Abstract: - Verbal communication is the most obvious instrument used to express our thoughts and ideas, considering only this part of speech without regarding its nonverbal part, may lead to overlooking important information of utterance or even misunderstanding it. The development of an automatic system for recognition of facial expressions is a rather difficult task. Such a system must perform automatically, accurately and in real-time by three main steps: face detection and localization, feature extraction and classification. The contributed paper deals with use of automatic system for recognition of facial expressions which have been being created for the Czech dialog system. A prototype of the recognition and understanding system was developed in the Department of Computer Science [8]. The proposed system is fully automatic, user-independent and real-time working. First experiments show that the speech recognition quality is increased by using automatic facial expression recognition system. The work presented in this paper was partly supported by the Grant Agency of Czech Republic under contract number 201/02/1553.

Key-Words: - Speech understanding, nonverbal communication, face detection and localization, feature extraction and classification.

1 Introduction
The information expressed by speech can be regarded to fall into three categories: linguistic, paralinguistic, and non-linguistic, though their boundaries may not always be clear. We define the linguistic information as symbolic information that is represented by a set of discrete symbols and rules for their combination. It can be represented either explicitly by written language, or easily and uniquely inferred from context. Paralinguistic information is defined as the information that is not inferable from the written counterpart but is added deliberately by the speaker to modify or supplement the linguistic information [9]. A written sentence can be uttered in various ways to different intentions, attitudes, and speaking styles which are under the conscious control of the speaker. Non-linguistic information concerns such factors as the age, gender, physical and emotional states of the speaker, etc. These factors are not directly related to the linguistic and paralinguistic contents of the utterances and, in general, cannot be controlled by the speaker, though it is possible for a speaker to control the way of speaking to intentionally convey an emotion, or to simulate an emotion, as it is done by actors. Understanding human emotions and their nonverbal messages is one of the most necessary and important skills for making the next generation of human-computer interfaces (HCI) easier, more natural and effective. Indeed, the first step toward an automatic emotion sensitive human-computer system having the ability to automatically detect users’ nonverbal signals is the development of an accurate and real-time automatic NVC analyzer. Such an analyzer must deal mainly with users’ facial expressions and paralanguage. The part of this problem leads to the consideration of the nonverbal communication in developed dialog system [10].

2 Nonverbal communication
Nonverbal communication has many functions in the communication process. By virtue of nonverbal communication, we simply express our emotions. In many cases we are able to exhibit our feelings by facial expression and gestures much more quickly than by using words. It regulates relationships and may support or replace verbal communication. On the other hand nonverbal communication has its disadvantages and seamy sides too. Difficulties may arise if communicators are unaware of the types of messages they are sending, and how the receiver is interpreting those messages. No dictionary can accurately classify nonverbal signals. Their meanings vary not only by culture and situation, but also by the degree of intention of their use. Many of them are ambiguous and could cause misunderstandings. Effective communication is the combined harmony of verbal and nonverbal actions. We can divide the functions of communication into four parts, see Fig.1. The first is called expressive function and it demonstrates outward our inner emotional state. The second is
signal function intended for sending simple or complex information. The third is the descriptive function serving for sending more complex and complicated information.

Main categories of nonverbal communication: facial expression, posture, gesture, proximity, gaze, paralanguage, touch, adornment.

2.1 Facial expression

The main objective of this thesis is analysis of systems for automatic recognition of facial expression. Facial expression carries most of our nonverbal meanings and often is considered as the most important category of nonverbal communication (by many experts 55-85 percent of NVC is exchanged by them). Although the human face is capable of creating 250,000 expressions, less than 100 sets of them constitute meaningful symbols. Three main categories of conversational signals have been identified: syntactic display - used to stress words, or clauses (raising or lowering eyebrows can be used to emphasize a word or clause), speaker displays: illustrate the ideas conveyed (“I don’t know” can be expressed by the corners of the mouth being pulled up or down), and listener comment display - used in response to an utterance (incredulity can be expressed with a longer duration of eyebrow raising). Facial expressions are generated by contractions of facial muscles (Fig. 2), which result in temporally features such as eye lids, eye brows, lips, skin texture, etc., and are often revealed by wrinkles and bulges. In general, facial actions last 250 ms to 5 seconds. In order to accurately measure and describe facial expressions, we need to know the location of facial actions, their intensity and their dynamics. It should be noted that measuring facial expressions and their intensity is a very difficult task, due to many problems such as their variation and appearance from one individual to another based on their age, ethnicity, sex, facial hair, cosmetic products, etc.

2.1 Paralanguage

Paralanguage directly accompanies verbal language. These “non-lexical” vocal communications (prosody) are considered to be a type of nonverbal communication. Aspects of the way in which a person speaks can give us much more information about him. It is perhaps the most difficult of the categories of nonverbal communication to comprehend due to its complexity, and contingency on the individual doing the communicating. It includes a number of sub-categories: rate, pitch, inflection, volume, quality, tempo, rhythm, resonance and enunciation. It is important to point out that something, which is heard and vocalized may vary and hold different meanings in different cultures. In the learning or attempt to send a message in a different language and culture, not knowing or using the correct paralanguage can cause conflict and misinterpretation.

