Chapter 5

Design of Emotion Recognition System

Dominik Uhrin, Pavol Partila, Jaroslav Frnda,
Lukas Sevcik, Miroslav Voznak and
Jerry Chun Wei Lin

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/116607

Abstract

The chapter deals with a speech emotion recognition system as a complex solution including a Czech speech database of emotion samples in a form of short sound records and the tool evaluating database samples by using subjective methods. The chapter also involves individual components of an emotion recognition system and shortly describes their functions. In order to create the database of emotion samples for learning and training of emotional classifier, it was necessary to extract short sound recordings from radio and TV broadcastings. In the second step, all records in emotion database were evaluated using our designed evaluation tool and results were automatically evaluated how they are credible and reliable and how they represent different states of emotions. As a result, three final databases were formed. The chapter also describes the idea of new potential model of a complex emotion recognition system as a whole unit.

Keywords: classifier, database, emotion, neural network, emotion, recognition system, sample

1. Introduction

There are many fields which require the information about the emotional state. Technological development puts more pressure on the greater accuracy and simplicity of communication between man and computer. Current applications use the speech as an input-output interface, and this trend is broadening increasingly. This type of interaction can develop two problems caused by the absence of information on the emotional state. The first one is an incorrect recognition of a sentence or a command from a person who is facing stress situation. The machine recognizes human speech differently than a human with hearing. The accuracy is affected by changes in the voice signal which are caused by stress in the vocal tract. The second problem

is an absence of emotional state regarding the machine speech of the loudspeaker. Typical application, such as Text-To-Speech, combines truly correct parts of speech sounds but on the other side, this signal does not contain any emotion. Such a speech influences the human and is synthetically unreliable.

2. State of the art

Psychological research confirmed that emotional state has an impact on human speech and also on the physiological state of body. Noticeable improvement has been made in a field of automatic classification of human emotion as well. These achievements have been attained by recognition techniques mainly in past few decades. In comparison to the past 10 or 20 years, contemporary computation power of processors has reached very different level. Thus, this new hardware allows us to real-time use of methods for emotion recognition. A lot of secondary information obtained from speech could not have been processed previously due to the lack of computation power and method used for the process. But the bad quality of training samples remained the significant issue. Nowadays, there is a lot of emotional recordings databases. However, a significant number of databases are created based on simulated emotions by actors instead of real-life emotions. On the other hand, quality of sound recordings is very high because of the use of studio recording. Therefore, the recordings do not contain any unnecessary noise. Creating such a type of database is much easier in comparison to the real emotion database samples. This kind of samples has to be manually cut out from sound recordings which contain real emotions and recording processing is much more time consuming. Working with simulated emotion recordings is simpler because each of them is labeled. The labels contain information on features like the kind of recorded emotion, gender of an actor, etc. The fact that actors are pronouncing mostly the same sentences also guarantees the same context of recordings.

These recordings are more efficient in terms of training emotional classifier. The following recording databases can be considered as some of the most known and recent ones: Humane [1], Emotional Prosody Speech and Transcripts, Danish Emotional Speech Corpus, Berlin Emotional Speech Database [2], Serbian Emotional Speech Database [3].

3. Methodology

The emotion recognition system can be divided into two parts: the first part is emotional classifier used for classification of emotion from a sound recording and the second part of the system is the emotion database intended for learning and training of emotional classifier. Emotional classifier is very important component of the emotion recognition system, and it is also the core of the system. Three parts of neural classifier are shortly described. The first part describes the process of a sound sample preparation, the second part discusses feature extraction from a sample, and the last part of emotional classifier describes a specific type of neural networks and the way how these networks work. **Figure 1** shows the block diagram

http://dx.doi.org/10.5772/116607

of emotional classifier. Three more subchapters are dedicated to emotion database. First subchapter deals with the creation of a sample database, its extraction and technical parameters. In the second part, we describe a tool for subjective evaluation. It describes certain parts of the tool and also the process of evaluation altogether with the processing of evaluation results. Last subchapter is dedicated to the future vision of a complex automated emotion recognition system and its use.

