Fuzzy Emotion Recognition in Natural Speech Dialogue

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

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

***Abstract*— This paper describes the realization of a natural speech dialogue for the robot head MEXI with focus on its emotion recognition. Specific for MEXI is that it can recognize emotions from natural speech and also produce natural speech output with emotional prosody. For recognizing emotions from the prosody of natural speech we use a fuzzy rule based approach. 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 in the 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 data base 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 reached average recognition rates of 84% in speaker-dependent mode and 60% in speaker-independent mode.**

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

I. Introduction

Emotions are an evident part of interactions between human beings. But also for interactions of humans with embodied or virtual agents emotions play a major role, since humans can never entirely switch off their emotions and usually also react emotionally in their communication with a robot or computer system. Therefore during the last years considerable effort was spent on developing emotional agents, that show artifical emotions in their behavior. A good overview of such socially interactive robots can for instance be found in [1]. When communicating with humans an agent's emotional behavior has to consider the current human emotions in order to react in a sensible way. Therefore it has to be able to recognize human emotions from facial expressions and/or natural speech. Accordingly we developed the robot head MEXI (Machine with Emotionally eXtended Intelligence) described shortly in Section III. MEXI recognizes emotions from facial expressions and from natural speech and shows artificial emotions by its facial expressions and speech utterances as well. The latter is described elsewhere [2], whereas the focus of this paper is on MEXI's emotion recognition from prosody of natural speech. MEXI's prosody based emotion recognition system, called PROSBER is described in Section IV. PROSBER allows switching between speaker-independent and speaker-dependent speech analysis in order to support emotion recognition also for people who rarely interact with MEXI like for instance at exhibitions. It uses a fuzzy based approach to distinguish the emotions anger, fear, sadness, happiness and a neutral emotional state. For each emotion a fuzzy rule system is automatically generated from a training data base of speech samples. Furthermore using this data base, up to six relevant speech features for each emotion are automatically determined from a set of twenty analysed features. This is important since our results show that the set of significant features differs considerably between the recognized emotions. By the restriction to six or less features on the one hand, overfitting is avoided and on the other hand analysis effort is decreased to support real-time emotion recognition. The automatic fuzzy model generation was a considerable advantage for the implementation of our systems and will also help to improve the system if for instance new training data is available. The fuzzy model generation is described in Section V. Afterwards in Section VI the results are summarized and a summary and outlook are given in Section VII.

II. Related Work

Emotions and their role in human-computer-interaction (HCI), also called affective computing [3] or KANSEI information processing [4], have become an important research area in the last years. 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 [5] 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 [6]. Low recognition rates of about 70% were reached by SpeakSoftly [7], a neural net based approach distinguishing five emotions, and Mercury [8], 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 [9] and the approaches of De Silva [10] and Dellaert [11]. 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 [12]. 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 (see Figure 1). 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 support audio output. These facilities allow MEXI to represent a variety of emotions like happiness, sadness or anger by its facial expressions, head movements and by its speech output.

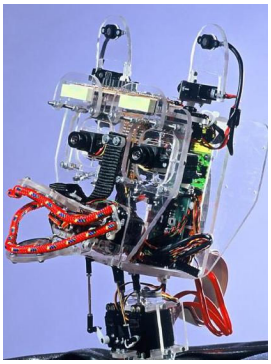

Fig. 1. The robot head MEXI

MEXI's software architecture (see Figure 2) is designed according to Nilsson's Triple-Tower Architecture, that distinguishes between perception, model and action tower [13]. The Perception component processes MEXI's visual inputs and natural language inputs. The Action Control component controls 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 allows MEXI to directly react to its visual and natural speech inputs received from its environment by corresponding head movements, facial expressions and natural speech output. Unlike many goal directed agents, MEXI has no internal model of its environment to plan and control its behavior but uses its internal state, representing the strength of 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. In a feedback loop MEXI's internal state is used by the Emotion Engine 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]).

In this paper we concentrate on the speech processing of MEXI, which allows MEXI to recognize emotions in natural speech inputs and generates outputs with a prosody corresponding to MEXI's emotional state. Figure 2 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 (PROSody Based Emotion Recognition). For Speech Recognition we use the commercially available software ViaVoice [14]. 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.

