COMBINING CATEGORICAL AND PRIMITIVES-BASED EMOTION RECOGNITION

M. Grimm[1], *E. Mower*[2], *K. Kroschel*[1], *and S. Narayanan*[2]

[1]Institut für Nachrichtentechnik (INT),
Universität Karlsruhe (TH), Karlsruhe, Germany,
Contact: `grimm@int.uni-karlsruhe.de`

[2]Speech Analysis and Interpretation Lab (SAIL)
University of Southern California (USC), Los Angeles, CA, USA

ABSTRACT

This paper brings together two current trends in emotion recognition: feature-based categorical classification and primitives-based dynamic emotion estimation. In this study, listeners rated a database of acted emotions using the three-dimensional emotion primitive space of *valence, activation,* and *dominance*.

The emotion primitives were estimated by a fuzzy logic classifier using acoustic features. The evaluation results were also used to calculate class representations of the emotion categories *happy, angry, sad,* and *neutral* in the three-dimensional emotion space. Speaker-dependent variations of the emotion clusters in the 3D emotion space were observed for *happy* sentences in particular.

The estimated emotion primitives were classified into the four classes using a *k*NN classifier. The recognition rate was 83.5% and thus significantly better than a direct classification from acoustic features. This study also provides a comparison of estimation errors of emotion primitives estimation and classification rates of emotion classification.

1. INTRODUCTION

Emotion recognition in speech plays an important role in man-machine-interaction and provides enriched descriptions for content and style-based speech data mining. In many cases, it is not only important, *what* a person says, but also *how* it is expressed. In recent years, the main focus of emotion recognition has been on the classification of utterances into a few coarse emotion categories, such as *happy, neutral, sad, angry* [1, 2]. In a valence-appraisal approach only binary emotion recognition has been studied, e.g. *negative vs. non-negative* [3, 4], or *negative vs. positive* [5]. There are some studies on how to represent emotions in a multi-dimensional emotion space (see Cowie [6] for an overview, [7, 8]). One powerful representation is in terms of the three emotional attributes ("primitives") namely *valence* (positive vs. negative), *activation* (excitation level high vs. low), and *dominance* (apparent srength or weakness of the speaker) [7].

In this paper we investigate the relations between categorical emotion classes and their realization in the three-dimensional emotion space. We present both a rule-based estimation system of the emotion primitives from acoustic features and a mapping from this three-dimensional emotion space to conventional emotion categories.

For some applications, individual emotion class probabilities are required, such as for instance, user frustration detection. For time-continuous emotion tracking, however, it is more reasonable to estimate generic emotion components. The benefit of clustering in the emotional primitive space is that such clustering lends itself to categorical emotion estimation while at the same time providing a basis for gradual and continuous automatic assessment of emotions.

To our knowledge, this paper reports on the first study that provides a direct comparison of categorical emotion recognition rates in terms of classical confusion matrices on the one hand, and emotion estimation errors expressed by distance measures in the 3D emotion space on the other hand.

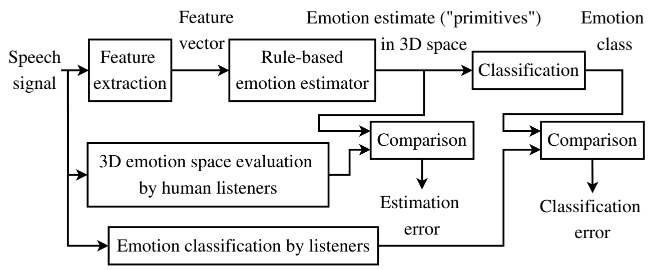

Fig. 1. System for categorical and dimensional emotion recognition described in this paper.

