RULE-BASED EMOTION CLASSIFICATION USING ACOUSTIC FEATURES

Michael Grimm, Kristian Kroschel

Universität Karlsruhe (TH), Institut für Nachrichtentechnik (INT)
Kaiserstr.12, 76128 Karlsruhe, Germany
{grimm,kroschel}@int.uni-karlsruhe.de

Abstract

In this paper we present current results in emotion classification based on features extracted from the speech signal and a Fuzzy Logic inference system. Our emotion recognition system uses the pitch and energy contour of the speech signal as basic features describing the emotional state of the speaker. Additional features are related to the speaking rate and spectral characteristics.
The classifier uses training data acquired from speakers expressing authentic emotions. All utterances are labelled according to a 3-dimensional emotion space representation by several human evaluators. Based thereon, we apply a rule-based fuzzy inference system which gives us an estimation of the emotional state expressed in each utterance. The rules are derived from the correlation between the acoustic features and the emotional content attested by the evaluators. In comparison to human evaluation consent, the recognition results show to be a promising basis for emotion recognition.

1. Introduction

Recognizing a human speaker's emotional state can be helpful in various contexts. The most promising one is probably the man-machine interaction, e.g. the communication between an assisting robot in the household and its human user. For the robot, emotion recognition and classification is an important step in understanding its environment.

But also for patient monitoring, emotion recognition might be helpful. An automatic classification of a patient's emotional expressions reveals much information on his/her pre-sent state.

In recent years, several works on human emotion recognition have been published. In most studies features are extracted from the acoustic signal of the person's speech, e.g. [1, 2, 3]. In addition, facial expression analysis is also commonly performed to estimate the emotion, e.g. [4, 5]. All current work assumes some simplifications of the nature of emotions - for example by working on artificial emotions performed by actors, or by restricting the emotion discrimination to few classes only. Some of these simplifications are reasonable, of course, since the generation, communication, and perception of human emotion is a very complex matter. In our emotion classification approach, however, we want to approximate in some sense the human emotion perception. Therefore we use only authentic emotion data consisting of recordings of non-professional speakers.
As a classifier, we chose a fuzzy inference system based on simple *IF ... THEN ...* rules. This is supported by the fact that human emotion perception is rather fuzzy in terms of distinction of single emotion realizations. A comparable approach was followed by Lee and Narayanan [6], who performed a classification into two classes *negative* and *non-negative* emotions.
The rule system is constructed using the correlation between acoustic features and the emotion contents for each utterance in a database. Since we work on authentic emotions of spontaneous speech, the emotional content is attested prior to this classification by human listeners' evaluation [7]. To assess our classifier, we compare the classifier output to this attested emotion by the evaluators.

The paper is organized as follows: Section 2 presents the basics of our emotion space approach as well as the evaluation output. Section 3 describes the acoustic features extracted from the speech signal. Section 4 introduces the database worked on as well as the parameters that are derived from the database in order to initialize the classifier. In section 5, the setup of the fuzzy logic classifier is presented with its elements fuzzification, inference, and defuzzification. Some emphasis is laid on the construction of the rule system. Section 6 presents significant results of the classification of the data in our database. Section 7 draws some conclusions and gives an outlook on related future work.

2. Emotion representation

As described in [7], we use a three-dimensional emotion space representation. Every emotion to be recognized consists of three components: *Valence, Activation*, and *Dominance*. *Valence (V)* describes how positive or negative an emotion is felt. The second dimension, *Activation (A)* expresses the degree of excitation. *Dominance (D)*

refers to how strong or weak the speaker appears.
The classifier output is compared to an emotion estimate based on the average ratings of 5 human evaluators, each weighted by their correlation to the mean value of all other evaluators. For evaluation, 5 text-free iconic representations of each emotion component are used [7]. To keep consistency with this assessment method, for each of the entities *Valence, Activation*, and *Dominance*, the scale of the classifier output ranges from 1 to 5 with 3 being the neutral value. The orientation of the scales is as follows: For *Valence*, 1 corresponds to a very positive, and 5 to a very negative emotion. For *Activation*, 1 corresponds to a very excited, and 5 to a very calm speaker. For *Dominance*, 1 corresponds to a very weak, and 5 to very strong speaker.

