Open access • Proceedings Article • DOI:10.1109/IROS.2005.1545341

Prosody based emotion recognition for MEXI — Source link

Anja Austermann, Natascha Esau, Lisa Kleinjohann, Bernd Kleinjohann

Institutions: University of Paderborn

Published on: 05 Dec 2005 - Intelligent Robots and Systems

Topics: Speaker recognition and Prosody

Prosody Based Emotion Recognition for MEXI

Anja Austermann
University of Paderborn
Paderborn, Germany
genetix@cyberspaces.de

Natascha Esau, Lisa Kleinjohann, Bernd Kleinjohann
C-LAB
University of Paderborn
Paderborn, Germany
lisa, bernd, nesau@c-lab.de

***Abstract*— This paper describes the emotion recognition from natural speech as realized for the robot head MEXI. We use a fuzzy logic approach for analysis of prosody in natural speech. Since MEXI often communicates with well known persons but also with unknown humans, for instance at exhibitions, we realized a speaker dependent mode as well as a speaker independent mode in our prosody based emotion recognition. A key point of our approach is that it automatically selects the most significant features from a set of twenty analyzed features based on a training database of speech samples. This is important according to our results, since the set of significant features differs considerably between the distinguished emotions. With our approach we reach average recognition rates of 84% in speaker dependent mode and 60% in speaker independent mode.**

***Index Terms*— Emotion recognition, prosody, fuzzy rules, robot head**

I. Introduction

During the last years the interest in believable agents has grown considerably as well in the software agent domain as in the robotics community. Often one of the major objectives for their design is the development of human like or human oriented interfaces in all domains where humans and robots or general IT-systems communicate. In Japan an entire design stream called KANSEI Information Processing [1] deals with such issues as to make IP systems more acceptable for humans from a subjective point of view. This was also a major objective when building the robot head MEXI, our Machine with Emotionally eXtended Intelligence (Fig. 1).

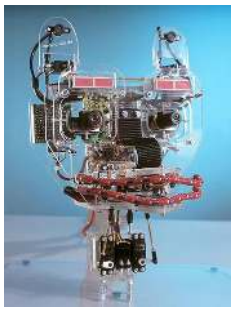

Fig. 1. The robot head MEXI

Since emotions are an evident part of interactions between human beings, they are indispensable also for a robot (head) in order to become believable in communicating with its human counterpart. MEXI's architecture as described in Section 3 was designed in such a way that MEXI should show attitudes in its communication behavior that humans would *interpret* as emotions or drives from their subjective point of view rather than exactly emulating them.

For this purpose MEXI on the one hand has to recognize emotions and on the other hand has to behave emotionally. MEXI recognizes emotions from facial expressions and from natural speech and shows artificial emotions by its facial expression and speech utterances as well. The latter is described elswhere [2], whereas the focus of this paper is on MEXI's emotion recognition from natural speech. For MEXI we developed a fuzzy logic approach for emotion recognition based on the prosody of natural speech, called PROSBER. Although PROSBER is designed for use in MEXI it can also be used as standalone version. In contrast to the majority of existing systems that work in a speaker dependent way in order to increase the recognition rate PROSBER allows the switching between speaker dependent and speaker independent mode. This allows to exploit the advantages of speaker dependent systems when users often interact with MEXI. But it also supports the emotion analysis for users who rarely interact with MEXI.

A key feature of our fuzzy logic emotion recognition approach is that it autonomously selects the most significant speech features (four to six features) from a set of twenty analyzed features based on a training database of speech samples. Thus, on the one hand overfitting is avoided and on the other hand analysis effort is decreased to support real-time emotion recognition. Furthermore the fuzzy classification rules are automatically generated from the training database, which was a considerable advantage for the implementation of our system and will help to improve the system if for instance new training data is available.

The remainder of the paper is structured as follows. Section 2 gives an overview of existing emotion recognition from speech. Sections 3 and 4 describe MEXI's overall architecture and the architecture of PROSBER while Section 5 outlines the principle ideas for fuzzy modeling of speech features and emotions used by PROSBER. Section 6 shows how the automatic generation of fuzzy rule systems is performed during the training phase. Section 7 summarizes the results obtained by our approach and Section 8 gives a summary and outlook.