The rest of the paper is organized as follows. Section 2 introduces the data we use. Section 3 describes both the human evaluation of emotional speech in a categorical way and in terms of the three emotion primitives. Section 4 presents details of estimating the three-dimensional emotion primitives from speech using a rule-based *fuzzy logic* classifier, and de-

	Angry	Happy	Neutral	Sad	Other
Angry	**80.3**	2.2	4.1	0.7	12.7
Happy	3.2	**75.6**	11.8	1.3	8.1
Neutral	1.2	0.4	**84.0**	11.8	2.6
Sad	0.3	0.6	6.3	**87.5**	5.3

Table 1. Confusion matrix of emotion class labeling of EMA corpus, in percent, by four human listeners ($\kappa = 0.48$).

	Valence	Activation	Dominance
Std. deviation	0.35	0.36	0.35
Correlation coeff.	0.63	0.79	0.75

Table 2. Standard deviation $\bar{\sigma}$ and correlation coefficient $\bar{r}$ for the emotion primitives evaluation of the EMA corpus by 18 human listeners, averaged over all speakers and all sentences.

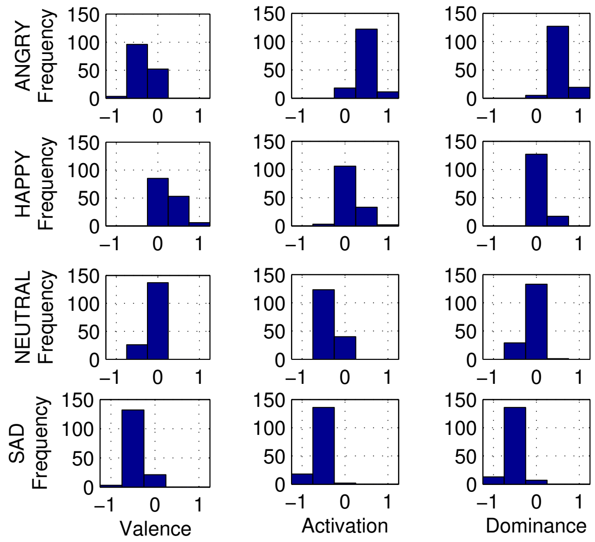

Fig. 2. Three-dimensional emotion evaluation of classes *angry, happy, neutral, sad* for the EMA corpus by 18 evaluators.

riving subsequent mapping to the emotion classes. The results are compared to those achieved by directly classifying acoustic features to emotion classes. Section 5 draws some conclusions and outlines future work.

2. DATA

For this study, we used the *EMA Corpus* [9]. In total, it contains 680 sentences of emotional speech, produced by one professional (f) and two non-professional (1f/1m) speakers. The female speakers produced 10, and the male speaker produced 14 sentences, each in the 4 different emotions *happy, angry, sad,* and *neutral* with 5 repetitions each [9]. All sentences are in English, spoken by native American English speakers.
The sampling frequency was 16 kHz, with 16 bit resolution.

3. EMOTION EVALUATION

3.1. Categorical emotion evaluation

The EMA database was evaluated by 4 native speakers of American English. They chose between the emotions *happy, angry, sad, neutral,* and *other*. On average, 81.8% of the acted emotions were recognized by the listeners. Table 1 shows the averaged confusion matrix of all three speakers in the database.
The evaluator agreement, corrected for chance agreement, was measured using the kappa statistics [10] with $\kappa \in [0,1]$. In our case, we got $\kappa = 0.48$ indicating moderate to high evaluator agreement.