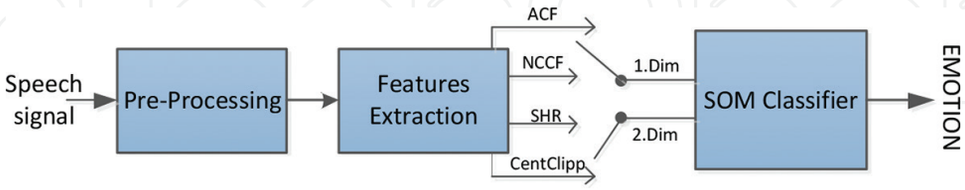

Figure 1. Speech emotion recognition system.

3.1. Preprocessing

Speech signal is stochastic by nature. However, the speech signal has a number of characteristics that may be considered as unwanted during the processing. This part of the chapter deals with a process called preprocessing—an important part of the digital speech signal processing. These few steps prepare the signal for subsequent extraction of signal parameters. The values of these parameters could be wrong without the preprocessing process. **Figure 2** shows the preprocessing. Speech signal digitizing and sound cards processing may also have side effects. Sound cards insert the DC component into the speech signal. It may not be suitable for the calculation of parameters such as signal energy or others. Therefore, removing of the DC component is part of the preprocessing. Unwanted DC offset is removed by subtracting the mean of each sample. In real-time applications that represent many cases, we do not have all the audio, which means that the true mean cannot be estimated. Thus, in real-time processing it is necessary to calculate the mean value for each sample. The mean value for the current sample can be determined based on the mean value of the previous sample. In the end, the DC component is removed by a simple subtraction of the mean value [4, 5]. The energy of the signal decreases with increasing frequency is another characteristic of the speech signal. Most of the speech signal energy is included in the first 300 Hz of spectra, which means that the information of the higher frequencies expires compared to higher energies from the bottom of the spectrum. Saving of the higher end of the spectrum is achieved by increasing the energy in the higher part of spectra artificially, which represents the second part of the preprocessing. Spectrum part energy increase is performed using pre-emphasis. As mentioned above, speech signal has a stochastic character. From a mathematical point of view, it is very difficult to find dependency and frequency in this signal. Because of that it is necessary to divide the speech signal into smaller parts called frames. Frame length is selected between 20 and 30 ms. This length is derived from the lag of the human vocal tract. Division of the signal is the third part of the preprocessing. Values of samples in neighboring frames may vary rapidly, therefore frame overlap is appropriate. The frame overlap is selected in half. Processing speech signals between the frames can have a

side effect because the edges of neighboring frames may have sharp transitions. It can have a bad influence mostly on speech processing and frequency analysis. The above-mentioned disadvantage can be removed by applying the window function on each frame. Many window functions are used in speech processing, and the choice depends on the characteristics of the following processing methods. Hamming window function is used in most such cases due to its suitable properties in both the time domain and frequency [5, 6].

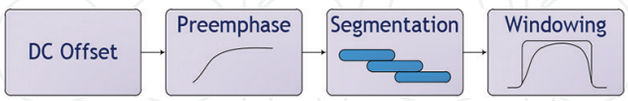

Figure 2. Preprocessing process.

3.2. Speech processing parameters

Volume, intonation, and tempo are speech characteristics that can be recognized by a human ear. In DSP (digital signal processing) and speech processing in particular, we use some other parameters which characterize speech signal and human vocal tract. Other two parameters signal energy and ZCR (zero crossing rate) are also important in speech processing. These parameters were used as a voice activity detector intended to eliminate silence or noise. Signal energy is characterized by intensity, and the human ear can sense it as a volume. Energy is influenced by the way of recording and digitizing speech, speaker distance from the microphone, and other features. Voiced and unvoiced parts can be separated using a sound energy profile. ZCR describes how many times the polarity of the signal changes, in other words how many times it crosses zero. This parameter can also carry information about F0 change. ZCR carries the information on both the voice activity and the energy [8, 9]. The fundamental frequency of vocal cords (F0) is one of the most important parameters within the speech processing because it carries a lot of information about the vocal tract, and thus also the basic features of man. Age, gender, speech errors, and emotional state of a man can be determined using this parameter. There are several methods of signal processing that enable to estimate the fundamental frequency. Human speech consists of voiced and unvoiced speech sounds. The vocal cords are almost completely open in the creation of voiceless phonemes. The basic tone does not arise with opened vocal cords, and therefore F0 can be calculated only from the nonvoiceless parts. Each of the methods used to calculate F0 has its advantages and disadvantages. The following methods can be used to calculate F0: auto-correlation function, normalized cross-correlation function, auto-correlation function with central clipping and sub-harmonic-to-harmonic ratio [9, 10].

