TOWARDS MORE REALITY IN THE RECOGNITION OF EMOTIONAL SPEECH

Björn Schuller[1], Dino Seppi[2], Anton Batliner[3], Andreas Maier[3], and Stefan Steidl[3]

[1]Institute for Human-Machine Communication, Technische Universität München, Germany
Schuller@IEEE.org
[2]ITC-irst, Trento, Italy, Seppi@itc.it
[3]Lehrstuhl für Mustererkennung, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
{Batliner, Andreas.Maier, Stefan.Steidl}@informatik.uni-erlangen.de

ABSTRACT

As automatic emotion recognition based on speech matures, new challenges can be faced. We therefore address the major aspects in view of potential applications in the field, to benchmark today's emotion recognition systems and bridge the gap between commercial interest and current performances: acted vs. spontaneous speech, realistic emotions, noise and microphone conditions, and speaker independence. Three different data-sets are used: the Berlin Emotional Speech Database, the Danish Emotional Speech Database, and the spontaneous AIBO Emotion Corpus. By using different feature types such as word- or turn-based statistics, manual versus forced alignment, and optimization techniques we show how to best cope with this demanding task and how noise addition or different microphone positions affect emotion recognition.

Index Terms— Emotion Recognition, Affective Computing, Noise Robustness, Spontaneous Emotions

1. INTRODUCTION

A number of approaches aiming at automatic recognition of emotion out of speech utterances have been presented over the last decade [1, 2]. As the accuracies reported increase with novel features introduced, optimized feature spaces, and classifier tuning, new challenges can be faced: recognition of emotion out of spontaneous field data independent of the speaker. Thereby noisy environments have to be considered, cf. [3, 4]. We discuss these aspects and show results for two popular acted sets recorded in studio conditions with step-wisely added noise, and for a spontaneous set recorded with a headset, a room microphone, and artificially reverberated speech. Different types of features relying on word- or turn-based statistical analysis with automatic versus manual annotation-based segmentation, their selection and classification are explained prior to the presentation of experimental results and their discussion.

2. DATABASES

2.1. Acted Data

In order to provide results on public corpora we firstly decided for the popular Danish Emotional Speech Corpus (DES) [5]. In this database the four emotions *anger*, *joy*, *sadness*, and *surprise* of the MPEG-4 set plus *neutrality* are contained. Four professional Danish actors, two of them female, simulated the word 'yes' and 'no', 9 sentences, and two text passages in each emotion. We split the text passages into single sentences and thereby obtain 414 phrases in total. The set was recorded in 16 bit, 20 kHz PCM-coding in a sound studio. 20 test-persons, 10 of them female, reclassified the samples in a perception test with an average accuracy of 67.3%.
As second well-known dataset to observe inter-set behavior we chose the Berlin Emotional Speech Database (EMO-DB, henceforth EMO) [6], which consists of 816 phrases in total. The emotion set resembles the "big six" of the MPEG-4 set consisting of *anger*, *disgust*, *fear*, *joy*, *sadness* and *surprise*, besides an exchange of *surprise* in favor of *boredom*, and added *neutrality*. 10 German sentences of emotionally undefined content have been acted in these emotions by 10 professional actors, 5 of them female. Throughout perception tests by 20 subjects, 494 phrases have been chosen that were classified as more than 60% natural and at least 80% clearly assignable. The database is recorded in 16 bit, 16 kHz under studio noise conditions. 84.3% recognition rate is reported for a human perception test.

2.2. Spontaneous Data

The database used is a German corpus with recordings of children communicating with Sony's AIBO pet robot; it is described in more detail in [7, 8, 9, 10]. The speech is spontaneous, because the children were not told to use specific instructions but to talk to the AIBO like they would talk to a friend. They were led to believe that the AIBO is responding to his or her commands, but the robot is actually being controlled by a human operator who causes the AIBO to perform a fixed, predetermined sequence of actions. Sometimes the AIBO behaved disobediently, by that provoking emotional reactions. The data was collected from 51 children (age 10 - 13, 21 male, 30 female) from two different schools ('MONT' and 'OHM'); the recordings took place in the resp. class-rooms. Speech was transmitted with a wireless head set (UT 14/20 TP SHURE UHF-series with microphone WH20TQG) and recorded with a DAT-recorder (sampling rate 48 kHz, quantization 16bit, down-sampled to 16 kHz). Five labelers (advanced students of linguistics) listened to the recordings and annotated independently from each other each word as neutral (default) or as belonging to one of ten other classes. We resort to majority voting (henceforth MV). If three or more labelers agree, the label is attributed to the word; in parentheses, the number of cases with MV is given: *joyful* (101), *surprised* (0), *emphatic* (2528), *helpless* (3), *touchy*, i.e., irritated (225), *angry* (84), *motherese* (1260), *bored* (11), *reprimanding* (310), *rest*, i.e. non-neutral, but not be-

longing to the other categories (3), *neutral* (39169). 4707 words had no MV; all in all, there were 48401 words. Some of the labels are very sparse. Therefore, we down-sampled *neutral* and *emphatic* and mapped *touchy* and *reprimanding*, together with *angry*, onto **A***ngry* as representing different but closely related kinds of negative attitude. (The initial letter is given boldfaced and recte; this letter will be used in the following for referring to these four cover classes. Note that now, **A***ngry* can consist, for instance, of two *touchy* and one *reprimanding* label; thus the number of **A***ngry* cases is far higher than the sum of *touchy*, *reprimanding*, and *angry* MV cases.). This more balanced 4-class problem consists of 1557 words for **A***ngry* (**A**), 1224 words for **M***otherese* (**M**), 1645 words for **E***mphatic* (**E**), and 1645 for **N***eutral* (**N**) [9]. Interlabeler correspondence is dealt with in [9]. These word-based labels were mapped onto turn-based labels yielding 868 **A** (21.7 %), 1347 **E** (33.7 %), 495 **M** (12.4 %), and 1280 **N** (32.0 %), summing up to 3990 (100 %) turn labels. The mapping procedure, the labels, the terminology, and more details are described fully in [10]. This set will be referred to as the 'AIBO Emotion Corpus' (AEC).

3. NOISE AND MICROPHONE CONDITION

3.1. Acted Data With Added Noise

We decided for controlled white noise addition to the samples of the DES and EMO. Such additive noise overlay is a common practice in general speech processing tasks, especially in speech and speaker recognition. While this approach does not take noise influences on the speaking style such as Lombard effect into account, it already forms a reasonable basis and partly covers scenarios as microphone mismatch, cellular/phone channels or voice coding effects. The SNR level is chosen in relative terms with respect to the level of each individual affective speech signal: an SNR of ∞ dB resembles clean speech, while 0 dB represents signal and noise mixed at even level. We investigate the effect of noise addition in 5 dB steps starting from clean speech, moving on to slightly noise overlaid 25dB SNR and terminating at heavily overlaid -10 dB, where the original sample is hardly understandable for a human listener. However, we aim only at investigation of acoustic feature analysis in search for emotional cues. Linguistic analysis is left aside as DES and EMO consist of predefined spoken content.

3.2. Reverberated Spontaneous Data and Room Microphone

The AEC database is available in three versions: the recordings were done with a close-talk microphone which was attached to the child's head as described in sec. 2.2; this version will be called 'close-talk' (CT). For documentary purposes the whole experiment was filmed with a video camera as well. The sound track of the film contains a lot of reverberation and background noises, since the camera's microphone is designed to record the whole scenery in the room, the child was not facing the microphone, and the camera was approximately 3m away from the child. This version will be called 'room microphone' (RM). The third version of the corpus was created using artificial reverberation: the data of the CT version were convoluted with different impulse responses recorded in a different room using multiple speaker positions (four positions arranged equidistantly on one of three concentric circles with the radii $r \in \{60cm, 120cm, 240cm\}$ and alternating echo durations $T_{60} \in \{250ms, 400ms\}$ spanning 180 °; for details refer to [11]). With each of the twelve responses $1/12$ th of the corpus was reverberated. This version will be called 'close-talk reverberated' (CTRV).

4. FEATURES AND CLASSIFICATION

For the experiments described in this paper we use two different feature sets, distinguished by the two labels FS1 (a systematically generated set taking a plethora of acoustic base-contours into account, used for all three databases) and FS2 (a compact knowledge-based prosodic set used for AEC). Roughly 4k acoustic features are obtained in total for FS1 and 25 for FS2. The aim of FS1 is to build a broad feature set for the subsequent feature selection process. N best features are selected per variant of database by applying closed-loop Sequential Forward Floating Search (SFFS) using the classifier's error as optimization criterion (roughly 100, for details see sec. 5). The sets are thereby reduced to save computation time by Gain Ratio-based open-loop-search to 1k prior to SFFS. FS2 was developed by manually extracting only those features which experience and literature proved to be more reliable. Thus we managed to keep this set reasonably limited. The following description sketches the peculiarities of both ensembles; for more in depth discussion of FS1 see [3], while for FS2 see [12]. Although the two sets are developed using different approaches, both have a common pre-processing phase in which selected base-contours are extracted that are well known to carry information about the emotional state of the speaker. The base-contours, similar to the Low-Level-Descriptors known from the MPEG-7 standard, are limited to deal exclusively with acoustic information (results using linguistic information are found in [10]). The original sampling frequency and quantization of the databases is kept, and each 10 ms a 20 ms frame is extracted by weighting the speech signal with a Hamming window-function. An explanation of the implemented base-contours in each feature-set follows:

contour	FS1	FS2
log-energy	√	√
pitch	√	√
harmonics-to-noise ratio	√	-
position, bandwidth & amplitude of formants 1-5	√	-
jitter and shimmer	√	-
16 MFCCs	√	-
spectral flux, centroid, 95%-roll-off	√	-

Table 1. Extracted acoustic base-contours.

Pitch and harmonics-to-noise ratio are based on auto-correlation function (pitch extraction is further improved by Dynamic Programming techniques), and formant analysis relies on Linear Predictive Coding, polynomial roots, and Dynamic Programming, too. Contours are always smoothed using symmetrical moving average low-pass filtering. Spectral features base on DFT-spectral coefficients after dB(A)-correction in order to better model human perception. FS1 aims at broad coverage of prosodic, articulatory, and voice quality aspects; conversely, the much more compact FS2 is exclusively based on supra-segmental prosodic features, as these should convey a relevant part of emotional state for spontaneous speech [10].
Besides different base-contours under analysis in FS1 and FS2 the functionals applied to the base-contours differ. These differences consist in the types of the functionals, which are listed below, but, above all, in the time intervals they are applied on. As far as the time-domain is concerned, the aforementioned functionals can be applied to the whole base-contour of the turn under analysis, or to chunks of the segmented turn. In the former case we obtain turn-level derived features, while in the latter we can distinguish according to the depth of the segmentation chosen. In this work we exploit word-based, chunk-based with variable length (VL), and turn-based (TL)

functional	FS1	FS2
mean & standard deviation	√	√
centroid	√	-
skewness & kurtosis	√	-
quartiles	√	-
ranges	√	-
extremes & relative positions	√	√
zero-crossing-rate	√	-
roll-off-points	√	-
on-/off-set & relative positions	-	√
linear regression coefficients & error	√	√
quadratic regression coefficients	√	-

Table 2. Applied functionals for acoustic feature calculation.

features. FS1 is obtained by adopting always turn-level segmentation: for each turn one final feature vector is extracted by applying the different functionals to the utterance. On the contrary, FS2 uses VL segmentation derived by dividing the base-contour into a certain number of segments proportional to the length of the turn. Furthermore, if a word-segmentation is at disposal, the FS2 feature set can be implemented by replacing VL by that word-based segmentation. This can be achieved in three different ways: through manual annotation (MA), by automatic forced alignment (FA) (this means that the correct *transcription* is at disposal), or by automatic speech recognition (AR). Furthermore, word segmentation allows to add syllable, word durations, and pause information to the feature set, as well as the normalization of the features depending on the word they refer to. The ASR that performed the word-level segmentation is described in [13]. It basically exploits 11 MFCCs plus the energy of the signal and relative delta coefficients for each 16ms frame. Phonemes were modeled by semi-continuous Hidden Markov Models, and an 1-gram language model has been adopted. The quality of the ASR segmentation is illustrated by the following word accuracies obtained by training and disjunctive testing with the same variant: 78.9% (CT), 68.2% (CTRV), 47.5% (RM). The presence of noise causes a considerable degradation of automatic segmentation methods, which are thus the most important source of performance decrease. Therefore, to study this influence, FS2 has been extracted in various versions, each one differing from the other solely by the type of segmentation adopted. As classifier we decided in favor of Random Forests (RF), as they led to slightly better results than Support Vector Machines (SVM) in [10]. In general they are popular due to their ability to create a set of highly accurate Decision Trees, handling of high number of features, and balance error in class population unbalanced data sets. In our configuration 100 trees are grown per forest.

5. EXPERIMENTS

Since datasets are sparse in the field of speech emotion recognition, an evaluation method which allows for training disjunctive tests on all samples seems favorable. As a general evaluation mean we therefore choose the popular j-fold stratified cross validation (SCV). Thereby speaker-independence (SI) is assured, by division into two stratified sub-folds of two different speaker groups (SI-SCV). In the case of DES, 2 vs. 2 speakers, in the case of EMO, 5 vs. 5 speakers - both in gender balance -, and in the case of AEC the schools MONT vs. OHM (see sec. 2.2) have been chosen.

5.1. Results for Acted Data

In tab. 3 we show the effect of white noise addition in various dB levels for the databases DES and EMO. All tests have been carried out using the full 4k identical feature set FS1 in order to focus on the direct effect of noise addition. Note that the use of SVM [3] yielded up to 10% higher accuracy depending on the noise condition. A

[%]	∞dB	20dB	10dB	0dB	-5dB	-10dB
	DES					
RR	53.5	51.3	46.6	44.3	43.7	41.6
CL	54.3	51.2	46.5	43.8	43.3	41.5
F	53.9	51.2	46.5	44.0	43.5	41.5
	EMO					
RR	72.3	71.7	67.6	64.5	64.3	62.9
CL	67.4	65.6	61.9	58.7	58.5	56.5
F	69.8	68.6	64.6	61.5	61.3	59.5

Table 3. Accuracies at selected SNR levels, databases DES and EMO using RF in a 2-fold SI-SCV, and 4k FS1 features. RR abbreviates recognition rate, CL mean class-wise RR, F uniformly weighted harmonic-mean ($2 \cdot RR \cdot CL/(RR+CL)$).

significant decrease in accuracy can be observed for each 5dB step. However, only selected steps are shown in the table due to space limitations.

In tab. 4 effects of feature selection on the accuracy for the databases DES and EMO-DB are depicted. N best thereby denotes the reduced feature set. Reduction helps to increase performance in most cases, but feature sets differ largely at the various noise levels and for the diverse databases. Here we present only the extrema of clean speech and highly noise overlaid -10 dB SNR samples.

Acc. [%]	DES	DES N best	EMO	EMO N best
∞ dB	53.5	57.1	72.3	72.5
-10 dB	41.6	49.4	62.9	66.8

Table 4. Accuracies at selected SNR levels, database DES and EMO using RF in a 2-fold SI-SCV and FS1 vs. N best FS1 features.

5.2. Results for Spontaneous Data

Tab. 5 shows our results for the spontaneous set AEC. Thereby the two generally different feature types introduced in sec. 4 are considered: apart from the analysis on turn-level TL (C1) with FS1 as for DES and EMO-DB , we consider also sub-turn-level analysis (C2-C6) with FS2 as described in sec. 4. This is possible, if a manual annotation MA exists. However, also segmentation by forced alignment in matched conditions FA and based on speech recognizer output AR are shown. If no segmentation is available (C3,C4) chunks of equal length are generated according to the number of words (plus two pause chunks, C3) or proportional to the turn length (C4). In all combinations C1-C6, CT is less challenging than CTRV, and RM proves to be the hardest recognition task. Especially FS1 seems to be heavily influenced by microphone and noise conditions, whereas FS2 seems to be almost invariant in this respect. Comparing C2-C6, word-based emotion recognition seems to be better if the provided segmentation is reliable (i.e. it is manually corrected), though the drop in accuracy is not dramatic (spanning from 2.8% - 4.3% absolute F-value comparing C2 with C6). To demonstrate this observed

[%]	C1	C2	C3	C4	C5	C6
Feature Set	FS1	FS2	FS2	FS2	FS2	FS2
Segments	TL	MA	VL	VL	FA	AR
Transcription	-	MA	MA	-	MA	AR
	Close Talk (CT)					
RR	51.3	53.5	51.7	49.6	49.2	50.0
CL	46.2	51.0	51.0	47.9	46.7	47.1
F	48.6	52.2	51.3	48.7	47.9	48.5
	Close Talk Reverberated (CTRV)					
RR	46.6	52.8	50.9	48.9	49.8	49.5
CL	43.1	50.6	50.5	48.7	47.3	48.3
F	44.8	51.7	50.7	48.8	48.5	48.9
	Room Microphone (RM)					
RR	40.0	52.0	50.3	48.6	49.3	47.0
CL	35.0	49.4	49.7	47.2	48.9	45.7
F	37.3	50.7	50.0	47.9	49.1	46.4

Table 5. Results database AEC using RF in a 2-fold SI-CV, close talk, close talk reverberated, and room microphone, diverse feature combinations C1-C6, MA manual annotation, VL variable length, TL turn-level, FA forced alignment, and AR recognizer output. FS1 features are reduced to 105 (CT), 90 (CTRV), 94 (RM).

loss in accuracy, Fig. 1 displays an exemplary section out of a longer turn which has been classified correctly as **E** in the MA condition (C2) but wrongly as **N** in the FA condition (C5) for CT. Two times, the vowel of *links* (*left*) which perceptually contributes heavily to the labeling as **E** has been neglected by the FA . Such erroneous segmentation is most likely the reason for the lower recognition rates in the FA condition.

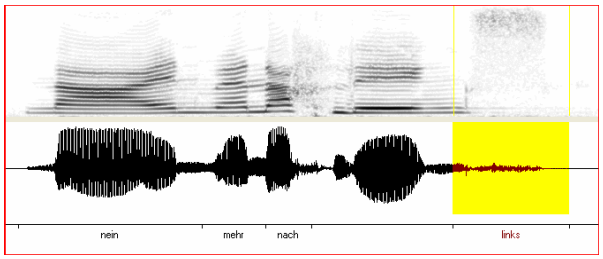

Fig. 1. Example of a corrupted segment boundary in forced alignment FA ('No! More to the left!'). Database AEC, CT.

6. DISCUSSION AND FUTURE WORK

The results presented in sec. 5, esp. in tab. 5, C6, clearly show the difficulty of speaker-independent recognition of spontaneous emotions compared to acted data recorded in studio-noise-conditions. Yet, emotion recognition seems to be less prone to noise than comparable speech processing tasks. If a robust word-segmentation is available, word-based statistics seem to be preferable. In general an adaptation to noise conditions by feature-set optimization proves highly advantageous. In future works we aim at in-depth feature type analysis in various noise and microphone conditions. Furthermore differences between automatically generated and expert-based feature-sets will be addressed. Finally, the influence of the classifier will be investigated.

7. ACKNOWLEDGMENTS

This work is a co-operation between several sites dealing with classification of emotional user states conveyed via speech. This initiative was taken in the European Network of Excellence HUMAINE under the name CEICES (**C**ombining **E**fforts for **I**mproving automatic **C**lassification of **E**motional user **S**tates). It was partly funded by the EU in the projects PF-STAR under grant IST-2001-37599 and HUMAINE under grant IST-2002-50742. The responsibility lies with the authors.

8. REFERENCES

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, January 2001.

[2] C. M. Lee and S. S. Narayanan, "Toward Detecting Emotions in Spoken Dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 293–303, 2005.

[3] B. Schuller, D. Arsic, F. Wallhoff, and G. Rigoll, "Emotion recognition in the noise applying large acoustic feature sets," in *Proc. Speech Prosody 2006*, Dresden, Germany, 2006, ISCA, p. no pagination.

[4] M. You, C. Chen, J. Bu, J. Liu, and J. Tao, "Emotion recognition from noisy speech," in *Proc. ICME 2006*, Toronto, Canada, 2006, IEEE, pp. 1653–1656.

[5] I. S. Engberg and A. V. Hansen, "Documentation of the danish emotional speech database DES," Tech. Rep., Aalborg, Denmark, 1996.

[6] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Proc. Interspeech 2005*, Lisbon, Portugal, 2005, ISCA, pp. 1517–1520.

[7] A. Batliner, S. Steidl, C. Hacker, E. Nöth, and H. Niemann, "Private emotions vs. social interaction - towards new dimensions in research on emotion," in *Proc. Workshop on Adapting the Interaction Style to Affective Factors, 10th Int. Conf. on User Modeling*, Edinburgh, 2005, p. no pagination.

[8] A. Batliner, S. Steidl, C. Hacker, E. Noeth, and H. Niemann, "Tales of tuning – prototyping for automatic classification of emotional user states," in *Proc. Interspeech 2005*, Lisbon, Portugal, 2005, pp. 489–492.

[9] S. Steidl, M. Levit, A. Batliner, E. Nöth, and H. Niemann, "Of all things the measure is man": Automatic classification of emotions and inter-labeler consistency," in *Proc. ICASSP 2005*, Philadelphia, U. S. A., 2005, pp. 317–320.

[10] 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 *Proc. IS-LTC 2006*, Ljubliana, 2006, pp. 240–245.

[11] A. Maier, C. Hacker, S. Steidl, E. Nöth, and H. Niemann, "Robust parallel speech recognition in multiple energy bands," in *Proc. Pattern Recognition 27th DAGM Synopsium*, Vienna, Austria, 2005, ISCA, pp. 133–140.

[12] A. Kießling, *Extraktion und Klassifikation prosodischer Merkmale in der automatischen Sprachverarbeitung*, Shaker, Aachen, 1997.

[13] G. Stemmer, *Modeling Variability in Speech Recognition*, Logos, Berlin, 2005.