Speech Emotion Perception by Human and Machine

Szabolcs Levente Tóth, David Sztahó, and Klára Vicsi

Laboratory of Speech Acoustics, Budapest University of Technology and Economics,
Department of Telecommunications and Media Informatics,
Stoczek u. 2, 1111 Budapest, Hungary
{toth.sz,vicsi,sztaho.david}@tmit.bme.hu

Abstract. The human speech contains and reflects information about the emotional state of the speaker. The importance of research of emotions is increasing in telematics, information technologies and even in health services. The research of the mean acoustical parameters of the emotions is a very complicated task. The emotions are mainly characterized by suprasegmental parameters, but other segmental factors can contribute to the perception of the emotions as well. These parameters are varying within one language, according to speakers etc. In the first part of our research work, human emotion perception was examined. Steps of creating an emotional speech database are presented. The database contains recordings of 3 Hungarian sentences with 8 basic emotions pronounced by nonprofessional speakers. Comparison of perception test results obtained with database recorded by nonprofessional speakers showed similar recognition results as an earlier perception test obtained with professional actors/actresses. It was also made clear, that a neutral sentence before listening to the expression of the emotion pronounced by the same speakers cannot help the perception of the emotion in a great extent. In the second part of our research work, an automatic emotion recognition system was developed. Statistical methods (HMM) were used to train different emotional models. The optimization of the recognition was done by changing the acoustic preprocessing parameters and the number of states of the Markov models.

Keywords: Emotion Recognition, Automatic Speech Recognition, Human Speech Perception, Speech Technology, Hidden Markov Models, MFCC.

1 Introduction

Linguistic and paralinguistic elements are both important parts of human-human communication. The additional meaning (emotions and attitude) represented in multimodal forms of communication helps to understand the contextual background. In human-machine communication the recognition and synthesis of emotions serves a various range of purposes: it helps to improve user acceptance level of communication applications (with special focus to the frustration emotion/attitude), it can support a parallel telecommunication channel (e.g. automatic generated emoticons in online text messaging), it helps to achieve more naturalistic virtual characters, it can even help researches in psychology. Multimodal emotion synthesis and recognition applications are widely used in user-orientated information technology.

A. Esposito et al. (Eds.): HH and HM Interaction 2007, LNAI 5042, pp. 213–224, 2008.

Researchers of emotion perception and recognition use mostly acted (simulated) emotional databases (e.g. the Berlin database [14], the Groningen corpus [10]or the Danish emotional speech database [7]. Natural emotional reactions are much harder to inspire; the number of truly spontaneous databases is limited [9]. One way of obtaining spontaneous emotional data is to record radio or television talk shows (but here the aired shows are already strongly edited), data from call centers, therapy sessions, etc. [1, 2, 5, 9]. Ethical problems may also come up with designing emotionally triggered situations. The difference between the much less accessible spontaneous emotional expressions and the acted emotions is a matter of discussion [4, 6].

It is widely known that prosody contains at least part of the emotional information in speech. Prosody can be divided into three components: intonation, tempo, and loudness, represented as pitch, duration and energy features of the speech signal [11], and according to our opinion, the tonal information, represented as spectra. The question, which main features contribute the most to understanding emotion information, is yet to be answered.

In the first part of our research, human speech emotion perception was examined. Our first aim was to decide whether the emotional database built from nonprofessional speakers' recordings can compete with professional actor/actress databases. Our second aim was to find out how much an emotionally neutral sentence of the same speaker contributes to the recognition of a sentence with emotional content.

In the second part of our research, automatic speech emotion recognition tests were carried out using a greatly expanded database. We were looking for the most important features for the automatic speech emotion recognition.