3.3. Self-organizing feature map

The emotional state classifier is based on self-organizing maps (SOM). These maps represent a specific type of neural networks with uncontrolled competitive learning. There are generally two-dimensional maps of neurons. The learning process of SOM is uncontrolled, which means that the input data do not need to know the output. In the process of learning, SOMs

http://dx.doi.org/10.5772/116607

determine for themselves, how to classify the inputs [7]. At the beginning of learning, the weights of all the inputs of neurons can be set randomly. Randomly selected input vectors are applied to neurons and then analyzed in order to find the one which is the most similar to an input. This neuron is called the winner. The weights of neighboring neurons are adjusted according to the following rule in Eq. (1). The equation describes the weight between neurons i and j for $t + 1$ iteration and input $x\_j\ (t)$ [10].

$$w_{ij}\ (t+1) = w_{ij}\ (t) + \gamma \left(x_j\ (t) + w_{ij}\ (t) \right) \tag{1}$$

3.4. Sample database creation

Next iteration means a new vector for input, finding new winner and changing weights between neurons again. When the learning process is completed, the map has a shape that represents the characters of input parameters. In order for the emotional classifier to be as precise as possible, sample database of real emotion for training and learning has to be created. Of course, we are speaking about Czech emotional sample database. It is very difficult to determine the emotion of sound recordings. So the precision of emotional classifier within determination of the emotion depends on how many real emotion samples it learns. A few hours recording from two Czech radio broadcastings was the first step to create real emotion database not simulated by actors. Some of the recordings have been available on the official web page archive of the third radio station, and some of the recordings of Czech television broadcasting were downloaded from share video portal YouTube. Out of television broadcasting, only sound part was cut for the creation of database samples. Parameters of database samples have to fulfill the following conditions:

- Sample had to have a duration from 6 to 6 seconds.
- Sample should not have contained environment noise.
- Sample had to contain human speech in a form of few words or a full sentence.
- Sample had to be uniquely named.

Name of the database sample consists of three parameters: the first one is state of the emotion. There are seven basic emotion states. Database samples have been made for four emotion states because, as for the rest of the states, it is difficult to find real emotion recordings, or it is hard to recognize it by using subjective methods (boredom, disgust, and fear) [1]. As an output format for database samples, waveform audio-file with 16-bit PCM (pulse code modulation) coding, monochannel and sample frequency at the level of 16 kHz have been used. These parameters are sufficient enough, taking into consideration that the source broadcasting recording has been recorded to MPEG-2 Audio Layer III audio file with a bit rate of 128 kbps. Some of the source recordings from which database samples have been obtained were recorded using VideoLAN Client media player. For editing and cutting of source recordings, software Audacity was used. By default, Audacity was unable to edit mp3 audio file; therefore, LAME Encode library had to be installed [11].

3.5. Emotional database formation

Next step to build the emotional database was the creation of the tool for evaluation of database samples from which emotional database has been created after evaluation. As methods for evaluation of emotion samples, the subjective one has been chosen. Subjective methods represent using people to evaluate a small amount of samples in this case. The web page represents the evaluation tool as a direct tool for evaluation connected to MySQL database in order to save the results of subjective evaluation. The web page consists of four pages: the first one is invitation page, it invites evaluating subject and gives it short instruction. The second page is the evaluating page. It is the core of whole evaluating tool. It consists of html5 audio player for playing database sound samples and rollout menu for selection of the state of emotion. Subject simply plays the recording selected by an algorithm, and consequently selects the emotion in rollout menu. The result is sent to and saved into MySQL database. The next two pages of the tool are the final page announcing the end of the process of evaluation to the subject, and the error page announcing to the subject that something went wrong during the evaluation process. The tool also consists of page used to insert samples to the system. It makes inserting much easier and less time consuming. Besides the web page, the tool also consists of MySQL database. As mentioned above, database was used to save the results of subjective evaluation. Database consists of two connected tables. These tables are shown in **Figure 3**. The first table provides information about individual emotional samples and contains four kinds of information. Number column represents how many times the sample has been loaded to audio player. Emotion column represents the first letter from English expression of a selected sample. Ref_id column represents a unique name for a sample. The second table provides the information about evaluation of sample, and it includes seven kinds of information. Meaning of first one, ref_id, is same as in table one.

