Integrating Information from Speech and Physiological Signals to Achieve Emotional Sensitivity

Jonghwa Kim, Elisabeth André, Matthias Rehm, Thurid Vogt, Johannes Wagner

Institute of Computer Science, University of Augsburg, Germany

{andre,kim}@informatik.uni-augsburg.de

Abstract

Recently, there has been a significant amount of work on the recognition of emotions from speech and biosignals. Most approaches to emotion recognition so far concentrate on a single modality and do not take advantage of the fact that an integrated multimodal analysis may help to resolve ambiguities and compensate for errors. In this paper, we describe various methods for fusing physiological and voice data at the feature-level and the decision-level as well as a hybrid integration scheme. The results of the integrated recognition approach are then compared with the individual recognition results from each modality.

1. Introduction

In the past few years, increasing attempts have been made to exploit emotional cues in man-machine communication. The driving force behind this work is the insight that a user interface is more likely to be accepted by the user if it is sensitive towards the user's affective states. A necessary prerequisite to realize such a behavior is the availability of robust methods for the recognition of emotions from various channels of expression as well as the context of interaction. Most research so far has focused on the analysis of a single modality or an integrated analysis of audio-visual data (see [1] for a comprehensive overview). In this paper, we will concentrate on speech in combination with physiological measures, a topic which has been largely neglected so far.

There are several advantages of using biosensor feedback in addition to 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 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? A straightforward approach is to compute a joint feature set for the single modalities for which several joint statistical pattern classifiers are tested. An alternative would be to fuse the recognition results at the decision-level based on the outputs of separate unimodal classifiers. Finally, we may combine feature-level and decision-level fusion by applying a hybrid integration scheme.

In this paper, we describe the results of a Wizard-of-Oz experiment we conducted in order to acquire a corpus of spontaneous vocal and physiological data that reveal information on the user's emotional state. We then use the corpus to evaluate the three different fusion methods and compare them with the unimodal recognition approaches. In particular, we are interested in the question of whether a bimodal analysis may indeed improve the accuracy of the recognition results.

2. Related Work

There is a vast body of literature on the automatic recognition of emotions with the aim to improve human-computer interaction. With labeled data collected from different modalities, most studies have used supervised pattern classification approaches for automatic emotion recognition systems. We briefly review some of these works.

With 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 the recognition systems. Mel-frequency cepstral coefficients (MFCC) are also extensively used feature in the literature to exploit dynamic variation along an utterance, for example, Nwe et al. [2] achieved an average accuracy of 66% for six basic emotions acted from two speakers by using 12-MFCC features as input to discrete hidden Markov model (HMM). A rule-based method for emotion recognition was proposed by Chen [3]. Particularly, 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 on vocal emotions without possible influence of linguistic/semantic content. Batliner et al. [4] 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, in the work [5] they showed that certain affective states may be recognized by using physiological measures including heart rate, skin conductivity, temperature, muscle activity and respiration velocity. Eight emotions deliberately elicited from a subject in multiple weeks were classified with overall accuracy of 81%. Nasoz et al. [6] used movie clips based on a study by Gross and Levenson [7] 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. [8] aimed at recognizing musical emotions by 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

the unimodal recognition system, many studies attempted to exploit the advantage of using multimodal information, especially by fusing audiovisual information. For example, De Silva and Ng [9] proposed a rule-based singular classification of audiovisual data recorded from two subjects into six emotion categories. Moreover, by comparing the outputs of each unimodal system, 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 [10], in which audio-visual data collected from five subjects was classified into 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. [11], which contains four HCI-related affective responses (confusion, interest, boredom, and frustration) in addition to seven basic emotions. 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 of these 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 recognition accuracies achieved in literature depend rather on the type of underlying database, whether the emotions were from acted, elicited or real-life situation, than the used algorithms and classification methods. Moreover, there has been scarcely any literature on emotion recognition by integrating biosignals and speech. 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. In particular, we will evaluate two fusion methods that combine the bimodal information at different levels of abstraction as well as a hybrid integration scheme.

3. Experimental Setting

As a test bed for our experiment, we implemented a system that was inspired by the quiz "Who wants to be a millionaire?". Questions along with options for answers were presented on a graphical display whose 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 user's 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.

The three test subjects of our experiment were all students - three males in their twenties. All subjects were native speakers of German, which was also the language for the experiment. 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 subjects interacted with the system, their physiological feedback was monitored by 4-channel biosensors to record electromyogram (EMG) at the nape of the neck, electrocardiogram (ECG), skin conductivity (SC) and respiration change (RSP). In addition, we recorded the interaction between the user and the quiz master and captured a visual impression of the user on video.

In the experimental setting, the agent is controlled through a Wizard-of-Oz interface by a human quiz master who guides the quiz, following a working script to evoke situations that lead to a certain emotional response. To achieve this, the quiz master may for example ask extremely difficult questions that nobody is supposed to know combined with a high loss of already gained money. 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.

The wizard's 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 [8]. The four phases of the experiment correspond to extreme positions on the axes of the emotional model (phase 1: low arousal, positive valence, phase 2: high arousal, positive valence, phase 3: low arousal, negative valence, phase 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 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 user's 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.

4. Recognition of Emotions from Speech and Biosignals

In the following, we first describe the employed unimodal methods to recognize emotions from speech and biosignals as well as three different fusion methods. To ease the integration of the recognition results, the same emotion model (see Section 3) is

applied to all unimodal classifiers.

4.1. Unimodal emotion recognition

Speech signals: First, all occurrences of speech were extracted from the videos. In order to synchronize the speech signals with window length of the physiological data, all speech signals belonging to the same question were analyzed together. As a consequence, we obtained about 60 speech segments per subject. Features were extracted following the same procedure as in [13] which was inspired by [14]: For every utterance, pitch, energy and 12 MFCCs as a function of time were calculated. From pitch, also the series of the minima and maxima, and of distances, magnitudes and steepness between adjacent extrema were obtained. Energy was treated like pitch, additionally using the first and second derivatives and their minima and maxima series. For the MFCCs, first and second derivatives were calculated with minima and maxima from all coefficients. Altogether, this amounted to 1280 features. Then, redundant features were removed by correlation-based feature subset selection [15]. This resulted in a new subset of 15-20 dimensional feature vectors. The numbers vary as for every task (speaker-dependent/independent) different features were removed. But a clear tendency towards 1 or 2 pitch minima or maxima related features and the rest being MFCC related features could be observed.

Physiological signals: All segments of physiological signal are firstly lowpass-filtered with pertinent cutoff-frequencies, which are empirically determined for each biosensor channel, in order to remove noisy samples. In the case of subject-dependent classification the baselines of 4-channel signals were calculated. Overall 26 features per segment were extracted, typical statistics such as mean and standard deviation as well as spectral/subband features from the periodic signals (ECG, RSP). Thus we have 26-dimensional feature vector per segment which varied between 10 to 115 seconds (on the average 42 seconds). Then all feature vectors were normalized using standard deviation and mean value.

For both speech and physiological signal, Fisher's linear discriminant function (LDF), a linear combination of components weighted by known prior probability, was used to classify four emotional states, i.e., positive/high, positive/low, negative/high, and negative/low. We note that the feature vectors extracted from speech and physiological signal might well be used to train dynamic models such as HMM, but we have chosen the LDF classifier because the feature vector of the two modalities consists of global-level features that are extracted at the same segment length for both signals. The sequential forward selection method (SFS) was employed to obtain a new subset of features which contains the most emotion-relevant features that maximize the performance of the classifier. Leave-one-out method was used to train and to test all classifiers.

4.2. Bimodal emotion recognition

We expect that the bimodal approach relying on the combination of vocal-physiological data may give better performance of emotion recognition system through complementary and redundant intereffects between the two modalities in the decision process. Three different fusion approaches were implemented to exploit the advantage of using two modalities for emotion recognition. In feature-level fusion, the features of both modalities are simply merged and provided as input to a single classifier. Thereby, we also attempted to extract the most significant features from the fused features by SFS to compare the results. In decision-level fusion, the outputs of two unimodal classifiers for speech and biosignals were integrated according to a given set of criteria. We employed posterior probability criteria used in [16]. As a further variation, we applied majority voting to the decision process, according to the recognition rates of each emotion from unimodal classifiers. Finally, we employed a new hybrid scheme of the two fusion methods in which the ouput of feature-level fusion is also fed as an auxiliary input to the decision-level fusion stage.

5. Results and Discussion

We classified the bimodal data subject-dependently (User A, User B, and User 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. Table 1 shows the classification results where selected features (by SFS) of both modalities were used as input to the classifiers. During our experiment, we could observe individual differences in the physiological and vocal expressions of the three test subjects. Indeed, as showed in Table 1, the emotions of User A and User C were more accurately recognized by using biosignals (77 % and 85%) than by their voice (68% and 76%) whereas it was inverse for the case of User B (75% for voice and 60% for biosignals). However, any suggestively dominant modality could not be observed in the results of subject-dependent classification, which may be used as a decision criterion in the decision-level fusion process to improve the recognition accuracy.

System	pos/ low	pos/ high	neg/ low	neg/ high	Quote
User A					
Bio Single	0.67	0.80	0.73	0.87	0.77
Speech Single	0.53	0.87	0.60	0.73	0.68
Feature Fusion	0.67	0.80	0.87	0.80	***0.78***
Decision Fusion	0.60	0.80	0.80	0.87	0.77
Hybrid Fusion	0.60	0.80	0.80	0.87	0.77
User B					
Bio Single	0.47	0.60	0.60	0.73	0.60
Speech Single	0.73	0.80	0.73	0.73	***0.75***
Feature Fusion	0.73	0.80	0.73	0.73	***0.75***
Decision Fusion	0.80	0.73	0.67	0.80	***0.75***
Hybrid Fusion	0.80	0.73	0.67	0.80	***0.75***
User C					
Bio Single	0.73	0.87	0.87	0.92	0.85
Speech Single	0.73	0.73	0.73	0.85	0.76
Feature Fusion	0.80	0.93	0.93	1.00	***0.92***
Decision Fusion	0.73	0.87	0.87	0.92	0.85
Hybrid Fusion	0.80	0.93	0.87	0.92	0.88
All					
Bio Single	0.58	0.57	0.36	0.61	0.53
Speech Single	0.38	0.59	0.60	0.52	0.52
Feature Fusion	0.56	0.64	0.71	0.73	***0.66***
Decision Fusion	0.51	0.70	0.47	0.61	057
Hybrid Fusion	0.51	0.68	0.53	0.68	0.60

Table 1: Summary of recognition results (rates) from uni- and bimodal systems with feature selection. Best results are printed in bold.

Moreover Table 1 shows that the performance of the unimodal systems varies not only from subject to subject, but also

System	User A	User B	User C	All
Bio Single	0.57	0.43	0.59	0.47
Speech Single	0.50	0.63	0.69	0.45
Feature Fusion	0.65	0.37	0.68	0.51
Decision Fusion	0.68	***0.63***	0.74	0.54
Hybrid Fusion	***0.70***	0.58	***0.78***	***0.55***

Table 2: Summary of recognition results (rates) from uni- and bimodal systems without feature selection.

for the single emotional states. For instance, classification of the physiological data shows a poor result of only 36% for neg/low while the single speech system achieved 60% for the same emotional state. In contrast, recognition accuracy of pos/low was just 38% for the speech system, but 58% for the physiological system. From our speech corpus, we obtained much better results (55%) than we achieved for the SmartKom corpus in an earlier experiment (41%) for four emotion categories (see [13]), which we regard as evidence that we succeeded quite well in inducing the intended emotions in the subject. Nevertheless, the results from biosignals were by far not as good as in the experiment we presented in our previous work [8] (up to 92% for four emotional states) using 4-channel biosignals. Here, we need to consider that in this work the subject was asked to put himself into a certain emotion while this process was supported by listening to music. In contrast, the participants in the experiment described in this paper were not told that we were interested in analyzing their emotions.