2 Examination of Speech Emotion Perception

2.1 Creating the Emotional Database

International literature of emotion research uses a number of basic sets of emotions: from just three emotions (depression, neutrality, suicidal state [9]) up to thirteen different kinds of emotions (like the [3, 16]) or even more. One of the most common sets is fear, anger, sadness and happiness. This categorization reflects the theory that a few universal types underlie the whole emotional life [6]. Another possible categorization is the one commonly used in psychology, linguistics and speech technology, also described in the MPEG-4 standard [15]: happiness, sadness, anger, surprised, and scorn/disgusted. The emotions in MPEG-4 standard were originally created to describe facial animation parameters (FAPs) of virtual characters. Our final selection of emotions is based on the six emotions described above, together with the nervous/excited emotion (because of its obvious role in telecommunication situations), and the emotionally neutral state.

The emotions recorded in our database were acted, forced emotions. Spontaneous speech emotions would have been a better base for a perceptual investigation, but they are much harder to acquire; not to mention the ethical dilemma of forcing people to produce genuine emotions. In most social environment it is highly impolite for a grownup human being to express emotions aloud in most circumstances. This dilemma is a common problem in emotional database projects, although a number of researches [4, 6] showed that a well-designed acted emotional database can play a useful role, too.

Table 1. Range of emotions examined

1.	happy
2.	sad
3.	angry
4.	surprised
5.	disgusted/scorn
6.	afraid
7.	nervous/excited
8.	neutrality

Table 2. Hungarian sentences without emotional meaning

Sentence in Hungarian	Translation
Kovács Katival szeretnék beszélni!	I would like to speak with Kate Smith, please.
A falatozóban sört, bort, üdítőitalokat és finom malacsültet lehetett kapni.	In the restaurant, you can get beer, wine, beverages and delicious pork.
A jövő hétvégén megyek el.	I leave the next weekend.

Building an emotional database, one must make sure that the emotional information carried by the speech should not be affected with the possible emotional information of the semantic layer already present in the speech. Therefore we decided to record sentences with the least possible emotional meaning, each one of the sentences spoken with all kinds of emotions. Three Hungarian sentences were selected for the speech database created for the speech emotion perception tests (Table 2).

The recorded speakers had to pronounciate three sentences with each emotion. Repetitions were carried out, if it was necessary. The first sentence is a commonly used phrase in telecommunication dialogues. The second sentence is often used in Hungarian speech databases because of its phonetic richness. The third sentence is lacking high frequency components (fricatives), which can be useful for a later spectral domain investigation.

7 speakers (3 women and 4 men) were selected for the subjective tests, each one of them making 8-8 emotional recordings of the 3 sentences. These speakers were workers of the laboratory and associates. The records were made under the same conditions in the Laboratory of Speech Acoustics. The recording equipment we used was PC with soundcards (Creative Audigy), and condensate microphones (Monacor). All speech was recorded at 44100 Hz sample rate, 16 bits and mono sound. The recordings were amplitude-normalized.

2.2 Speech Emotion Perception Tests

The human speech emotion recognition tests were carried out using software we built, ran on both portable and desktop computers. 13 listeners were contributing to the research as subjects for the perception tests. The listeners had to listen to series of

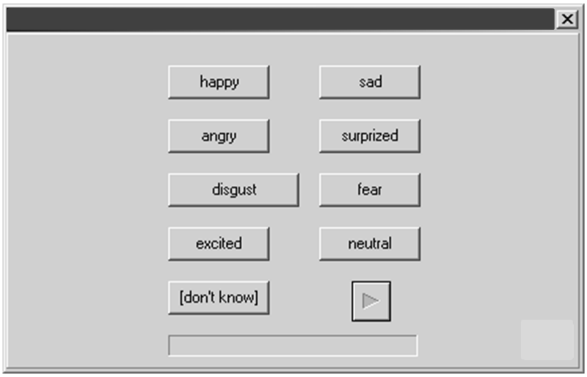

Fig. 1. User interface of the software used for subjective testing

emotional speech recordings through headphones, and after each emotional recording, they had to choose, which emotion they recognized. The speech recordings were played back using software developed for subjective testing purposes (Figure 1).