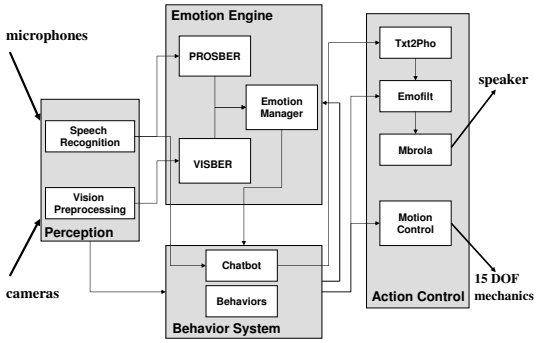

Fig. 2. Architecture of MEXI

For generating answers MEXI uses the slighlty extended commercially available Chatbot ALICE [15], 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 [17], 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. FUZZY EMOTION RECOGNITION IN PROSBER

PROSBER is a fuzzy rule based system for emotion recognition from natural speech. It takes single sentences as input and classifies them into five 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 Figure 3. During 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 unknown audio data. 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.

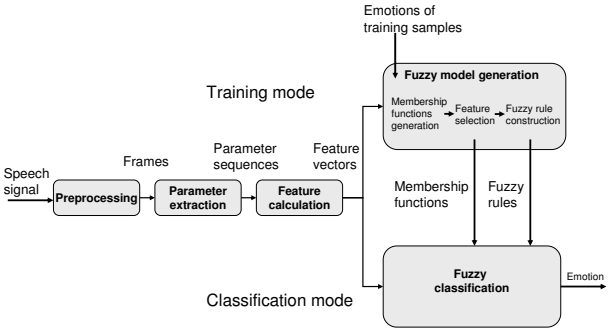

Fig. 3. Architecture of PROSBER

Preprocessing: The audio data is collected by recording speech signals through the 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.

Parameter extraction: For each frame PROSBER extracts different acoustic parameters. From these afterwards the features for the determination of the speakers emotion will be computed. 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.

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 the dynamic course of 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 are calculated. Furthermore, the smoothed parameters are used to determine the statistical information from the dynamic process of the speech/pause rate, speech speed, jitter and shimmer.

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.

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. First these features are fuzzified and then evaluated by means of fuzzy rules. Afterwards, the output for each emotion is defuzzified using the center of gravity (COG) method. The computed degrees for each emotion are then compared by PROSBER and the strongest emotion is returned as recognized. Figure 4 shows the principle structure of the fuzzy classification.

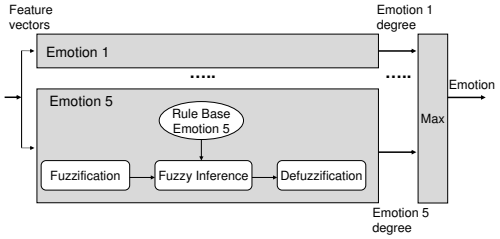

Fig. 4. Principle structure of fuzzy classification

V. AUTOMATIC GENERATION OF FUZZY MODELS

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. In our case inputs are the different features and outputs are the emotions. In order to reduce the time needed for the fuzzy evaluations and to reach a real-time communication behavior for MEXI, we chose simple triangular and trapezoidal membership functions and distinguish five levels $verylow$, low, $medium$, $high$ and $veryhigh$ for both features and emotions. Figure 5 shows the principle shape of the membership functions. Their exact shape depends on the training samples and is determined automatically as described in Subsection A for features and Subsection C for emotions.

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 [18].

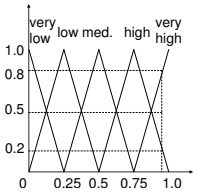

Fig. 5. Principle shape of membership functions for emotions

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 then 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 $verylow$, low, etc. Figure 6 shows four training samples (two for happiness, one for anger and one for sadness) and the sorted feature lists for pitch mean, pitch variance and jitter.

The center of the membership function $verylow$ 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.

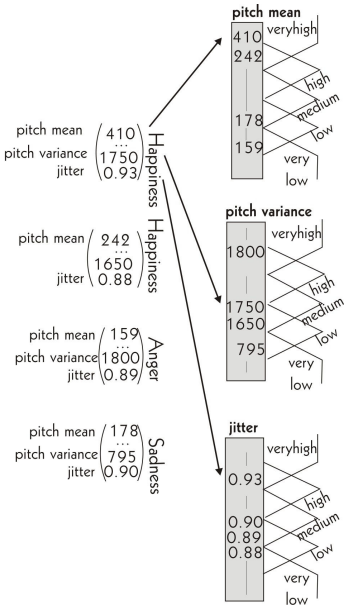

Fig. 6. Generation of membership functions

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 determined as the center of its right neighbour. There are two exceptions from this rule: The membership function of the term $verylow$ begins at 0 with the value 1 and the membership function of the term $veryhigh$ remains 1 from its center to positive infinity. The membership functions are shown at the right side of each list in Figure 6.

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 $verylow$, 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 $verylow$, low, $medium$, $high$ or $veryhigh$ 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 of which 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 Figure 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.