3.2. Three-dimensional emotion evaluation

We adopt the appraisal-power concept of emotion space from Kehrein [7] using the three dimensions of emotion attributes *valence, activation,* and *dominance*. The EMA corpus was evaluated by 18 evaluators along the three dimensions. For emotion assessment, a text-free evaluation tool based on Self Assessment Manikins (SAMs) [11] was used. For each of the emotion components, the evaluators had to choose one out of five given iconic images depicting the level of the attribute.
In contrast to [11], we chose the axes scaled to the range $[-1,+1]$. For better comparison, they are oriented from negative to positive (*valence*), from calm to excited (*activation*), and from weak to strong (*dominance*).
As a counterpart to the categorical classification confusion matrix above, the standard deviation of the evaluations was calculated. On average over all 680 sentences, it was between 0.35 and 0.36 for all all primitives; see Tab. 2 for details.
The evaluators show moderate to high inter-evaluator agreement as can be derived from Pearson's correlation coefficient. On average, the correlation coefficient between an evaluator's rating and the mean value of all other evaluators was 0.63, 0.79 and 0.75 for *valence*, *activation*, and *dominance*, respectively. Agreement on *valence* was not as high as on the other dimensions, c.f. Tab. 2. This might be due to the overall more narrow distribution of *valence* values in the database resulting in a greater effect for disagreement on particular sentences.
In the following, only those 614 sentences (90%) of the EMA database were used that had been evaluated with a deviation of not more than one manikin. Thereby, the average standard deviation was slightly reduced to 0.34, 0.36, and 0.34, respectively.

3.3. Emotion classes in three-dimensional emotion space

To study the relationship between emotion categories and their location in emotion space, we analyzed the three-dimensional evaluation of the EMA database for each emotion separately. Fig. 2 shows the emotion space distri-

Emotion	EMA corpus Valence	Activation	Dominance	Schröder *et al.*[1] Evaluation	Activation	Power	Cowie *et al.*[2] Evaluation	Activation
Angry	-0.35 ±0.17	0.46 ±0.18	0.53 ±0.14	-0.35	0.35	-0.34	-0.70	0.65
Happy	0.31 ±0.17	0.16±0.15	0.12 ±0.10	0.40	0.29	0.13	0.54	0.48
Neutral	-0.16 ±0.09	-0.32±0.09	-0.14±0.10	0	0	0	0	0
Sad	-0.43 ±0.12	-0.57 ±0.13	-0.54±0.15	-0.43	-0.09	-0.55	-0.80	-0.15

Table 3. Comparison of emotion class centroids in the 3D emotion space.

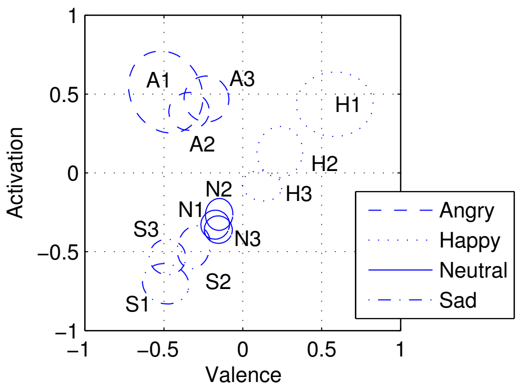

Fig. 3. Emotion classes in the *valence-activation* plane, as a result of the 3D attribute evaluation. For each of the 3 speakers, 4 emotion subspaces are shown as calculated from the mean values of all 18 evaluators.

butions for each class for all three speakers as a result of 18 evaluator's decisions.

Angry was found to be very negative, very excited and very strong. *Happy* was moderately positive and excited. *Neutral* in our database was moderately negative, and moderately calm and weak, showing the greatest values of standard deviation. These speaker-dependent *neutral* values could be used as a baseline for the emotion recognition. *Sad* was found to be negative, calm and weak, forming an antipode to *happy*.

The emotion evaluation using the primitives also reveals why the human listeners' recognition of the category *happy* was worse than other categories (c.f. Tab. 1): The perceived values of *valence* are only moderately positive (c.f. Fig. 2, second row, left column).

We calculated the centroids and the covariances for each emotion class. As a result of the 3D emotion space evaluation, each class was found to be concentrated in an individual subspace of the emotion space. Since *activation* and *dominance* were highly correlated ($r = 0.9$), Fig. 3 shows the projection of the 2σ-regions on the *valence-activation* plane. The high correlation might be due to the selected emotions in this database (e.g. there was no *fear* emotion which would probably have positive *activation* but negative *dominance* values).

[1]Original values in the range of [-100,+100] were scaled to the range [-1,+1] for comparison (c.f. Tab. 3 in [8]).

[2]Given values were read from Fig. 3 in [12] and are only approximate.

