Neural Comput & Applic (2000)9:290–296

Neural
Computing
& Applications

Emotion Recognition in Speech Using Neural Networks

J. Nicholson, K. Takahashi and R. Nakatsu

ATR Media Integration & Communications Research Laboratories Kyoto, Japan

Emotion recognition in speech is a topic on which little research has been done to-date. In this paper, we discuss why emotion recognition in speech is a significant and applicable research topic, and present a system for emotion recognition using one-class-in-one neural networks. By using a large database of phoneme balanced words, our system is speaker- and context-independent. We achieve a recognition rate of approximately 50% when testing eight emotions.

Keywords: Context independence; Emotion recognition; Neural networks; Speaker independence; Speech

1. Introduction

Non-verbal information plays an essential part in human communication. In addition to the meaning conveyed in spoken language, the manner in which the words are spoken conveys a great deal of information. Spoken text can have several different meanings, depending on how it is said. For example, with the word 'really' in English, a speaker can ask a question, express either admiration or disbelief, or make a definitive statement. An understanding of text alone cannot successfully interpret the meaning of a spoken utterance.

Emotion recognition in speech has many potential applications. One current use is in interactive movies [1]. The state-of-the-art in speech recognition and natural language understanding is not yet at a level where understanding of spontaneously uttered speech can be easily integrated into many systems. However, emotion understanding can allow characters in an interactive movie to react to the emotions conveyed by a participant's utterances.

Another possible use of emotion recognition is as an aid to speech understanding. Speech understanding has traditionally treated emotion as 'noise' that detracts from understanding the text of an utterance. It is possible that by recognising the emotions in speech, one would be able to 'subtract' them from the speech and improve the performance of speech understanding systems.

Finally, emotion recognition systems for speech could serve as a kind of 'emotional translator'. Emotions are often portrayed differently in different cultures and languages. For example, one type of intonation which indicates admiration in Japanese can indicate disbelief in English. A method of translating emotions, in addition to words, between languages can help improve international communication.

Although emotions are an important aspect of human communication, little research has been done in this field thus far. With this and its potential uses in mind, we have developed a system using neural networks for the recognition of emotions in speech. This paper discusses our motivations for investigating emotion recognition in speech, the design and implementation of our system, and some of the results that we have obtained thus far.

2. System Design

2.1. Emotions

2.1.1. Conscious versus unconscious emotion. Emotions conveyed in speech can be grouped into two main categories: consciously expressed emotions;

Correspondence and offprint requests to: Ryohei Nakatsu, ATR Media Integration & Communications Research Laboratories, 2–2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619–0288 Japan Email: nakatsu@mic.atr.co.jp

and unconsciously expressed emotions. Consciously expressed emotions are generally more obvious than unconsciously expressed emotions. For example, when someone raises their voice in speaking, they are often consciously expressing that they are angry. However, in other cases, the only indication of a person trying to conceal their anger might be a slight terseness to their words.

Consciously expressed emotions are easier for humans to recognise [2], and significantly easier to gather data on. Therefore, this study is limited itself to the recognition of emotions that are consciously and purposefully expressed by the speaker. We expect to be able to expand our methodology to unconscious or concealed emotions as well in future work.

2.1.2. Classification of emotions. How to classify emotions is an interesting but difficult issue. Researchers on emotion recognition differ on the number of categories and the kinds of categories to use. Some classification systems that have been used include:

neutrality, joy, boredom, sadness, anger, fear, indignation [3]
anger, fear, sadness, joy, disgust [4]
neutrality, happiness, sadness, anger, fear, boredom, disgust [5]
fear, anger, sadness, happiness [6]

We dealt with four emotional states (anger, sadness, happiness and cheerfulness) in our previous study; based on examining these examples, and on the consideration that increasing the number of recognisable emotional states is effective for achieving interaction between humans and computers, we have selected the following eight emotional states to use in this study:

joy, teasing, fear, sadness, disgust, anger, surprise, neutral

2.2. Speaker and Context Independence

Speaker independence is an important part of speech and emotion recognition. A speaker-dependent system requires a training period for each new speaker before the system is able to function at a reasonable level. On the other hand, a speaker-independent system will tend to have a lower level of accuracy, since it is not finely tuned for each speaker. However, eliminating the need for training sessions with each new speaker seems well worth the resultant loss in accuracy. By carrying out initial training using a number of different speakers, our system has become speaker-independent.

Context independence is also a desirable quality in emotion recognition systems. A system that can determine the emotion in an utterance regardless of the context or text of the utterance is considered context-independent. A context-dependent system would require language understanding in order to determine the context. Our system achieves context independence by using a large set of phoneme balanced words as its training set.

Of course, there are other important factors, such as the effect of social and cultural differences. However, these are difficult issues that will require long-term research. Therefore, in this research, we deal with emotions contained in the utterances spoken only by Japanese speakers. Under this restriction, we tried to achieve speaker-independent and context-independent emotion recognition.

2.3. Processing Flow

The processing flow of our system is illustrated in Fig. 1. The process is divided into two main parts: speech processing and emotion recognition.

A speech input (an utterance) is input into the speech processing part. First, the speech features for that utterance are calculated. Next, the utterance is divided into a number of speech periods. Finally, for each speech period the speech features are extracted, and features for the utterance are compiled into a feature vector.

The feature vector is then input into the emotion

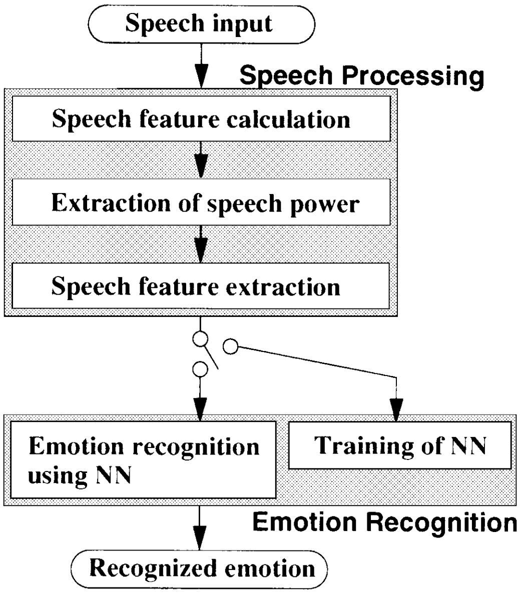
Speech input
Speech Processing
Speech feature calculation
Extraction of speech power
Speech feature extraction
Emotion recognition using NN
Training of NN
Emotion Recognition
Recognized emotion

Fig. 1. Processing flow.

recognition part. In the training stage, the feature vector is used to train the neural network using backpropagation. In the recognition stage, the feature vector is applied to the already trained network, and the result is a recognised emotion.

These steps are explained further in the following sections.

2.4. Speech Features

2.4.1. The choice of speech features. There are two main types of speech features: phonetic features and prosodic features. Phonetic features deal with the types of sounds involved in speech, such as vowels and consonants, and their pronunciation. Prosodic features deal with the more musical aspects of speech, such as rising or falling tones and accents or stresses.

Speech understanding traditionally uses only phonetic features. One viewpoint on emotion recognition states that, in the light of this, emotion recognition should focus on only prosodic features. Another view is that prosodic features and phonetic features are intertwined in expressing emotion, and that it is impossible to express emotion using only prosodic features. Following the latter perspective, we examine both prosodic features and phonetic features in this study.

The features that we examine are

prosodic features:
Speech power (P)
Pitch (p)

phonetic features:
12 LPC parameters ($c_1, c_2, c_3, \ldots c_{12}$)
Delta LPC parameter (d)

Here, the LPC (Linear Predictive Coding) parameters are obtained in the LPC analysis [7] of the speech, and the delta LPC parameter calculated from the LPC parameters expresses a time variable feature of the speech spectrum.

2.4.2. Speech feature extraction. The speech features must be extracted from each utterance for the emotion recognition training or testing. Figure 2 illustrates this process.

The first step is to determine the beginning and end points of an utterance. Speech power is compared with a predetermined power threshold. When the speech power first exceeds this threshold for several consecutive frames, this marks the beginning of the utterance. The utterance ends when the speech power drops below this threshold for several frames. By requiring that the speech power be above or below the threshold for several frames before beginning or ending the utterance, short fluctuations due to random noise are neglected.