II. Related work

Emotions and their role in human-computer-interaction (HCI), also called affective computing [3] or KANSEI information processing [1], have become an important research area in the last years. Since the beginning of the nineties several approaches for emulation and recognition of emotions were realized including emotions in HCI. Robot heads [4], [5] and virtual avatars [6] emulate emotions by their facial expressions and speech utterances. Communication systems analyze the facial expressions of their human counterpart in order to find out their current emotional state and since the end of the nineties also approaches for emotion recognition in natural language have been developed. Most of the systems we investigated work in a speaker dependent mode with recognition rates from about 70% to 95 %. An example for a speaker independent system is ASSESS [7] which has a recognition rate of about 55%. However this decrease in the recognition rate for speaker independent systems does not surprise, since even humans could hardly reach emotion recognition rates of 60% from natural language for unknown speakers [8]. Low recognition rates of about 70% were reached by SpeakSoftly [9], a neural net based approach distinguishing five emotions, and Mercury [10], a statistical approach (called maximum a posteriori probability) distinguishing two emotions. Medium recognition rates of about 80% for distinction of four up to seven emotions are reached by RAMSES [11] and the approaches of De Silva [12] and Dellaert [13]. The first two use hidden Markov models whereas Dellaert uses a statistical approach based on an extended K-nearest-neighbors-clustering. For distinguishing two emotions Verbmobil, a neural net approach, reaches considerable recognition rates of about 90%. The highest recognition rate of about 95% was reached by the Sony study using a combination of decision trees and rule systems for distinguishing four emotions [14]. The promising results obtained by the rule based approach in combination with decision trees lastly convinced us to develop our fuzzy rule based approach, although fuzzy logic according to our knowledge up to now was only successfully used for emotion recognition from facial expressions and not for speech based analysis.

III. Overview of MEXI

MEXI realizes an embodied interface that communicates to humans in a way that humans recognize as human or animal like. The robot head can show artifical emotions using its facial expressions, head movements and by its speech output. MEXI is equipped with two cameras and two microphones. MEXI has 15 degrees of freedom (DOF), that are controlled via model craft servo motors and pulse width modulated (PWM) signals. Speakers in MEXI's mouth allow audio output. These facilities allow MEXI to represent a variety of emotions like joy, sadness or anger.

MEXI's software architecture (Fig. 2) is designed according to Nilssons Triple-Tower Architecture, that distinguishes between perception, model and action tower [15]. The Perception component processes MEXI's visual inputs and natural language inputs. The Action Control component controls the actors, i.e. the servo motors for the above mentioned 15 DOF, and the speech synthesis. The Behavior System determines MEXI's behavior in a purely reactive manner. That would allow MEXI to directly react on its visual and natural speech inputs received from its environment by corresponding head movements, facial expressions and natural speech output. The Behavior System is further controlled and configured by MEXI's Emotion Engine that handles its current internal state made up of emotions and drives. Unlike many goal directed agents MEXI has no internal model of its environment to plan and control its behavior but uses its emotions and drives for that purpose. In principle MEXI has two objectives that determine its actions. One is to feel positive emotions and to avoid negative ones. The second objective is to keep its drives at a comfortable (homeostatic) level. The Emotion Engine is responsible for maintaining MEXI's internal state represented by the current strength of emotions and drives. This internal state is used in a feedback loop to configure the Behavior System in such a way that appropriate behaviors are selected in order to meet the two objectives stated above (see [2]). Thus, MEXI does not simply imitate its human counterpart but acts in a pro-active manner in order to "feel good" and satisfy its drives.

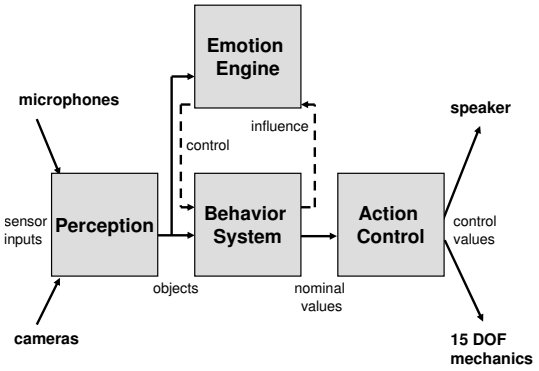

Fig. 2. MEXI's Software Architecture (Overview)

In this paper we concentrate on the speech processing of MEXI. The speech processing allows users to communicate with MEXI in natural language. Besides the visual impressions also the input sentences influence MEXI's emotional state. On the one hand this emotional state is reflected in MEXI's facial expressions and movements. On the other hand the content of an answer MEXI generates and the prosody used for pronouncing the answer are adapted to MEXI's emotional state.

