České vysoké učení technické v Praze
Fakulta elektrotechnická

Czech Technical University in Prague
Faculty of Electrical Engineering

Doc. Ing. Jana Tučková, CSc.

Aplikace umělých neuronových sítí při zpracování řeči

Artificial Neural Network Applications in Speech Processing

Summary

In the present study, I deal with two principal thematic subjects. The first concentrates from the outset on the modelling of prosody parameters for a synthetic form of Czech, while the second is oriented towards the analysis of disordered children's speech with the diagnosis of developmental dysphasia (DLD). In both cases, problems of speech processing are solved by artificial neural networks. The results ensuing from both research projects were discussed in 94 contributions to impacted or peer-reviewed journals or in conference presentations (57 of them being international journals or conferences). The main idea and results have become part of a monograph „Selected applications of the artificial neural networks in signal processing“ [18] (in Czech).

The book is an introduction to the theory and application of the selected paradigm, the most widespread paradigm of artificial neural networks (ANN). Basic information about theory and specific applications forms the main content of the book. The first ANN group discussed is the multilayer neural network (MLNN) with BPG learning algorithm; basic, fast and optimised learning algorithms, the analysis of parameter choice and ANN real applications are described. Kohonen Self-Organising Maps, Supervised SOM and the LVQ form are the second ANN group. Applications focused on signal processing, speech analysis and TTS synthesis form the major part of this book. The book should be a motivation for seeking new paths to the solution of the task, and for facilitating orientation in the selection of the most suitable architecture and learning algorithm.

The research has been carried out in cooperation with several research institutes. In the past, these have been two bodies of the Academy of Sciences of the Czech Republic – the Institute of Computer Science and the Institute of Radio Engineering and Electronics (now Institute of Photonics and Electronics), and the Institute of Phonetics of the Faculty of Arts, Charles University in Prague. At present, the important collaborators are the Department of Cybernetics of the Faculty of Applied Sciences (FAV) at the University of West Bohemia (UWB) in Pilsen, Czech Republic and Department of Child Neurology of the Motol University Hospital in Prague.

Souhrn

V této studii se zabývám dvěma hlavními tématickými okruhy. První z nich je zaměřen na modelování prozodických parametrů pro syntézu češtiny, druhý pak na analýzu narušené řeči dětí, kterým byla diagnostikována vývojová dysfázie. V obou případech se úlohy zpracování řečového signálu řeší pomocí umělých neuronových sítí. Shrnují výsledky výzkumu. Tyto výsledky plynoucí z řešení projektů z obou oblastí byly zpracovány v 94 publikacích v časopisech a na konferencích, z toho v 57 zahraničních. Rozhodující části se staly součástí monografie „Vybrané aplikace umělých neuronových sítí při zpracování signálů“ [18].

Kniha je úvodem do teorie a aplikací vybraných nejrozšířenějších paradigmat UNS. Podává základní informace o teorii dvou typů UNS a seznamuje s jejich některými aplikacemi. Je to vícevrstvá neuronová síť s učením BPG - základní, rychlé a optimalizované učení, rozbor volby parametrů a algoritmů učení neuronové sítě při řešení reálných úloh. Druhým typem jsou varianty Kohonenovy samoorganizace, SOM s učitelem a LVQ. Aplikace jsou zaměřeny na zpracování signálů, filtraci šumu, rozpoznání a syntézu řeči z textu. Kniha má být motivací k hledání nových způsobů řešení úloh a usnadnit orientaci při výběru vhodné architektury a algoritmu učení. O tom pojednává i tato studie.

Výzkum probíhá ve spolupráci s několika dalšími výzkumnými pracovišti. V minulých letech to byly především pracoviště Akademie věd České republiky, a to Ústav informatiky (ÚI AV ČR) a Ústav radiotechniky a elektroniky (dnesní Ústav fotoniky)(ÚRE AV ČR), a Fonetický ústav Filosofické fakulty Univerzity Karlovy (FF UK) v Praze. Dnes jsou to především pracoviště Katedry kybernetiky (KKY) Fakulty aplikovaných věd (FAV) na Západočeské univerzitě v Plzni (ZČU) a Kliniky dětské neurologie UK, 2.LF.

Klíčová slova

Neuronové sítě, analýza řeči, modelování prozodie, metody optimalizace neuronové sítě, klestění, SOM, SOM s učitelem, narušená dětská řeč, vývojová dysfázie.

Keywords

Neural networks, speech analysis, prosody modelling, neural net optimisation method, pruning, SOM, supervised SOM, disordered children's speech, developmental dysphasia.

Contents

1 Introduction

The nervous system and the brain are ranked among the most crucial components of a living organism, with particularly great influence on the quality of human life. Their activity has great influence on quality of human live. For this reason, the neural networks of biology have become the inspiration for computer modelling of their features and modelling of their function. Many problems in technology, medicine, and the natural and social sciences still remain unsolved: it is the complexity of solutions, the importance of time, and the considerable quantity of data required for processing that 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 increasing requirements through simple principles.

Today, it is possible to obtain many professional software products with a focus on ANN, though it still remains necessary to known when, where and how it should be applied for productivity and achievement of results that are comparable to or better than those achieved by conventional methods. At the present time, ANN applicattions have proven successful, but we must gain further knowledge of its advantages and disadvantages. We must establish its best conditions for use, and understand the nexus of mutual relations between the individual elements of the investigated system for correct interpretation of the results.

It is perhaps now evident that ANN applications are significant in research and application when conventional methods malfunction or prove to be too complicated. 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 will be a highly useful source of help, and the results thus acquired could be of a higher quality than those found with standard methods. Still, it is necessary to bear in mind that conventional methods do allow for better results in many areas of research, particularly wherever numerical accuracy is important.

Presented in the following text are two projects in which ANN proved to be one of the best solution. The first is the modelling of two basic prosodic parameters for synthesis of the Czech language; the second is analysis of children's disordered speech. In each project, a solution was attempted using a different type of neural net. MLNN with the fast back-propagation learning algorithm with a moment, an adaptive learning rate and feed-forward recall was use for prosody modelling. Kohonen's self-organizing maps (KSOM) were used for analyais of children's speech. In particular, the text of this contribution was derived from three publications:

- [16] *Tučková, J., Šebesta, V.: The Prosody Optimisation of the Czech Language Synthesizer. In: Neural Network World, vol.18, No.4, 2008, pp.291-308. ISSN 1210-0552.*
- [17] *Tučková, J., Komárek, V.: Effectiveness of Speech Analysis by Self-Organizing Maps in Children with Developmental Language Disorders. In: Neuroendocrinology Letters, vol. 29, No. 6, Nov/Dec 2008, ISSN 0172-780X*
- [18] *Tučková, J.: Selected applications of the artificial neural networks at the signal processing. (Book in the Czech – Vybrané aplikace umělých neuronových sítí při zpracování signálů). Nakladatelství ČVUT, Praha, 2009, ISBN 978-80-01-04229-8.*

2 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, more and more favour is being paid to emotional speech, a matter that our research team hopes to address in the future.

The speech corpus is composed from a written text and its corresponding speech signal, both which parts 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 served as a testing set, also used for the monitoring of the training process. Presently, we have at our disposal three corpuses of male and female speech and two corpuses containing children's speech. All corpuses are invariably complemented. Some experiments in speech signal pre-processing focused on the search for suitable methods of parametrization, which are the principles for neural net input data creation.

The first small corpus of speech data for this research has been created through the careful choice of available sentences by experts in phonetics. It consists of only 25 sentences (partly indicative, interrogative and warning). All types of sentences must be included into the corpus as uniformly as possible. The text of the sentences was read by one professional male TV announcer. 18 sentences (78 words) with 465 input vectors after transcription make up the training set and another 7 sentences (19 words) with 136 input vectors have been used as a testing set. The first experiments were performed on the Czech-language synthesizer, designed and built in the IREE AS CR.

The larger corpus is composed from a special professional domain - from the Czech radio station Radiojournal News (weather forecast). The sentences are read by a professional speaker, one female and one male. This extensive speech corpus has also been used for the synthesizer at WBU in Pilsen [7]. Here, the corpus is also divided into the training and test sets. The training set contains 103 indicative sentences (978 words) and 6212 input vectors (after transcription). The test set comprises 9 indicative sentences (127 words) and 861 input vectors (after transcription).

Currently, no childrens voice database is available. Our team has created a speech database of children with developmental dysphasia and a comparative database of healthy children. First, the speech corpus of 72 healthy children (44 female and 28 male, age from 4 to 10 years) is composed from speech recorded in kindergartens and the first level of elementary school. The second part of children's speech is continuously being expanded at the hospital. In our first experiments, we examined 35 children who were referred for suspected developmental dysphasia (DLD) to the hospital child neurology centre, community neurologists and speech therapists. Of them, 28 (aged from 3 years 5 months to 9 years 1 month) fulfilled the diagnostic criteria of DLD. All children underwent an overnight sleep video-EEG and logopedic/psychological evaluation.

The utterance texts are compiled in a paediatric neurological clinic by neurological specialists in the course of medical treatment. The same text is used by the healthy children for comparative analysis. The text (phonemes, syllables, words and sentence) is read aloud by an assistant (for healthy children) or by a psychologist (for patients) and the children repeat the text. Identical conditions for all age categories must be adhered to. The speech of each child is recorded as a wav-file and is subsequently segmented.

Voice recording was performed in real settings with high noise level. The noise reduction necessary for conventional methods could cause irreversible information loss, though this problem is minor when artificial neural networks are applied. The second problem deals with the speech evolution of children. The speech quality was strongly influenced by the childrens emotional tone: the children were afraid or were shy. All these characteristics and problems associated with children's speech can be overcome through the use of artificial neural networks. Methods based on the ANN are robust enough for the minimization of these effects.

3 Neural network applications

It is necessary to bear in mind that dozens, perhaps even hundreds of real ANN applications, exist. It is rare to find an industry in which no experiments, at the very least, had beenperformed with ANN applications. In many instances, ANN are used as interfaces between real situations with real data and machine or computer. Problems with a sulution of this situations, such as influence of noise, different light conditions, ageing and failure of devices, are usually addressed by standard methods. Yet in spite of a high level of sophistication in terms of working time, computing and technological methods, situations, often arise in which the real environment cannot be eliminated. In such a situation, it is most suitable to apply an ANN method, as in this case. Most of current ANN applications are based on MLNN and on supervised learning. Increasingly, attention is being devoted to self-organizing maps, and the new variants of this approach, which augment number-specific properties and eliminate deficiencies, have caused great advances for ANN in many branches. For our purposes, we can group application tasks into two major categories: classification (including diagnostics) and prediction. Both categories will be demonstrated in the following text.

3.1 Prosody modelling by ANN

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 dependence on the specific attributes of different languages: e.g. [12] for English, [13] for German, [2] for French, [8] 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, Riedi), including the author of this study, use the 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 speakers individual physiologies and mental states, on 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 signal in automatic input data creation: for example, the phenomenon known as prominence, which demonstrates the different weights of stress of any sentence [13], cannot be differentiated. 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 trained ANN will be used for the prosody modelling, allowing it to be as natural as possible.

TTS system is a very complicated complex of problems starting with the modelling of the human speech organs, continuing with the formulation of transcription and prosody algorithms and ending with the implementation of the resulting system functioning in real time. TTS enables the full synthesis of an arbitrary text, irrespective of the length of the text and the diversity of the themes. Naturalness of speech is considerably dependent on the implementation of prosodic features. Additionally, the investigation of the manner of articulation of a speech sound sequence in abstract linguistic terms of intonation (melody), rhythm, and loudness has its counterpart phonetically manifested in acoustic properties. In short, there are phonetic aspects of prosody in addition to the phonological ones most commonly linked to acoustics, two highly important properties being the fundamental frequency (F_0) and the duration (Du) of the speech sound.

In conventional TTS techniques, control over prosody is granted by rules. As is common knowledge, every national language can be described in terms of specific grammatical rules. Nonetheless, rule-based knowledge representations alone cannot be used for the natural flow of speech. For one, the generation of rules is very tedious; for another, the result is deterministic and therefore more or less unnatural. One other option is the use of applied ANN for prosody modelling. Prosody is very important for any kind of synthetic speech. The prosodic parameters depend on the speaker's physiology and his mental state, on the uttered speech segments and on the universal phonetic properties. Improper prosody is namely one of the differences between natural and synthetic speech. Since, our effort is to minimize these differences, it is necessary to choose optimal phonetic and phonologic parameters to influence the naturalness of speech.

With regard to the exact parameters, however, phonetic experts are far from adhering to s single opinion. After many discussions and experiments, we chose 10 basic parameters for a description of the focus phoneme (see Tab.1). The same parameters were used for a coarticulation (see subsection 3.1.3) in position „Left“ and „Right“ in Tab.1. If the prosody of the synthesizer is controlled by ANN, an optimisation of the ANN topology is necessary.

A multi-layer neural network with one hidden layer was applied for the determination of prosody parameters F_0 and Du. The number of neurons in the input layer is given by the important language parameters which are needed for characterization of the Czech language (see Tab.1). The ANN outputs are fundamental frequency F_0

PROPERTIES	POSITION		
	Left	Focus	Right
Silent pause ident.	P1	P11	P21
Stress unit ident.	P2	P12	P22
Syllable nucleus ident.	P3	P13	P23
Punctuation mark ident.	P4	P14	P24
Phoneme identification	P5	P15	P25
Height of vowel	P6	P16	P26
Length of vowel	P7	P17	P27
Voice of consonant	P8	P18	P28
Creation mode / consonant	P9	P19	P29
# of phonemes /word	P10	P20	P30

Table 1: The characteristic properties of the Czech language for the training of fundamental frequency and duration.

and segment (phoneme) duration Du. The target values of prosodic parameters were extracted from the natural speech signal.

Several types of training algorithm may be applied, e.g. the resilient back-propagation learning algorithm (in the first experiments) respectively the fast back-propagation learning algorithm with a moment and an adaptive learning rate. The other type of algorithm (using heuristic techniques) with feed-forward recall has been determined as optimal. The transfer functions were a sigmoid function in the hidden layer and a linear function in the output layer. The optimal number of training iterations (epochs in MATLAB) was determined in the training process by the increment of sum square errors of test patterns. Small uniform random values from the interval $(-1; +1)$ were used for weight initialization. The results are the real values of F_0 in $[kHz]$ and Du in $[ms]$. In the back-propagation algorithm, the batch mode was used.

3.1.1 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 units) and phonetic transcription of sentences from the speech corpus is done in the phase of pre-processing. The speech signal was labelled by hand in our case, but an automatic approach based on ANN is also under construction.

Our first experiments [20] documented a considerable improvement in the naturalness of synthetic speech; however, this approach required completion of the input feature values by hand, a procedure highly time-consuming for extensive files. It is, therefore, necessary to improve the prosody by other approaches that use only automatically classified features (input parameters).

The changes of fundamental frequency F_0 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 with its emotional timbre. Specific types of sentences (e.g. two types of questions: yes-or-no questions and why-questions, indicative sentence etc.) can be distinguished by the intonation. The duration of speech units is also related to the speed of the speech.

3.1.2 Optimisation of the architecture MLNN

The topology of the ANN is also dependent on the number of input neurons that represent the most important speech parameters. The method of pruning of the ANN based on several approaches (GUHA method, sensitivities of the synaptic weights, etc.) is suitable tool for reducing the ANN structure.

As mentioned earlier, it is possible to find relevant phonetic and phonologic parameters for prosody modelling, by ANN particularly in the event that the prosody of the synthesizer is controlled by an artificial neural network. Because the numerical algorithm for the determination of important input parameters (so called markers) is not known, either the neural net pruning procedure or data mining procedures may be used for a relevant input parameters determination.

The GUHA method (General Unary Hypotheses Automaton), see [1], generates and evaluates the hypotheses between the conjunctions of input and output parameters (the characteristic properties are input parameters, the prosody parameters create ANN output). According to the number of accepted hypotheses, we can determine the importance of individual parameters. The entire process of features selection is described in [10], [11], and the number of input parameters was reduced from thirty to eighteen.

Next possible approach for the determination of important input parameters of a system is the comparison of absolute values of output signals with small changes of input signals. This approach first

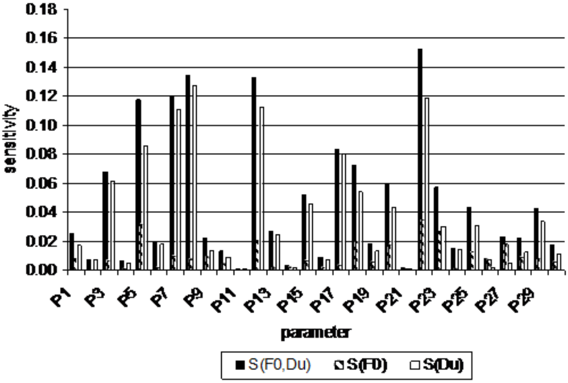

Figure 1: Sensitivity of fundamental frequency and phoneme duration on input parameters.

appeared in the field of linear electronic circuit in the seventies but was later used in many other domains, e.g., in data mining. We can define the sensitivity of the n-th output y_n to the small changes of the k-th input x_k:

$$S_{nk} = S\ (y_n, x_k) = \lim_{\Delta x_k \to 0} \frac{\Delta y_n}{\Delta x_k} \tag{1}$$

The main problem with the practical use of this approach in numerical calculations is the determination of optimal value Δx_k. If it is too small, the influence of the rounding errors can be significant, and if it is too large, the calculated value of (1) will not be correct in the working point under consideration. In our case $n = 1, 2$ (fundamental frequency, duration of speech unit) and $k = 1, \ldots, 30$ (characteristic features of the Czech language - see Tab.1). The parameters with the minimal sum of S_{nk}.

$$S(y_n, x_{missed}) = \min_k \sum_{n=1}^{N} S(y_n, x_k) \tag{2}$$

can be dismissed because their influence on the output signal is minimal (see Fig.1). One further possibility is to create two neural networks for prosody control, the first one for the fundamental frequency and the second for the duration of the speech unit.

The next method for data mining is a standard network pruning method based on the optimisation of the number of hidden neurons [14], which minimizes network redundancy. The choice of excluded neurons can be based on several different strategies; for our attempts, we have selected the method based on the exclusion of neurons having the minimal absolute value of the sum of all input weights. After this process the number of hidden neurons is set at 22 neurons (from 25 neurons on the start). All methods are used for improvement of the generalization ability.

3.1.3 Coarticulation problem

The coarticulation problem is very important in speech processing. For the term „coarticulation“, we mean the influence of the previous and the following phonemes on the current phoneme. We have analyzed the errors between the target and calculated values of F_0 and Du from the point of view of the different context of speech units. The context of three phonemes combinations CCC, VVC, VCV, CVV, VCC, CCV, and CVC (C = consonant, V = vowel) were analyzed for the determination of a further improvement of prosody. After this research, the following facts can be extracted:

- The values of the error function (difference between the target and the output values) for the coarticulation C-C-V and V-C-V are increasing in order (from max. error to min. error) – nasals, explosives, semi-explosives, fricatives for C-C-V and explosives, nasals, semi-explosives, fricatives for V-C-V.
- Errors of fundamental frequency of vocals follow the vocalic triangle, known by the experts in phonetics. Max. error was obtained for the vowel **i**, min.error for the vowel **u** and for the diftong **ou**. Generally, the greatest errors were recognized in the case of C-i-C coarticulation.
- The results confirmed the opinion of the phonetic experts [6] that the influence of the neighbouring consonants on vowels in coarticulation C-V-C is not very important.