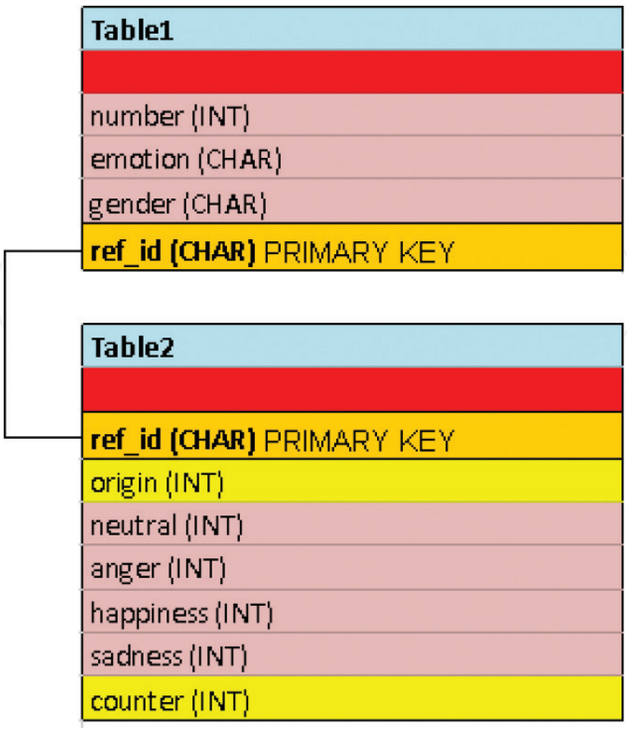

Figure 3. MySQL database tables.

Origin column shows how many times the same emotion has been selected by the subject for sample as was originally selected by the author during the process of creating database samples. Information included in columns three to six represents the emotional state selected by the subject during the process of evaluation. And finally the last column, counter, shows the total number of sample evaluation. Logically, the web page contains also scripts to perform queries by using mysql_query() with SELECT or UPDATE queries. It was also necessary to create a customs function for loading and saving data from or to the database named database_load(). Furthermore, the web page uses POST and GET forms to obtain data from previous page loads. It was necessary to create custom functions generate_sample() and audio(), too. The first function is used to generate the name of the sample that is loaded into html5 audio player using second function audio(). Using htlm5 audio player is much easier. Flash or any other plug-in does not have to be installed. This all is the part of web browser, and it uses much less computation performance. Custom functions were created to insert the obtained results of sample evaluation to database tables. In the first version of this tool, it was necessary to manually process results by exporting tables with results and use external scripts to achieve statistic results. But in this modified version, we created tool function evaluate_results() that uses MySQL scripts to automatically processed evaluation results. The diagram of processing from evaluation till final database is shown in **Figure 4**.

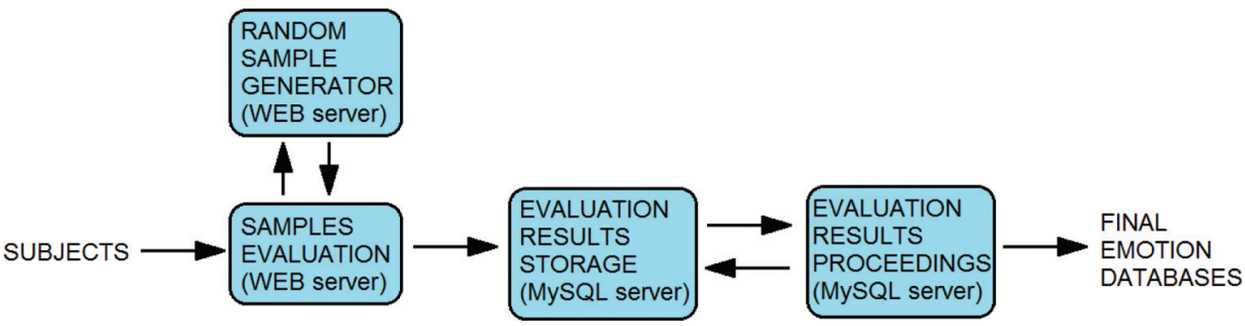

Figure 4. Sample evaluation processing diagram.

3.6. Vision of future

Emotion recognition of Czech spoken language or the attempts to recognize the emotion from Czech spoken words are not at beginnings of their development. Thus, the actual status of an emotion recognition system does not allow the automatized system to be created for emotion recognition which could be used as real application. It is caused by the fact that emotional classifier has abilities to determine emotion variety with precision about of 70–75%. But this is not applicable in real life. The level of precision can be increased by choosing and also modifying the method of learning emotion states, as well as by using real, not acted, emotion samples to learn the emotion. Due to this we are trying, but it remains the close future. For the future when the precision level of emotion determination will be useable in real application, the model of the automated recognition system should appear same as indicated in **Figure 5**. For example, in call centers, employees can be divided into groups. Each group

would take care of customers with different emotional state. Before the customer would be forwarded to call center employee, they would be asked to repeat some sentence or to answer a question.

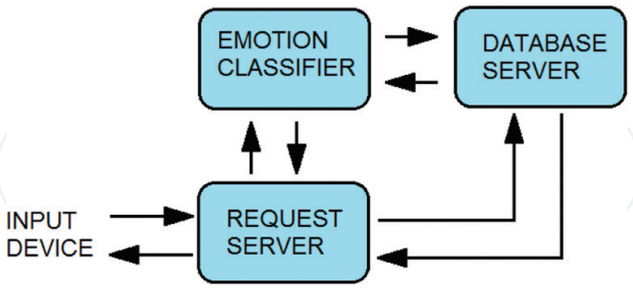

Figure 5. Automatized emotion recognition system.

Based on the results of emotion analysis of their voice made by automatized emotion recognition system, the customer would be then forwarded to the assigned employee. All this would be done real time. Police special force could use this automatized system as well during the negotiation with a kidnapper or terrorist. Their voice could be analyzed for emotion state, which could help the police to make the right steps when solving the problem. The above-mentioned examples are only few from many possible real applications. In general, the automatized system receives the request by a network from some device (such as PC, smart phone, or server) to recognize the emotion from the voice. First, request is registered by the request server and recorded to the database server. Afterward, the file is sent to emotional classifier for analysis. The classifier takes file through all the components described above and sends the result to the both database and request servers. This would result in the customer device from which request was received. It could be a potential scheme for the future automatized emotion recognition system. Emotional classifiers can also be used during the evaluation of samples instead of subjective evaluation methods in the future. It should be easier and less time consuming, but it is a sound of the future.

4. Results

The evaluation of database samples was made by subjects represented by students in age range from 18 to 26 years. The selection rate of each emotion kind for random sample is listed in **Table 1**. As mentioned before, the automatized system was used for sample evaluation. The veracity value was determined for each sample from the database as well. **Table 2** shows first five samples of database with percentage of veracity, state of emotion, and a level of veracity. After determining the veracity value, three levels of veracity were assigned to samples: low, medium, and high. Levels have been set to the following ranges: a low level with a range from 0 to 75, a medium level range from 75 to 90, and a high level percentage range from 90 up to 100. Three final emotion databases have been created based on these three levels. Names of these emotion database samples were formed under

conditions, as shown in **Figure 6**. First database with high veracity of samples is suitable for learning of neural emotional classifier. This classifier has been developed by Mr. Partila at our university [8, 9] and its components are shortly described in the chapter. Second database of samples with medium veracity is suitable for training of neural classifier and verifying of its learning skills. The last database of emotion samples with low level of veracity is formed by samples that contain mixed emotions. As for this emotion database, it was difficult to determine the emotion state of samples. In order to determine the emotion state better, more evaluations of samples have to be performed, or some samples are simply not suitable for it.

Emotion	Probability for one sample (%)
Neutral	34
Anger	28
Happiness	20
Sadness	18

Table 1. Probability with what emotion was selected.

Evaluation veracity (%)	Emotion kind	Level of veracity
71.34	Neutral	Low
38.21	Anger	Low
48.77	Neutral	Low
82.85	Happiness	1. Medium
82.85	Neutral	Medium

Table 2. Veracity table for first five samples with emotion kind determined.

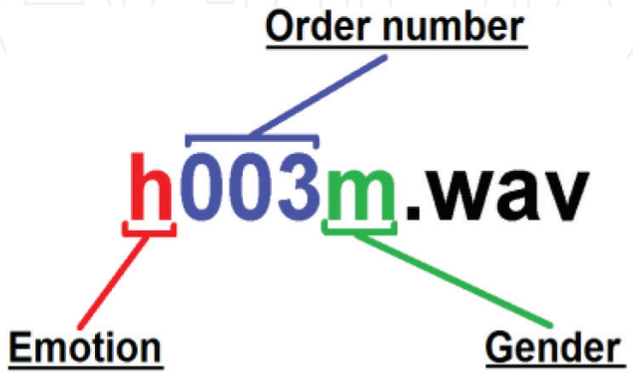

Figure 6. Database sample name.