3 Automatic Nonverbal Communication Analyzer

As noted previously, nonverbal communication plays a crucial role in man-to-man interaction. Also, it is clear that correct and effective recognition of spontaneous speech without the use of nonverbal communication (facial expression, paralanguage, etc.) is impossible. Likewise, being aware of how the user is receiving a piece of information provided would be very valuable. Being able to know when the user needs more feedback, by observing cues about his emotional state has many advantages. The human sensory system uses multimodal analysis of multiple communication channels (HCI). This human ability could be our goal for developing an emotion-sensitive and intelligent multimodal HCI. For developing such a system, we need an automatic multimodal nonverbal communication analyzer (ANCA). The performance of a multimodal system (e.g., our nonverbal communication analyzer) does not depend only on the number and types of integrated modalities (sight, sound, touch, etc.). Another important issue is how the data carried by multiple channels should be fused to achieve high performance in recognizing NVC. In general, fusion of multimodal information has always...
been a topic of discussion, because the technique of how these modalities are fused plays an essential role.

Fig. 3 Automatic bimodal audiovisual ANCA for speech recognition

3.1 Automatic Bimodal Audiovisual ANCA for Speech Recognition
An automatic nonverbal communication analyzer needs to mainly deal with users’ facial expressions and users’ paralanguage. The logical flow of such a system can look like Fig. 3.

3.1.1 Audio Input
The audio input carries various kinds of information, and considering only the verbal part without regarding the manner in which it was spoken and analyzing facial expressions, might lead to overlooking important information of utterance or even misunderstand it. According to our experiments, the accuracy of automated speech recognition, which is about 80 to 90 percent for neutrally spoken words, tends to drop to 50 to 60 percent if it concerns emotional speech. The auditory features estimated from the audio signal are: pitch (the fundamental frequency of the acoustic signal delimited by the rate at which vocal cords vibrate); intensity (the vocal energy); speech rate (the number of words spoken in a time interval); pitch contour (pitch variations described in terms of geometric patterns); phonetic features (features that deal with the types of sounds involved in speech, such as vowels and consonants and their pronunciation).

3.1.1 Visual Input
The significant role of facial expressions convinces us to use visual input to process and analyze them. The facial expression recognition problem can be divided into the following three partial problems: face detection; facial feature extraction; facial expression classification.

In spite of significant advances of computer vision in recent years, developing robust and accurate facial expression recognition in an automatic way and in real-time is still very problematic and at present belongs to one of the greatest dreams and most active areas in the computer vision. All existing systems have their advantages and disadvantages. Many of them are not real-time or if are real-time, they are not fully automatic. In addition, the systems, which are real-time and fully automatic are very limited or are not accurate. Another large problem is that there is a strong need to generate an equivalent to the FERET (Face Recognition Technology) database of facial images for face recognition specifically for facial expression recognition.

Approaches for extraction and representation of facial features can be categorized according to several factors. In general, we can distinguish them as local or holistic. Local approaches deal with features, which are prone to change with facial expressions. Holistic approaches deal with the face as a whole.

4 System ARFE (Automatic Recognition of Facial Expression)
The most important feature and advantage of ARFE in comparison to many other existing systems is that ARFE is a fully automatic system, which is a necessary condition for using it in the field of HCI. Furthermore, ARFE is based on state of the art approaches, is a multi-user system (see Fig. 4) and has two working modes: static images and dynamic images (photo and video) mode. This is possible due to the fact, that each frame is processed and classified separately. Another distinction is that the ARFE does not deal with explicit detection and alignment of internal facial features, which spares processing time important for real-time applications.

ARFE, see Fig. 5, is a fully automatic user independent real-time facial expression recognition system.

Fig. 4 ARFE - static images mode
The system automatically detects frontal faces in complex backgrounds and makes classification for each found face. The only requirements of the system are frontal faces, a good illumination condition and acceptable light direction. In other words, faces should not contain shadows and must be well lighted. After a face was detected in an arbitrary image, which could be a digitized video signal or a digitized image, the face finder returns the coordinates of a square box around the face. Furthermore, the found face is normalized as follows. First, the content of each face box is converted from RGB to the grayscale, furthermore, rescaled to 76 × 76, non-symmetrically cropped to 40 × 56 and rescaled to 48 × 48. Finally the gray-level image is rescaled to the range 0 – 255.

