systems Evaluation*, Las Palmas, 2002, pp. 33–37.

[16] 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*, A. Paiva, R. Prada, and R. W. Picard, Eds. Berlin-Heidelberg: Springer, 2007, pp. 488–500.

[17] 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.

[18] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK book (v3.4)*. Cambridge, UK: Cambridge University Press, 2006.

[19] F. Eyben, M. Wöllmer, and B. Schuller, "openEAR – Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit," in *Proc. Affective Computing and Intelligent Interaction (ACII)*. Amsterdam, The Netherlands: IEEE, 2009.