The listener impression is better in the case of a large amount of small differences than in the case of a small amount of large differences.

The most complex evaluation is the listening test, but it is very subjective and cannot be described by an objective metric.

3.2 Kohonen's SOM application to children's speech analysis

Kohonen's Self-Organizing Features Map (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 search for interdependences and joint properties in a set submit patterns. T. Kohonen was inspired by the self-organising procedure in a human brain, by its adaptation and learning ability [5]. Every SOM is usually used for data compression.

3.2.1 Principle of a functioning of KSOM

Multidimensional input data are transformed into a decreasing-dimensional space during the iterative learning procedure. At the basis of this method lies the fact that a human brain creates a map with specific areas, the areas that concentrate and treat different impulses. One executive layer coordinates particular input vector elements (created by investigate properties or characterizations) with all executive neurons. The search of the minimal distance between the input vectors and coordinate of the neurons in the map is expressed by a basic mathematic formula. The Euclidean distance is most frequently used. The neuron with minimal distance is called „winner of the competitions“, coming nearest in the coded patterns. The principle is one of competitive learning. A neighbourhood of the winner is created in the Kohonen algorithm. All similar input vectors are updated, and a cluster created of all input vectors with common properties. Then, they are allocated on the map and indicate the number of dominant properties in one training epoch; clusters can point to movement in the input data and „to regrade“ any characterization into different groups in the course repetition. Both cases fit into category „classifications“.

The unified distance matrix or U-matrix (see Fig. 2) is a representation of the KSOM that visualizes the distance between the neurons and its neighbors. The KSOM neurons are represented by hexagonal cells (in our experiment). The distance between the adjacent neu-

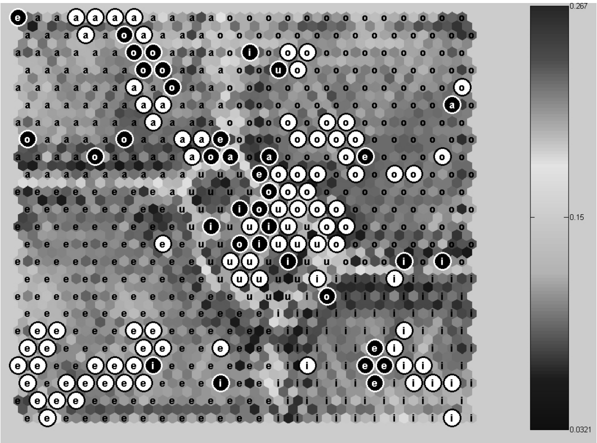

Figure 2: Visualization of vowel classification of one children patient by U-matrix.

rons is calculated and presented with different colors. Dark colors between neurons correspond to a large 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. A new SOM variant has been in use for vowel 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 (in our case, a text known to us in its pronunciation and thus its phonetic classes).

The SSOM consists of m units located on a regular, low-dimensional grid of map units. The map unit positions on the regular grid are fixed; each map unit is connected to a number of neighboring map units with a neighborhood relation. Supervised learning means that the input vector is formed of two parts, x_0 and x_c, where $x_0 = [x_{01}, x_{02}, \ldots, x_{0n}]^T$, $x_0 \in \Re^n$ is an original input vector of dimension n and $x_c = [x_{c1}, x_{c2}, \ldots, x_{ck}]^T$, $x_c \in \Re^k$, k is assigned as known class of x_c (supervisor) in a training set (indication of vowels in our experiments). Each element of vector x_c represents one of k classes. A new vector $x = [x_0, x_c]^T \in \Re^{n+k}$ will have a dimension $n + k$, which is valid for a prototype vector $m = [m_1, m_2, \ldots, m_{n+k}]^T$, $m \in \Re^{n+k}$ as well. During the classification of an unknown input vector x, only its x_0 part was compared with the corresponding part of the prototype vectors. The class of each unit (neuron) is found by taking maximum over these added elements, and a label is give accordingly (Fig.3).

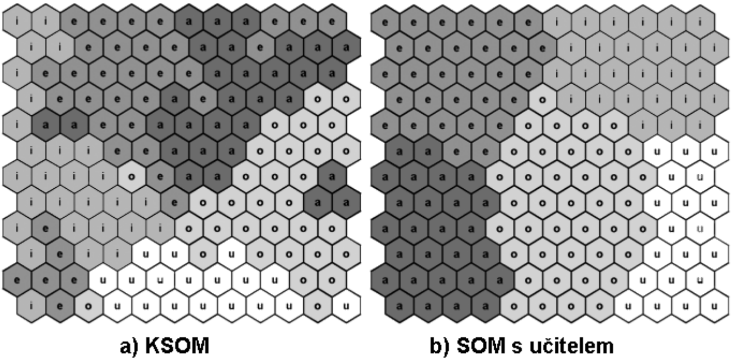

Figure 3: Comparison of standard Kohonen SOM and supervised SOM.

Spoken speech is a time-dependent sequence of phonemes, making the reason why it is necessary to process the input data to ANN in a batch form: this method is significantly faster and does not require any specification of a learning-rate factor (in comparison with the incremental learning algorithm, which is a commonly used algorithm in ANN training). New prototype vectors are calculated as a weighted average of the input vectors, where the weight of each input vectors is the neighborhood function value $h_{i,m^*(j)}$ at its winner $m^*(j)$:

$$m_i(t+1) = \frac{\sum_{j=1}^{N} h_{i,m^*(j)}(t)\, x_j}{\sum_{j=1}^{N} h_{i,m^*(j)}(t)} \tag{3}$$

where t is the number of iteration, x_j is the input vector, N is the number of input vectors. The most usual neighborhood function is the Gaussian one:

$$h_{ij}(t) = exp\left(\frac{-\|r_j - r_i\|^2}{2\sigma^2(t)}\right),\ \|r_j - r_i\| \leq \sigma(t) \tag{4}$$

$$h_{ij}(t) = 0,\ \|r_j - r_i\| > \sigma(t) \tag{5}$$

$r_j, r_i \in \Re^2$ are the location vectors of units j and i in map for $2 - D$. A parameter $\sigma(t)$, the neighborhood radius, defines the width of the kernel, usually a smoothly decreasing function of time. We use the batch map (see [5]) for our experiments. This method is defined as an iterative process in which a number of input vectors $\mathbf{x}$ are classified into the respective V_i regions first. Secondly, new prototype vectors $\mathbf{m}$ are computed as weighted averages of all training samples:

$$\mathbf{m}_i(t+1) = \frac{\sum_{i=1}^{m} h_{ij}(t) N_i \overline{\mathbf{x}_i}}{\sum_{j=1}^{n} h_{ij}(t) N_i} \tag{6}$$

where x_i is an input vector, n is the number of input vectors, m is the number of units, N_i is the number of input vectors in the Voronoi set V_i:

$$V_j = \{\mathbf{x} \in \Re^N \mid \|\ \mathbf{x} - \mathbf{m}_j\ \| \leq \|\ \mathbf{x} - \mathbf{m}_k(t)\ \|\ \forall k \neq j\} \tag{7}$$

$$\overline{\mathbf{x}_i} = \frac{1}{N_i} \sum_{\mathbf{x} \in V_i} \mathbf{x} \tag{8}$$

is the mean of the vector $\mathbf{x}$ in the Voronoi set V_i. The value of the neighborhood function between map units $\mathbf{m}_j$ and $\mathbf{b}_i$ is sign as $\mathbf{h}_{bij}$, the winner (denoted also as the best-matching prototype - BMU) to the input vector $\mathbf{x}_i$ is computed by the following equation.

$$\mathbf{b}_i = arg \min_j \{\|\mathbf{x}_i - \mathbf{m}_j\|\} \tag{9}$$

This method has used for training of particular maps for each patient (see Fig.5, in page 24).

3.2.2 Disordered speech analysis

One area where researchers are applying tested mathematical engineering methods is that of helping people with different forms of disabilities. Our research in this area is focused on searching for the relation between clinical and electrophysiological symptoms of children with developmental dysphasia. Sleep EEG and speech analyses are the primary areas under discussion, while the finding of methods acceptable for improvement of the diagnosis and determination of therapeutic procedures is the research topic. Developmental dysphasia is defined as a disruption in the ability to acquire normal language skills adequately to age, while at the same time displaying intact peripheral hearing, normal intelligence and absence of behavioral disorder or negative social factors.

From the point of view of the impairment of speech, differences between healthy and ill children were registered by speech signal analysis. Our method involves clustering the pattern characteristics visible by the allocation of the vowels respectively by changes in allocation of the vowels pronounced by the patients. The success of the process of analysis is definitely dependent on the precision of the labeling of the natural speech signal in the database.

A preference for self-organizing maps (SOM) has been assumed from the nature of our problem. For many real problems, the target values for all the patterns of the database are not known. Nor do we know all the characteristics of the patterns. In such a case, a selection of one form of unsupervised learning - clustering is suitable. One of the symptoms of the children with a disorder indicating developmental dysphasia is a malfunction of perception and impairment of speech, yet we cannot say which one is affected, nor to what degree. The standard and supervised SOM for the training maps by healthy children were compared. The ability of vowel classification and allocation in the map as a vocalic triangle is investigated in neural network applications. We have started from the hypothesis that it involves the disorder of movement of vocal organs in articulation in the case of developmental dysphasia, influencing the formant generation. The vowel mapping of patients is different in comparison to the vowel mapping of healthy children.

One of the goals of our research is to create a software pack with a user-friendly interface for medical doctors or other medical staff.

3.3 Software

The present trends of processing large continuous speech databases require the automation of the methods used. All analyses and experiments were performed by the computational system MATLAB.

The original software Speech Laboratory (SL) package (see [9]) was used for the creation of user-friendly application of the neural networks in the prosody modelling of synthetic speech. This package based on NN-Toolbox of MATLAB can serve for automation of the database creation, neural network training and graphical processing. The application of ANN in SL is divided into two steps. In the first step, ANN is created and trained by the input data. Prosodic parameters have to be extracted from the acoustic wav-files. The extracted parameters create target vectors for the training process. In the second step, ANN is used as a projection of input data extracted from the text into the outputs. These outputs directly represent prosodic parameters F_0 and Du.

The synthesizer of the Czech ARTIC prepared by the Dept. of Cybernetics of the University of West Bohemia in Pilsen ([7]) is used for results verification, within the framework of general cooperation on the grant project. The listening-opinion tests by MOS (Mean Opinion Score) are employed for difference evaluation between particular variants of the synthetic speech (see [19]).

The software, called SOM Toolbox©, was applied in second experiment topic. 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, and is available free of charge under the General Public License from ([21]). For the project, new special M-files, which should be a part of supporting program package, were created (see [15]).

3.4 Results

Results of the research in both ANN applications, which are at the center of our attention and which we discussed in previous parts of this paper, show better effects (in the case of some parameters) than standard methods. The following text summarizes these findings.

The comparison of all three pruning methods of prosody modelling (sensitivity analysis, GUHA method and classical pruning) is collected in Fig.4. The utilization of the ANN and data mining principles for prosody control brings about a serious improvement of synthetic speech quality. The results achieved have not, until now, been optimal, especially for F_0.

The next improvement can be expected by the more precise determination of target values of prosodic parameters from the speech corpuses. All results, achieved by the analysis and mathematical tools must be verified by listening tests. This is the way to achieve not only understandable but also natural synthetic speech. The corpus of the pronounced speech signal together with the written text must be available for the solution of the speech processing tasks. As we have told above, the language is a complicated complex of attributes and it is very difficult to create a really representative speech corpus. On the other hand, the results of speech modelling are strongly dependent on the quality of the training set. According to our previous results the selection of input speech parameters, recommended by language experts and suitable for the classical prosody modelling methods for prosodz (e.g. by rules) is not optimal for prosody control, based on the ANN utilization without serious modification. It is particularly valid for the fundamental frequency modelling.

In the experiments describing disordered speech analysis, we analyzed the vowel mapping. Speech analysis has been performed for 12 children (3 girls and 9 boys) in the course of three- or four-month periods. After each period, the same utterances are recorded and analyzed. Six of them are on medication, while the others have only visited a speech therapist and a control clinical examination. An example of the cluster visualization in the maps which represents the vowels distribution of an alternative child patient in comparison with the vowel distribution of 55 healthy children (35 girls and 20 boys) is showed in Fig.5. The children indicated for vowel classification in this contribution are between the ages of 7 and 10. The patient in example was 7 years old.

White units indicate the successful classifications from the map trained by the speech data of healthy children, black units represent classification errors. Their number and location in the map change after each recording depending on the change of the state of health of pacients. Likewise, the ability for good pronunciation depends on age.

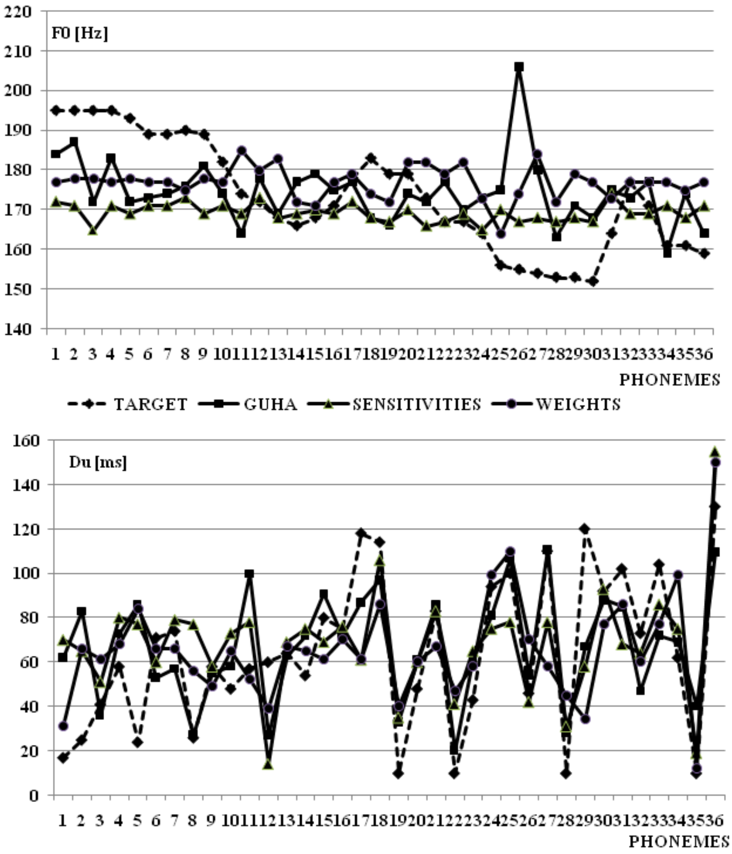

Figure 4: Comparison of different approaches of prosody parameters determination – the sentence „Weather forecasting for the night and tomorrow“. Above: Values of the individual F_0 contours. Below: Values of the individual Du contours.

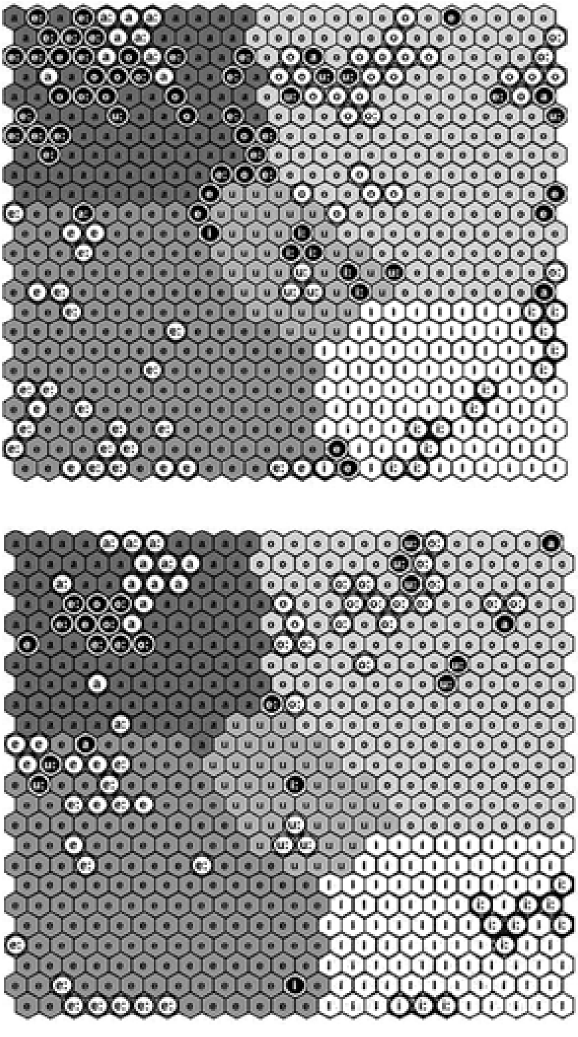

Figure 5: Exemple of the classification of the vowels of the healthy and ill children indicated. Above: before therapy - 1st recording. Below: after third part of therapy - 4th recording.

The aim of medical therapy is to achieve a minimum of wrong classifications. Data analysis is aggravated by the following fact: ill children are not able to pronounce some vowels (the monitored children have displayed problems with the pronunciation „e“, „i“, at certain times with „u“). The obtained results are confirmed by psychological evaluation of patients and by the results of the sleep EEG.

4 Conclusion and future work

I would like to summarize the knowledge from neural network research described in the previous text and draw attention to the possibility of using ANN in the future in research and pedagogy.

The conclusions to be presented ensue from our previous experiences with ANN applications in speech processing. It is possible to appreciate that an exclusively mathematical approach to the evaluation, i.e. a simple mean square error between the target and the predicted fundamental frequency and duration contours, is not the best indicator of the perceived naturalness of speech. It is necessary to judge synthetic sentences by listening, yet there remain significant differences between the mathematical results and the listening tests. For one, 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 resulting speech has turned out better in the case of prosody control by ANN with the use of optimal topology after pruning in comparison with the previous state.

The second topic involves an original method for the intensity of speech defect monitoring in child patients with developmental dysphasia. We were drawing upon a body of knowledge consisting of phonetics, acoustics and ANN applications. The SOM were chosen for solving part of the project. New variants of the SSOM were tested theoretically and experimentally after the first experiments with Kohonens SOM.

Our effort in future work will be focused on both the ANN application domain described here (prosody modelling and disorder speech analysis). We have started research on emotions and their influence on prosody, as well as greater naturalness of speech. We will concentrate on deeper analysis of child speech, mainly devoting attention to longer speech units (syllables, multisyllabic words) and the inability to formulate multisyllabic words (three and four syllables) or phoneme overlap faults, which are other symptoms of developmental dysphasia. Also, verbal dyspraxia, i.e. an obvious clumsiness in word repetition, is mentioned in [5]. The processing of speech signals is complicated by the effect of the real environment (non-professional speakers, high noise in the environment if the speech was recorded in ordinary rooms). The second problem that we have to solve is the fact that we are analyzing childrens speech. Often, its own specific development is not terminated for a particular age group, or the quality of the utterances is strongly influenced by emotion. Also, we have at our disposal only a small amount of speech data, especially for patients, even though a permanent database is kept of child speakers. The size of the database of healthy child speech is also limited by the possibilities of data recording in preschool and primary school institutions, especially with respect to the concern over parent permissions. We assume that it would be necessary to open a sizable screening project during preventive medical check-ups of small children.

The self-organizing maps are favorable for persons without an engineering background, primarily for the ability to visualize higher-dimensional data samples in a low-dimensional display. One of the goals of our research is to create a software pack with a user-friendly interface for doctors or other medical staff.

The considerable student interest in neural network applications was the impetus for the writing of the monograph [18] and teaching text for the training course in Algorithms and Structures of Neurocomputers. In future, more caution will be devoted to biomedical applications, though students have displayed an interest not only in signal processing, but also in data mining. Particular tasks will be solved within framework of the semester, bachelor's and master's projects, as so far. Large-scale and more complicated problems are to form part of the PhD. dissertation. New research results are immediately brought into pedagogic practice. Students working in the field of ANN are concentrated in the laboratory LANNA [4].

Reference

[1] Hajek P., Sochorova A., Zvarova J.: GUHA for personal computers., Computational Statistics and Data Analysis, Vol.19, 1995, North Holland, pp. 149-153.

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

[3] Kohonen T, Hynninen J, Kangas J, Laaksonen J.: SOM_ PAK: The Self-Organizing Map Program Package. Helsinki University of Technology, Lab. of CIS. Available via anonymous ftp at internet address cochlea.hut.fi (130.233.168.48).

[4] Laboratory of Artificial Neural Network Applications (LANNA), http://ajatubar.feld.cvut.cz/lanna/

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

[6] Palkova, Z.: Phonetics and phonologics of the Czech language (in Czech: Fonetika a fonologie češtiny). Univerzita Karlova-Praha, 1994, ISBN: 80-7066-843-1.

[7] Psutka,J., Muller,L., Matousek,J., Radová, V.: We talk Czech with computer (in Czech), Academia Praha,2006, ISBN 80-200-0203-0.

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

[9] Santarius, J., Tihelka, J.: Prosody Modelling of Synthetic Speech. Proc. of the Int. WSP ECMS 2003. TU Liberec , vol.1, 89-92. ISBN 80-7083-708-X.

[10] Sebesta,V., Tuckova,J.: Optimisation of Artificial Neural Network Topology applied in the Prosody Control in Text-to-Speech Synthesis. Theory and practice of informatics: Proc. of 27th Annual Conf. on Current Trends in Theory and Practice

of Informatics SOFSEM 2000, November 25-December 2, 2000, Milovy, Czech Republic, ISBN: 3-540-41348-0 Springer-Verlag Berlin Heidelberg New York, ISSN: 0302-9743.

[11] Sebesta,V., Tuckova,J.: Optimisation of Artificial Neural Network Topology Applied in the Prosody Control in Text-to-Speech Synthesis. In: ICANNGA2001, Prague, Avril 2001, pp. 420-430, ISBN:3-540-41348-0

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

[13] Traber, C.: 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, 287-304. Elsevier Science Publishers,1992.

[14] Tučková, J., Šebesta, V.: Influence of Language Parameters Selection on the Coarticulation of the Phonemes for Prosody Training in TTS by Neural Networks. In.: Proc. of the Int. Conf. on "Artificial Neural Nets and Genetic Algorithms (ICANNGA 2003)". Ed.:David W.Pearson, Nigel C.Steele, Rudolf F.Albrecht. April 2003, Roanne, France pp.85-90, ISBN: 3-211-00743-1 Springer-Verlag Wien-New York.

[15] Tučková, J., Zetocha,P.: Speech analysis of children with developmental dysphasia by Supervised SOM. Int. Journal on Neural and Mass-Parallel Computing and Information Systems „Neural Network World“, Ed. M.Novak, ICS AS CR and CTU, FTS. 16/6: 533-545. ISSN 1210-0552.

[16] Tučková, J., Šebesta,V.: The Prosody Optimisation of the Czech Language Synthesizer. In: Int. Journal on Neural and Mass-Parallel Computing and Information Systems „Neural Network World“, Ed. M.Novak, ICS AS CR and CTU,FTS, vol., No.4, 2008, pp.291-308 . ISSN 1210-0552.

[17] Tučková, J., Komárek, V.: Effectiveness of Speech Analysis by Self-Organizing Maps in Children with Developmental Language Disorders. In: Neuroendocrinology Letters. Ed.: Peter G.

Fedor-Freybergh. Society of Integrated Sciences, vol. 29, No. 6, Nov/Dec 2008, ISSN 0172-780X.

[18] Tučková,J.: Selected applications of the artificial neural networks at the signal processing. (in Czech – Vybrané aplikace umělých neuronových sítí při zpracování signálů). Nakladatelství ČVUT, Praha, 2009, ISBN 978-80-01-04229-8.

[19] Tučková,J., Holub,J., Duběda,T.: Technical and Phonetic Aspects of Speech Quality Assessment: the Case of Prosody Synthesis. Eds: Anna Esposito, Robert Vích. In: Book from COST 2102 Conf. in Prague 2008, Lecture Notes in Computer Science, Springer, 2009 (in press).

[20] Tuckova, J., Vich, R., 1997. Fundamental Frequency Modelling by Neural Nets in Czech Text-to-Speech Synthesis. Proc.of the IASTED Int.Conf. Signal and Image Processing SIP'97, New Orleans, USA, pp.85-87.

[21] Vesanto, J., Himberg, J., Alhoniemi, E., Parhankangas, J.: SOM Toolbox for Matlab 5, SOM Toolbox Team, Helsinki Univesity of Technology, Finland, Homepage of SOM Toolbox: www.cis.hut.fi/projects/somtoolbox

5 Curriculum Vitae

Doc. Ing. Jana Tučková, CSc.

Date of birth: August 16, 1950, Prague

Education:

1974 Ing.(M.S. equivalent), Faculty of Electrical Engineering, CTU in Prague

1980 CSc.(PhD equivalent), Faculty of Electrical Engineering, CTU in Prague

1993 PGS, EPFL Lausanne, Switzerland „Biological and artificial neural networks“(Attestation in graduation of a CPI – Cours Postgrade en Informatique Technique „Réseaux de neurones biologiques et artificiels“), thesis „Quantification vectorielle du signal parole par un réseau de Kohonen“. Garant : prof. Martin Hasler

1997 Doc.(Associate Professor), Faculty of Electrical Engineering, CTU in Prague (in Telecommunication Technology), Habilitation: „Artificial Neural Networks and Isolated Word Recognition.“

Professional experiences

1974–1976 Institute of Research of a Communication Engineering, PhD student(interruption of PhD study for severe injury)

1976 Department of Circuit Theory, the Faculty of Electrical Engineering, CTU in Prague – traineeship

1977 –1981 FEE CTU in Prague, Department of Circuit Theory – Assistant

1978–1979 FEE CTU in Prague – pedagogic course for assistant

1978–1990 Research secretary of grant of Czechoslovak III-3-1, III-6-2 a III-7-6

1981–1997 FEE CTU in Prague, Department of Circuit Theory – Assistant Professor

from 1997 FEE CTU in Prague, Department of Circuit Theory – Associate Professor

1992 Academic stay abroad from EU in TIK CIRC EPFL, Lousanne, Switzerland

1993 Swiss Federal Institute of Technology (EPFL), Dept.of Electrical Engineering. Lousanne, Switzerland. Consultation, defence of postgraduate course thesis ”Biological and artificial neural networks”. Short mission.

1994 Technical University Dresden, Dept. of Technical Acoustics. Scientific Mission – partnership in project „Processing of Speech Signals“

1995–2001 Institute of Radioengineering and Electronics of the Academy of Sciences of the Czech Republic, research worker (part-time work load).

1995 TU Dresden, Dept. of Technical Acoustics. Short Scientific Mission (within the frame FEE CTU)

1996 TU Dresden, Dept. of Technical Acoustics. Short Scientific Mission (within the frame IREE CAS CZ)

1998 ETH Curych, Switzerland. Short Scientific Mission, COST 258.

2002-2006 Technical University of Liberec – ANN consultant, lesdership-specialist of the PhD thesis, PhD lectures

Pedagogical activity:
Lectures:

- Special Structures of the Digital Systems - FEE CTU of Prague (master study programme)
- Algorithms and Structures of neurocomputers - structure study – FEE CTU of Prague (master study programme)

- Introduction in the Artificial Neural Network. University of Zilina, Faculty of Electrical Engineering, Slovakia (master study programme)
- Neural network applications in signal processing, University of Zilina, Faculty of Electrical Engineering, Slovakia (separate lectures)
- Neural Network Applications in the Speech Processing, Brno University of Technology (distance PhD study, tutorial)
- Theorie of the Neural Networks, Neural Networks, Technical University of Liberec, Faculty of Mechatronics and Interdisciplinary Engineering Studies – doctoral study programme, tutorial, examination
- Biological Signal Processing by Artificial Neural Networks. Partial lecture within the scope of Biological Signals. Faculty of Electrical Engineering, CTU in Prague (bachelor study programme)

Training course, laboratory exercise,
computer course:

- Special Structures of the Digital Systems (master study programme),
- Algorithms and Structures of neurocomputers (ASN) (master study programme),
- Electrical Filters (master study programme),
- Electrical circuits 2(EO2) and 3(EO3)(bachelor study programme),
- Classification of the Biological Sognals by Self-Organizing Maps (partial computer laboratory within the scope of Biological Signals) (bachelor study programme)

Supervising of graduate student: 34 master thesis in total, 13 master thesis in 1999-2009, 9 bachelor projects in total, 8 bachelor projects in 1999–2008.

Supervising of PhD students: 1 succesfull, 2 after state doctoral examination, currently 3 PhD students, supervisor-consultant for 3 PhD student.

Activities: Reviewing and reader's reports for international and national technical journals and publishers (Signal Image and Video Processing, Neural Network World Journal, Journal of Electrical Engineering), for the Czech Science Foundation and Czech Accreditation Institute. Reader review of The Handbook of Technology Management (3 Volume Set, Hossein Bidgoli, Editor-in-Chief, John Wiley& Sons, Inc. 2009, Hoboken, N.J. 07030. Volume III, Management Support Systems, Electronic Commerce, Legal and Security Considerations. Telecommunications and Networking an Management Support Systems, 178. Artificial Neural Networks). Reader review of monographs from the authors: Novak, M. and others – Artificial Neural Networks., C.H.Beck. Reviewer of the habilitation thesis for Associated Professor in Brno University of Technology, TU in Liberec, WBU in Pilsen. Participation on 6 research grants (in 2 grants as joint applicant in CTU) of the Czech Science Foundation, Ministry of Education, Ministry of Health and COST. PhD State examination committee – member in UWB in Pilsen, TU in Liberec, Institut National des Science Appliquées (INSA) de Toulouse (in French). President or member of MSc. State examin. committee in University of Žilina, Slovakia, in FEE CTU in Prague (in Electronics and Biomedical Engineering). President of BSc. State examination committee in CTU in Prague, FEE. Reviewer of 31 papers for Int. Conferences.

Research areas: Applications of ANN, especially on a speech processing – the modelling of synthetic speech prosody (data mining methods for ANN optimisation), the new research field of speech signal analysis for medical purposes (part of joint research program with the Clinic of Paediatrie Neurology, Faculty of Medicine, Charles University in Prague). Leadership of LANNA (the Laboratory of Artificial Neural Network Applications). Master degree and Phd students from the LANNA participate in both research areas.

Awards:

- Best Presentation in Session 9.2 „Speech Processing“, IJCNN, Washington, DC, USA, 1999.

- Best Poster Award „Best of Conference“, IJCNN, Washington, DC, USA, 1999.
- Best Presentation in „Signal and System Control“ session, IEEE-ICIT 2004, Hammamet, Tunisko.
- Minister of Health of the Czech Republic Award for research grant solution „Computer Analyses of Speech and Overnight EEG in Children“ in 2008 – joint applicant in CTU.

Memberships in professional committees: The INNS (Int. Neural Networks Society), ELSNET-STN (as „Language and Speech Technology Expert“), the programme and technical committees of IASTED (International Association of Science and Technology for Development), the Final Opposition Procedure Council of the Int. Co-operation German-Czecht Projekt „Bilingual Speech Synthesis German/Czech - Synthesis Inventoires“.

Publications:

125 contributions from 1975, from that 27 publications focused on the theory of linear circuits and active filters, 98 publications concentrating on the domain of theory and applications of the artificial neural networks (from 1993). They are: 1 book-monograph, 1 book chapters, 1 lecture note, 59 articles in international journals or conferences, 12 articles in national journals or conferences, 22 invited lectures, 2 unpublished thesis.

Public Acceptance

- Czech Television – ČT 24 (ČT 2) Planet - science: Neural networks.
- Internet journal „3pod“: Artificial neural networks help even medical doctors.