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

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

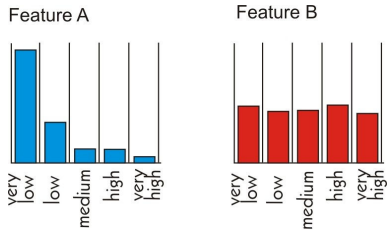

Fig. 7. Feature A: good indicator for the prevalence of emotion e, Feature B: bad indicator for the prevalence of emotion e.

features, make up the fuzzy model of an emotion. The rule generation is also performed separately for each emotion $e_i, i = 1, ..., 5$. Each rule takes the fuzzified features $f_{j_i}, j_i = 1, ..., n_i, 4 \leq n_i \leq 6$, as input and produces a fuzzy emotion value e_i as output. The rules for emotion e_i have the form:

IF f_{1_i} *IS verylow AND ...* f_{n_i} *IS verylow THEN* e_i *is veryhigh*
IF f_{1_i} *IS verylow AND ...* f_{n_i} *IS low THEN* e_i *is veryhigh*
IF f_{1_i} *IS verylow AND ...* f_{n_i} *IS medium THEN* e_i *is medium*
...

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 $verylow$, rules that have a relevance value which belongs to the highest part are assigned the conclusion $veryhigh$. The other fuzzy output values are assigned accordingly. The computation of the rule conclusions is shown schematically in Figure 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^{n_i+1} rules for emotion e_i is generated. Thus, as a last step of the algorithm, rules can be joined using the OR conjunction in order to reduce the number of rules that the fuzzy system has to work with.

VI. 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

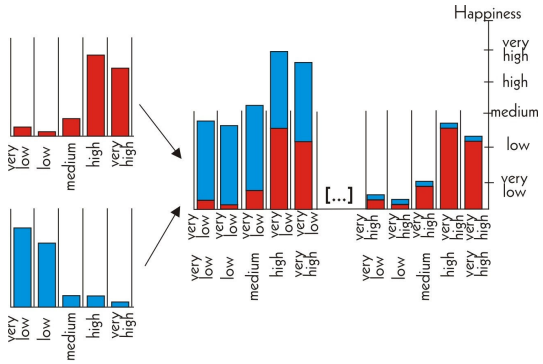

Fig. 8. Fuzzy rule generation

[19] is used, which works with rule systems written in the standardized FCL-format [20]. 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 training 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 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 the different emotions in speaker-dependent and speaker-independent mode.

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

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. Table II shows as an example the six best features which were chosen by the feature selection for speaker-dependent recognition.

TABLE II
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 III shows, that the emotions that were best recognized were sadness and anger. This is true for speaker-independent recognition as well as speaker-dependent recognition.

Emotions that were hard to recognize, 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 [21].

TABLE III
CONFUSION-MATRIX FOR SPEAKER-INDEPENDENT RECOGNITION

	happin.	sadn.	anger	fear	neutral
happin.	23%	6%	38%	19%	4%
sadn.	0%	84%	4%	4%	8%
anger	8%	0%	84%	4%	4%
fear	8%	28%	12%	42%	12%
neutral	4%	24%	8%	4%	64%

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 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 [19], increases noticeably when using six 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.

VII. SUMMARY AND OUTLOOK

We presented the fuzzy rule based system PROSBER that recognizes emotions from the prosody of natural language. From a set of about twenty analyzed speech features it automatically selects up to six most important features to generate a rule system for each emotion to recognize. Thus, recognition complexity is reduced in order to support a real-time dialogue between MEXI and its human counter part. PROSBER reaches recognition rates of 84% in speaker-dependent mode and 60% in speaker-independent mode. In speaker-dependent mode its performance is comparable to existing systems. In speaker-independent mode it is similar to humans as reported by psychologists [6]. In order to avoid ambiguities of about 5%, which arise in speaker-dependent mode, we plan to combine the evaluation of facial expressions and natural speech in MEXI. This would certainly also increase the recognition rate. Furthermore, it would be interesting to investigate the sequence of emotions over time and whether certain emotions are likely to appear in direct sequence or not.

REFERENCES

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

[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] 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.

[5] 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.

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

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

[8] A. Boozer, ”‘Characterization of Emotional Speech in Human-Computer-Dialogues”’, M.Sc Thesis, MIT, 2003.

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

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

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

[12] 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.

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

[14] IBM ViaVoice, http://www-306.ibm.com/software/voice/viavoice/

[15] Alicebot, http://www.alicebot.org/

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

[17] MBROLA, http://tcts.fpms.ac.be/synthesis/mbrola.html

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

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

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

[21] V. A. Petrushin, “Creating Emotion Recognition Agents for Speech Signal”, in Socially Intelligent Agents, K. Dautenhahn, A. H. Bond, L. Canamaro, B. Edmonds (eds.), Kluwer Academic Publishers, 2002, pp. 77-84.