In general, after cropping, the distance between pupils is 30 pixels on average. Distance between pupils for effective facial expression recognition should be from 20 to 100 pixels. It is clear, that working with such a feature vector can bring many unpleasant conditions like bigger computational cost. The feature selection step results in a choice of the subset of those features, which best describe our classification classes. The classification is performed also by Adaboost, which combines several weak classifiers (in our case a naive Bayesian classifier) to get a strong and accurate one. For searching the face of an unknown size in the whole input image, we can move the search window across the image several times at different scales, and check each location using our resultant classifier. The basic classifiers are designed to be easily resizable in order to be able to find the objects of interest at different sizes. Viola’s detection system uses three kinds of features [5]. A two-rectangle feature is the difference between the sum of the pixels within two rectangular regions. A three-rectangle feature computes the sum within two outside rectangles subtracted from the sum in a center rectangle. Finally a four-rectangle feature computes the difference between diagonal pairs of rectangles. The kernels of Gabor filters are similar to the 2D receptive field profiles of the mammalian cortical simple cells, see Fig. 6, exhibit strong characteristics of spatial locality and orientation selectivity, and are optimally localized in space and frequency domains [2]. The cells of a pair of adjacent cells defined by similar frequencies have certain symmetries. One has even and the other odd symmetry. This allows model both receptive fields of such a pair of cells by a complex-valued function such as Gabor filters, see Fig. 7.

For implementation of Gabor wavelength ARFE uses a slightly modified version of Gabor filters [7]. For each filter, the complex Gabor transform of the image is made up of sine and cosine parts corresponding to even and odd-phased filtering functions. For representation of the facial image, ARFE uses bank of 24 Gabor filters of 8 orientations, in increments of π/8 and 5 spatial frequencies in half octave increments with σ = π and kmax = π/2. By using convolution theorem, ARFE reduces the convolution processing time to only several milliseconds, which is very important for our real time application. The Gabor representation of the facial image results in a feature vector of 92,160 elements. There exist
several feature selection techniques [3]. ARFE uses Adaboost, which is not only a strong classifier, but also it can be used as a strong feature selection technique. It attempts to select most relevant features, by finding features that generate least error when classifying the training set. ARFE using Adaboost selected 1120 most relevant features, which resulted in a processing speed increase. It is interesting, that most of features selected by Adaboost are from the first and last scale of Gabor filters.

5 Experimental Implementation
For the purpose of experimental implementation of ARFE, a small training set has been created. The dataset consists of 24 adult volunteers and 1 infant. None of the subjects wore eyeglasses. Some of the subjects had hair covering their foreheads, no subject wore caps, or had makeup on their brows, eyelids or lips. The subjects included both male (56%) and female (44%). The important condition was maximum illumination with a minimum of facial shadows. The primary idea was to ask each volunteer to look at some examples of all 6+1 facial expressions (happy, fear, anger, disgust, sadness, surprise and neutral) and try to copy them. Also the primary idea was that each expression has to be repeated several times and the best one is chosen for the training set. Unfortunately, in reality this was different. During this data gathering, we have come to know that people are able to exhibit their neutral expression with hardly or no effort. Also, we have come to know that they are able to simply smile for many minutes without any pauses. But other expressions such as a fear expression, disgust expression or angry expression are very difficult and maybe impossible for people without theatrical experiences. Therefore sometimes a simple facial expression recording lasts more than a half hour for one person in place of only two minutes planned. All these problems resulted in the fact, that in the present day, the gathered training set provides only three acceptable facial expressions: neutral, happy and surprise. Unfortunately, also not all of the surprise expressions are really perfect surprise expressions. The final training set contains 75 images, 25 images per expression from 25 volunteers. If you would like to try ARFE and get acceptable results, your experiments must me done under the following conditions (see Fig. 8), only the frontal view of the facial image must be analyzed. As was mentioned previously, the present day ARFE distinguishes only the following expressions (classes): neutral, happy and surprise expressions. If we will consider that each of these expressions (barring neutral) has a beginning, apex and ending duration, then ARFE, due to the training set properties, is able to classify correctly only the apex part of the expression’s duration.

6 Conclusion
The development of an effective and efficient spontaneous speech recognition system without analyzing users’ nonverbal signals is impossible. It is well known that speech and nonverbal communication complement each other, and when used together, can create a powerful system.
We have presented an automatic user-independent real-time facial expression recognition system, called ARFE. The most important feature and advantage of ARFE in comparison with many other existing systems is that ARFE is fully automatic. It is a necessary condition for using it in the dialog system. Furthermore, ARFE is based on state of the art approaches, is a multi-user system (see Fig. 5) and has two working modes: static images and dynamic images (photo and video) mode.
We have described the structure of ARFE and have seen
that the most important stages of the system are: the face localization, the Gabor wavelet representation of the facial image and the classification stage performed by Adaboost. We have seen that Gabor filters are very powerful for deriving distinguishable features for describing the facial image. Furthermore, we were able to see how easy the implementation of the state of the art classification method Adaboost is. Furthermore, the power of Adaboost is comparable with first rate Support Vector Machines.

Obtained results are interesting and at least show that designing of a fully automatic facial expression system in a constrained environment in the present day is possible. One of the factors, which put down the accuracy of the current version of ARFE is the used face detection and localization system provided by OpenCV. This system provides an excellent face detection system. But unfortunately, the face localization performed by this system is not accurate enough for an automatic facial expression recognition system and without any doubt is in need of an improvement.

Also we can expect that a problem could come into existence if we would deal with expressions like fear, where the changes in the user’s face are very tenuuous. Probably the Gabor representation of the facial image with the used parameters will not be able to derive enough distinguishable features for such expressions. This problem could be solved by a combination of the present day ARFE and a local approach dealing directly with facial features.

References: