Proceedings of the International Multiconference on Computer Science and Information Technology pp. 141–147
ISBN 978-83-60810-27-9
ISSN 1896-7094

Emotional Speech Analysis using Artificial Neural Networks

Jana Tuckova
Czech Technical University in Prague
Faculty of Electrical Engineering
Technicka 2, 166 27 Prague 6, Czech Republic
E-mail: tuckova@fel.cvut.cz

Martin Sramka
Czech Technical University in Prague
Faculty of Electrical Engineering
Technicka 2, 166 27 Prague 6, Czech Republic
E-mail: sramkma2@fel.cvut.cz

***Abstract*—In the present text, we deal with the problem of classification of speech emotion. Problems of speech processing are addressed through the use of artificial neural networks (ANN). The results can be use for two research projects - for prosody modelling and for analysis of disordered speech. The first ANN topology discussed is the multilayer neural network (MLNN) with the BPG learning algorithm, while the supervised SOM (SSOM) are the second ANN topology. Our aim is to verify the various of knowledge from phonetics and ANN but also to try to classify speech signals which are described by musical theory. Finally, one solution is given for this problem which is supplemented with a proof.**

I. Introduction

MANY problems in technology, medicine, and the natural and social sciences still remain unsolved: the complexity of solutions, the importance of time, and the considerable quantity of data required for processing form the real cause of the situation. Seeking help through new information technology is highly appropriate; and one such method is through the development of artificial neural networks (ANN). Success in the application of ANN depends on thorough knowledge of their function, which cuts across a wide range of academic disciplines – mathematics, numerous technical fields, physiology, medicine, phonetics, phonology, linguistics and social sciences. Initially, the ANN paradigm was regarded as a cure-all for many problems, yet simultaneously was often disparaged by its detractors for its inability to solve increasingly high requirements through the use of simple principles. The robustness of the solution for real methods by ANN is a great advantage, for example, in the area of noise signal processing. In this case, ANN should be a highly useful source of help, and the results thus acquired could be of a higher quality than those found with standard methods. The research goal described in this contribution was to verify an ability to classify optional speech through the use of ANN. We use three approaches for comparison of results. First, a frequency dependence and statistical parameters are created from input data, while the second approach is based on music theory (see [8]) and the final approach is a combination of both cases.

This research was supported by grant GACR No. 102/09/0989.

The contribution has two parts. A brief notice about some publications from international researchers which concerns emotional speech, basic information about emotions, and the specific ANN applied to the experiments create the first part of the text. The second part is dedicated to the results of the experiments themselves.

A. Classification of emotions in publications of international researchers

Much research around the world is engaged in the processing of emotional speech. Specific projects differ in the number and type of classified emotions, acoustic characteristics, the type of classifiers, and precision. Classification of three emotions (sadness, anger and neutral state) for human-computer communication are described in [13]. Fundamental frequency F_0, voice intensity and cepstral coefficients were the input characteristics. Classification success was $64\,\%$ with the classification of the five classes (for anger, pleasure, sadness, surprose, and neutral state) described in [14] and [15]. Data from Danish Emotional Speech were tested by the Bayes classifier and classification success was $54\,\%$ (for both gender), $61.1\,\%$ for males and $57.1\,\%$ for females respectively. Also the five emotions are classified in [12] (fear, pleasure, sadness, anger, and neutrale state). A Gausse SVM (Support Vector Machine) algorithm was applied with a $55\,\%$ success rate. A comparison of the SVM, RBF (Radial Basis Function), kNN (k-Nearest Neighbours), Naive Bayes and MLNN (two hidden layers with 15 neurons) is described in [9]. The success for the five classes was $81\,\%$. A description of the five emotional states (pleasure, sadness, fear, anger, and neutral state) is undertaken in [11]. An algorithm is based on relationship of a height note versus the 12 half tones of the melodic scale. The latest publication is closest to our methods described in this contribution.

II. Solving the problems and an applied methods

Automatic speech synthesis is an interdisciplinary part of artificial intelligence, drawing upon knowledge from acoustics, phonetics, phonology, linguistics, physiology, psychology, signal processing and informatics for a successful solution. Many research teams around the world are engaged in the modelling of the prosody of synthetic speech. This problem

must be solved with relation to the specific attributes of different languages: e.g. [4] for English, [5] for German, [1] for French, [3] for Japanese for example. A majority of prosody control systems are based on the implementation of grammatical rules e.g. realised by decision trees, but some researchers (Sejnowski, Traber), including the authors of this study, use neural networks for prosody modelling. Different input parameters with a significant impact on speech prosody have to be used for neural network training in different languages. As a result, it is very difficult, indeed nearly impossible, to compare the results of prosody controllers for different languages. The most complex evaluation is the listening test, but it is very subjective and cannot be described by an objective metric. A reason for this difficulty is that prosody is deeply affected by the speaker's individual physiologies and mental states, as well as by the uttered speech segments and the universal phonetic properties. The influence of the phonological and phonetic properties of the Czech language, the influence of the quality, size of the speech database, and the influence of the synthesizer type all need to be explored. Furthermore, it is not possible to make complete use of all the information extracted from natural speech signals in automatic input data creation. Our research has taken as its central focus the question of prosody modelling for Text-to-Speech (TTS) Synthesis. A text and its speech signal will be used for the training process of ANN, and only the text and the trained ANN will be used for prosody modelling, allowing it to be as natural as possible. Previously, processing speech had a neutral character, yet in recent months research has concentrated its attention on emotional speech.

A. Basic information about speech emotions

Emotion is a mental state of a living organism accompanied by motive and glandular activities. Emotions are classified according to their psychological aspect. As a result, the term "emotion" represents physiologic disturbance, shock or attack. The second category – attitude – represents a behaviour and a chronic state. Feigned and active emotions have different manners of their division. A physiological reaction (change of cardiac rate and blood pressure, whiteness or redness) is linked to opposite emotions (anger, fear, pleasure, and sadness). Hence it is impossible use this physiological reaction as ANN input features separately. However, it is possible to use prosody characteristics, such as timbre, intensity and rhythm. These is a change of a fundamental frequency f_0, range of fundamental frequency, change of a formant location etc.

The melody, i.e. change of a haight of voice in a sentence, is very important from the point of view of a communication. Expressive changes of melody are important indicators for an emotional and voluntary attitude of a speaker (more in [17]).

B. Classification of emotions by ANN

ANN was used for classification of emotions. A multilayer neural network with one hidden layer was one of the methods applied for the classification of speech emotions. The number of neurons in the input layer is given by the key linguistic parameters which are needed for characterization of the Czech language. The ANN outputs are the various classes of emotions. The target values of prosodic parameters were extracted from the natural speech signal. Many learning algorithms for feed-forward neural networks are based on the gradient descent algorithm. Usually, they have a poor convergence rate and depend on input parameters which characterize specific problems. No theoretical basis for choosing optimal parameters for ANN training exists, but the values of these parameters are often crucial for the success of the training. Therefore we decided to use a Scaled Conjugate Gradient (SCG) algorithm with superlinear convergence rate. SCG belongs to the class of Conjugate Gradient Methods, which shows superlinear convergence for most problems; further description is offered in [19].

Kohonen's Self-Organizing Features Map (KSOM) was the second ANN which was applied for the solution of emotion classification. KSOM is a form of ANN that is trained by unsupervised learning rules, i.e. without target (required) values. It is an iterative process based on the clustering method; cluster analysis methods searching for interdependences and joint properties in a set of submitted patterns. A new SOM variant has been used for emotion classification, namely the supervised Self-Organizing Map (SSOM), which combines aspects of the vector quantization method with the topology-preserving ordering of the quantization vectors. The algorithm of the SSOM represents a very effective method of classification, but only for well-known input data or for well-known classes of input data.

C. Corpus creation

For testing and refining the ANN, it is necessary to create a speech corpus of sentences and, through pre-processing of the corpus, to prepare input data for the network's training and testing. In general, corpuses of natural speech have been created through careful choice from among a wide variety of different neutral sentences. Currently, no emotional speech database is available. As a result, an emotional speech corpus and database for ANN training had to be created for our research. The sentences was read by professional actors, two female and one male. Speech recording was materialized in a recording studio with a professional equipment (format "wav," sampling frequence $44.1kHz$, $24bit$).

The speech corpus is composed of a written text and its corresponding speech signal, both of which will be used for the training of ANN. The compound corpus was divided into two parts, the first set used for training and the second part serving as a testing set, also used for the monitoring of the training process.

Utterances were realised for four types of emotions: anger, boredom, pleasure and sadness – see Table I and Table II.

D. Input data creation

The success of prosody control is clearly dependent on the labelling of the natural speech signal in the database. The labelling (determination of boundaries between speech

TABLE I
DATABASE OF UTTERANCES – IN CZECH

Words (in Czech)	Words – translation
Jé.	*Whoah.*
Má?	*Got it?*
Nevím.	*I don't know.*
Vidíš?	*See you?*
Povídej!	*Tell me!*
Poezie.	*Poetry.*

TABLE II
DATABASE OF UTTERANCES – TRANSLATION INTO ENGLISH

Sentences (in Czech)	Sentences – translation
To mi nevadí.	*I don't mind.*
Neumím to vysvětlit.	*I don't know to explain this.*
To bude světový rekord.	*It will be a world record.*
Jak se ti to líbí?	*How do you like it?*
Podívej se na nebe!	*Look up at the heavens!*
Až přijdeš uvidíš.	*When you come, you'll see.*

units) and phonetic transcription of sentences from the speech corpus is done in the phase of pre-processing. The changes of fundamental frequency F_0, formant frequency F_i, $i = 1, \ldots, 4$ and duration Du of phonemes during the voicing of sentences create the melody of the sentence (its intonation). Intonation is also related to the meaning of the sentence and its emotional timbre.

Recorded emotion speech was subjectively evaluated by four persons. The final database contained 720 patterns (360 patterns for one-word sentences and 360 patterns for multi-word sentences).

III. EXPERIMENTS

All analyses and experiments described in this contribution were performed through use of the computational system MATLAB with NN-toolbox [16] and SOM Toolbox. SOM Toolbox was developed in the Laboratory of Information and Computer Science (CIS) in the Helsinki University of Technology and it is built using the MATLAB script language. The SOM Toolbox contains functions for creation, visualization and analysis of the Self-Organizing Maps. The Toolbox is available free of charge under the General Public License from ([7]). For the projects from the domain of the speech processing by ANN (which are being addressed by our university's department of Circuits Theory), new special M-files, which should be a part of the supporting program package, were created.

MLNN and SOM were applied particularly to the utterances from Table I and Table II. The results from MLNN training are concentrated into the so-called matrix of replacement, where "class 1" is specified as anger, "class 2" is specified as boredom, "class 3" is specified as pleasure and "class 4" is specified as sadness. The database for ANN training obtained 216 patterns, for validation 72 patterns and for testing as many as 72 patterns.

TABLE III
INPUT PARAMETERS – TIME AND FREQUENCY DOMAIN

Time domain
Arithmetic average of absolute value
Standard deviation
Maximum
Minimum
Frequency domain
Fundamental frequency F_0
Formant frequency $F_1, \ldots F_4$

The unified distance matrix or U-matrix is a representation of the KSOM that visualizes the distance between the neurons and their neighbors. The KSOM neurons are represented by hexagonal cells (in our experiment). The distance between the adjacent neurons is calculated and presented in different colors. Darker colors between neurons correspond to a larger distance and thus represent a difference between the values in the input space. Light colors between the neurons mean that the vectors are close to each other in the input space. Light areas represent clusters (classes) and dark areas represent cluster boundaries (more in [2]). The size of the map was 15x15, while quantization (QE) and topographic (TE) errors of the map were also computed.

A. Method I: The patterns based on time and frequency characteristics

Nine patterns for MLNN training are created through the characteristics of the time and frequency domains (see Table III). The hidden layer was 20 neurons, while the output layer was 4 neurons. The number of training epoch was 56 resp. 53 for one-word sentences resp. multiword sentences.

B. Method II: The patterns based on musical theory.

The second presented method is based on the idea of the musical interval: the frequency difference between a specific n-tone and reference tone. E.g. quint is ratio of the fifth tone divided by the first tone, with a numerical value of 1.498. The ratios of the musical intervals are shown in Table IV.

TABLE IV
FREQUENCY RATIOS OF THE MUSICAL INTERVALS

Interval	Variant	Frequency ratios
first		1,000
second	minor	1,059
	major	1,122
third	minor	1,189
	major	1,260
fourth		1,335
fifth		1,498
sixth	minor	1,587
	major	1,682
seventh	minor	1,782
	major	1,888
octave		2,000

TABLE V
COMPARISON OF THE RESULTS

Method	one-word sentences		multiword sentences	
	MLNN [%]	SOM [QE/TE]	MLNN [%]	SOM [QE/TE]
I	88.7	0.185 / 0.02	77.8	0.184 / 0.017
II	70.4	0.274 / 0.014	65.3	0.275 / 0.017
III	85.9	0.431 / 0.011	84.7	0.439 / 0.006

This method of the parametrization of the utterances consists in the description of the signal patterns based on the musical intervals or more precisely on their frequency ratios. The reference frequency, i.e. the fundamental frequency in our case, is given by the choices in each utterance feature. We use autocorrelation function. The frequency ratios are compared with the music intervals and input vector for MLNN training with 20 values is computed. The hidden layer was 35 neurons, the output layer was 4 neurons. The number of the training epoch was 75 for one-word sentences resp. 84 for multiword sentences.

C. Method III: Combination of both previous approaches

The third method was the combination of both previous methods. 29 patterns for MLNN training are created by 20 values containing the ratios respective to the music intervals and 9 values describing the acoustic qualities of the utterance feature. The hidden layer was 55 neurons, while the output layer was 4 neurons. The number of the training epoch was 57 for one-word sentences resp. 106 for multiword sentences.

IV. RESULTS OF EXPERIMENTS

The results of experiments are shown in the following table and figures.

This table summarizes the success rate of emotional classification for all three described methods. For the MLNN approach the first method (based on acoustic parameters) was best for the one-word utteraces, but difference between one-word and multiword utterances are the greatest (9.9 %). Success for the second method (based on music theory) is worse for both type of utterances, but the difference between them is smaller (5.1 %). The third method (combination) is the best of them for the multiword utteraces, while additionally the differences between one-word and multiword utterances are absolutely smallest (1.2 %). With the SOM approach, the determining of SOM quality is complicated. We were monitoring the quality of learning by topographic and quantization errors for the comparison of methods. The topographic error (TE) predicates the conservation of data topology between input and output space. The quantization error (QE) reflects the accuracy of the mapping (it relates to the number of the input matrix elements and the size of the map).

The success of the SOM training decreases when the number of map units is larger than the number of training samples, which may be the main problem in our SOM approach. In our experimets we made use of the uniform size of maps

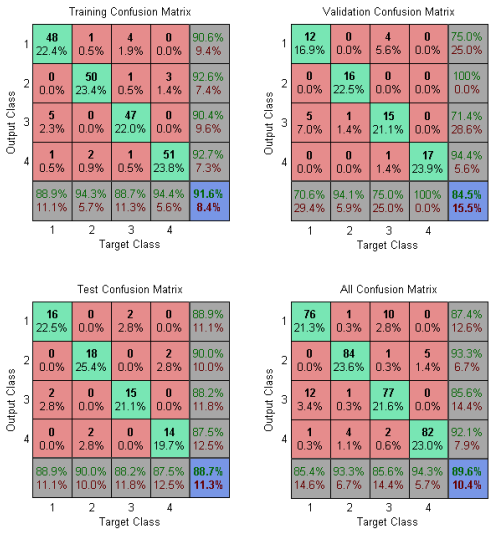

Fig. 1. Method I: Matrix of changes for one-word sentences – time and frequency domain parameters

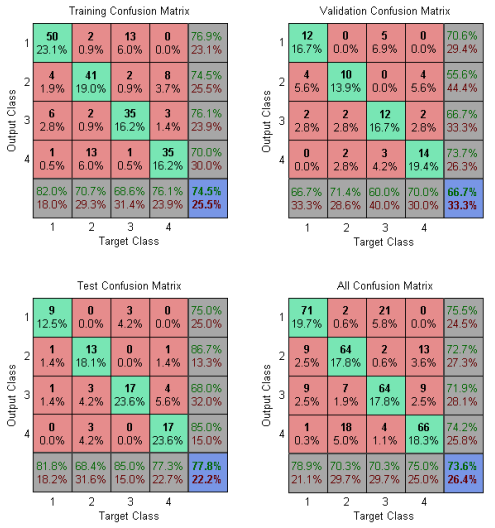

Fig. 2. Method I: Matrix of changes for multiword sentences – time and frequency domain parameters

for the comparison of all three methods. Our results show progressive values of the quantization errors in dependence to number of training data features, whereas a decreasing value for topographic error shows a very good ability of the classification.

We can see the matrix of changes for the MLNN classifier in Figure 1, 2. These figures summarize the first described method. We can observe the replacement of the emotion classification between active emotions, i.e. pleasure – anger, and between passive emotions, i.e. sadness – tedium. The results from the second method are shown in Figure 3, 4, the results from the third method are demonstrated in Figure 5, 6. Just as in the Method I, the worst score is for passive emotions (Method II and Method III).

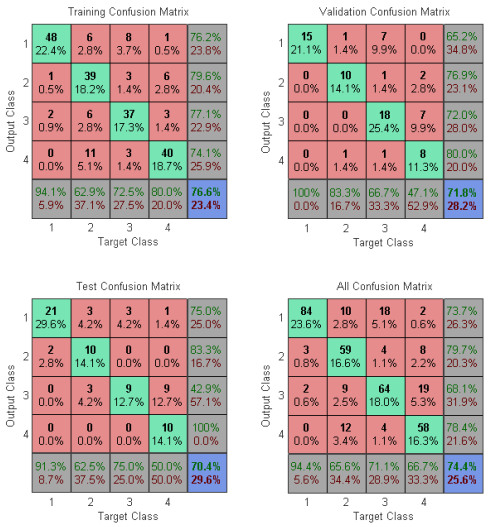

Fig. 3. Method II: Matrix of changes for one-word sentences – time and frequency domain parameters

Fig. 4. Method II: Matrix of changes for multiword sentences – time and frequency domain parameters

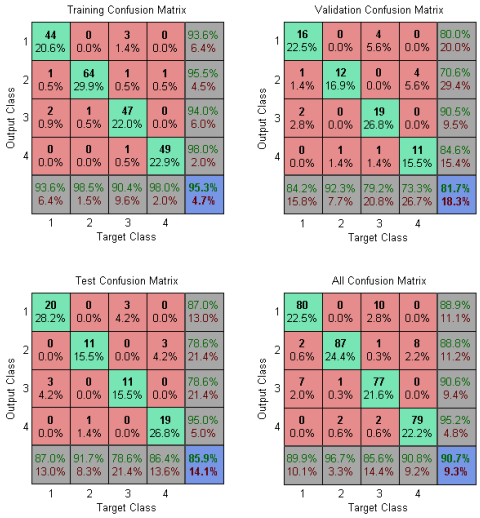

Fig. 5. Method III: Matrix of changes for one-word sentences – time and frequency domain parameters

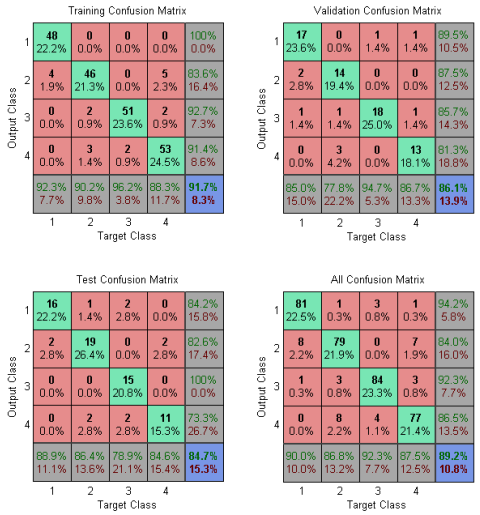

Fig. 6. Method III: Matrix of changes for multiword sentences – time and frequency domain parameters

The unified distance matrix or U-matrix is a representation of the KSOM that visualizes the distance between the neurons and their neighbors. The KSOM neurons are represented by hexagonal cells (in our experiment) marked by 'H' for anger, 'N' for tedium, 'R' for pleasure and 'S' for sadness. Each cell is marked also by a character for class, by real classified font and number registered pattterns.

The distance between the adjacent neurons is calculated and presented with different colors. Dark colors between neurons correspond to a larger distance and thus represent a difference between the values in the input space. Light colors between the neurons means that the vectors are close to each other in the input space. Light areas represent clusters and dark areas represent cluster boundaries.

The U-matrix represents emotion classes based on the parametrization by Method I are visualize in Figure 7, resp. 8, by Method II in Figure 9, resp. 10 and by Method III in Figure 11 and 12. The matrix is divided into four parts respective particular emotions. The topographic error for both type of utterances (one-word and multiwords) is lowest from all three methods (see on the Table V). This result documents the availability to apply the method based on the combination standard and music theory.

V. Conclusion and Future Work

We have established differentiation between the mathematical results and the listening tests. It is necessary to judge recording of emotional speech which is determined for database creation and the resulting synthetic sentences

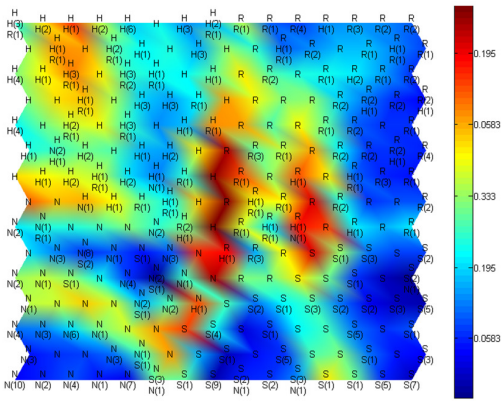

Fig. 7. Method I: U-matrix for one-word sentences

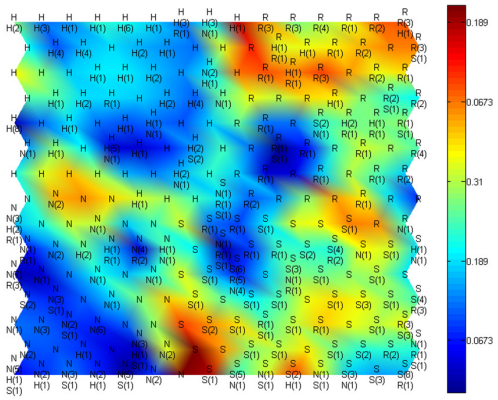

Fig. 10. Method II: U-matrix for multi-word sentences

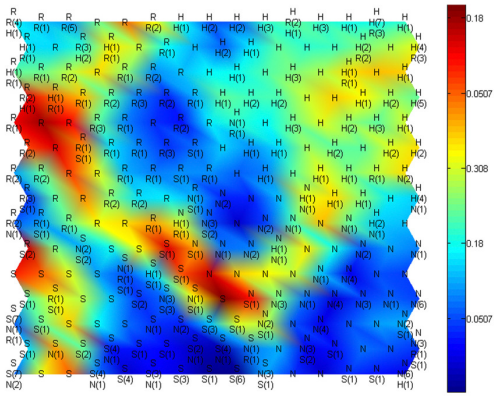

Fig. 8. Method I: U-matrix for multi-word sentences

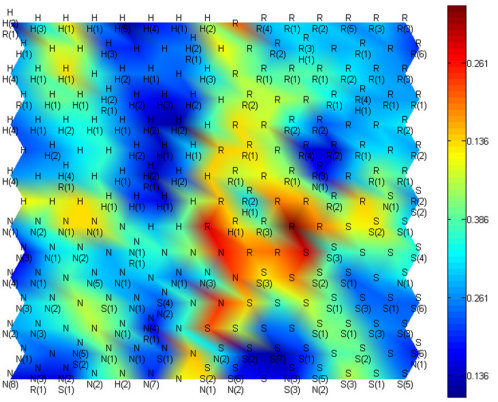

Fig. 11. Method III: U-matrix for one-word sentences

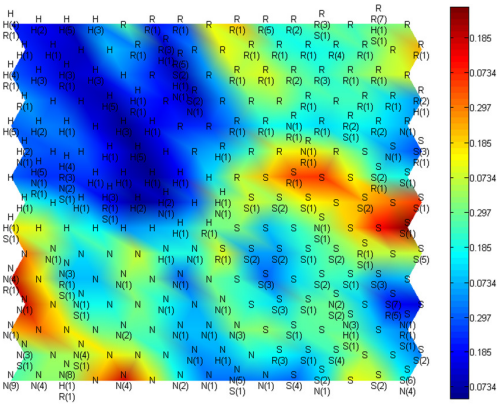

Fig. 9. Method II: U-matrix for one-word sentences

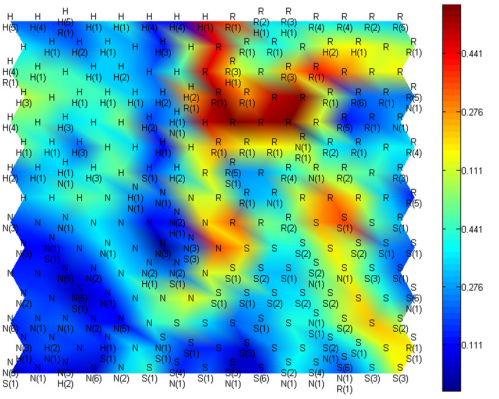

Fig. 12. Method III: U-matrix for multi-word sentences

by listening. The physical properties of the acoustic wave, which are perceived as sounds, are transformed several times: first in the organ of hearing, later at the emergence of neural excitement, and last in the cerebral analysis. Therefore, the sounds perceived in listening tests do not correlate completely with the objective properties of the acoustic patterns. The listening tests must consequently form a part of experiments. We want to conclude by establishing when to use MLNN or SSOM on the base of acquired events and on the base of augmentation, of the number of experiments.

Our effort in future work will focus on our lack of knowledge regarding the possibilities of ANN application in prosody modelling and children's disordered speech analysis. These different domain of the application influence the database creation. Long sentences (called multiword sentences in this contribution) are more acceptable for prosody modelling, yet

the database created by one-word sentences is suitable for the analysis of children's disordered speech (often a speech malfunction is manifested in an inability to pronounce whole sentences). We are going to apply results from the described experiments with emotional speech to the improvement of synthetic speech naturalness, but also to the domain of neurodevelopmental disturbances (above all, developmental dysphasia).

REFERENCES

[1] E. Keller, S. Werner " Automatic Intonation Extraction and Generation for French." 14th CALICO Annual Symposium. ISBN 1-890127-01-9, West Point. NY, 1997.

[2] T. Kohonen *Self-Organizing Maps.* Ed.:Huang, T. S., Kohonen, T. ,Schroeder, M. R., 3rd ed.Springer-Verlag Berlin, 2001, ISBN 3-540-67921-9.

[3] Z. Sagisaka, T. Yamashita and Y. Kokenawa " Generation and perception of F0 markedness for communicative speech synthesis." *Speech Communication*, 2005, Vol. 46, Issues 3–4, pp. 376–384.

[4] T. J. Sejnowski, C. R. Rosenberg " NETtalk: A parallel network that learns to read aloud". *Technical Report JHU/EECS-86/01, The Johns Hopkins University Technical Report*, 1986.

[5] C. Traber " F0 generation with a database of natural F0 patterns and with a neural network." G.Bailly,C.Benoit, and T.R. Sawallis, ed., *Talking Machines: Theories, Models, and Design*, pp. 287–304. Elsevier Science Publishers,1992.

[6] J. Tuckova, V. Sebesta "The Prosody Optimisation of the Czech Language Synthesizer." *Int. Journal on Neural and Mass-Parallel Computing and Information Systems "Neural Network World"*, Ed. M. Novak, ICS AS CR and CTU, FTS, vol. 4, 2008, pp. 291–308 . ISSN 1210-0552.

[7] J. Vesanto, J. Himberg, E. Alhoniemi and J. Parhankangas "SOM Toolbox for Matlab 5," SOM Toolbox Team, Helsinki Univesity of Technology, Finland, 2000, ISBN 951-22-4951-0. Homepage of SOM Toolbox: www.cis.hut.fi/projects/somtoolbox

[8] M. Cerny, *Influence of the speech signal parametrization to prosody modelling. (in Czech)*, Diploma work, Prague, CTU FEE 2009.

[9] Z. Xiao, E. Dellandrea, WW. Dou, and L. Chen "Multi-stage classification of emotional speech motivated by a dimensional emotion model,"*Multimedia Tools and Applications journal*, Springer Netherlands, vol. 46, Nu 1/January, pp. 119–145, ISSN 1380-7501.

[10] M. Shami, W. Verhelst, "Automatic Classification of Expressiveness in Speech: A Multi-corpus Study". *In Speaker Classification Ii: Selected Projects, C. Müller*, Ed. Lecture Notes In Artificial Intelligence, vol. 4441. Springer–Verlag, Berlin, Heidelberg, 2007, pp. 43–56.

[11] A. M. Mahmoud, W. H. Hassan, "Determinism in speech pitch relation to emotion". *Proceedings of the 2nd international Conference on interaction Sciences: information Technology, Culture and Human* Seoul, Korea, November 24–26, 2009, vol. 403, ACM, New York, NY, pp. 32–37.

[12] S. McGilloway, R. Cowie, Ed. Cowie, S. Gielen, M. Westerdijk, S. Stroeve, "Approaching automatic recognition of emotion from voice: a rough benchmark," *Proceedings of the ISCA workshop on Speech and Emotion*, pp. 207–212, Newcastle, Northern Ireland, 2000.

[13] T. Polzin, A. Waibel, "Emotion-sensitive human-computer interfaces," *Proc.of the ISCA workshop on Speech and Emotion*, pp. 201–206, Newcastle, Northern Ireland, 2000.

[14] D. Ververidis, C. Koltopoulos, "Automatic speech classification to five emotional states based on gender information," *Proc. of 12th European Signal Processing Conference*, pp. 341–344, Austria, 2004.

[15] D. Ververidis, C. Koltopoulos, I. Pitas, "Automatic emotional speech classification". *Proc. of ICASSP 2004*, pp. 593–596, Montreal, Canada, 2004.

[16] *MATLAB Help* version 2009a Natick, Massachusetts: The MathWorks Inc., 2009.

[17] M. Krcmova, *Phonetics and phonology* in Czech *Fonetika a fonologie* [online]. Brno : Masarykova univerzita, 2008 [cit. 2010-04-04]. http://is.muni.cz/elportal/?id=766384. ISSN 1802-128X.

[18] M. F. Møller, "A scaled conjugate gradient algorithm for fast supervised learning."*NEURAL NETWORKS*, vol. 6, nu 4, pp. 525–533, 1993.

[19] http://www.mathworks.com/access/helpdesk_r13/help/toolbox/nnet/trainscg.html