Fig. 3 shows how the speech processing is integrated into MEXI's overall architecture. The Emotion Engine receives spoken sentences as audio files from the component Speech Recognition and analyzes their prosody via PROSBER (PROSsody Based Emotion Recognition). Visual inputs undergo the Vision Preprocessing before they are analyzed

by VISBER (VISion Based Emotion Recognition) for their emotional content. The Emotion Manager is responsible for maintaining MEXI's overall emotional state that is calculated from both, audio and visual, inputs.

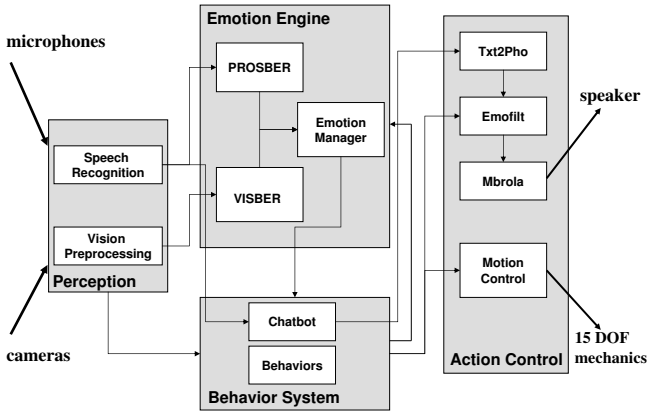

Fig. 3. Architecture of MEXI

For generating answers MEXI uses a slighlty extended commercially available Chatbot, that receives the textual representation of input sentences from the Speech Recognition and generates textual output. The content of the sentences generated by the chatbot is influenced by MEXI's Emotion Manager. If MEXI is happy for instance, by chance corresponding output sentences like "I am happy" are generated. In order to deliver these sentences in emotional speech to the human listener they are first transformed into an audio file by Txt2Pho. Afterwards the audio data is adapted by Emofilt [16] to represent the prosodic features corresponding to MEXI's current emotional state. If MEXI is happy, for instance the velocity of the speech ouput is increased and more syllables are emphasized to create a more vivid prosody. This audio file is then processed by the freely available speech synthesis component Mbrola, such that MEXI speaks the generated answer sentence in natural speech with the respective prosody. The components Behaviors and Motion Control are responsible for generating MEXI's facial expression and movements according to its current emotional state.

In the following we describe the prosody based emotion recognition system PROSBER in more detail.

IV. Architecture of PROSBER

PROSBER is a fuzzy rule based system for emotion recognition from natural speech. It takes single sentences as input and classifies them into the emotion categories happiness, sadness, anger, fear and neutral. PROSBER automatically generates the fuzzy models for emotion recognition. Accordingly two working modes are distinguished, training and recognition, as depicted in Fig. 4. During the training the training samples with well-known emotion values are used to create the fuzzy models for the individual emotions. These fuzzy models are used in the emotion recognition process to classify the unknown audio data. Fig. 4 shows the emotion recognition approach and automatic training of the fuzzy models. The training works similar to the emotion recognition in four steps with one major difference: Instead of the fuzzy classification as fourth step the fuzzy model generation takes place. These steps are described below.

A. Preprocessing

The audio data is collected by recording speech signals with a microphone and stored as file in the wave format. The audio file is divided into frames, in this case short signal cutouts of 32ms length. The frames are passed to the parameter extraction.

B. Parameter extraction

For each frame PROSBER extracts different acoustic parameters. Particularly important information for the emotion recognition is generated from the fundamental frequency and energy time progression of the speech signal. The speed and pause differentiations of the speech signal and its power spectrum are as well important. Therefore for each frame the values of the fundamental frequency, energy, jitter and shimmer as well as the power spectrum and the speech/pause time are determined. The frames are processed completely by the parameter extraction, so that in each case a parameter sequence that describes the dynamics of the individual parameters in the speech signal is passed to the feature calculation.

C. Feature calculation

The dynamic course of the individual parameters in the speech signal cannot be processed directly by the fuzzy classification, because fuzzy models do not work with time-dependent data. Therefore the feature calculation summarizes the parameter sequences by statistical analysis. In addition a smoothing of the computed data is necessary for fundamental frequency, in order to filter out outliers and noise. From the smoothed parameter values the average values and variances of the fundamental frequency and energy and the statistic information extracted from the dynamic course of speech/pause rate, speech speed, jitter and shimmer is determined.

D. Fuzzy model generation

Besides the features extracted from the training samples the fuzzy model generation gets the associated emotion for each sample. From this data it calculates the membership functions for every feature. Afterwards the n best features for each emotion are selected (n usually is set to values between 4 and 6). As a last step for each emotion a separate fuzzy rule system is generated.

E. Fuzzy classification

In the recognition working mode the five generated fuzzy rule systems are used to classify to which degree the actual speech sample belongs to each of the five emotions. For this purpose each fuzzy emotion model gets the relevant features selected for the respective emotion during the training, evaluates these features by means of fuzzy rules and writes out a value, which indicates the degree to which this emotion is contained in the spoken sentence. The computed degrees

for each emotion are then compared by PROSBER and the strongest emotion is returned as recognized.

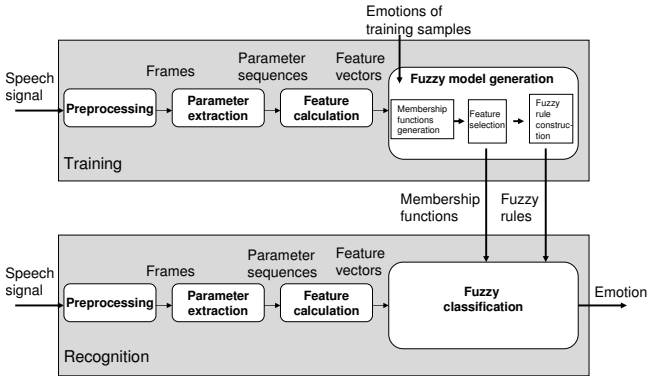

Fig. 4. Architecture of PROSBER

V. Fuzzy modelling of emotions

Fuzzy models consist of the membership functions, that represent fuzzy sets of input and output variables, and a fuzzy rule system, which describes relations between these variables.

Therefore to define the fuzzy emotion model at first the feature values must be transformed into fuzzy inputs by membership functions. Membership functions can be conveniently used to represent linguistic expressions like approximately zero, small, large, etc. For fuzzy representation of features we use simple parameterized fuzzy sets like triangular and trapezoidal functions. A typical fuzzy partition with five different linguistic expressions is shown in Fig. 5 (left side).

In the next step the different levels of output variables are defined by specifying the fuzzy sets for an emotion. In particular emotion models in psychology distinguish two, five and ten levels of emotion intensity. With more than two levels it is already possible to return different degrees of modelled emotions. We chose fivel levels corresponding to the terms in Fig. 5 (right side). Similar to the input variables we use a simple triangular function for representation of emotions as well. This relative simple representation of features and emotions was selected for MEXI in order to reduce the time needed for the fuzzy evaluations and to reach a real-time communication behavior for MEXI.

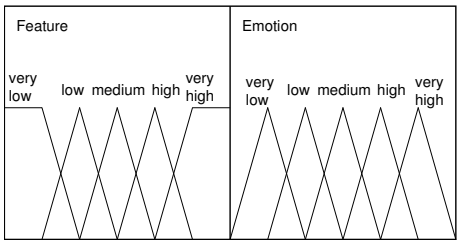

Fig. 5. Fuzzy sets for features and emotions

The last step is to define a fuzzy rule system for each emotion. The fuzzy rules represent a relation between the input and output variables of each fuzzy model and are expressed in the form of IF-THEN statements. Let $f_{i1}, ..., f_{in}$ denote the input features selected for the i_{th} emotion e_i, $i = 1, ..., 5$, then the fuzzy rules for the i_{th} fuzzy emotion model have the form:

$$R_i^l\text{: } \textit{If } f_{i1} \textit{ is } F_{i1}^l \textit{ and ..., } f_{in} \textit{ is } F_{in}^l\textit{, then } e_i \textit{ is } E_i^l,$$

where $l = 1, ..., M$ and M is the number of rules. $F_{ik}^1, ..., F_{ik}^l$ are the fuzzy sets of the k_{th} feature selected for the i_{th} emotion and $E_i^1, ..., E_i^l$ are the fuzzy sets of the i_{th} emotion (see also Section VI-C).

VI. Fuzzy model generation

The learning algorithm that is used to train the fuzzy models for emotion recognition from speech is an adapted version of the "Fuzzy Grid"-algorithm described in [17].
The algorithm consists of three consecutive steps: First, the membership functions for every feature are generated. Afterwards, the best features for every emotion are selected and the algorithm generates the fuzzy rule system for each emotion.

A. Generation of membership functions

For all training samples the values of the different features are inserted into sorted lists which are used to model the possible values of every feature by triangular membership functions for the fuzzy-terms "very low", "low", "medium", "high" and "very high".
The center of the membership function "very low" is set to the value which separates the lowest 16% of the values from the rest. The center of the membership function "low" separates the lower 33% from the rest. The centers of the remaining membership functions are calculated accordingly.

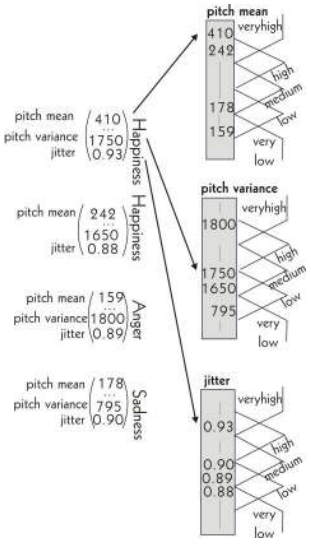

Fig. 6. Generation of membership functions

After calculating the centers of all membership functions, the starting point for each membership function is determined

as the center of its left neighbour. The ending point of a membership function is accordingly determined as the center of its right neighbour. There are two exceptions from this rule: The membership function of the term "very low" begins at 0 with the value 1 and the membership function of the term "very high" remains 1 from its center to positive infinity. Fig. 6 shows the generation of the membership functions. On the left, there are the feature vectors extracted from four training samples. The values from the feature vectors are inserted in sorted lists which are represented by the grey boxes in the middle of the image. The membership functions, which are generated using the sorted lists, are shown on the right side of each list.

B. Feature selection

The feature selection is executed for each emotion separately. The process begins with the generation of histograms for every emotion and every feature, which count the frequency by which the values of a certain feature in the training samples of one emotion fall into the categories "very low", "low", etc. Thus, a histogram of the emotion "happiness" and the feature "mean pitch" contains information about how many training samples, which belong to the emotion "happiness", have a "very low", "low", "medium", "high" or "very high" mean pitch.

The category or categories to which a training sample belongs in the histogram of a certain feature, is calculated by applying the membership functions provided by the first step of the algorithm. A training sample falls into the category where the degree of membership is highest.

A second approach does not only increment the category belonging to the term with the highest membership value by one but increments all categories where the corresponding fuzzy term has a degree of membership greater than zero by their degree of membership. This approach was implemented because just incrementing the category with highest membership value by one discards the information about the concrete degree of belongingness to the different categories. Especially when there are many samples with only little difference between the highest and the second highest membership value, this approach is expected to lead to better results than just selecting the category with the highest degree of membership.

After inserting all training samples belonging to one emotion into the histograms, the most informative features can be selected as follows: Features which contain only little information about the prevalence of an emotion, are represented by histograms which contain roughly equal numbers of training samples in every category. Good features for distinguishing an emotion are represented by histograms that contain categories with distinct peaks. This fact is shown in Fig. 7. The left histogram shows a feature which contains much information on the prevalence of emotion e. The right histogram shows a feature which does not contain much information that can be used for distinguishing emotion e from other emotions.

Thus, the quality Q_f of a feature is modelled by equation 1 where h_{fev} is the histogram category's value corresponding to the feature f, the emotion e and the fuzzy term v.

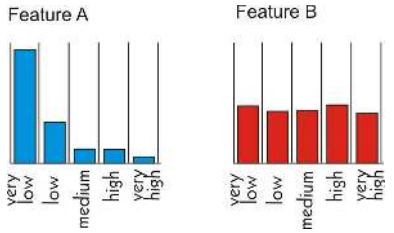

Fig. 7. Feature A: good indicator for the prevalence of emotion e, Feature B: bad indicator for the prevalence of emotion e.

$$Q_f = \frac{\max\ h_{fev}}{\sum_{v \in terms(f)} h_{fev}} \tag{1}$$

C. Fuzzy rule construction

The features, which were chosen by the feature selection are used in the third step to create the fuzzy rules which, together with the membership functions of the selected features, make up the fuzzy model of an emotion. The rule generation is performed separately for each emotion. In each fuzzy model rules for every combination of features and fuzzy terms are generated. These are examples of possible rule premises in the fuzzy model of an emotion:

IF pitchmean IS verylow AND energymax IS verylow AND ... THEN...

IF pitchmean IS verylow AND energymax IS low AND ... THEN ...

IF pitchmean IS verylow AND energymax IS medium AND ... THEN ...

The conclusions of all rules are generated using the histograms provided by the second step of the algorithm. For each rule the histogram values which correspond to the different terms that the rule consists of are cumulated. The cumulated value can be interpreted as the relevance of the corresponding rule for the prevalence of a certain emotion.

After calculating the relevance of all rules, the maximum relevance value has to be determined in order to calculate the boundaries for the conclusions "emotion IS very low", "emotion IS low" etc. First, the distance between the minimum and the maximum relevance value is divided into five equal-sized parts. Rules that have a relevance value which belongs to the lowest part are assigned the conclusion "very low", rules, that have a "relevance value" which belongs to the highest part are assigned the conclusion "very high". The other fuzzy output values are assigned accordingly. The computation of the rule conclusions is shown schematically in Fig. 8. On the left hand side of the figure, two histograms for different features are shown. The combined histogram on the right side of the picture shows the relevance values of all combinations of categories of both features' histograms.

As the original Fuzzy Grid-Algorithm uses only the "'AND"'-conjunction, a large number of $5^{numberOfFeatures}$ rules is generated. Thus, as a last step of the algorithm, rules

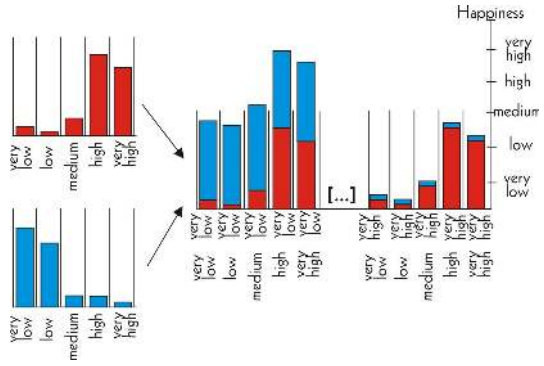

Fig. 8. Fuzzy rule generation

can be joined using the "'OR"' conjunction in order to reduce the number of rules that the fuzzy system has to work with.

VII. Implementation and results

The speech recognition system itself is implemented in C++ for performance reasons. The training algorithm is written in Java. For fuzzy classification, the FFLL-Library [18] is used, which works with rule systems written in the standardized FCL-format [19]. The system puts several restrictions on the way, fuzzy rule systems may look like. The most important restriction for this application is the use of "'AND"' as the only possible conjunction of fuzzy terms in a rule. That means, fuzzy rules cannot be summarized by "'OR"', as long as the program uses FFLL for fuzzy classification.

The database for speaker dependent recognition consists of 280 samples recorded from two female and two male speakers. The database for speaker independent recognition contains 260 samples overall which were also recorded from two female and two male speakers.

The speaker independent evaluation was performed under speaker closed but vocabulary open conditions. Due to the limited number of speakers training and evaluation speakers were the same but their utterances differed.

The best average recognition rates for speaker dependent recognition were achieved with a feature selection choosing six features and using histograms which are built up from fuzzy values. In this case, the algorithm recognizes 84% of the test samples correctly with 5% ambiguity. That is, for 5% of the test samples two or more fuzzy models reach the same maximum value.

For speaker independent recognition, best recognition rates are achieved using six features and histograms consisting of integer values. In this case, 60% of the test samples were recognized correctly without any ambiguous choices.

Table I shows the average recognition rates for speaker independent and speaker dependent recognition and the recognition rates for the different emotions.

The features chosen in speaker dependent mode vary only slightly across the different training speakers. Generalization to unseen evaluation speakers is mainly based on the system's

TABLE I

Average recognition rates for different emotions

	speaker independent	speaker dependent
happiness	59.7	84.3
sadness	23.1	69.6
anger	84.6	92.8
fear	42.3	83.9
neutral	64.0	92.9

feature selection which chooses the least speaker dependent features during the speaker independent training. Features that are stronlgy speaker dependent (i. e. their typical values vary strongly between different speakers) will not show one distinct maximum category in the histograms (see Fig. 7) during the speaker independent training and thus will be considered as less important by the feature selection.

The most important features, which were most often chosen by the feature selection, are derived from pitch and energy. Jitter, a feature that has not been chosen for emotion recognition from speech by most researchers, was also chosen as an important indicator for several emotions by the feature selection. Tables II and III show the six best features which were chosen by the feature selection for speaker dependent and speaker independent recognition.

TABLE II

Best features for speaker independent recognition

emotion	best features
happiness	medium frequency ratio, high frequency ratio, pitch variance, jitter, low frequency ratio, pitch mean
sadness	energy mean, energy minimum, energy variance, energy maximum, jitter, pitch variance
anger	pitch variance, fraction of falling pitch segments, energy variance, energy mean, energy maximum, pause ratio
fear	energy range, high frequency ratio, energy minimum, fraction of rising pitch segments, low frequency ratio, medium frequency ratio
neutral	average pitch slope, fraction of falling pitch segments, energy mean, energy maximum, pitch mean, fraction of rising pitch segments

As table IV shows, emotions that were best recognized by the emotion recognition were sadness and anger. This is true for speaker independent recognition as well as speaker dependent recognition.

Emotions that were hard to recogize, especially for speaker independent recognition, were fear and happiness. Happiness was most often confused with anger while fear was most often taken for sadness, which agrees with psychological research, examining human emotion recognition ability.

The system is able to recognize emotions in near real-time if only four or five features are chosen by the feature selection process. With an average sample length of 2.5s the system

TABLE III

BEST FEATURES FOR SPEAKER DEPENDENT RECOGNITION

emotion	best features
happiness	pitch variance, pitch minimum, average pitch slope, fraction of falling pitch segments, low frequency ratio, high frequency ratio
sadness	jitter, low frequency ratio, energy maximum, pitch minimum, energy range, pitch mean
anger	pitch minimum, energy minimum, pitch maximum, jitter, medium frequency ratio, pitch range
fear	pitch minimum, average pitch slope, fraction of falling pitch segments, energy minimum, jitter, high frequency ratio
neutral	energy mean, energy minimum, average pitch slope, energy variance, pitch mean, pitch maximum

TABLE IV

CONFUSION-MATRIX FOR SPEAKER INDEPENDENT RECOGNITION

	happiness	sadness	anger	fear	neutral
happiness	23%	6%	38%	19%	4%
sadness	0%	84%	4%	4%	8%
anger	8%	0%	84%	4%	4%
fear	8%	28%	12%	42%	12%
neutral	4%	24%	8%	4%	64%

needs an average computation time of 1.98s for four and 3.39s for five features on a Pentium IV running at 2,6 GHz. Because of the large number of rules when using six or more features, the time which is needed to initialize the fuzzy system in FFLL, increases noticeably when using s ix features. In this case, the algorithm takes an average of 10 seconds to compute the emotional value of the samples. The time which is needed for computation can be reduced to less than 1.5 seconds for four, five and six features by initializing the fuzzy rule system in advance.

VIII. SUMMARY AND OUTLOOK

In this paper a fuzzy rule based approach for emotion recognition from spoken natural language for the robot head MEXI was introduced. With our fuzzy grid learning algorithm we reached average recognition rates of 84% in speaker dependent mode and 60% in speaker independent mode. In the speaker independent mode the recognition performance thus is similar to humans as reported by psychologists [8]. In the speaker dependent mode our results are comparable to existing approaches. In the speaker independent mode the system recognized emotions unambiguously and in the speaker dependent mode only 5% of the results were ambiguous (more than one emotion was identified by the classification). This difference is due to the use of fuzzy values in the emotion specific histograms for each feature which according to our experience causes more ambiguities. Nevertheless, we decided to use it in the speaker dependent mode, since the overall classification showed better results in this case. For eliminating ambiguities it would be interesting to investigate the influence of recent emotions recognized in the past and whether certain emotions are likely to appear in direct sequence or not. Also the consideration of keywords might give additional hints on the prevalence of a certain emotion. Furthermore, if facial expressions recognized by MEXI are evaluated in conjunction with natural speech, this would certainly decrease the ambiguity. Hence, our realization proofed that fuzzy logic, which showed its applicability already for facial expression classification, can also successfully be used for the emotional classification of natural speech. Our next steps in the development of MEXI and its emotion recognition capabilities will be the integration of emotion recognition from speech and from facial expressions.

REFERENCES

[1] S. Hashimoto, "KANSEI as the third target of Information Processing and Related Topics", Proc. Of Intl. Workshop on Kansei Technology of Emotion, pp 101-104, 1997.

[2] N. Esau, B. Kleinjohann, L. Kleinjohann, D. Stichling "MEXI: Machine with Emotionally eXtended Intelligence - A Software Architecture for Behavior Based Handling of Emotions and Drives ", In Proceedings of Int. Conf. on Hybrid Intelligent Systems (HIS03), Melbourne, Australia, 2003.

[3] R. W. Picard, "Affective Computing", MIT Press, 1997.

[4] C. Breazeal, "Affective Interaction between Humans and Robots", in J. Kelemen and P. Sosik (eds.), Proc. of ECAL 01, Prague, pp. 582-591, Springer 2001.

[5] T. Fong, I. Nourbakhsh, K. Dautenhahn, "A Survey of Socially Interactive Robots", Robotics and Autonomous Systems, 2003.

[6] S. Kopp and I.Wachsmuth, "A knowledge-based approach for lifelike gesture animation", in W. Horn, editor, ECAI 2000 Proceedings of the 14th European Conference on Artificial Intelligence, pages 661-667, Amsterdam, 2000, IOS Press.

[7] S. McGilloway, R. Cowie, E. Douglas-Cowie, S. Gielen, M. Westerdijk, S. Stroeve, "Approaching Automatic Recognition of Emotion from Voice: A Rough Benchmark", in ISCA Workshop on Speech and Emotion, Belfast 2000.

[8] K. Scherer, "Vocal communication of emotion: A review of research paradigms", in Speech Communication, 40(2003), 227-256, Elsevier 2003.

[9] V.A. Petrushin, "Emotion in Speech: Recognition and Application to Call Centers", Proceedings of the 1999 Conference on Artificial Neural Networks in Engineering, 1999.

[10] A. Boozer, "Characterization of Emotional Speech in Human-Computer-Dialogues", M.Sc Thesis, MIT, 2003.

[11] A. Nogueiras, A. Moreno, A. Bonafonte, J. B. Marino, "Speech Emotion Recognition Using Hidden Markov Models", In EUROSPEECH-2001, 2679-2682, 2001.

[12] T. L. Nwe, S. Foo, S. Wei; L. De Silva, "Speech emotion recognition using hidden Markov models", Speech communication 41,4, 2003.

[13] F. Dellaert, T. Polzin, A. Waibel, "Recognizing Emotion in Speech", Proceedings of the ICSLP-96, 1996.

[14] P.-Y. Oudeyer, "'The Production and Recognition of Emotions in Speech: Features and Algorithms"', International Journal of Human Computer Interaction, 59(1-2):157-183 2003. Special issue on Affective Computing.

[15] N. J. Nilsson, "Artificial Intelligence - A New Synthesis", Morgan Kaufmann Publishers, 1998.

[16] F. Burkhardt, "Simulation of emotional speech with speech synthesis methods", (in German), Shaker, 2001.

[17] H. Ishibuchi, T. Nakashima, "A Study on Generating Fuzzy Classification Rules Using Histograms", Knowledge based Intelligent electronic Systems, Bd. 1, 1998.

[18] Free Fuzzy Logic Library, http://ffll.sourceforge.net/

[19] International Electrotechnical Comission (IEC), IEC 1131 - PROGRAMMABLE CONTROLLERS Part 7 - Fuzzy Control Programming, 1997.