The centroids and covariances for each emotion class varied for different speakers, as shown in Fig. 3. In particular, the *happy* sentences were perceived significantly different for the individual speakers. The average values of the class centroids are given in Tab. 3.

We compared these results to values found in the literature [12, 8]. Cowie *et al.* use a 2D evaluation-activation space [12], while Schröder *et al.* use a 3D evaluation-activation-power space [8]. They derive their results from evaluation tests using the Feeltrace tool [13] and an additional word lexicon for the power values. Apart from *neutral*, which they *define* to be (0,0,0), the results are similar but not identical, c.f. Tab. 3. The differences might be caused by the different data and evaluation systems used. To our knowledge, the values given in Tab. 3 are the first comparable results achieved by the use of the text-free emotion evaluation method based on SAMs.

4. EMOTION CLASSIFICATION

For emotion classification, we extract 20 prosodic and 26 spectral features from the speech signal [14]. The prosodic features include statistical parameters of both the pitch and the energy contour, as well as timing related features. The spectral features are the mean values and standard deviation of 13 MFCC's.

4.1. Categorical emotion classification

For comparison, we performed a direct emotion classification based on the 46 features extracted from the acoustic signal. The feature vector dimension was reduced to 17 using PCA and an eigenvalue threshold of 0.01. As a classifier we implemented a Mahalanobis distance classifier using the covariance matrices calculated from the training data. On average, the recognition rate of this multiple classification task was 54%. Using a k-Nearest Neighbor (kNN) classifier and a Euclidean distance measure improved the recognition rate to 58% ($k = 5$). We chose the same classifier for the mapping from 3D emotion estimates to emotion classes (Sec. 4.3).

	Valence		Activation		Dominance	
	EV	*CC*	*EV*	*CC*	*EV*	*CC*
Error $\bar{e}$	0.21	0.21	0.17	0.15	0.18	0.17
Correlation $\bar{r}$	0.67	0.70	0.89	0.90	0.85	0.87

Table 4. Emotion estimation error and correlation results using fuzzy logic, mean values for all speakers. The references are the average evaluator rating (*EV*) and the class centroids (*CC*), respectively.

4.2. Three-dimensional emotion estimation

We use a fuzzy logic inference system to estimate the three emotion attributes from speech features. Fuzzy logic was also applied in the emotion recognition context by Lee *et al.* [15] and Huang *et al.* [16], but not to estimate the aforementioned 3 emotion components. Fuzzy logic is a reasonable choice because of the fuzzy nature of emotion description and perception [16].
The rules in our inference engine are derived from the correlation between the acoustic features and the emotion reference as described in Sec. 3.2. The greatest correlation coefficients are found between energy features as well as some spectral features and *activation* and *dominance* ($r > 0.8$). Correlation to *valence* varies significantly for the different speakers.

For the fuzzy logic system, we use 3 membership functions for both input and output. Each acoustic feature is processed to membership grades of the linguistic variables *low, medium,* and *high*. *Valence* is represented by the linguistic variables *negative, neutral*, and *positive*. Similarly, *activation* is represented by *calm, neutral,* and *excited*. *Dominance* is represented by *weak, neutral*, and *strong*. For each linguistic input feature variable we define 3 rules that relate the fuzzy input variables to the fuzzy output variables. The details of aggregation, implication and defuzzification are reported in [14]. As a result of the defuzzification we get one estimate for each of the emotion primitives. We scale the results by a constant factor of 1.64 to map the range of the defuzzification output, [-0.61, +0.61], to the initial range of [-1, +1]. Fig. 4 shows the estimates, projected onto the *valence-activation* plane.
The emotion estimates were compared to the emotion reference. We considered two different references, (1) the evaluators' mean rating, taken individually for each sentence (*EV*), and (2) the class centroids of the underlying acted emotion as computed in Sec. 3.3 (*CC*). Overall, we observed a mean error of 0.28 when compared to either the evaluators' mean or the speaker-dependent class centroids, as computed above. The details for each emotion component are shown in Tab. 4.
The mean correlation between the emotion estimates and the reference was 0.80 when the reference was the evaluators' rating, and 0.82 when the reference was given by the class centroids. The details for each emotion component are also reported in Tab. 4. These results indicate a low estimation error compared to the standard deviation achieved by human labeling. The correlation between estimates and reference is high. Probably the estimation results for the class centroids as a reference are better because the evaluators' agreement is only moderate, c.f. Tab. 2.

	Angry	Happy	Neutral	Sad
Angry	**91.9**	2.0	4.3	1.9
Happy	18.8	**80.5**	0.7	0.0
Neutral	0.7	0.0	**85.4**	13.9
Sad	0.0	0.0	26.6	**73.4**

Table 5. Confusion matrix of emotion classification from three-dimensional emotion components using a *k*NN classifier ($k = 7$).

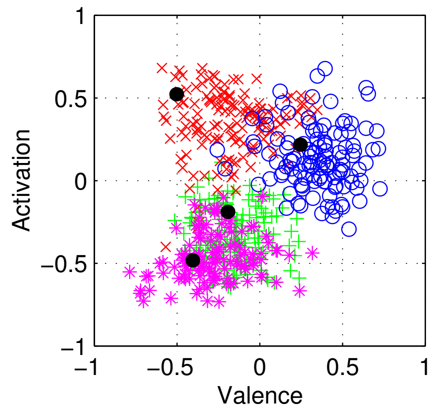

Fig. 4. Emotion class distribution of the primitive estimates in *valence-activation* plane. The emotion classes are *angry* (x), *happy* (o), *neutral* (+), and *sad* (*). The initial class centroids of the evaluation are included for comparison (•).

4.3. Emotion classification from the three-dimensional estimates of emotion primitives

As a final step, we classified the three-dimensional emotion attribute estimates into the 4 emotion classes. This procedure allows for a comparison of the calculated estimation errors to "classical" recognition rates. For classification we used a Mahalanobis distance-based classifier and the class centroids and covariances as calculated in Sec. 3.3. An average recognition rate of 73.3% was achieved for the classification of 4 emotions. Using a *k*NN classifier (with Leave-One-Out cross validation) improved the results significantly. The mean recognition rate was 83.5% using the best parameter set of $k = 7$ and the 3D emotion estimates based on the evaluator ratings (*EV*). The recognition rate was 81.2% when the 3D emotion estimates were based on the class centroids (*CC*). The confusion matrix of the best results is shown in Tab. 5.
The classification errors are mainly due to *neutral* - *sad* mis-

classifications. These reflect the estimated class distributions (c.f. Fig. 4) and the confusion seen in the linsteners' evaluations.

5. CONCLUSION AND OUTLOOK

In this paper we investigated the feasability of emotion recognition based on three primitive attributes, *valence, activation* and *dominance*. We compared classification based on continuous characterization of emotional attributes to direct classification into emotion categories. The 3D emotion estimation is particularly suited for time-continuous emotion estimation of natural, and therefore value-continuous, emotions. Using a database of acted emotions *angry, happy, neutral,* and *sad* we demonstrated the performance of the 3D emotion recognition method.

The standard deviation of 3D evaluation was found to be moderately low ($\bar{\sigma} = 0.35$), where the range of value of the primitives was $[-1, +1]$. The correlation between different evaluators was moderate to high ($0.6 < r < 0.8$). We showed that the emotion classes form separable subspaces in the emotion primitive space, as a function of the speaker. The significant speaker-dependency in the centroids of the emotion class *happy*, e.g., stresses the fact that just one category label for all "happy" utterances is not enough to capture emotion intensities or dynamics.

The 3D emotion components were automatically estimated using a fuzzy logic inference system. On average, the estimation error was 0.28 and thus even slightly below the evaluators' standard deviation. The correlation to the reference was higher than human agreement ($0.7 \leq r \leq 0.9$). Both assessment and estimation was better for *activation* and *dominance* than for *valence*.

For comparison, the 3D estimates were classified into the four emotion classes, achieving a recognition rate of 83.5%. This was significantly higher than a direct classification of the acoustic features into four classes. Note that both classifiers were distance-based *k*NN classifiers. Since for our data with defined emotion categories we calculated the estimation error and the recognition rate, these results can serve as a rule of thumb for future research on authentic emotions of spontaneous speech, where it is not possible to calculate emotion recognition rates due to the gradual nature of authentic emotions.

There are many applications which would benefit from estimating gradual variation of emotion values as presented in this study. The estimation of emotion primitives lends itself to dynamic representations of emotions and the ability to adapt the emotion baseline to individual speakers.

In future work, both the speaker dependency and the listener (evaluator) dependency of emotion should be considered in the classification methods. More sophisticated features and classifiers will further improve the recognition results.

6. ACKNOWLEDGMENT

This work was supported by grants of the German Academic Exchange Service (DAAD) and the Collaborative Research Center (SFB) 588 of the Deutsche Forschungsgemeinschaft.

7. REFERENCES

[1] S. Yildirim et al., "An acoustic study of emotions expressed in speech," in *Proc. ICSLP*, Jeju Island, Korea, October 2004, pp. 2193–2196.

[2] D. Ververidis and C. Kotropoulos, "Automatic speech classification to five emotional states based on gender information," in *Proc. EUSIPCO*, 2004, pp. 341–344.

[3] C.M. Lee, S. Narayanan, and R. Pieraccini, "Recognition of negative emotions from the speech signal," in *Proc. IEEE Wsh. ASRU*, Trento, Italy, 2001.

[4] J. Liscombe, G. Riccardi, and D. Hakkani-Tür, "Using context to improve emotion detection in spoken dialog systems," in *Proc. Eurospeech*, 2005, pp. 1845–1848.

[5] L. Vidrascu and L. Devillers, "Detection of real-life emotions in call centers," in *Proc. Eurospeech*, 2005.

[6] R. Cowie and R.R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech Communication*, vol. 40, pp. 5–32, 2003.

[7] R. Kehrein, "The prosody of authentic emotions," in *Proc. Speech Prosody Conf.*, 2002, pp. 423–426.

[8] M. Schröder, Roddy Cowie, Ellen Douglas-Cowie, Machiel Westerdijk, and Stan Gielen, "Acoustic correlates of emotion dimensions in view of speech synthesis," in *Proc. Eurospeech*, 2001, vol. 1, pp. 87–90.

[9] S. Lee, S. Yildirim, A. Kazemzadeh, and S. Narayanan, "An articulatory study of emotional speech production," in *Proc. Eurospeech*, 2005, pp. 497–500.

[10] J. Carletta, "Assessing agreement on classification tasks: the kappa statistic," *Comput. Linguist.*, vol. 22, no. 2, pp. 249–254, 1996.

[11] M. Grimm and K. Kroschel, "Evaluation of natural emotions using self assessment manikins," in *Proc. ASRU*, San Juan, Puerto Rico, December 2005, pp. 381–385.

[12] R. Cowie et al., "What a neural net needs to know about emotion words," *Proc. CSCC*, pp. 5311–5316, 1999.

[13] R. Cowie et al., "'FEELTRACE': An instrument for recording perceived emotion in real time," in *Proc. ISCA Wsh. on Speech and Emotion*, 2000, pp. 19–24.

[14] M. Grimm and K. Kroschel, "Rule-based emotion classification using acoustic features," in *Proc. Int. Conf. on Telemedicine and Multimedia Communication*, 2005.

[15] C.M. Lee and S. Narayanan, "Emotion recognition using a data-driven fuzzy inference system," in *Proc. Eurospeech*, Geneva, 2003, pp. 157–160.

[16] C.-F. Huang and M. Akagi, "A multi-layer fuzzy logical model for emotional speech perception," in *Proc. Eurospeech*, Lisbon, Portugal, 2005, pp. 417–420.