The software statistically mixed the playing order of the recordings. Because of that, subjects never listened to a recording of the same speaker successively, they could not get used to one or another person's individual way of expressing emotions. The perception tests began with instructions and a few samples of the emotional speech to make the subjects familiar with the test setup. In the first sequence, the test subjects listened to a sequence of emotional sentences alone. In the second sequence they listened to emotional sentences just after a neutral sentence of the same speaker and sentence type. In order to check the learning effect, a third sequence of shuffled recordings was also played (once more, without the reference sentences). The listening tests lasted about 30-40 minutes; therefore we enrolled two short brakes to avoid the effects of fatigue.

2.3 Discussion of the Perception Test Results

Perception test results are shown in Figures 2-4. During the emotion perception tests, one of our aims was to decide whether the emotionally neutral sentence before the emotional one helps the human emotion recognition. Figure 2 shows the emotional recognition of the previously described three test-phases. The light gray columns represent recognition results of the emotional sentences alone (first sequence of listening), the darker columns stand for results of emotional sentences after neutral ones (second turn), and the white bars represent the results of emotional sentences alone again (third listening). The average and standard deviation values presented in the legend of the figure were calculated according to the different speakers.

The test results show that the listeners can recognize all of the emotions better together with emotionally neutral sentences, but the order of magnitude of the standard deviations makes the improvements less significant. This result suggests that the emotional sentence alone may give almost all the information necessary for the recogni tion of its emotion. The learning effect does not influence the result, because results

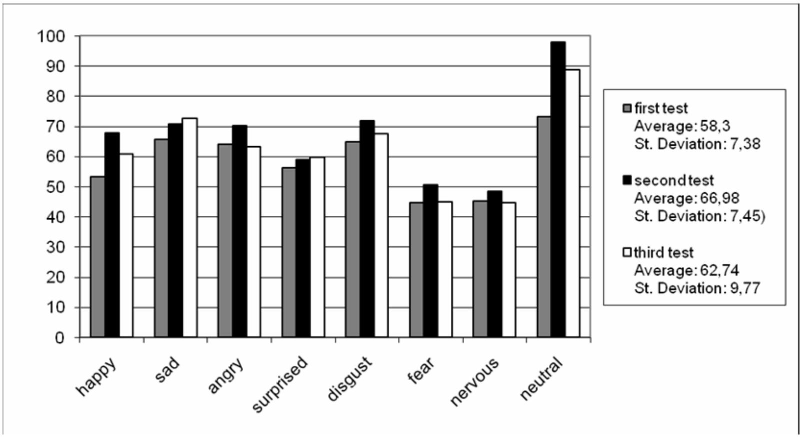

Fig. 2. Emotional perception test results; in percentage

of the third listening series do not differ from the first one in a big extent, considering the order of magnitude of the standard deviation.

The best recognized emotions were sad, disgust/scorn, and angry. These emotions are mostly categorized as strongly negative emotions [6], but so is fear, which was the least recognized emotion in our perception research. The next level below the best recognized three emotions includes surprised and happy. Fear and nervous were the least recognized emotions during the speech emotion perception tests. Altogether, it is clear, that at this rate of human speech emotion recognition (around 60%); it will be a hard task to create automatic speech emotion recognition system with much higher recognition results. It is also obvious from the results, how much the semantic information improves human emotion recognition.

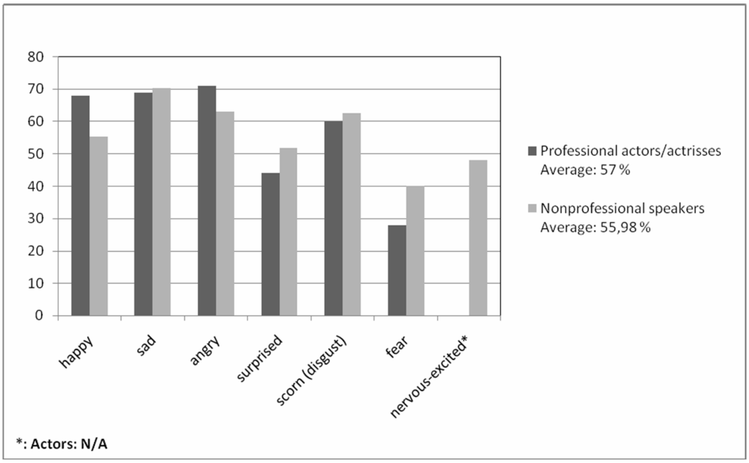

Fig. 3. Comparison of recognition rates of the corresponding emotions, with actors and nonprofessional speakers; in percentage

The results of our subjective listening tests were compared to a former Hungarian emotion perception research result [8]. The speech recordings of the former research were made with professional actors and actresses. The purpose of perception tests in that research was to help evaluation of a Hungarian emotional speech synthesis. Our research and [8] had a common set of seven emotions. Five different sentences were recorded from two actresses and one actor. The comparison is shown in Figure 3.

It was interesting for us to see that our perception test results obtained with nonprofessional speakers does not differ systematically from the other one with actors. Happiness and anger was recognizable better in case of professional actors, all the other emotions were recognized better with nonprofessional speakers, But the difference is not considerable. This knowledge is useful for planning new emotional databases in the future, for the expenses of the actors can be eliminated [8], [12].

3 Automatic Emotion Recognition

Discrete recognition in general means to estimate, which communicable, discrete symbol (in our case: emotion) the examined section of the signal (speech) contains. The estimation should be based on certain measurable properties of the infinitely diverse, continuous speech signal. In the automatic recognition task, the recognizable shape should first be characterized by measurable parameters. This step is often called feature extraction, because the preprocessing intends to create the most efficient parameters for the recognition. The aim of feature extraction is to select those features, which can be a base of a good classification. Unfortunately, currently, we do not know exactly, which speech parameters characterize the emotions. In general, fundamental frequency and energy (intensity) time courses are the most commonly used features in speech emotion research.

In the literature, there are detailed experiments looking for the most efficient parameters [11], [17]. For example, Hozjan and Kacic described two basic approaches in feature extraction tasks in emotion recognition. The time-domain and spectral features of the signal can be examined in a longer or in a shorter time frame. The long-term features' frame length is characteristically in the time range of the phrases, while the short-term frames are of smaller range, like the length of words, morphemes or phonemes, or even smaller-than-phoneme time range.

Human speech emotion perception test resulted around 60% recognition rate, Therefore, we do not expect to get much better automatic speech emotion recognition results.

3.1 Database

Preliminary tests showed that the size of the database was not sufficient to train an automatic recognizer; therefore it was greatly expanded with new recordings. The expanded database contained speech of a wider range of speakers as in the perception tests: besides the original recordings of the researchers and students of the laboratory, 30 more people joined to expand the emotional speech database: students from other faculties, e.g. physiologists, teachers, lawyers, musicians and many more were included. In view of the former subjective test results, the new recordings were elaborately selected by listening tests; thus the worst recognized recordings were not included in the final database. The listening tests were carried out by a small group of the researchers; altogether 133 recordings were excluded from the training and testing databases. The expanded database contained emotional speech recordings from 37 speakers. The new database contained altogether 888 recordings (37 speakers, 3 sentences, 8 emotions), from which 755 were selected to train the recognizer. In average, every sentence was recorded 31 times with each emotional state.

3.2 Recognition Experiment

The emotional recognizer was built using HTK [13] for training Hidden Markov Models. The feature extraction was done by the prosodic acoustic preprocessor unit of a segmentator and recognizer developed in our laboratory [19], computing the smoothed fundamental frequency and intensity values. The classification was prepared by the framework of HTK. The fundamental frequency and intensity values were calculated from the input speech signal with 10 ms frame rate, fundamental frequency with 75 ms, intensity with 150 ms window. Fundamental frequency values were corrected by median filtering. Delta and delta-delta parameters were obtained by derivating fundamental frequency and intensity; using both time and frequency domain derivating. The spectral vector-coefficients were obtained with the commonly used Bark-scale filtering.

The average length of the spoken sentences was 2 to 3 seconds; therefore the number of training vectors per recordings was about 200-300. The HMM tries to divide the frames (coming in every 10 ms) by 2 less than the number of states. Defining the number of states in the Markov-model is a basic question during optimization. Number of states set too low or too high has an influence on the basic functioning of the recognizer.

Numerous optimization experiments were done to obtain the best emotion recognition parameters. In the first experiments we looked for the optimal number of states of the HMM emotion models. In the second part, spectral domain features were added to the recognizer's input vectors. Trouble with spectral domain processing is, that phonetic variability in sentences causes a problem calculating the long-term spectrum of different sentences (especially the fricatives, because of their wide spread spectrum). In accordance with some sources, another problem is that, although the third and following formants may be defining the emotional content of speech, the higher second formant of certain phonemes (e.g. in the phoneme "i") can have a higher average frequency, then the third formant of other phonemes (e.g. "a"). Therefore, there is no way to differentiate by formant frequencies without segmenting and recognizing the corresponding segments. If the speech is unknown, with no fixed vocabulary, this information is hard to get and will cause another important source of error in the emotion recognition procedure.

3.3 Test Results and Evaluation

Four sets of training data were separated:

- all three sentences,
- the first sentence,
- the second sentence, and
- the third sentence.

32 recordings of each sentence, of the best 4 speaker (selected with subjective tests) served as testing data.

After the preliminary selection of recordings it turned out that the amount of training data was insufficient for training the third sentence alone with 8 emotions. Therefore training the recognizer with the third sentence type was not possible.

The results of our preliminary tests showed that the size of the training database was clearly insufficient. The further expansion of the database would probably lead to better recognition results, but for that we have not had the resources. Therefore, it was decided to carry on with the automatic speech emotion recognition experiments with the reduced set of the four best recognized emotions (Figure 4).

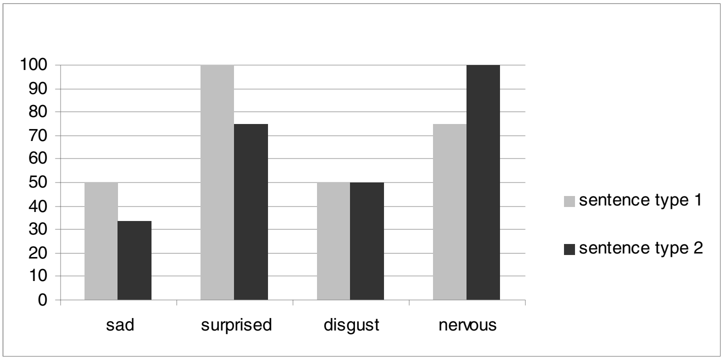

Fig. 4. Recognition of 4 emotions (surprised, disgust, nervous, sad), recognizer trained with two sentences separately (Kovács Katival szeretnék beszélni! – A falatozóban sört, bort, üdítőitalokat és finom malacsültet lehet kapni.). The input training vectors were composed of the values of fundamental frequency and energy, together with their first and second derivates.

Figure 4. shows that the results of the recognizer trained with the reduced set of emotions were in the same order of magnitude as the subjective (human) recognition results.

In the first optimization experiment fundamental frequency and intensity features together with their first and second derivates were extracted as acoustical preprocessing (14 dimensional vectors, 7 for intensity and 7 for fundamental frequency). The state numbers were changed from 3 up to 17. The best results were obtained when the state number was 13, that is the state number we used in the recognition experiments

In the second experiment we added spectral domain information to the training data. The spectral parameters were calculated with 25 ms window size, with an HTK implementation of discrete cosine-transformation of Mel-scale cepstral filter coefficients.

Comparison of the automatic speech emotion recognition results, recognizer trained with fundamental frequency, energy, and with or without spectral features is shown in Figure 5.

Adding spectral features to fundamental frequency and intensity clearly improved recognition rates.

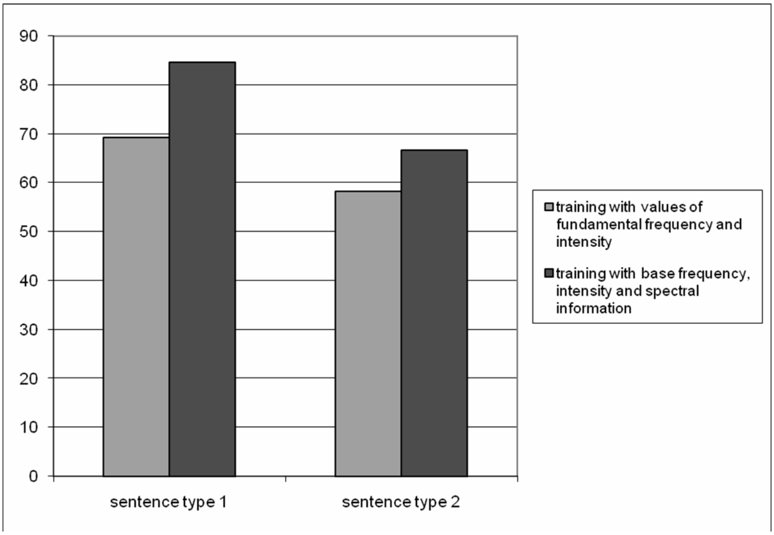

Fig. 5. Recognition of 4 emotions (surprised, disgust, nervous, sad), recognizer trained with the two sentences separately. Training data composed of fundamental frequency and energy (light grey bars) and fundamental frequency, energy and spectral values.

In these experiments, the recognizer was trained separately with two different types of sentences. We also trained the recognizer with the two different types of sentences together, but the recognition was much weaker. While the first sentence is a short stating sentence, the second one is a complex sentence with enumerations. The different grammatical structure probably forces the speaker to choose different suprasegmental expressions. The automatic recognizer performs better, when only separate grammatical type of sentence serves for training data. Multiple types of training sentences clearly prevent the recognizer from creating good classes. We suppose that this is the reason, why separate training resulted better recognition rates.

By a thorough examination of the vowel duration, fundamental frequency and energy time-courses of the speech signals, it was found that these three prosodic parameters depend on the linguistic type of the sentences in a big extent. Examples of vowel duration and fundamental frequency time-courses of sentence type 1. and 2. are shown in Figures 6-9.

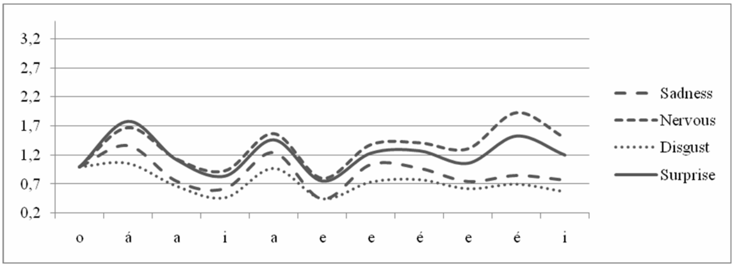

Fig. 6. Vowel durations of the second sentence, normalized with the value of the first vowel's duration, for each emotion

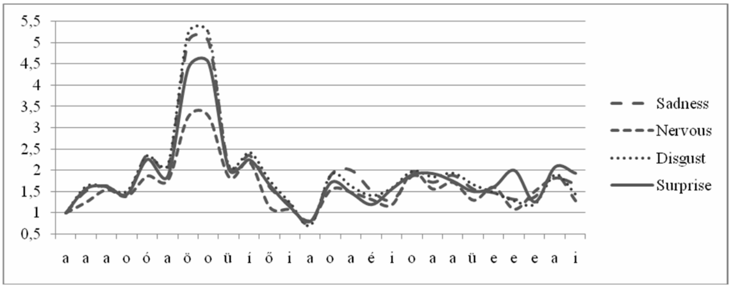

Fig. 7. Vowel durations of the second sentence, normalized with the value of the first vowel's duration, for each emotion

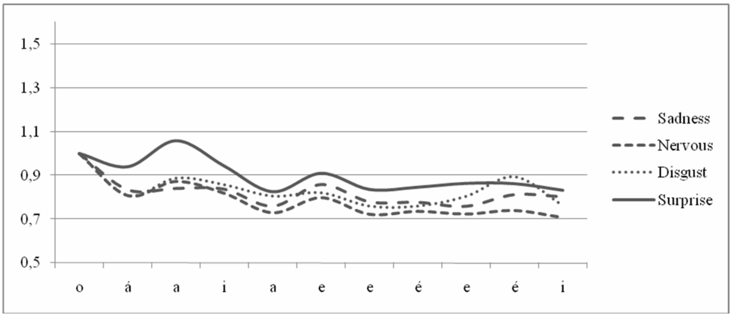

Fig. 8. Fundamental frequency time-course of the first sentence, normalized with the first frequency value, for each emotion

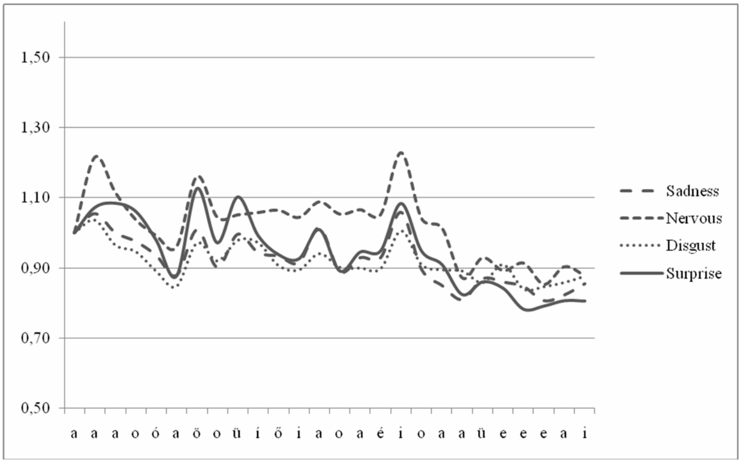

Fig. 9. Fundamental frequency time-course of the second sentence, normalized with the first frequency value, for each emotion

It was observed, that according to the emotions, the time courses of these prosodic parameters were changing in parallel for most part of the sentences. Special places were found, where theses parallelisms were mixed up, e.g. the second part of the second sentence. These places were found in different locations, according to the linguistic type of sentences [18]. The semantic meaning of all of the sentences were carefully selected not to affect the emotional meaning, therefore, we suppose that the basic grammatical structure of the sentences is involved with the location of the emotional expression.

4 Conclusions

The article shows early steps in development of an emotion recognition system capable of recognizing eight basic emotions. The number of these emotions was reduced during the development of the automatic recognizer to the four best recognized emotions: sad, surprised, disgusted/scorn and nervous/excited. The first aim of our research was to learn about the processes of human emotion recognition. Our second aim was to construct an emotion recognizer and optimize its important parameters.

In the human emotion perception experiment, we did not get significantly different results with emotional speech recordings of nonprofessional and professional speakers. The test results show that the listeners cannot recognize the sentences with emotional meaning better together with sentences without emotional meaning in a great extent. This result suggests that a sentence with emotional meaning alone may give all the information necessary for emotion recognition.

In general, fundamental frequencies and intensities time courses were the most commonly used features for the expression of emotions, both in the field of speech recognition and synthesis. In our automatic emotion recognition experiment, it was found, that adding spectral information greatly improves the recognition results.

