ARCHIVES OF ACOUSTICS
32, 1, 177–186 (2007)

APPLICATION OF NEW ACOUSTIC PARAMETERS IN ANN-AIDED PATHOLOGICAL SPEECH DIAGNOSIS

Joanna SZALENIEC$^{(1)}$, Maciej MODRZEJEWSKI$^{(1)}$,
Maciej SZALENIEC$^{(2)}$, Wiesław WSZOŁEK$^{(3)}$

$^{(1)}$Jagiellonian University, Collegium Medicum
Chair of Otolaryngology
Śniadeckich 2, 31-501 Kraków, Poland

$^{(2)}$Institute of Catalysis and Surface Chemistry, JLBEC
Niezapominajek 8, 30-239 Kraków, Poland

$^{(3)}$AGH University of Science and Technology
Mickiewicza 30, 30-059 Kraków, Poland
e-mail: asiat@agh.edu.pl

(*received October 16, 2006; accepted December 1, 2006*)

Most diseases of the vocal tract cause changes in the voice quality. Acoustic analysis of the speech signal is a widely used, noninvasive, objective and low-cost method of laryngeal pathology recognition and classification. There have been numerous attempts [1–3] to develop an automatic system which could aid the laryngological diagnosis. The goal of the presented research is to verify, whether the new approach to the acoustic analysis and parameters introduced in the Voice Analysis and Screening System (VASS 3.0 [4]) such as turbulence noise index (TNI) and normalized first harmonic energy (NFHE), can improve the effectiveness of automated diagnosis. The automated diagnosis was performed using Artificial Neural Networks (ANN). Multilayer perceptron and radial basis function neural networks of various architectures were trained to classify between pathologic and non-pathologic voices, while the parameters computed with VASS were used as input data. Preliminary results show that the Voice Analysis and Screening System coupled with ANN can be a highly effective tool for ANN-aided pathological speech diagnosis.

Keywords: speech analysis, pathological speech, speech recognition, neural networks, surgical treatment.

1. Introduction

Pathological processes that affect the vocal tract in most cases cause changes in the speech production process, which can be heard as abnormal voice (dysphonia). These changes are often the first, isolated and therefore very important symptom in early stages of larynx pathologies. It should be emphasized, that voice problems such as hoarseness are frequently underestimated by the patients. Moreover, they often cannot be properly diagnosed on the most accessible primary health care level, since they require an expert laryngologist and expensive professional equipment. Consequently, in many cases the correct diagnosis and treatment is introduced in advanced stages of the disease, often when it is no longer possible to cure the patient. There are many diseases which can be characterized by the sequence described above, but it is evident that the most important problem in this field is early detection of the larynx cancer. It is well known that at early stages, the disease can be cured by means of minimally invasive methods, while advanced stages of cancer often require more aggressive, crippling treatment, including permanent loss of the ability to speak (laryngectomy, i. e. excision of the larynx). The mortality rate in advanced stage is also significantly higher than in early stages (the 5-year survival rate equals 80–90% in the lowest Stage I and decreases to 30–40% in the highest Stage IV) [5]. An easily accessible, low-cost and noninvasive method of laryngological pre-diagnosis, based on acoustic signal analysis and advanced processing of the phonological data, could therefore improve the detection and treatment of larynx diseases.

There have been many attempts to create a reliable computer system, which could distinguish between patients without serious vocal tract problems and those who need a consequent laryngological diagnosis and treatment. A noninvasive, automated method of voice diagnosis is based on advanced acoustic analysis of the speech signal and using of artificial intelligence methods. Acoustic parameters which can be extracted from the signal reflect such changes in the voice as loss of power, changes in the pitch, constriction of the voice range (displacement towards lower frequency), addition of noises, etc. which are important from the medical point of view [6].

Although the range of parameters which can nowadays be calculated for a speech sample is very wide, it is not always possible to determine, what changes (or patterns of changes) are distinctive for larynx pathologies. Therefore, artificial neural networks (ANN) are an effective and frequently applied tool for acoustic analysis based voice diagnosis. The ANN learn to classify cases during a process of training, where acoustic parameters for pathological and non-pathological speech are presented.

The main purpose of the research projects mentioned above was to increase the accuracy of pathology detection and eliminate the most dangerous error – classification of a patient with laryngeal disease as a normal speaker [7].

The presented work is based on the use of Voice Analysis and Screening System (VASS) – a new computer system for acoustic analysis of pathological voice signals and screening of laryngeal diseases, which introduces a novel approach to the acoustic analysis of the speech signal and introduces new parameters for estimation of the turbulent noise and breathiness in the voice. The most detailed description of such system is given in the paper [6]

The acoustic parameters extracted by means of VASS were used as input parameters in all previous (preliminary) analyses, which were performed for evaluation of the usefulness of acoustical parameters as an element of diagnosis, prognosis and treatment control in various laryngological problems [8, 9]. In this paper we use the same parameters as the input signals for the artificial neural networks, because we hope the combination of a good description of the object under consideration (speech articulation system) by means of the selected parameters combined with the well-known power of neural network-based modeling ability, can provide an effective and accurate preliminary diagnosis of the larynx.

2. Material and methods

2.1. Speech samples

The speech samples were obtained from 32 patients hospitalized in the Chair of Otolaryngology, Collegium Medicum UJ. In this group 12 patients presented various larynx diseases (vocal fold cancer, vocal fold polyp, chronic laryngitis) and 20 were a control group with no laryngeal pathology (the latter patients were hospitalized because of other pathologies that did not involve the larynx). The presence or absence of the laryngeal pathology was proved by ENT examination, in several cases followed by computer tomography and/or histopathological examination of the pathological tissue.

All patients pronounced Polish vowel /a/ repeated three times in a sustained manner at comfortable levels of pitch and volume, which corresponded to their conversational natural voice. All recordings were carried out in the same idealized acoustical environment (sound-treated room in the Chair of Otolaryngology, Collegium Medicum UJ, where the measured noise level was below 32 dB). The equipment used for recording consisted of a microphone G.R.A.S 40 AF, a preamplifier Norsonic 1201, an amplifier G.R.A.S. 12AA with gain of 40 dB, a professional digital audio tape recorder HHB PDR 1000 with dynamical range min. 80 dB, two channel real time sound analyser Nor 840 for fast inspection of the samples, and PC computer for data analysis and database collection. During the digitalization process the sampling rate was 44100 Hz and the sample quantization resolution was 16 bit without amplitude compression. The recorded samples were converted into *.wav files with the Samplitude Project V5.55 program [10]. From the three consecutive vowels pronounced by a patient, only the quasi-stationary state was extracted, after this the concatenation the samples was performed (with guarantying of the signal continuity conditions) and the combined signal was used as a single voice sample.

2.2. Voice analysis and screening system

For analysis of the speech signal, the Voice Analysis and Screening System (VASS) was used [4]. The new approach introduced in the system consists in tracing all glottal cycles by means of a cross-correlation detector. Basing on the so determined beginning and duration of all glottal cycles, shimmer, jitter, harmonics-to-noise ratio and other widely used acoustic parameters are calculated. New parameters introduced in VASS are TNI for estimation of the turbulent noise in voice signals and NFHE for the "breathy" voice characterization [6]. Other parameters calculated in VASS include: standard deviation of the fundamental frequency, maximal, minimal and mean fundamental frequency (in this work replaced with NF$_0$, that is a new parameter which we introduce to measure the fundamental frequency changeability – see Table 1), pitch perturbation quotient (jitter) and amplitude perturbation quotient (shimmer), harmonic-to-noise ratio (HNR) calculated by Yumoto's method, HNR in frequency domain, HNR in frequency domain by Qi's, normalized noise energy (NNE), amplitude and frequency of the most intensive amplitude and frequency tremors. More detailed description of the parameters is presented in Table 1.

Table 1. Acoustic parameters used as input data for the ANN.

Parameter (abbr.)	Parameter description
SD F$_0$	Standard deviation of the fundamental frequency.
NF$_0$	Measure of the fundamental frequency changeability. $\mathrm{NF}_0 = \frac{F_{0\,\max} - F_{0\,\min}}{F_{0m}}$, $F_{0\,\max}$ – maximal fundamental frequency, $F_{0\,\min}$ – minimal fundamental frequency, $F_{0\mathrm{m}}$ – mean fundamental frequency.
Pitch Perturbation Quotient (jitter)	$\mathrm{PPQ} = \frac{N \sum_{i=2}^{N-1} \left\lvert \frac{P_{i-1}+P_i+P_{i+1}}{3} - P_i \right\rvert}{(N-2)\sum_{i=1}^{N} P_i}$, P_i – pitch of the i-th glottal cycle, N – number of cycles contained in the considered segment of the voice signal [11, 12, 13]. This method for calculation reduces the influence of slow amplitude and pitch variations (tremors) due to non-pathological factors (physical overtension) [6].
Amplitude Perturbation Quotient (shimmer)	$\mathrm{APQ} = \frac{N \sum_{i=2}^{N-1} \left\lvert \frac{A_{i-1}+A_i+A_{i+1}}{3} - A_i \right\rvert}{(N-2)\sum_{i=1}^{N} A_i}$, A_i – amplitude of the i-th glottal cycle, N – number of cycles (as above).
Turbulence Noise Index (TNI)	Ratio of the turbulent noise energy to the total energy of the voice signal: $\mathrm{TNI} = 100\left(1 - \frac{1}{N-1}\sum_{i=1}^{N-1} R(p_i,T_i)\right)$, $R(p_i,T_i) = \frac{\sum_{t=0}^{T_i} s_i(p_i+t)s(p_i+T_i+t)}{\sqrt{\sum_{t=o}^{T_i} s_i^2(p_i+t)\sum_{t=o}^{T_i} s^2(p_i+T_i+t)}}$, $R(p_i,T_i)$ – correlation coefficient between two consequent glottal cycles of the signal with beginning p_i and a duration of T_i, N – number of cycles contained in the segment [14].

HNR by Yumoto's method	Ratio of harmonic energy to noise energy, informative for the hoarseness of the voice signal. It accounts for the turbulent noise in the voice and for the modulation noise due to the jitter and shimmer [15, 16, 17].
HNR in frequency domain	As above, but counted in frequency domain after FFT transform.
HNR in frequency domain by Qi's	A modification of Yumoto's method that consists in normalisation of the spectrum [18].
Normalized Noise Energy (NNE)	$\mathrm{NNE} = 10\log_{10}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{E_{Ni}}{E_{Hi}+E_{Ni}}\right)[\mathrm{dB}],$ E_H – harmonics energy, E_N – noise energy
TNtSR	Turbulence Noise-to-Signal Ratio.
DH	Degree of Hoarseness in spectral domains.
Normalized First Harmonic Energy	Ratio of the amplitude of the first harmonic from the power spectrum to the total energy of the rest of harmonics in the 4 kHz frequency band, used instead of the most frequently used ratio between the amplitudes of the first and second harmonics [19]. The parameter reflects non-simultaneous lengthwise closing of vocal fold resulting in so-called 'breathy' phonation.
Fr AT, A AT, Fr FT, A FT	Frequency and amplitude of the most intensive amplitude and frequency tremor, respectively. Tremors are slow fundamental frequency and amplitude variations (of frequency 15–20 Hz), due to physical over-tension [6].

2.3. Artificial neural networks

The acoustic parameters listed above were used as input signals for artificial neural networks, which were designed, implemented, optimized, learned and performed using professional program Statistica Neural Networks 6.0 [20].

We decided to use two types of popular neural networks architectures: multilayer perceptron (with three-layer architecture) and radial basis functions feed-forward neural network. Every network under consideration was formed with the following numbers of elements constituting its structure:

- 4–16 neurons in the input layer,
- 5–15 neurons in the hidden layer,
- 1 neuron in the output layer.

The single output neuron, which is the only common feature for all of the tested models, was expected to provide binary output (1 or 0 for samples recognized as pathological or non-pathological, respectively). However, the structure described above includes two levels of freedom – the number of neurons in the input layer and the number of neurons in the hidden layer. The values of these parameters had to be determined by means of special goal oriented experiments.

