Emotion Recognition Using Physiological and Speech Signal in Short-Term Observation

Jonghwa Kim and Elisabeth André

Institute of Computer Science
University of Augsburg, Germany
{Kim, Andre}@informatik.uni-augsburg.de

Abstract. Recently, there has been a significant amount of work on the recognition of emotions from visual, verbal or physiological information. Most approaches to emotion recognition so far concentrate, however, on a single modality while work on the integration of multimodal information, in particular on fusing physiological signals with verbal or visual data, is scarce. In this paper, we analyze various methods for fusing physiological and vocal information and compare the recognition results of the bimodal recognition approach with the results of the unimodal approach.

1 Introduction

Recent work by Picard and others [1] has aroused considerable awareness for the role of emotions in human-computer interaction. Indeed, there is evidence that human-computer interaction is more likely to be accepted by the user if it is sensitive towards the user's affective states. An important prerequisite to realize affective interfaces is a reliable emotion recognition system which guarantees acceptable recognition accuracy, is robust against artefacts, and easily adapts to pragmatic constraints. Most research so far has focused on the analysis of a single modality or an integrated analysis of audio-visual information (see [2] for a comprehensive overview). On the one hand, the integration of multiple modalities raises the expectation of higher recognition rates compared to those obtained from a single modality. On the other hand, more complex classification problems arise.

In this paper, we concentrate on the integration of physiological measures (biosignals) and speech signals for emotion recognition based on short-term observations. Several advantages can be expected when combining biosensor feedback with affective speech. First of all, biosensors allow us to continuously gather information on the users' affective state while the analysis of emotions from speech should only be triggered when the microphone receives speech signals from the user. Secondly, it is much harder for the user to deliberately manipulate biofeedfack than external channels of expression which allows us to largely circumvent the artifact of social masking. Finally, an integrated analysis of biosignals and speech may help to resolve ambiguities and compensate for errors.

When combining multiple modalities, the following questions arise: (1) How to handle conflicting cases between the single modalities? For instance, a user

E. André et al. (Eds.): PIT 2006, LNAI 4021, pp. 53–64, 2006.

may consciously or unconsciously conceal his/her real emotions by external channels of expression, but still reveal them by internal channels of expression. (2) At which level of abstraction should the single modalities be fused in order to increase the accuracy of the recognition results? (3) How should the window sizes of different modalities be synchronized when same emotional cues in the modalities occur with a time discrepancy?

In the next section, we discuss selected previous work. Section 3 reports on the data set we used and describes the features we extracted from 5-channel biosignal and speech signal. Several classification methods are presented including feature-level fusion, decision-level fusion, and a hybrid fusion scheme. In Section 4, we analyze the classification results with respect to the effect of bimodal integration. We conclude this work with a short outlook on future work.

2 Related Work

There is a vast body of literature on the automatic recognition of emotions. With labelled data collected from different modalities, most studies rely on supervised pattern classification approaches for automatic emotion recognition.

Following the long tradition of speech analysis in signal processing, many efforts were taken to recognize affective states from vocal information. As emotion-specific contents in speech, suprasegmental prosodic features including intensity, pitch, and duration of utterance have been widely used in recognition systems. To exploit the dynamic variation along an utterance, Mel-frequency cepstral coefficients (MFCC) are extensively employed. For example, Nwe and colleagues [3] achieved an average accuracy of 66% for six emotions acted by two speakers using 12 MFCC features as input to a discrete hidden Markov model (HMM). A rule-based method for emotion recognition was proposed by Chen [4]. The data used in this work contained two foreign languages (Spanish and Sinhala) for the judges who did not comprehend either language and were therefore able to make their judgment based on vocal expression without being influenced by linguistic/semantic content. Batliner et al. [5] achieved about 40% for a 4-class problem with elicited emotions in spontaneous speech.