On the other hand, it was also found, that recognition results with separately trained sentence-types are much better than training all sentence types together. Examination of certain features of the speech signal showed why the recognition result was improved.

In the future, increased training database is necessary to clear out the nature of speech emotions better, and to improve the results of the automatic recognition.

References

1. Amir, N., Ziv, S., Cohen, R.: Characteristics of authentic anger in Hebrew speech. In: Proceedings of Eurospeech 2003, Geneva, pp. 713–716 (2003)
2. Ang, J., Dhillon, R., Krupski, A., Shipberg, E., Stockle, A.: Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In: Proceedings of ICLSP 2002, Denver, Colorado, pp. 2037–2040 (2002)
3. Banse, R., Scherer, K.: Acoustic profiles in vocal emotion expression. J. Pers. Social Psychol. 70(3), 614–636 (1996)
4. Bänziger, T., Scherer, K.R.: Using Actor Portrayals to Systematically Study Multimodal Emotion Expression: The GEMEP Corpus. In: Affective Computing and Intelligent Interaction, pp. 476–487. Springer, Berlin (2007)

5. Batliner, A., Fischer, K., Huber, R., Spilker, J., Noeth, E.: How to find trouble in communication. Speech Commun. 40, 117–143 (2003)
6. Douglas-Cowie, E., Campbell, N., Cowie, R., Roach, P.: Emotional speech: towards a new generation of databases. Speech Communication 40, 33–60 (2003)
7. Engberg, I.S., Hansen, A.V., Andersen, O., Dalsgaard, P.: Design, recording and verification of a Danish Emotional Speech Database. In: Proceedings of the Eurospeech 1997, Rhodes, Greece (1997)
8. Fék, M., Olaszy, G., Szabó, J., Németh, G., Gordos, G.: Érzelem kifejezése gépi beszéddel. Beszédkutatás 2005. MTA-NyTI, pp. 134–144 (2005)
9. France, D., Shiavi, R., Silverman, S., Silverman, M., Wilkes, D.: Acoustical properties of speech as indicators of depression and suicidal risk. IEEE Trans. Biomed. Engng. 47(7), 829–837 (2000)
10. Groningen corpus S0020 ELRA, http://www.icp.inpg.fr/ELRA
11. Hozjan, V., Kacic, Z.: A rule-based emotion-dependent feature extraction method for emotion analysis from speech. The Journal of the Acoustical Society of America 119(5), 3109–3120 (2006)
12. Hozjan, V., Kacic, Z.: Context-Independent Multilingual Emotion Recognition from Speech Signals. International Journal of Speech Technology 6, 311–320 (2003)
13. HTK: HMM Toolkit; Speech Recognition Toolkit, http://htk.eng.cam.ac.uk/docs/docs
14. Kienast, M., Sendlmeier, W.F.: Acoustical analysis of spectral and temporal changes in emotional speech. In: Proceedings of the ISCA ITRW on Speech and Emotion, Newcastle, September 5–7, pp. 92–97 (2000)
15. Mozziconacci, S.: Speech Variability and Emotion: Production and Perception. Technical University of Eindhoven, Proefschrift (1998)
16. MPEG-4: ISO/IEC 14496 standard (1999), http://www.iec.ch
17. Navas, E., Hernáez, I., Luengo, I.: An Objective and Subjective Study of the Role of Semantics and Prosodic Features in Building Corpora for Emotional TTS. IEEE transactions on audio, speech, and language processing 14(4) (July 2006)
18. Sztahó, D.: Emotion perception and automatic emotion recognition in Hungarian. In: Proceedings of Scientific Conference for Students, BME (2007), http://www.vik.bme.hu/tdk
19. Vicsi, K., Szaszák, G.: Automatic Segmentation for Continuous Speech on Word Level Based Suprasegmental Features. International Journal of Speech Technology (April 2006)