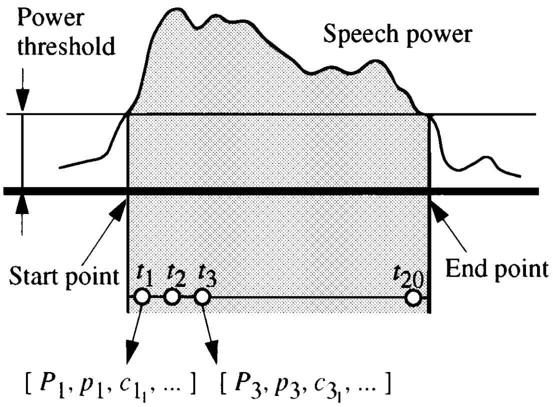

Fig. 2. Speech feature extraction.

Once the start and end points of an utterance have been determined, the utterance is divided into 20 intervals of an equal length in time. Let these 20 intervals be expressed as the vectors $\boldsymbol{f}_1$, $\boldsymbol{f}_2$, $\boldsymbol{f}_3, \ldots \boldsymbol{f}_{20}$. Each of these 20 vectors in turn consists of the 15 speech feature parameters for that interval, (P, p, c_1, $c_2, \ldots c_{12}$, d). The 20 speech vectors of 15 parameters compose a feature matrix of 300 values, or can be flattened out into a 300-length feature vector:

$$\boldsymbol{F}_V = [\boldsymbol{f}_1, \boldsymbol{f}_2, \boldsymbol{f}_3, \ldots \boldsymbol{f}_{20}] \tag{1}$$

where

$$\boldsymbol{f}_n = [P_n, p_n, c_{n1}, c_{n2}, \ldots c_{n12}, d_n] \tag{2}$$

This feature vector $\boldsymbol{F}_v$ is then used as input for the emotion recognition stage.

2.5. Neural Network Architecture

2.5.1. Emotion recognition. The emotion recognition stage of the processing flow is shown in Fig. 3. The network is actually composed of eight sub-neural networks, with one network for each of the eight emotions that are examined. This type of network is called a One-Class-in-One Neural network (OCON) [7]. The feature vector (300 speech features for each utterance) is input into each of the eight sub-neural networks. The output from each sub-network is a value ($v_1, v_2, \ldots, v_8$), representing the likelihood that the utterance corresponds to that sub-network's emotion. Decision logic selects the 'best' emotion based on these values.

2.5.2. Sub-neural networks. Figure 4 illustrates the configuration of each of the eight sub-neural

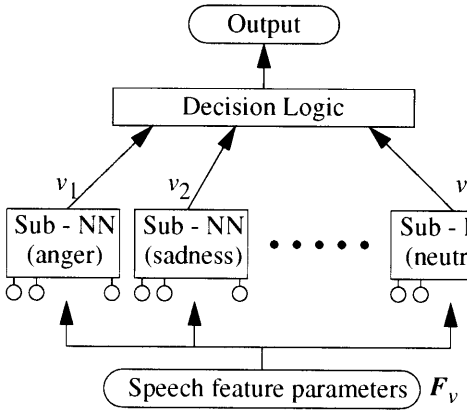

Fig. 3. Emotion recognition stage.

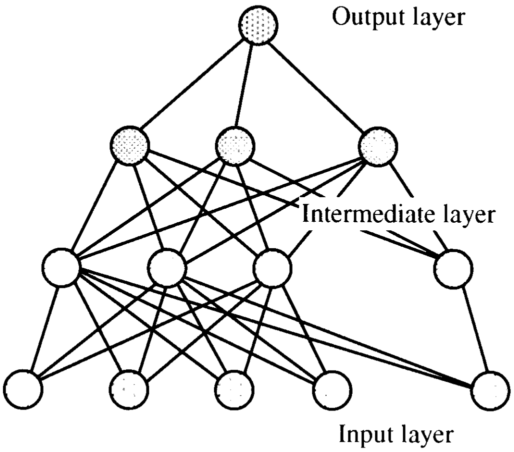

Fig. 4. Sub-neural network configuration.

networks. Each network consists of a four-layer neural network. The input layer takes the 300 speech feature values for each utterance. There are two intermediate layers. The output layer is a single node with an analogue value between 0 and 1. The training of each sub-neural network was done by a backpropagation algorithm, where the maximum allowable error was 0.001 and the training epoch was limited to 1.6×10^4.

Using separate sub-neural networks for the eight emotions allows each network to be adjusted separately. Through testing, it was found that negative emotions, such as anger or sadness, were easy to recognise, but that positive emotions, such as joy, are harder to recognise. With this architecture, we are able to easily change each network separately without redesigning the entire system.

3. Emotion Recognition Experiment

3.1. The Speech Database

We have collected a large speech database for use in the training and testing of this system. In gathering data, we first brought radio actors into the studio and recorded them portraying each of the eight emotions. Since we are only interested (at this stage) in recognising consciously expressed emotions, we had each of our subjects listen to the recordings of the voice actors and try to imitate them.

A total of 100 speakers, 50 male and 50 female native Japanese speakers, served as subjects. Training the database on a large number of subjects produced speaker-independent results.

Each subject uttered a list of 100 Japanese words eight times, one time for each of the eight emotions. Some examples of the words are ‘ikioi’ (force), ‘omoshiroi’ (interesting) and ‘juuichigatsu’ (November). This list of words is phoneme-balanced, meaning that each of the different phonemes of Japanese speech is equally represented within the list. The words are read separately, with no surrounding context. By using a large phoneme-balanced list of context-free words as data, the system achieves context independence.

3.2. Training and Testing Methods

The networks were trained separately for male and female data. Furthermore, the network was trained several different times using different numbers of training subjects, in order to gauge the number of subjects necessary for significant results.

After training, the network is tested using both open and closed testing. In closed testing, the network is tested using the same set of data on which it was trained. In open testing, the network is tested using the remaining data that was not used for training. For example, we can train a network using Speakers #1 and #2. In this case, closed testing uses Speakers #1 and #2, and open testing uses Speaker #3 through Speaker #50.

To optimise the topology of the sub-neural networks, we carried out the training of the network and the closed testing with a small database (10 subjects). Table 1 shows the recognition results. As shown in Table 1, a maximum rate of the closed testing was reached when four-layer networks (16, 4 intermediates nodes) were used. Eight sub-neural networks were optimised in the same way, but the sub-neural networks were grouped into two sets in order to improve performance: (joy, teasing, fear

Table 1. Optimisation results of the number of layers and number of intermediate nodes (10 subjects).

Network		Closed results
3 layer	300-8-1	62.9
	300-32-1	63.6
4 layer	300-8-2-1	66.6
	300-16-4-1	68.1
	300-32-8-1	67.9

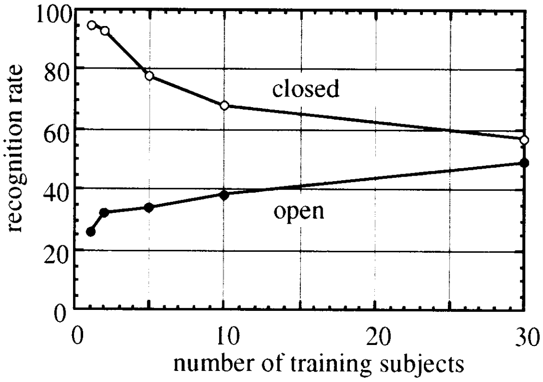

Fig. 5. Recognition results for male data.

and neutral) and (sadness, disgust, anger, surprise). For the first set of emotions, we used a neural network topology of 300, 32, 8, 1 for the input, intermediate #1, intermediate #2, and output layers, respectively. The second set used a topology of 300, 16, 4, 1.

3.3. Results and Discussion

The recognition rates for open and closed testing after training with 1, 2, 5, 10 and 30 subjects are shown in graphical format in Fig. 5 (male data) and Fig. 6 (female data), and in tabular format in Table 2. Each data point is the average of the recognition rate for all eight emotions, then averaged over different training sets. For example, the data point for closed testing on female data with one subject in the training set is the average of the closed recognition rates of several different neural networks, each trained with a different subject as the training set. The average recognition rate for a given network is the average of the recognition rates for the eight emotions using that network.

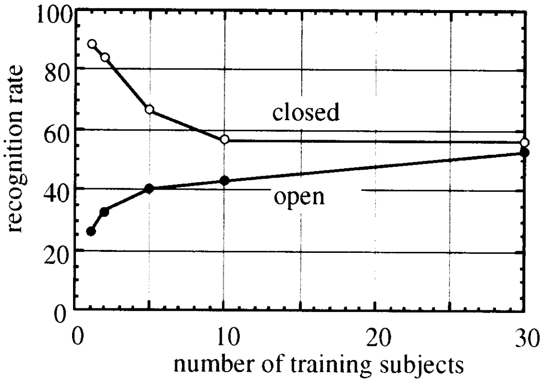

Fig. 6. Recognition results for female data.