Relatively little attention has been paid so far to physiological signals for emotion recognition compared to other channels of expression. A significant series of work has been conducted by Picard and colleagues at MIT Lab. For example, they showed that certain affective states may be recognized by using physiological measures including heart rate, skin conductivity, temperature, muscle activity and respiration velocity [1]. Eight emotions deliberately elicited from a subject in multiple weeks were classified with an overall accuracy of 81%. Nasoz et al. [6] used movie clips to elicit target emotions from 29 subjects and achieved the best recognition accuracy (83%) by applying the Marquardt Backpropagation algorithm. More recently, Wagner et al. [7] presented an approach to the recognition of emotions elicited by music using 4-channel biosignals which were recorded while the subject was listening to music songs, and reached an overall recognition accuracy of 92% for a 4-class problem.

In order to improve the recognition accuracy obtained from unimodal recognition systems, many studies attempted to exploit the advantage of using multimodal information, especially by fusing audio-visual information. For example, De Silva and Ng [8] proposed a rule-based singular classification of audio-visual data recorded from two subjects into six emotion categories. Moreover, they observed that some emotions are easier to identify with audio, such as sadness and fear, and others with video, such as anger and happiness. Using decision-level fusion in bimodal recognition system, a recognition rate of 72% has been reported. A set of singular classification methods was proposed by Chen and Huang [9], in which audio-visual data collected from five subjects was classified into the Ekman's six basic emotions (happiness, sadness, disgust, fear, anger, and surprise). They could improve the performance of decision-level fusion by considering the dominant modality, determined by empirical studies, in case significant discrepancy between the outputs of each unimodal classifier has been observed. Recently, a large-scale audio-visual database was collected by Zeng et al. [10], which contains five HCI-related affective responses (confusion, interest, boredom, and frustration) in addition to seven affects (the six basic emotions + neutral). To classify the 11 emotions subject-dependently, they used the SNoW (Sparse Network of Winnow) classifier with Naive Bayes as the update rule and achieved a recognition accuracy of almost 90% through bimodal fusion while the unimodal classifiers yielded only 45-56%.

Most previous studies have shown that the performance of emotion recognition systems can be improved by the use of audio-visual information. However, it should be noted that the achieved recognition rates depend rather on the type of the underlying database, whether the emotions were from acted, elicited or real-life situation, than the used algorithms and classification methods. Moreover, apart from our previous work [11], work on the integration of biosignals and speech is rare. In this paper, we will investigate in how far the robustness of an emotion recognition system can be increased by integrating both vocal and physiological cues. We will evaluate two fusion methods that combine bimodal information at different levels of abstraction as well as a hybrid integration scheme. Particularly we focus on shorter observations compared to or earlier work.

3 Methodology

3.1 Dataset

We use the same Quiz data set as in our prior work [11]. The dataset contains speech (sampled with 48Kz/16Bit), physiological (using 6-channel biosensors[1]), and visual information from three male German-speaking subjects in their twenties.

To acquire a corpus of spontaneous vocal and physiological emotions, we used a slightly modified version of the quiz "Who wants to be a millionaire?". Questions along with options for answers were presented on a graphical display whose

[1] ECG (electrocardiogram), BVP (blood volume pulse), EMG (electromyogram), RSP (respiration), SC (skin conductivity), Temp (finger temperature).

design was inspired by the corresponding quiz shows on German TV. In order to make sure that we got a sufficient amount of speech data, the subjects were not offered any letters as abbreviations for the single options (as very common in quiz shows on TV), but were forced to produce longer utterances. Furthermore, the users current score was indicated as well as the amount of money s/he may win or loose depending on whether his/er answer is correct or not. Each of the session took about 45 minutes to complete. The subjects were equipped with a directed microphone to interact with a virtual quiz master via spoken natural language utterances. The virtual quiz master was represented by a disembodied voice using the AT&T Natural Voices speech synthesizer. While the users interacted with the system, their bio and speech signals as well as the interaction with the quiz master were recorded.