3. Feature extraction

Currently, $M = 46$ acoustic features are extracted from the speech signal. They are listed in the first column of Tab.1. They can be divided into the groups of 9 pitch related features, 5 speaking rate related features, 6 volume related features, and 26 spectral features:

Pitch related features: f0 mean value (*f0_mean*), standard deviation (*f0_std*), median (*f0_median*), minimum (*f0_min*), and maximum (*f0_max*), 25% and 75% quantiles (*f0_quartlow, f0_quartup*), difference between f0 maximum and minimum (*f0_range*), difference of quartiles (*f0_quartrange*)

Speaking rate related features: ratio between the durations of unvoiced and voiced segments (*pause_to_speech_ratio*), average duration of voiced segments (*speech_duration_mean*), standard deviation of duration of voiced segments (*speech_duration_std*), average duration of unvoiced segments (*pause_duration_mean*), and standard deviation of duration of voiced segments (*pause_duration_std*)

Volume related features: volume mean (*vol_mean*), standard deviation (*vol_std*), maximum (*vol_max*), 25% and 75% quantiles (*vol_quartlow, vol_quartup*), and difference of quartiles (*vol_quartrange*)

Spectral features: mean value and standard deviation of 13 Mel Frequency Cepstral Coefficients (MFCC) (*mfcc01_mean* to *mfcc13_mean* and *mfcc01_std* to *mfcc13_std*)

All features can be extracted directly from the speech signal without the need of an automatic speech recognition (ASR) unit.
The pitch related features are based on an estimation of the pitch contour for voiced segments of the utterance. This is obtained using the autocorrelation method [8]. The volume based features are derived from the envelope of the time-domain signal.

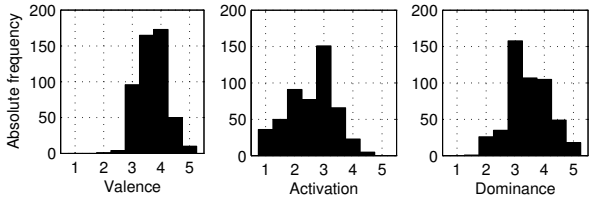

Figure 1: Histogram of emotion occurances in database

As it can be assumed that the features are partly correlated, a *Principal Component Analysis (PCA)* should be applied to obtain uncorrelated features. The computational complexity could be reduced by using a new basis of $\tilde{M} \leq M$ basis vectors, using eigenvalues greater than a given threshold only [9]. This could reduce the number of (transformed) features to $\tilde{M}$, keeping the most significant ones only.

4. Data

4.1. Data acquisition

The database used for this emotion recognition study consists of emotional speech of guests in a TV talk-show. Among the 19 speakers are both male and female subjects. In total, the database contains 499 sentences of emotional speech in German.

The signals are sampled using a sampling rate of 16 kHz and a resolution of 16 bit. The range of emotions in this database, attested by the evaluators, can be seen in Fig.1. A wide range of emotions is covered by the database, especially for the *Activation* and *Dominance* components. *Valence*, however, seems not to be captured in all of its possibilities since the greater part of the database lies in the range of $[2.5, 4.5]$ which signifies neutral or negative emotions. This is due to the topics in talk-shows: In most cases, they cover personal problems of the guests, like family issues or troublesome wife-husband relations. In addition, it has to be mentioned that averaging over several human evaluators of course results in emotion values concentrating more around the neutral values than around the extreme values like 1 or 5.

4.2. Correlation of features

For each of the features m defined in section 3, the correlation between the feature value v_m and each of the emotion components $x^{(i)}$, $i \in \{V, A, D\}$ was calculated. These values are a measure of how the emotion output of the classifier should follow the input features. Therefore, these correlation coefficients have a major influence on the set-up of the classifier (c.f. subsection 5.2).
Let $v_{m,n}$ be the value of the m^{th} feature of the n^{th} signal in the database, $m = 1, \ldots, M$ and $n = 1, \ldots, N$,

respectively. Also, let $x_n^{(i)}$ be the i^{th} emotion component, $i \in \{V, A, D\}$ of the emotion x attested to signal n by the evaluators. Then the correlation $\rho_m^{(i)} = \rho(v_m, x^{(i)})$ between the two sequences $\{v_{m,n}\}_{n=1,\ldots,N}$ and $\{x_n^{(i)}\}_{n=1,\ldots,N}$ can be calculated according to

$$\rho_m^{(i)} = \frac{\sum_{n=1}^{N} (v_{m,n} - \bar{v}_m)\left(x_n^{(i)} - \bar{x}^{(i)}\right)}{\sqrt{\sum_{n=1}^{N} (v_{m,n} - \bar{v}_m)^2}\sqrt{\sum_{n=1}^{N}\left(x_n^{(i)} - \bar{x}^{(i)}\right)^2}}, \quad (1)$$

where mean subtraction was accounted for.

Tab. 1 shows these correlation coefficients for all features and all emotion components. It can be seen that most of the features actually do have a correlation significantly different from 0. Pitch and volume related features show highest correlation, followed by the spectral features. Speech rate related features show the smallest correlation to the emotion components. For example, the correlation between feature *vol_mean* and emotion component *Activation* is $\rho_{10}^{(A)} = -0.77$, and the correlation between feature *pause_duration_mean* and *Valence* is $\rho_{19}^{(V)} = -0.06$.

The emotion entities *Activation* and *Dominance* in general show higher absolute values of correlation than *Valence*.
These correlation coefficients are an essential basis for the initialization of the classifier described below.

5. Classification

As a classifier, we apply a rule-based *fuzzy inference system (FIS)* consisting of the elements *fuzzification, inference,* and *defuzzification* [9]. A separate classifier is constructed for each of the emotion components $x^{(i)}$, $i \in \{V, A, D\}$. Fuzzy logic is used since, in general, all emotion descriptions are fuzzy and vague.
Therefore we switch over to fuzzy description terms for the emotion components (see Fig.2):

$$\begin{aligned} x^{(V)} &\rightarrow B_l^{(V)} \in \mathcal{B}^{(V)} = \{positive, neutral, negative\} \\ x^{(A)} &\rightarrow B_l^{(A)} \in \mathcal{B}^{(A)} = \{excited, neutral, calm\} \\ x^{(D)} &\rightarrow B_l^{(D)} \in \mathcal{B}^{(D)} = \{weak, neutral, strong\} \end{aligned} \quad (2)$$

It is not before the last step of defuzzification that these fuzzy terms are transformed back into crisp numeric values along the scales $x^{(i)} \in [1, 5],\ i \in \{V, A, D\}$.

In the following subsections, each element of the fuzzy inference system is described.

5.1. Fuzzification

First of all, the crisp input variables are fuzzified. These input variables are the M features extracted from speech.

Table 1: *Correlation between acoustic features and emotion components*

m	Feature	Emotion component *V*	*A*	*D*
1	*f0_mean*	0.44	-0.59	0.54
2	*f0_std*	0.20	-0.42	0.40
3	*f0_median*	0.46	-0.58	0.53
4	*f0_min*	0.08	-0.17	0.13
5	*f0_max*	0.15	-0.33	0.30
6	*f0_quartlow*	0.44	-0.51	0.45
7	*f0_quartup*	0.44	-0.61	0.56
8	*f0_range*	0.12	-0.26	0.24
9	*f0_quartrange*	0.25	-0.49	0.47
10	*vol_mean*	0.39	-0.77	0.75
11	*vol_std*	0.22	-0.64	0.66
12	*vol_quartlow*	0.41	-0.63	0.58
13	*vol_quartup*	0.32	-0.73	0.73
14	*vol_max*	0.26	-0.60	0.60
15	*vol_quartrange*	0.19	-0.57	0.60
16	*pause_to_speech_ratio*	-0.10	0.36	-0.38
17	*speech_duration_mean*	0.15	-0.26	0.27
18	*speech_duration_std*	0.12	-0.22	0.22
19	*pause_duration_mean*	-0.06	0.30	-0.28
20	*pause_duration_std*	-0.08	0.26	-0.26
21	*mfcc01_mean*	0.34	-0.73	0.71
22	*mfcc01_std*	-0.20	0.21	-0.17
23	*mfcc02_mean*	-0.36	0.31	-0.26
24	*mfcc02_std*	-0.26	0.41	-0.37
25	*mfcc03_mean*	-0.21	0.39	-0.37
26	*mfcc03_std*	0.06	-0.20	0.23
27	*mfcc04_mean*	-0.08	0.11	-0.11
28	*mfcc04_std*	0.09	-0.22	0.28
29	*mfcc05_mean*	-0.13	0.21	-0.18
30	*mfcc05_std*	0.08	-0.25	0.25
31	*mfcc06_mean*	0.03	0.03	-0.03
32	*mfcc06_std*	0.15	-0.29	0.29
33	*mfcc07_mean*	0.11	0.05	-0.06
34	*mfcc07_std*	-0.06	-0.07	0.06
35	*mfcc08_mean*	0.00	0.06	-0.08
36	*mfcc08_std*	0.11	-0.24	0.25
37	*mfcc09_mean*	-0.11	0.13	-0.13
38	*mfcc09_std*	0.17	-0.35	0.36
39	*mfcc10_mean*	-0.08	0.06	-0.06
40	*mfcc10_std*	0.18	-0.34	0.35
41	*mfcc11_mean*	-0.05	0.06	-0.08
42	*mfcc11_std*	0.32	-0.45	0.42
43	*mfcc12_mean*	-0.12	0.17	-0.17
44	*mfcc12_std*	0.36	-0.52	0.49
45	*mfcc13_mean*	-0.24	0.30	-0.27
46	*mfcc13_std*	0.39	-0.54	0.51

They are transformed to *linguistic variables* with distinct *membership grades*.
We use the fuzzy set

$$\mathcal{A} = \{low, medium, high\} \quad (3)$$

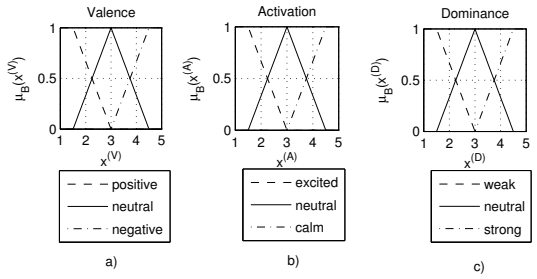

Figure 2: Membership functions of emotion components

of $K = 3$ linguistic variables for the fuzzification of each feature. In contrast to *crisp* logic, each feature may now be both *low* and *medium*, e.g., to a certain membership grade.
The membership grade $\mu_{k,m}$ relating feature m to the linguistic variable $A_k \in \mathcal{A}$ is defined by evaluating the *membership function* $\mu_{A_k,m}(\alpha)$ at the point of the feature value $\alpha = v_m$:

$$\mu_{k,m} = \mu_{A_k,m}(v_m), \tag{4}$$

for $1 \leq m \leq M = 46, 1 \leq k \leq K = 3$.
The membership functions $\mu_{A_k,m}(\alpha)$ are different for each acoustic feature. They are piecewise linear functions, defined by few parameters that are extracted from our training database. If $V_m = \{v_{m,1}, \ldots, v_{m,N}\}$ is the set of the feature values $v_{m,n}$ of all signals in the database, $n = 1, \ldots, N$, we calculate the 10% and 90% quantiles of V_m. This *10-90 range* determines the edges of the K membership functions of v_m. The 10-90 range instead of the complete value range of the feature values is chosen in order to neglect single outliers.

Example: Let us take the first feature *f0_mean* as an example. 10% of all *f0_mean* values in the database are smaller than 175 Hz, and 90% of all *f0_mean* values are smaller than 300 Hz. Therefore we define the membership functions as shown in Fig.3.
If we want to classify a signal whose first feature value is $v_1 = 220$ Hz, this would result in the three membership grades

$$\begin{array}{lcll} \mu_{1,1} = \mu_{A_1}(v_1) & = & 0.27, & \text{for } A_1 = low \\ \mu_{2,1} = \mu_{A_2}(v_1) & = & 0.73, & \text{for } A_2 = medium \\ \mu_{3,1} = \mu_{A_3}(v_1) & = & 0, & \text{for } A_3 = high. \end{array} \tag{5}$$

This is indicated in Fig.3 by the thin lines.

5.2. Rule system

The rule system is the core element of the fuzzy logic classifier. It defines how the fuzzified input variables $A_k \in \mathcal{A}$ are related to the fuzzy output variables $B_l^{(i)} \in \mathcal{B}^{(i)}$. We apply simple *IF... THEN...* rules, always linking one *premise* to one *conclusion*:

$$\text{IF } v_m \text{ is } A_k \text{ THEN } x^{(i)} \text{ is } B_l^{(i)}. \tag{6}$$

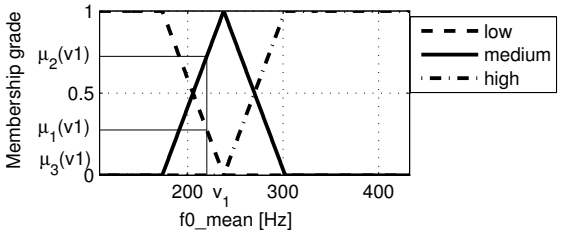

Figure 3: Fuzzy membership functions *low, medium,* and *high* of feature *f0_mean*. The on-/offset of the functions is defined by the quantiles $Q_{10\%} = 175$ Hz and $Q_{90\%} = 300$ Hz, respectively. The thin line indicates the example of $v_1 = 220$ Hz.

Since there are $K = 3$ linguistic variables A_k for each of the $M = 46$ features v_m, we get $K \cdot M = 138$ rules for each of the emotion component $x^{(i)}, i \in \{V, A, D\}$. This is the total number of rules for *all* linguistic variables of $x^{(i)}$. As we use the $L = K = 3$ linguistic variables $B_l^{(i)}, 1 \leq l \leq L$, for the description of $x^{(i)}$, we have $M = 46$ rules for each of the fuzzy output variables $B_l^{(i)}$.
Note that the k^{th} input variable A_k does not necessarily imply a conclusion of the k^{th} output variable $B_k^{(i)}$. The sign of the correlation coefficient $\rho_m^{(i)}$ as stated in (1) defines which output variable $B_l^{(i)}$ is used in the rule:

$$l = \begin{cases} k, & \rho_m^{(i)} \geq 0 \\ K - k + 1, & \rho_m^{(i)} < 0 \end{cases}. \tag{7}$$

The application of the rules is as follows: For each rule relating feature m to the output variable $B_l^{(i)}$, a *degree of support* $H_{l,m}^{(i)}$ of the conclusion is calculated. Since we use only one premise in each rule, this can be calculated as

$$H_{l,m}^{(i)} = \mu_{k,m}, \tag{8}$$

where l and k are related by (7).

Rule weights. As described in section 4.2, the correlation between the acoustic features v_m and the emotion components $x^{(i)}$ varies significantly for the different features. Therefore the correlation coefficients $\rho_m^{(i)}$ as shown in Tab.1 are used as weight factors for the rules. As a consequence, the degrees of support are multiplied by these weights,

$$\tilde{H}_{l,m}^{(i)} = H_{l,m}^{(i)} \cdot \rho_m^{(i)}. \tag{9}$$

Example: The first row of Tab.1 shows the correlation between the feature $m = 1$, *f0_mean*, and the emotion components $x^{(i)}, i \in \{V, A, D\}$. For this example, we use $\rho_1^{(V)} = 0.44$ and $\rho_1^{(A)} = -0.59$.
The rules according to (6) and (7), relating the fuzzy input

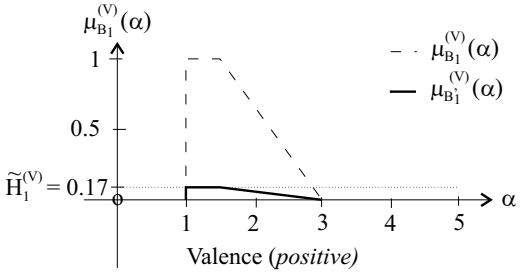

Figure 4: Implication of example value $\tilde{H}_1^{(V)} = 0.17$ to the output membership function of $B_1^{(V)} =$ *positive*

set (3) to the fuzzy output set (2), can be defined as

$$\begin{array}{llllllll}
\text{IF} & v_1 & \text{is} & low & \text{THEN} & x^{(V)} & \text{is} & positive \\
\text{IF} & v_1 & \text{is} & medium & \text{THEN} & x^{(V)} & \text{is} & neutral \\
\text{IF} & v_1 & \text{is} & high & \text{THEN} & x^{(V)} & \text{is} & negative \\
\text{IF} & v_1 & \text{is} & low & \text{THEN} & x^{(A)} & \text{is} & calm \\
\text{IF} & v_1 & \text{is} & medium & \text{THEN} & x^{(A)} & \text{is} & neutral \\
\text{IF} & v_1 & \text{is} & high & \text{THEN} & x^{(A)} & \text{is} & excited.
\end{array} \tag{10}$$

Applying these rules to the membership grades (5) of the fuzzified feature $v_1 = 220$ Hz yields the following degrees of support:

$$\begin{aligned}
\tilde{H}_{1,1}^{(V)} &= 0.27 \cdot 0.44 = 0.1188 \\
\tilde{H}_{2,1}^{(V)} &= 0.73 \cdot 0.44 = 0.3212 \\
\tilde{H}_{3,1}^{(V)} &= 0 \cdot 0.44 = 0 \\
\tilde{H}_{3,1}^{(A)} &= 0.27 \cdot 0.59 = 0.1593 \\
\tilde{H}_{2,1}^{(A)} &= 0.73 \cdot 0.59 = 0.4307 \\
\tilde{H}_{1,1}^{(A)} &= 0 \cdot 0.59 = 0.
\end{aligned} \tag{11}$$

5.3. Aggregation

In the next step, the rules of all features are combined for each output variable. We fuse the M rules of each feature by applying a *maximum* operator. This is done separately for each output variable $B_l^{(i)} \in \mathcal{B}^{(i)}$ of all components $i \in \{V, A, D\}$. Thus the output membership is determined by the greatest degree of support,

$$\tilde{H}_l^{(i)} = \max_{1 \le m \le M} \left\{ \tilde{H}_{l,m}^{(i)} \right\}. \tag{12}$$

This means that the weighted rules are aggregated by an operator of *OR* characteristic.

Example: To continue the previous example, we need to extend it to at least two features, v_1 and v_2. If we want to aggregate all rules of the fuzzy output variable $B_1^{(V)} =$ *positive*, we therefore need to know the degrees of support $\tilde{H}_{1,1}^{(V)} = 0.1188$ and $\tilde{H}_{1,2}^{(V)} = 0.85 \cdot 0.2 = 0.17$,

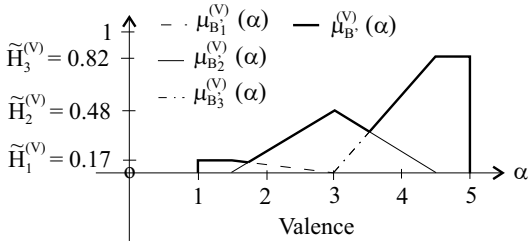

Figure 5: Accumulation of the three output membership functions $\mu_{B_1'}^{(V)}, \mu_{B_2'}^{(V)}$, and $\mu_{B_3'}^{(V)}$ for *Valence* (example values)

where $\mu_{1,2} = 0.85$ was assumed and $\rho_2^{(V)} = 0.2$ was read from Tab.1.
This yields the output membership

$$\tilde{H}_1^{(V)} = \max \left\{ \tilde{H}_{1,1}^{(V)}, \tilde{H}_{1,2}^{(V)} \right\} = 0.17. \tag{13}$$

5.4. Implication

Implication defines how the output membership function is affected by the conclusions of each aggregated rule. At this point of inference the degrees of support $\tilde{H}_l^{(i)}$ of each output variable $B_l^{(i)}$ are mapped on the output membership functions $\mu_{B_l}^{(i)}(\alpha)$. These output membership functions are shown in Fig.2. The implication therefore determines how the general output membership functions are transformed due to the features of a test signal.
We use *product* implication. Therefore we keep the shape of the output membership functions in Fig.2 and scale them by multiplying them with the degrees of support,

$$\mu_{B_l'}^{(i)}(\alpha) = \tilde{H}_l^{(i)} \cdot \mu_{B_l}^{(i)}(\alpha). \tag{14}$$

Example: We have calculated the degree of support of the fuzzy output variable $B_1^{(V)} =$ *positive*, $\tilde{H}_1^{(V)} = 0.17$. Therefore the dashed membership function in Fig.2a) is scaled by the factor 0.17 to represent the value *positive* for the given test signal, see Fig.4.

5.5. Accumulation

As a last step within the inference part, the $L = 3$ linguistic variables $B_l^{(i)}, 1 \le l \le L$, are fused for each emotion component. To obtain one membership contour only, the three membership functions of the output variables are accumulated using *maximum* method.

$$\mu_{B'}^{(i)}(\alpha) = \max_{1 \le l \le L} \left\{ \mu_{B_l'}^{(i)}(\alpha) \right\} \quad \forall \alpha \tag{15}$$

Example: The three output membership functions $\mu_{B_1}^{(V)}, \mu_{B_2}^{(V)}$, and $\mu_{B_3}^{(V)}$ for *Valence* are accumulated for the example values $\tilde{H}_1^{(V)} = 0.17$, $\tilde{H}_2^{(V)} = 0.48$, and $\tilde{H}_3^{(V)} = 0.82$. The results are indicated by the thick line in Fig.5.

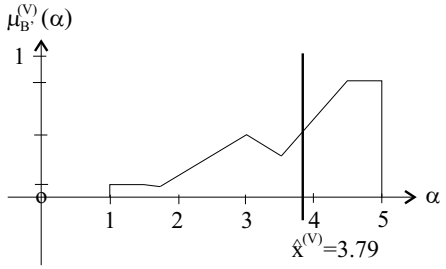

Figure 6: Defuzzification of the output membership function $\mu_{B'}^{(V)}$ for *Valence*, using centroid method (example values)

5.6. Defuzzification

Finally, from the membership contour $\mu_{B'}^{(i)}(\alpha)$ of each emotion component, a crisp estimate has to be calculated. This is achieved using *centroid* method. The centroid $\hat{x}^{(i)}$ of the accumulated membership function $\mu_{B'}^{(i)}(\alpha)$ yields the estimate of the emotion component $i \in \{V, A, D\}$. It can be calculated as

$$\hat{x}^{(i)} = \frac{\int_1^5 \alpha \cdot \mu_{B'}^{(i)}(\alpha) d\alpha}{\int_1^5 \mu_{B'}^{(i)}(\alpha) d\alpha}. \qquad (16)$$

The classifier output therefore consists of 3 estimates of the 3 emotion components: $\hat{x}^{(V)}, \hat{x}^{(A)}, \hat{x}^{(D)}$.

Example: The output membership function $\mu_{B'}^{(V)}$ of the previous example is defuzzified applying centroid method. This yields the emotion component estimate $\hat{x}^{(V)} = 3.79$ as can be seen in Fig.6.

6. Results

The fuzzy inference system presented above was tested on the spontaneous speech database described in section 4. Thus, the rule system was initialized using 499 signals of which both the acoustic features and the emotional content attested by the evaluators was available. Afterwards, each of the sentences was tested as an unknown, new signal.

For each of the test signals, the acoustic features explained in section 3 were extracted from the acoustic signal. These were taken as a crisp input for the fuzzy inference system. The classifier output finally was compared to the emotion attested by the listeners.
For each emotion entity, we calculated the classification error $e^{(i)} = |x^{(i)} - \hat{x}^{(i)}|$. Since the complete database of $N = 499$ utterances was tested, we can also calculate the mean error

$$E^{(i)} = \frac{1}{N} \sum_{n=1}^{N} e_n^{(i)}. \qquad (17)$$

The results are summarized in Tab.2. The classification error obtained was $E^{(V)} = 0.81$ for *Valence*, $E^{(A)} = 0.57$ for *Activation*, and $E^{(D)} = 0.57$ for *Dominance*. These values are to be seen in the scale of our emotion space $x^{(i)} \in [1, 5]$, c.f. sec. 2. Therefore these errors are already in the range of good values, at least for *Activation* and *Dominance*. If we consider that human evaluation agreement, expressed in terms of standard deviation, amounts to average values of 0.6 to 0.8 [7], the results show that our automatic emotion classifier performs comparably well.

Table 2: *Classification results of the fuzzy logic classifier*

	Valence	*Activation*	*Dominance*
Mean error $E^{(i)}$	0.81 ± 0.43	0.57 ± 0.42	0.58 ± 0.40
Correlation coefficient	0.47	0.78	0.75

In addition to the mean error, the correlation coefficient between the true emotion and the emotion estimates was calculated. For the three emotion components, it amounts to 0.47 for *Valence*, and to 0.6 for *Activation* and *Dominance*, respectively. Since these correlation coefficients are significantly greater than 0, it can be concluded that there is indeed a positive correlation between our classifier results and the emotion attested by the evaluators. Again, *Activation* and *Dominance* seem to be better to classify than *Valence*.
Fig.7 shows a direct comparison of the rated emotions and the estimates of the classifier. It can be seen that the majority of the classifier results is located around the neutral values of 3 in the range of moderate emotions, $\hat{x}^{(i)} \in [2, 4]$. This is due to the inference engine giving more weight on a moderate output than on an extreme one. All parameters of the classifier have an impact on the emotion estimates, especially the shape of the membership functions and the defuzzification operator.

It has to be mentioned also, that the speech database used for this study is very demanding. It contains many different speakers. There was no distinction between male and female speakers. There is always some background noise. The nature of spontaneous speech also implies some non-speech parts in the acoustic signal.

The classifier was also tested on a database of one female speaker only ($N = 103$), still containing authentic emotions only. By using the same membership functions and the same methods of implication, aggregation, and defuzzification, a mean error of 0.41 (V), 0.31 (A), and 0.42 (D) was achieved [10]. This indicates that the results of speaker independent emotion classification might also be further improved.

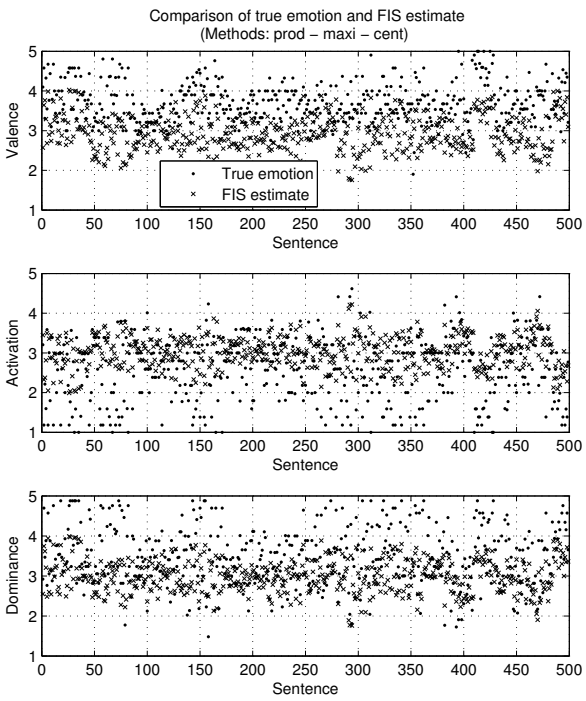

Figure 7: Comparison of estimated emotion and "true" emotion attested by listeners for each signal in the database

7. Conclusion and outlook

In this paper we presented an emotion classifier estimating the three emotion components *Valence*, *Activation* and *Dominance* of an spontaneous emotional speech utterance. We introduced a fuzzy inference system based on rules derived from the correlation between acoustic features extracted from the speech signal and emotion component values attested by human listeners.

The fuzzy logic classification showed to be a promising basis for automatic emotion classification. The classification error is in the range of human evaluation performance. Furthermore, the results are at least moderately correlated with the true emotion.

In following studies, the single parameters of the classifier should be varied to get an optimal performance. Different membership function shapes should be tested, too.

Also acoustic emotion recognition could be combined with other modalities, like facial expression recognition. But especially in the context of patient monitoring, other signals might also be used: blood pressure, heart beat rate, respiratory rate. In a similar manner to the presented work, some further rules related to features extracted from the additional modalities could be defined. This might further improve the emotion recognition results.

8. Acknowledgements

This work is supported by the *Sonderforschungsbereich (SFB) No. 588: "Humanoide Roboter - Lernende und kooperierende multimodale Roboter"* of the *Deutsche Forschungsgesellschaft (DFG)*.

Thanks to Celine Hernandez who contributed a lot to the classifier implementation and testing.

9. References

[1] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech," in *Proc. ICSLP*, Philadelphia (PA), USA, 1996, vol. 3, pp. 1970–1973.

[2] A. Batliner, R. Huber, H. Niemann, E. Nöth, J. Spilker, and K. Fischer, *The recognition of emotion*, pp. 122–130, Springer, Berlin, 2000, in: Wahlster, W. (Ed.), Verbmobil - Foundations of Speech-to-Speech Translations.

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

[4] N. Tsapatsoulis, A. Raouzaiou, S. Kollias, R. Cowie, and E. Douglas-Cowie, *Emotion Recognition and Synthesis Based on MPEG-4 FAPs*, John Wiley & Sons Ltd, Chichester, UK, 2002, in: Pandzic, I.S. and R. Forchheimer (Eds.), MPEG-4 Facial Animation.

[5] N. Sebe, M.S. Lew, I. Cohen, Y. Sun, T. Gevers, and T.S. Huang, "Authentic facial expression analysis," in *Proc. International Conference on Automatic Face and Gesture Recognition (FG'04)*, Seoul, Korea, 2004, pp. 517–522.

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

[7] M. Grimm and K. Kroschel, "Evaluation of natural emotions using self assessment manikins," Submitted for publication to IEEE Wsh. ASRU'05, Cancun, Mexico, 2005.

[8] D. O'Shaughnessy, *Speech Communications - Human and Machine*, John Wiley & Sons Inc, 1999.

[9] K. Kroschel, *Statistische Informationstechnik: Signal- und Mustererkennung, Parameter- und Signalschätzung*, Springer Verlag Berlin, 4th edition, 2004.

[10] C. Hernandez, "Einsatz von Fuzzy Logic zur Erkennung von Emotionen in der Sprache," Studienarbeit, Universität Karlsruhe (TH), Germany, 2005.