Both graphs have similar characteristics. For a small number of training subjects, the closed recognition rates are very high, but the open recognition rates are very low. This is because the network is very finely tuned for that one speaker, and cannot effectively deal with a general case. As the number of speakers used for training increases, the network becomes more general and the open recognition rates increase; however, the closed recognition rates decrease. The open and closed recognition rates appear to converge around a 50% recognition rate. With a network trained using 30 speakers, we reached this recognition rate of 50%. These results indicate that an emotion recognition rate of 50% can be achieved in the speaker-independent mode if we have enough speakers for training.

As a reference for comparing the results of our method using OCON, we studied both an All-Class-in-One Neural Network (ACON) (shown in Fig. 7) and a single-layer neural network using the Learning Vector Quantisation (LVQ) method [9] (shown in Fig. 8).

The ACON used 300, 40, 16 and 8 neurons in the input, intermediate #1, intermediate #2, and output layer, respectively. The training of the ACON was carried out according to the backpropagation algorithm, where a momentum parameter was used. The resultant output is given by selecting the class where the output layer neuron has the highest value.

In the LVQ method, 300 input nodes, 14,400 codebook vectors and eight output clusters were used. The feature vector is compared with all of the codebook vectors. Then, the smallest of the Euclid-

Table 2. Open vs closed results for male and female data.

Number of subjects	closed		open	
	male	female	male	female
1	95.7	88.4	26.4	25.7
2	92.3	83.8	32.7	33.1
5	77.8	66.3	34.1	40.0
10	68.1	56.7	38.5	43.0
30	57.4	56.9	49.1	52.9

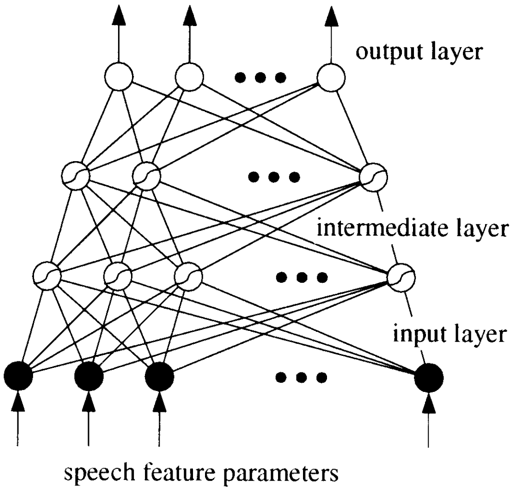

Fig. 7. ACON as an alternative strategy.

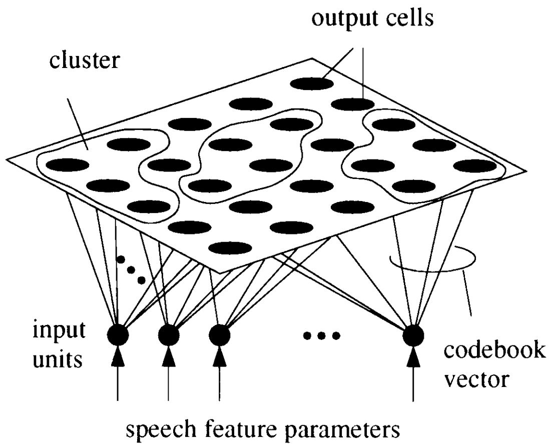

Fig. 8. LVQ as an alternative strategy.

ian distances is made to define the best matching node, and the feature vector is determined to belong to the same class as the nearest codebook vector. The training of the codebook vector $\boldsymbol{w}$ is performed using the LVQ algorithm. At first, the initial values of the codebook vector can be random, but are subsequently trained using the self-organising map algorithm as follows:

$$\begin{aligned}\boldsymbol{w}_i(t+1) = \boldsymbol{w}_i(t) + \exp[-0.5\, \|\boldsymbol{r}_c - \boldsymbol{r}_i\|^2 \\ \times \{\sigma_0(1 - t/T_e)\}^{-2}]\epsilon_{s0}(1 - t/T_e) \\ \times \{\boldsymbol{x} - \boldsymbol{w}_i(t)\}\end{aligned} \tag{3}$$