The first and most important parameter under consideration was therefore the dimension (and content) of the input vector. In the first step all of the parameters used as input data were scaled linear within the 0-1 range to eliminate potential domination of some parameters over the other.

The next, more complicated and confusing step included selection of the input parameters. It is very well known (and evident) that using more input signals provides better and more complete information about the process passed to the neural network working as a model of such process. It means that the neural model of any object under consideration can be more precisely and better adapted to the real process, and this result is very profitable. In the discussed case the richest model (fed by more input parameters) can better reproduce the speech articulation process and can better help in differentiation between the normal and pathological situations. On the other hand it is also known that too many input parameters cause always increasing difficulties during the learning process, delaying the model tuning and sometimes leading up to the instability of the learning iterations and total loosing of the generalization abilities in the trained neural model.

In our research we tried to find out, how many input parameters are necessary for proper functioning of the neural network model of the speech articulation process, and – what was much more difficult – which from the posed input parameters (describing and measuring different aspects of the analyzed speech signal) are absolutely necessary to build an adequate model. The latter problem has not been solved entirely in the presented work and requires further investigation, but the results of preliminary trials are described below.

The second level of freedom during the neural network forming is always the number of neurons in the hidden layer. If this number is too big, the network can not solve the problem because its "intelligence" is too low. Nevertheless, a too big number of hidden neurons (which theoretically produces a "too smart" network) is also not good, because limited amount of information, given by so-called "learning set" (see below) can not determine all necessary values of connection parameters ("synoptic weight") inside the rich internal structure of the network, which often makes its behavior unpredictable and the whole network – useless. The theory of neural networks until now gives no precise solutions, answering the question: How many hidden neurons do we need for such collection of modeled data? It brings about an indispensable next series of time consuming experiments, during which one must change the structure of the network, giving more or less hidden neurons, evaluating after full learning process the quality of the obtained solutions for every structure under consideration.

Not only the network architecture but also the learning algorithms (error backpropagation, quick-backpropagation, conjugated gradient descent, Levenberg–Marquadt), were adjusted experimentally.

The experiments were conducted with the help of the Statistica Neural Networks Intelligent Problem Solver (IPS). The IPS selects the best network structures with search algorithms that use state-of-the-art techniques to determine the selection of inputs, the number of hidden units, and other key factors in the network design. As many as 2000 experiments with different designs were conducted, and the best networks were selected. Afterwards, the most promising of the developed models were learned to obtain the best achievable performance. The best results will be discussed below.

The learning process was performed using such assumptions: the input data were divided at random into three groups: 19 samples formed the training set, 7 and 6 cases were used as the validation and test set respectively.

The learning coefficient was fixed on the value 0.1 (without changes during the learning process) and the momentum coefficient was selected on the value 0.3. Both values were selected basing on preliminary experiments.

3. Results

The best networks developed in the experiments are presented in Table 2.

Table 2. Other developed networks (examples).

No.	Structure	Learning error	Validation error	Testing error
1	RBF 16:16-4-1:1	0.3389	0.5387	0.5783
2	RBF 13:13-5-1:1	0.2965	0.5493	0.5215
3	RBF 13:13-7-1:1	0.2555	0.4821	0.4537
4	RBF 13:13-4-1:1	0.3425	0.4093	0.3936
5	MLP 16:16-10-1:1	0.4187	0.2222	2.1032
6	MLP 16:16-10-1:1	0.4641	0.1888	1.3847
7	MLP 16:16-10-1:1	0.4239	0.1763	1.6367
8	MLP 16:16-10-1:1	0.4122	0.0886	1.5568
9	MLP 16:16-10-1:1	0.5253	0.4406	0.5762
10	MLP 16:16-9-1:1	0.4735	0.3717	0.5794
11	MLP 13:13-8-1:1	0.3160	0.3898	0.5855
12	MLP 16:16-14-1:1	0.1707	0.3661	0.4699
13	MLP 16:16-12-1:1	0.0004	0.3516	1.6081
14	MLP 16:16-14-1:1	0.0051	0.2783	0.7650

The best of the developed models was a three-layer perceptron with 16 neurons in the input layer, 7 neurons in the hidden layer and 1 neuron in the output layer. The structure of the network is shown in Fig. 1.

The learning algorithm which proved to be most efficient in this model was a combination of backpropagation (applied during first 50 epochs) and Levenberg–Marquadt (applied during last 50 epochs). The process of learning is presented on the training error graph (Fig. 2).

The learning error of the trained network was 0.007773, while the validation and testing errors were $2.53 \cdot 10^{-11}$ and $1.62 \cdot 10^{-12}$ respectively. The network classified all cases correctly, which means that the specificity and sensitivity of the classification was 100%.

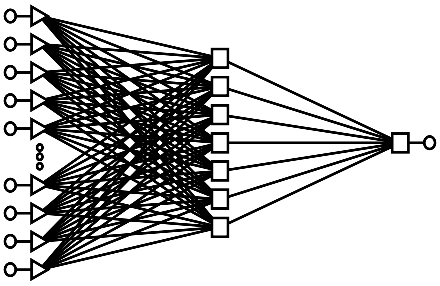

Fig. 1. Structure of the best developed multilayer perceptron.

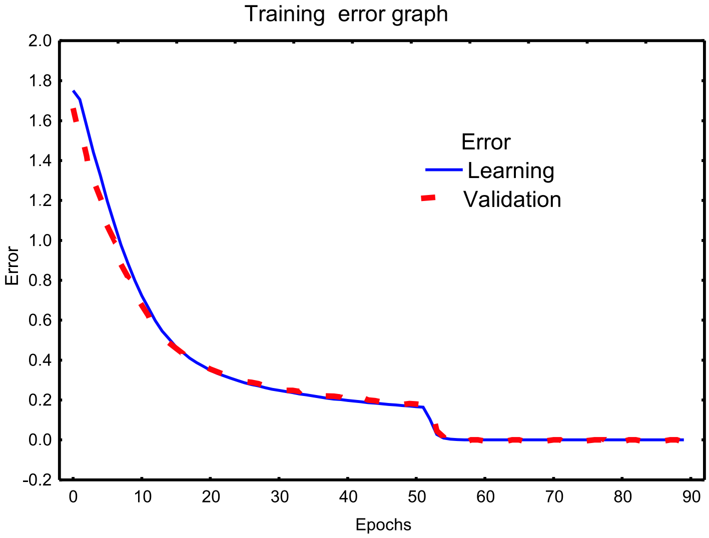

Fig. 2. Training error graph for the best developed network.

4. Discussion

The artificial neural network presented in this work achieved very good performance in classification between pathological and non-pathological speech in the samples used as the learning, validation and training set. However, there are several reasons why the research into this problem still needs to be developed.

The preliminary research project which was presented above involved a group of patients with advanced larynx pathologies. Further research is necessary to determine whether it is possible to achieve satisfactory performance of the ANN-based classification for patients at early stages of the disease. This problem is most crucial in the larynx cancer which, as mentioned before, can be cured easily in the beginning of its progression, but is characterized by high mortality when the process is advanced.

Secondly, the ANN trained to classify between pathological and non-pathological speech, would be a useful tool only if it could be used easily by any general practitioner.

As it was mentioned earlier, the voice signal applied in the presented work was recorded in a sound-treated room with noise level below 32 dB. Such conditions are hardly available in the primary health care. It is therefore necessary to prove whether the ANN can produce accurate diagnoses for voice samples recorded with a background noise level typical for a general practitioner's office.

Other authors have also made attempts to provide a possibility of preliminary voice diagnosis without personal contact with a physician. These attempts included assessment of the speech signal transmitted by the public telephone network [21]. Further research that will follow the project presented in this paper will include training of the ANN to classify the voice samples processed in a telephone network or even an internet microphone.

5. Conclusions

The very high rate of accurate predictions achieved by the ANN applied in the research suggests that the novel approach to acoustic analysis of the speech signal may improve the effectiveness of automated laryngological pre-diagnosis. The artificial neural network proved to be a promising, cheap and convenient tool for voice pathology recognition. Possible application of the presented results includes preliminary laryngological diagnosis on the primary health care level, "remote" voice diagnosis via internet or telephone network and screening (early detection) for larynx diseases, especially cancer. A widely available, easy-to-use computer system, providing sensitive and specific voice diagnosis, could improve the results of laryngological treatment.