The quiz experiment was designed in a Wizard-Of-Oz fashion where the quiz agent who presents the quiz is controlled by a human quiz master who guides the actual course of the quiz, following a working script to evoke situations that lead to a certain emotional response. The wizard was allowed to freely type utterances, but also had access to a set of macros that contain pre-defined questions or comments which made it easier for the human wizard to follow the script and to get reproducible situations (see Fig. 1). The wizards working script can be roughly divided into four situations which serve to induce certain emotional states in the user. We make use of a dimensional emotion model which characterizes emotions in terms of the two continuous dimensions of arousal and valence (see [12]). Arousal refers to the intensity of an emotional response. Valence determines whether an emotion is positive or negative and to what degree. Apart from the ease of describing emotional states that cannot be distributed into clear-cut fixed categories, the two dimensions valence and arousal are well suited for emotion recognition. The four phases of the experiment correspond to extreme positions on the axes of the emotion model: (1) low arousal, positive valence, (2) high arousal, positive valence, (3) low arousal, negative valence and (4) high arousal, negative valence.

First, the users are offered a set of very easy questions every user is supposed to know to achieve equal conditions for all of them. This phase is characterized by a slight increase of the score and gentle appraisal of the agent and serves to

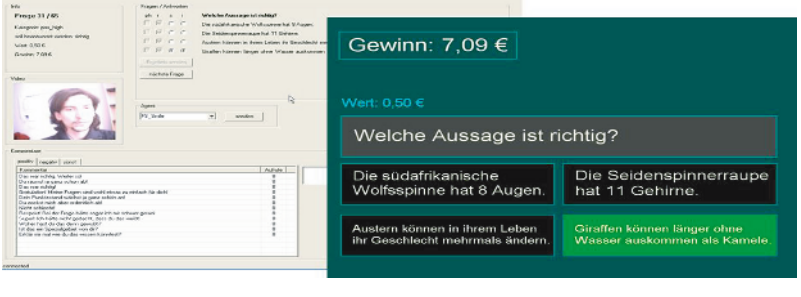

Fig. 1. Interface for the wizard (left) and for the user (right)

induce an emotional state of positive valence and low arousal in the user. In phase 2, the user is confronted with extremely difficult questions nobody is supposed to know. Whatever option the user chooses, the agent pretends the users answer is correct so that the user gets the feeling that s/he hits the right option just by chance. In order to evoke high arousal and positive valence, this phase leads to a high gain of money. During the third phase, we try to stress the user by a mix of solvable and difficult questions that lead, however, not to a drastic loss of money. Furthermore, the agent provides boring information related to the topics addressed in the questions. Thus, the phase should lead to negative valence and low arousal. Finally, the user gets frustrated by unsolvable questions. Whatever option the user chooses, the agent always pretends the answer is wrong resulting in a high loss of money. Furthermore, we include simple questions for which we offer similar-sounding options. The user is supposed to choose the right option, but we make him/er believe that the speech recognizer is not working properly and deliberately select the wrong option. This phase is intended to evoke high arousal and negative valence.

3.2 Synchronized Segmentation of the Bimodal Signals

In our previous work [11], we segmented and labelled the data based on the four experimental phases taking into account that the agreement between coders annotating material of everyday emotions is usually not very high [13]. All speech and physiologial signals that may be interpreted as a response to the same question have been segmented into one chunk and labelled with the emotion corresponding to the experimental phase in which they occurred.

For the analysis described in this paper, the segmentation and labelling was refined by two expert labellers considering the situative context as well as the audio-visual expression of the subjects. In this way, we tried to handle cases where we did not succeed in eliciting the intended emotion. To segment speech and physiological data, we started from verbal phrases. The borders of the segments for both modalities were chosen to lie in the middle of two verbal phrases so that they cover the same time span. For the analysis of speech, we only consider the part of the segment when the verbal phrase occurs while for the analysis of physiological data the complete segment is taken. As a consequence, the observations for speech are usually shorter than the observations for the physiological data, but the length of the corresponding segments is the same which facilitates the later fusion proces. In total, we got 343 samples for classification (343 x 6 channels = 2058 segments in total) from the data set. Based on the four phases of the experiments, our labellers relied on dimensional rating (i.e. labelling within the 4 quadrants of the 2D emotion model). Disagreements between the ratings of the two labellers were discussed and resolved after the annotation process.

Fig. 2 shows a sample segmentation for data from the used channels. The length of the observations varies from 2 to 6 seconds for the speech and from 3

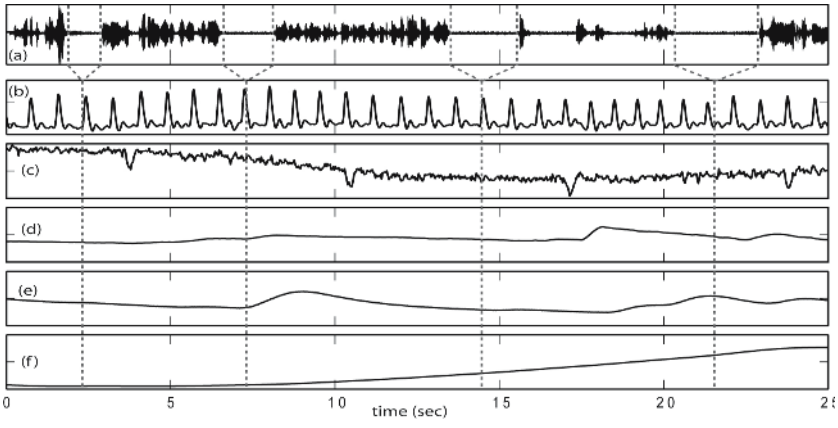

Fig. 2. Segmentation of bimodal signals based on verbal phrases: (a) speech, (b) BVP, (c) EMG, (d) RSP, (e) SC, (f) Temp

to 15 seconds for the biosignals. That is the observations are rather short-term compared to previous studies that start from a segment length between 50 and 300 seconds[2] [15].

3.3 Feature Extraction

An essential step in pattern classification is to extract class-relevant features (preferrably in a compressed form) from the raw signal. Moreover the classification of short-term observations requires more reasonable treatments in signal processing stages, e.g. extracting spectral features in biosignals (containing very low frequencies) within limited bandwidth due to the very short window size.

From physiological data: To remove noisy signals, all segments of the 5-channel biosignals (BVP, EMG, SC, RSP, Temp) are lowpass-filtered using pertinent cut-off frequencies that are empirically determined for each biosensor channel. Differing from [11], we employ the BVP signal instead of the ECG signal and use the Temp signal as an additional channel from the data set. Generally the ECG is measured by using electrodes which do need a firm skin contact, whereas the BVP is measured by using a photoplethysmograph. Hence, using the BVP signal has some advantages such as robustness against motion artefacts during recording process and stable baseline in the signal flow. From the raw signal, we first calculated the 8 subband spectral powers using the conventional 512 points short-time Fourier transform (STFT). To capture the irregularity and the local

[2] Haag et al. [14] used 2 seconds observation of 6-channel biosignals and classified arousal and valence by using a range of specified distance. However, their observation length might be difficult to be compared to our synchronized segmentation of bimodal signals. Moreover using such short length of segment restricts range of usable features, e.g. spectrum features and HRV. They used a limited feature set including 7 fundamental features from each channel for the classification.

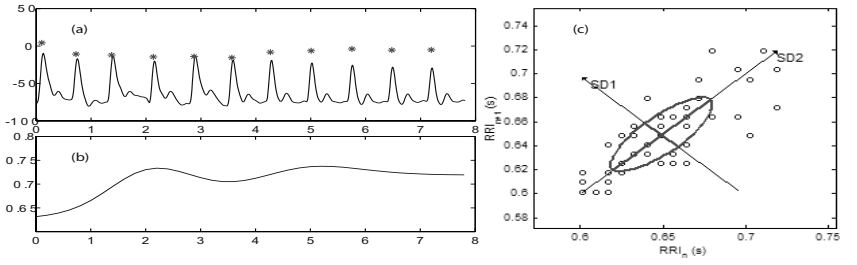

Fig. 3. Example of BVP Analysis: (a) detected pulse interbeats, (b) interpolated PRV like series, (c) Poincaré plot of the PRV

spectral distribution, the spectral entropy is calculated from each subband by converting the spectrum into a PMF-like (Probability Mass Function) form.

Heart rate variability (HRV) is the most frequently used characteristic of the heart activity in biomedical engineering to assess cardiac health. Using the QRS detection algorithm of Pan and Tompkins [16], the HRV like time series (we refer to as PRV)[3] is obtained and typical statistics (mean value, standard deviation, slope, etc.) are calculated from the time series. By calculating the standard deviations in different distances of pulse-pulse interbeats, we also added the Poincaré geometry in the feature set to capture the nature of pulse interval fluctuations. Figure 3 shows an example plot of the geometry. Lastly from the spectrum of the PRV time series, power spectrum densities (PSD) from three subbands are calculated from the ranges of VLF(0-0.04Hz), LF(0.05-0.15 Hz), and HF (0.16-0.4 Hz), respectively and the ratio of LF/HF. Since the RSP signal is quasi periodic we calculated similar types of features like the BVP features including the typical statistics, except for the geometric features and the PSDs. After appropriate detrending the signals using mean value and lowpass filter, we calculated the BRV (time series of the breathing rates) by detecting the peaks using the maxima ranks within zero-crossing. From the SC and EMG signal respectively we calculated 10 features including the mean value, standard deviation, and mean values of first and second derivations. Particularly because of the nature of the signal, the EMG signal required additional pre-processing, such as deep smoothing. The number of transient changes (occurrences) within 4 seconds in SC and EMG signals are calculated from two low-passed signals, very low-passed (SC: 0.08 Hz, EMG: 0.3 Hz) and low-passed signals (SC:0.2 Hz, EMG: 0.8 Hz) respectively. From the Temp signals, three statistical features are calculated: mean value, standard deviation, and ratio of max/min. Finally, we obtained a total of 77 features from the 5-channel biosignals.

From the speech signal: For all segments, the conventional statistics in time domain are calculated, such as mean, absolute extremum, root mean square, standard deviation, energy/power, intensity in dB etc. In frequency domain, three spectrum contents are obtained using the STFT; pitches using a window

[3] Strictly speaking, it is the pulse rate variability (PRV) we use when relying on the BVP instead of the ECG signal.

length of 40 ms,energy spectrum, and formant object using a window length of 25 ms. In addition, 10 MFCCs from each segment are calculated using a window length of 15 ms. From pitch and energy spectrum, also the series of the minima and maxima, and of the distances, magnitudes and steepness between adjacent extrema were obtained. For the MFCCs, we first exponentiated the cepstral coefficients to obtain non-negative values and calculated the spectral entropy as in the case of the biosignal in order to capture the distribution of cepstral energy. From each feature content above, we tried to extract single features (i.e., mean, standard deviation, mean of first and second derivative) representing characteristics (i.e., variance and slop) of each time series vector of spectrum, instead of taking all feature vectors. As a result, we obtained a total of 61 features from the speech segments.

3.4 Feature Selection and Classification

In the next step, we tried to determine which features are most relevant to differentiate each affective state. Reducing the dimension of the feature space has two advantages. First of all, the computational costs are lowered and secondly the removal of noisy information may lead to a better separation of the classes. In all cases, we achieved indeed considerably higher accuracy rates (an increase of about 30 %) when applying sequential backward selection (SBS) to reduce the set of features. Of course, the success of the selection process heavily depends on the employed classifier. Several features were selected by SBS for all three subjects, e.g., the subband spectral entropy from BVP, the number of occurrences in SC and EMG, and the mean values of the MFCCs in the speech features. However, due to the small number of subjects, these findings should not be generalized.

After testing several classification schemes, such as kNN (k-nearest neighbour), MLP (multilayer perception), and LDA (Linear discriminant analysis), we have chosen the LDA classifier which gave the highest accuracy in our case and which we already used for emotion recognition from physiological data in [7].To combine multiple modalities, we need to decide at which level the single modalities should be fused. A straightforward approach is to simply merge the features calculated from each modality (feature-level). An alternative would be to fuse the recognition results at the decision-level based on the outputs of separate unimodal classifiers (decision-level). Finally, we may combine both methods by applying a

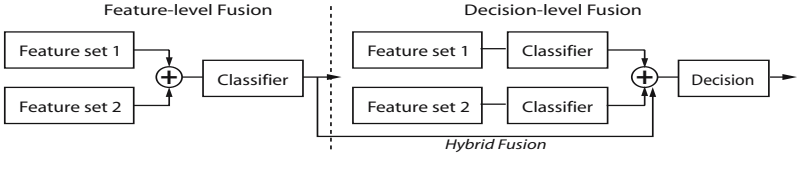

Fig. 4. Considered fusion schemes for integrating bimodal information

Table 1. Recognition results in rates (1.0=100% accuracy) achieved by using SBS, LDA, and leave-one-out cross validation

System	**high/pos**	**high/neg**	**low/neg**	**low/pos**	***Average***
Subject A					
Biosignal	0.95	0.92	0.86	0.85	0.90
Speech signal	0.64	0.75	0.67	0.78	0.71
Feature Fusion	0.91	0.92	1.00	0.85	**0.92**
Decision Fusion	0.64	0.54	0.76	0.67	0.65
Hybrid Fusion	0.86	0.54	0.57	0.59	0.64
Subject B					
Biosignal	0.50	0.79	0.71	0.45	0.61
Speech Single	0.76	0.56	0.74	0.72	0.70
Feature Fusion	0.71	0.56	0.94	0.79	**0.75**
Decision Fusion	0.59	0.68	0.82	0.69	0.70
Hybrid Fusion	0.65	0.64	0.82	0.83	0.73
Subject C					
Bio Single	0.52	0.79	0.70	0.52	0.63
Speech Single	0.55	0.77	0.66	0.71	0.67
Feature Fusion	0.50	0.67	0.84	0.74	**0.69**
Decision Fusion	0.32	0.77	0.74	0.64	0.62
Hybrid Fusion	0.40	0.73	0.86	0.71	0.68
All: Subject-independent					
Bio Single	0.43	0.53	0.54	0.52	0.51
Speech Single	0.40	0.53	0.70	0.53	0.54
Feature Fusion	0.46	0.57	0.63	0.56	**0.55**
Decision Fusion	0.34	0.50	0.70	0.54	0.52
Hybrid Fusion	0.41	0.51	0.70	0.55	0.54

hybrid integration scheme (see Figure 4). We performed both feature-level fusion and decision-level fusion using LDA in combination with SBS. Feature-level fusion is performed by merging the calculated features from each modality into one cumulative structure, selecting the relevant features using SBS, and feeding them to the LDA classifier. Decision-level fusion caters for integrating asynchronous, but temporally correlated modalities. Each modality is first classified independently by the LDA classifier, and the final decision is obtained by fusing the output from the modality-specific classification processes. Three criteria, maximum, average, and product (see [17]) were applied to evaluate the posterior probabilities of the unimodal classifiers at the decision stage. As a further variation of decision-level fusion, we employed a new hybrid scheme of the two fusion methods in which the output of feature-level fusion is also fedas an auxiliary input to the decision-level fusion stage. In Table 1 the best results are summarized that we achieved by the classification schemes we described above. We classified the bimodal data subject-dependently (Subject A, B, and C) and subject-independently (All) since this gave us a deeper insight on what terms the multimodal systems could improve the results of unimodal emotion recognition.

4 Analysis of Results

Table 1 shows that the performance of the unimodal systems varies not only from subject to subject, but also for the single modalities. During our experiment, we could observe individual differences in the physiological and vocal expressions of the three test subjects (see Table 1). As shown in Table 1, the emotions of subject A were more accurately recognized by using biosignals (90 %) than by his voice (71 %) whereas it was inverse for subject B and C (70 % and 67 % for voice and 61 % and 63 % for biosignals). In particular for subject A, the difference between the accuracies of the two modalities is sizable. However, no suggestively dominant modality could be observed in the results of subject-dependent classification in general, which may be used as a decision criterion in the decision-level fusion process to improve the recognition accuracy. Different accuracy rates were also obtained by using the single fusion methods. Overall, we obtained the best results from feature-level fusion. Generally, feature-level fusion is more appropriate for combining modalities with analogous characteristics. For instance, we got an acceptable recognition accuracy of 92 % for subject A when using feature-level fusion which considerably went down, however, when using decision-level or hybrid fusion. As our data show, a high accuracy obtained from one modality may be declined by a relatively low accuracy from another modality when fusing data at the decision level. This observation may indicate the limitations of the decision-level fusion scheme we used, which is based on to a pure arithmetic evaluation of the posterior probabilities at the decision stage rather than a parametric assessment process. Actually, the design of optimal strategies for decision-level fusion, such as the integration of a parametric refinement stage, is still an open research issue. As expected, the accuracy rates for subject-independent classification were not comparable to those obtained for subject-dependent classification. Figure 5 illustrates

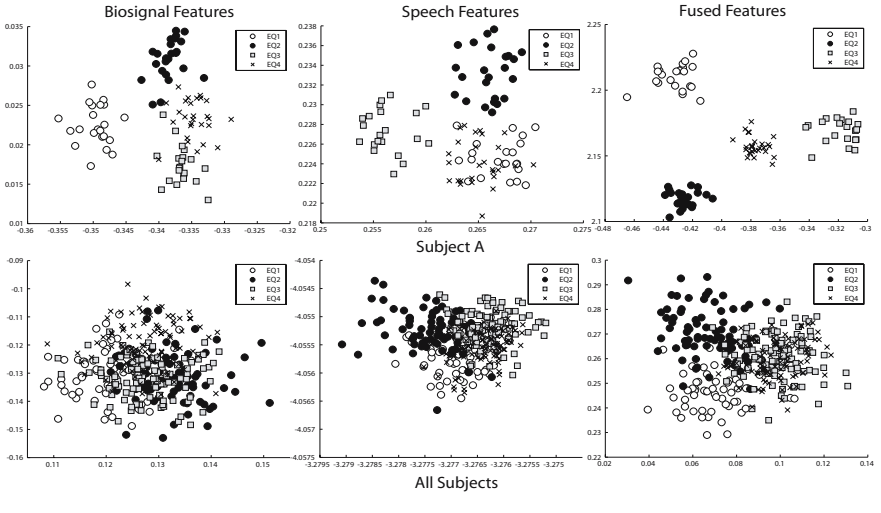

Fig. 5. Fisher projection examples for Subject A and all subjects (person-independent)

examples of Fisher projection which is often used to preview the distribution of the features. Obviously, merging the features of all subjects does not refine the information related to target emotions, but rather leads to scattered class boundaries.

5 Conclusion

In this paper, we treated all stages of emotion analysis, from data collection to classification using short-term observations, and evaluated several fusion methods as well as a hybrid decision scheme. We also compared the results from multimodal classification with the unimodal results. As in our earlier work [11] where we relied on longer observation phases and a different set of features, the best results were obtained by feature-level fusion in combination with feature selection. In this case, not only user-dependent, but also user-independent emotion classification could be improved compared to the unimodal methods.

We did not achieve the same high gains that were achieved for audio-visual data which seems to indicate that speech and physiological data contain less complementary information. Furthermore, in a natural setting like ours, we cannot exclude that the subjects are inconsistent in their emotional expression. Inconsistencies are less likely to occur in scenarios where actors are asked to deliberately express emotions via speech and mimics which explains why fusion algorithms lead to a greater increase of the recognition rate in this case. Ambiguities in emotional expressions are also reflected by work on corpus annotation. For instance, Cowie and colleagues [13] noticed that the agreement between human coders labelling multimodal corpora of everyday emotions was lower when considering both audio and video than when relying on a single modality.

Furthermore some important problems are pointed out, such as the use of posterior probabilities when fusing information with high disparity in accuracy. Most of the existing classifiers used in the literature are generalized methods based on statistics or estimating linear regression of given data. Such classifiers may not be able to capture emotion-specific features and to apply self-adapting decision rules that consider contextual information, for instance. Therefore, the design of an emotion-specific classification scheme is one of the most important issues for the future, and this issue becomes even more critical when classifying combined multimodal observations. To overcome these problems, we need to develop a multilayer fusion scheme with parametric refinement stages in each decision layer.

Acknowledgements

We would like to thank Olena Kuzik for her help with the annotation of the bimodal corpus. All stages from feature extraction to classification are implemented using Matlab/Statistics Toolbox (www.mathworks.com), except for speech feature calculation using Praat (www.praat.org).

References

1. Picard, R., Vyzas, E., Healy, J.: Toward machine emotional intelligence: Analysis of affective physiological state. IEEE Trans. Pattern Anal. and Machine Intell. **23** (2001) 1175–1191
2. Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.G.: Emotion recognition in human-computer interaction. IEEE Signal Processing Mag. **18** (2001) 32–80
3. Nwe, T.L., Wei, F.S., Silva, L.C.D.: Speech based emotion classification. In: IEEE Region 10 International Conference on Electrical Electronic Technology. Volume 1. (2001) 297–301
4. Chen, L.S.: Joint processing of audio-visual information for the recognition of emotional expression in human-computer interaction. PhD thesis, University of Illinois at Urbana-Champaign, Dept. of Electrical Engineering (2000)
5. Batliner, A., Zeissler, V., Frank, C., Adelhardt, J., Shi, R.P., Nöth, E.: We are not amused-but how do you know? user states in a multi-modal dialogue system. In: EUROSPEECH'03, Geneva (2003) 733–736
6. Nasoz, F., Alvarez, K., Lisetti, C., Finkelstein, N.: Emotion recognition from physiological signals for presence technologies. International Journal of Cognition, Technology, and Work - Special Issue on Presence **6(1)** (2003)
7. Wagner, J., Kim, J., André, E.: From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification. In: ICME'05, Amsterdam (2005)
8. De Silva, L.C., Ng, P.C.: Bimodal emotion recognition. In: IEEE International Conf. on Automatic Face and Gesture Recognition. (2000) 332–335
9. Chen, L.S., Huang, T.S.: Emotional expressions in audiovisual human computer interaction. In: ICME-2000. (2000) 423–426
10. Zeng, Z., Tu, J., Liu, M., Zhang, T., Rizzolo, N., Zhang, Z., Huang, T.S., Roth, D., Levinson, S.: Bimodal HCI-related affect recognition. In: ICMI 2004. (2004)
11. Kim, J., André, E., Rehm, M., Vogt, T., Wagner, J.: Integrating information from speech and physiological signals to achieve emotional sensitivity. In: INTERSPEECH-2005, Lisbon, Portugal (2005) 809–812
12. Lang, P.: The emotion probe: Studies of motivation and attention. American Psychologist **50(5)** (1995) 372–385
13. Douglas-Cowie, E., Devillers, L., Martin, J.C., Cowie, R., Savvidou, S., Abrilian, S., Cox, C.: Multimodal Databases of Everyday Emotion: Facing up to Complexity. In: InterSpeech, Lisbon (2005)
14. Haag, A., Goronzy, S., Schaich, P., Williams, J.: Emotion recognition using bio-sensors: First step towards an automatic system. In: Affective Dialogue Systems, Tutorial and Research Workshop, Kloster Irsee, Germany (2004)
15. Kim, K.H., Bang, S.W., Kim, S.R.: Emotion recognition system using short-term monitoring of physiological signals. Med Biol Eng Comput **42** (2004) 419–27
16. Pan, J., Tompkins, W.: A real-time qrs detection algorithm. IEEE Trans. Biomed. Eng. **32** (1985)
17. Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C.H., Kazemzaden, A., Lee, S., Neumann, U., Narayanan, S.: Analysis of emotion recognition using facial expression, speech and multimodal information. In: ICMI'04, State College, Pennsylvania, USA (2004) 205–211