5. Conclusion

This chapter focused on the emotional classifier as the complex thing including creation of training and learning database. Furthermore, the chapter try to solve the problem of samples preprocessing and also the feature extracting process, which are both important for the next procedure of emotion recognition. Last but not least, it describes functioning of self-organizing feature maps. The SOM classifier has the lowest error rate, and thus the best resolving power between normal and stress emotional states. The tool for subjective sample evaluation has been upgraded for automatized result evaluation that made it easier and less time consuming. All created samples have been evaluated by the specific group of subjects and based on the results, three final databases were formed. Two of them are usable for learning and training the classifier. As for further development, automatic evaluation of samples is an option to be used instead of subjective evaluation in a form of neural classifier.

Acknowledgements

This publication was created within the project Support of VŠB-TUO activities with China with financial support from the Moravian-Silesian Region and partially was supported by the grant SGS reg. no. No. SP2016/170 conducted at VSB-Technical University of Ostrava, Czech Republic.

Author details

Dominik Uhrin[1,*], Pavol Partila[1], Jaroslav Frnda[1], Lukas Sevcik[1], Miroslav Voznak[1] and Jerry Chun Wei Lin[2]

*Address all correspondence to: dominik.uhrin@vsb.cz

1 Department of Telecommunications, Faculty of Electrical Engineering and Computer Science, VSB-Technical University of Ostrava, Ostrava, Poruba, Czech Republic

2 School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, China

References

[1] Ververidis, D., Kotropoulos, C., 2010. A review of emotional speech databases. *In Proc. Panhellenic Conference on Informatics (PCI)*, pp. 560–574.

[2] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B., 2003. A Database of German Emotional Speech. *In Interspeech 2005 – Eurospeech, 9th European Conference on Speech Communication and Technology*, pp. 1517–1520.

[3] Jovičić, S.T., Kašić, Z., Đorđević, M., Rajković, M., 2004. Serbian emotional speech database: design, processing and evaluation. *In 9th Conference Speech and Computer (SPECOM 2004)*, pp. 77–81.

[4] Nicholson, J., Takahashi, K., Nakatsu, R., 2006. Emotion Recognition in Speech Using Neural Networks. *In Neural Computing & Applications*, Volume 9, Issue 4, Springer Verlag, pp. 290–296.

[5] Partila, P., Voznak, M., Mikulec, M., Zdralek, J., 2012. Fundamental Frequency Extraction Method using Central Clipping and its Importance for the Classification of Emotional State. *In Advances in Electrical and Electronic Engineering*, Volume 10, Issue 4, pp. 270–275.

[6] Kasi, K., Zahorian, S. A., 2002. Yet Another Algorithm for Pitch Tracking. *In IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol.1, pp. 361–364, IEEE.

[7] Roussinov, D., Chen, H. A., 1998. Scalable Self organizing Map Algorithm for Textual Classification. *A Neural Network Approach to Thesaurus Generation. In Communication Cognition and Artificial Intelligence*, Volume 15, pp. 81–111.

[8] Lee, C.M., Narayanan, S.S., 2005. Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*, Volume 13, Issue 2, pp. 293–303.

[9] Voznak, M., Partila, P., Mehic, M., Jakovlev, S., 2013. Recognizing Emotions from Human Speech Using 2-D Neural Classifier and Influence the Selection of Input Parameters on its Accuracy. *In 21st Telecommunications Forum Telfor*, art no. 6716272, pp. 482–485.

[10] Partila, P., Voznak, M., 2013. Speech Emotions Recognition Using 2-D Neural Classifier, *Nostradamus 2013: Prediction, Modeling and Analysis of Complex Systems, Advances in Intelligent Systems and Computing* Volume 210, Springer International Publishing, pp. 221–231.

[11] Uhrin, D., Partila, P., Voznak, M., Chmelikova, Z., Hlozak, M., Orcik, L. 2014. Design and implementation of Czech database of speech emotions. *22nd Telecommunications Forum, TELFOR 2014 - Proceedings of Papers*, art. no. 7034463, pp. 529–532.

IntechOpen