Here, $\boldsymbol{x}$ is the input vector (i.e. feature vector $\boldsymbol{F}_v$), t is the discrete-time coordinate, ϵ_{s0} is the learning-rate coefficient, T_e is the total number of iterations, $\boldsymbol{r}_c$ and $\boldsymbol{r}_i$ are the location vectors of nodes c and i, respectively, and σ_0 is the coefficient defining the width of the kernel. Next, the codebook vectors are tuned using the following LVQ algorithm:

$$\begin{aligned}&\boldsymbol{w}_c(t+1) = \boldsymbol{w}_c(t) + \epsilon_{l0}(1 - t/T_e)\ \{\boldsymbol{x} - \boldsymbol{w}_c(t)\}, \\ &\text{if } \boldsymbol{x} \text{ is classified correctly} \\ &\boldsymbol{w}_c(t+1) = \boldsymbol{w}_c(t) - \epsilon_{l0}(1 - t/T_e)\ \{\boldsymbol{x} - \boldsymbol{w}_c(t)\}, \\ &\text{if the classification of } \boldsymbol{x} \text{ is incorrect} \\ &\boldsymbol{w}_i(t+1) = \boldsymbol{w}_i(t), \text{ if } i \text{ does not equal } c\end{aligned} \tag{4}$$

Here, ϵ_{l0} is the learning-rate coefficient.

Table 3 shows the recognition results when each network is trained using 30 female speakers. As shown here, the open recognition results of the OCON are similar to those of the ACON, and the open recognition rate achieved by the OCON is higher than that of the LVQ method. These similar results indicate that nonlinear mapping of multi-layer neural networks is feasible for the recognition of emotions.

There are two main ways in which we worked to improve the recognition of the neural network. One way was to raise the standards of convergence for a network, resulting in a longer training time.

The second way was to tune each of the sub-networks for their corresponding emotions. It quickly became apparent that some emotions were much easier to recognise than others. The recognition rates for sadness, disgust, fear and surprise were uniformly higher than the recognition rates for joy, teasing, fear or neutral. To attempt to compensate for this difference, we grouped the emotions into two sets, and used a different neural network topology with a different number of nodes in the intermediate layer for each set, as mentioned previously. Further study may lead to a different network topology for each of the eight emotions under consideration.

4. Conclusions

This paper has proposed a speaker- and context-independent system for emotion recognition in speech using neural networks. We have designed and implemented a one-class-in-one network for emotion recognition. Using a large database of phoneme-balanced Japanese words read by speakers

Table 3. Comparison of different neural network structures (30 subjects).

Network	Closed results	Open results
OCON	56.93	52.87
ACON	71.99	57.18
LVQ	72.33	33.32

consciously trying to portray an emotion, we trained and tested this system. We achieved a recognition rate of approximately 50%. For comparison, two different types of neural networks were also trained using the same data, and similar results were obtained.

The results obtained in this study demonstrate that emotion recognition in speech is feasible, and that neural networks are well suited for this task.

There is still more work that needs to be done in the field of emotion recognition in speech. An examination of the speech features used in this study may allow the number of features to be reduced. In addition, further trials with different topologies of neural networks may also help improve performance.

Acknowledgements. We would like to thank all those who contributed their time and effort into this research at ATR. We would especially like to thank Alexandre Solomide, whose work as an intern at ATR was the basis for this report.

References

1. Nakatsu R, Tosa N, Ochi T. Construction of an interactive movie system for multi-person participation. Proc Int Conf on Multimedia Computing and Systems 1998; 228–232
2. Murray IR, Arnott JL. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. J Acoustical Soc Am 1993; 93(2): 1097–1108.
3. Mozziaconacci S. Pitch variations and emotions in speech. Proc ICPhS 95 1995; 1: 178
4. Sheren KR. How emotion is expressed in speech and singing. Proc ICPhS 95 1995; 3: 90
5. Klasmeyer G, Sendlneier WF. Objective voice parameters to characterise the emotional content in speech. Proc ICPhS 95 1995; 1: 182
6. McGilloway S, Cowie R, Cowie ED. Prosodic signs of emotion in speech: Preliminary results from a new technique for automated statistical analysis. Proc ICPhS 95 1995; 1: 250
7. Markel JM, Gray AH. Linear Prediction of Speech. Springer-Verlag, 1976
8. Obaidat MS. Editorial: Artificial neural networks to systems, man, and cybernetics: Characteristics, structures, and applications. IEEE Trans Systems, Man and Cybernetics – Part B: Cybernetics 1998: 28(4): 489–495
9. Kohonen T. Self-Organizing Maps. Springer-Verlag, 1995