Tables 1 and 2 show that the overall recognition rate for the bimodal system surpassed or was equal to the results of the best unimodal system regardless of whether features were selected beforehand or not. When performing feature selection, the best results were obtained for feature-level fusion (see Table 1). When classifying the data subject-independently, feature-level fusion led to an improvement from 53% to 66%. When omitting feature selection, hybrid feature selection yielded the best results for User A and C. For User B, we obtained the best results using decision-level fusion.

6. Conclusion

In this paper, we presented a Wizard-of-Oz study we conducted to acquire a multimodal corpus of emotional speech and biosignals. An analysis of the corpus revealed great individual differences in the degree of expression for the single modalities which emphasizes the added value of approaches relying on more than one modality. We evaluated several fusion methods as well as a hybrid recognition scheme and compared them with the unimodal recognition methods. 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. Our experiment is based on the assumption that we actually succeeded in eliciting the intended emotional states during the complete phases of the experiment. A comparison of the results with the outcome we got for the SmartKom corpus seems to indicate that this objective was achieved quite well. In the future, we will refine our analysis by studying the subjects' reactions individually which might lead to higher recognition rates.

7. Acknowledgements

This work was partially funded by a grant from the DFG in the graduate program 256 and by the EU Network of Excellence Humaine. We would like to thank Nikolaus Bee for his help with the implementation and conduction of the experiment.

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 Mag.*, vol. 18, pp. 32–80, 2001.

[2] T. L. Nwe, F. S. Wei, and L. C. D. Silva, "Speech based emotion classification," in *IEEE Region 10 International Conference on Electrical Electronic Technology*, vol. 1, Aug. 2001, pp. 297–301.

[3] L. S. Chen, "Joint processing of audio-visual information for the recognition of emotional expression in human-computer interaction," Ph.D. dissertation, University of Illinois at Urbana-Champaign, Dept. of Electrical Engineering, 2000.

[4] A. Batliner, V. Zeissler, C. Frank, J. Adelhardt, R. P. Shi, and E. Nöth, "We are not amused-but how do you know? user states in a multi-modal dialogue system," in *EUROSPEECH'03*, Geneva, 2003, pp. 733–736.

[5] R. Picard, E. Vyzas, and J. Healy, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 23, no. 10, pp. 1175–1191, 2001.

[6] F. Nasoz, K. Alvarez, C. Lisetti, and N. Finkelstein, "Emotion recognition from physiological signals for presence technologies," *International Journal of Cognition, Technology, and Work - Special Issue on Presence*, vol. 6(1), 2003.

[7] J. J. Gross and R. W. Levenson, "Emotion elicitation using films," *Cognition and Emotion*, vol. 9, pp. 87–108, 1995.

[8] J. Wagner, J. Kim, and E. André, "From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification," in *ICME'05*, Amsterdam, July 2005.

[9] L. C. De Silva and P. C. Ng, "Bimodal emotion recognition," in *IEEE International Conf. on Automatic Face and Gesture Recognition*, March 2000, pp. 332–335.

[10] L. S. Chen and T. S. Huang, "Emotional expressions in audiovisual human computer interaction," in *International Conf. on Multimedia and Expo (ICME)*, 2000, pp. 423–426.

[11] Z. Zeng, J. Tu, M. Liu, T. Zhang, N. Rizzolo, Z. Zhang, T. S. Huang, D. Roth, and S. Levinson, "Bimodal HCI-related affect recognition," in *ICMI'04, The 6th International Conf. on Multimodal Interfaces*, Oct. 2004.

[12] P. Lang, "The emotion probe: Studies of motivation and attention," *American Psychologist*, vol. 50(5), pp. 372–385, 1995.

[13] T. Vogt and E. André, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition," in *ICME'05*, Amsterdam, 2005.

[14] P.-Y. Oudeyer, "The production and recognition of emotions in speech: features and algorithms," *International Journal of Human-Computer Studies*, vol. 59, no. 1-2, pp. 157–183, 2003.

[15] M. A. Hall, "Correlation-based feature subset selection for machine learning," Master's thesis, U. of Waikato, New Zealand, 1998.

[16] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. H. Lee, A. Kazemzaden, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expression, speech and multimodal information," in *ICMI'04*, State College, Pennsylvania, USA, Oct. 2004, pp. 205–211.