References

[1] Umapathy K., Krishnan S., Parsa V., Jamieson D. G., *Discrimination of pathological voices using a time-frequency approach*, IEEE Trans. Biomed. Eng., **52**, 3, 421–430 (2005).

[2] Godino–Llorente J. I., Gomez–Vilda P., *Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors*, IEEE Trans. Biomed. Eng., **51**, 2, 380–384 (2004).

[3] Cheolwoo J., Daehyun K., Soogeon W., *Classification of Pathological Speech into Normal/Benign/Malignant State*, Eurospeech'99, 399–402, 1999.

[4] Boyanov B., Mitev P., Hadjitodorov S., *Voice Analysis and Screening System VASS 3.0. Center on biomedical research*, Bulgarian Academy of Sciences 2000.

[5] Bull P. D., *Lecture Notes on Diseases of the Ear, Nose and Throat*, Via Medica, Gdańsk 1999.

[6] Hadjitodorov S., Mitev P., *A computer system for acoustic analysis of pathological voices and laryngeal diseases screening*, Medical Engineering and Physics, **24**, 419–429 (2002).

[7] Hadjitodorov S., Boyanov B., Teston B., *Laryngeal pathology detection by means of class-specific neural maps*, IEEE Trans Inf. Technol. Biomed., **4**, 1, 69–73 (2000).

[8] SZALENIEC J., MODRZEJEWSKI M., WSZOŁEK W., *Application of acoustic analysis of speech signal for evaluation of intubation-related damages of the speech organ*, 3rd International Workshop MAVEBA Proceedings 2003, pp. 269–272, 2003.

[9] SZALENIEC J., MODRZEJEWSKI M., WSZOŁEK W., *Research on the influence of endotracheal intubation on the speech signal. Speech analysis, synthesis and recognition in technology, linguistics and medicine*, Materiały konferencji naukowej, Szczyrk 23–26.09.2003, pp. 127–133, 2005.

[10] TOST T., HERBERGER T., FLEMMING G., HIRCHE H., HEISE T., BENSCH J., MUEHLE V., Samplitude Project V5.55. SEK'D 2000.

[11] KOIKE Y., *Application of some acoustic measures for the evaluation of laryngeal dysfunction*, Studia Phonologica, **7**, 45–50 (1971).

[12] KOIKE Y., *Acoustic measures for detecting laryngeal pathology*, Acta Otolaryngol., **84**, 105–117 (1977).

[13] TAKAHASHI H., KOIKE Y., *Some perceptual dimensions and acoustic correlates of pathological voices*, Acta Otolaryngol., **338** (Suppl), 2–24 (1975).

[14] MITEV P., HADJITODOROV S., *A method for turbulent noise estimation in voiced signals*, Med. Biol. Eng. Comput., **38**, 625–631 (2000).

[15] YUMOTO E., GOULD W., BAER T., *The harmonics-to-noise ratio as an index of the degree of hoarseness*, J. Acoust. Soc. Am., **71**, 1544–1550 (1982).

[16] YUMOTO E., *The quantitative evaluation of hoarseness. A new harmonics to noise ratio method*, Arch. Otolaryngol., **109**, 48–52 (1983).

[17] AWAN S., FRENKEL M., *Improvements in estimating the harmonics-to-noise ratio of voice*, J. Voice, **8**, 255–259 (1994).

[18] QI Y., WEINBERG B., BI N., HESS W., *Minimising the effect of period determination on the computation of amplitude perturbation in voice*, J. Acoust. Soc. Am., **97**, 2525–2532 (1995).

[19] HILLENBRAND J., HOUDE R., *Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech.*, J. Speech Hear Res., **39**, 311–321 (1996).

[20] StatSoft, Inc. STATISTICA (data analysis software system), version 6 (2001).

[21] CHEOLWOO JO, KWANGIN KIM, SOOGEON WANG, *Screening of pathological voice from ARS using neural networks*, Proceedings of Maveba 2001, vol. 1, 96–97, 2001.