Acoustic and Biomedical Engineering

Signal Processing and Analysis of Pathological Speech Using Artificial Intelligence and Learning Systems Methods

W. Wszołek[a], A. Izworski[b] and G. Izworski[a]
AGH — University of Science and Technology
[a]Faculty of Mechanical Engineering and Robotics
[b] Faculty of Automatics and Biomedical Engineering
al. A. Mickiewicza 30, 30-059 Krakow, Poland

In this paper, selected results are presented of research which is carried on for over a decade and covers valuation of chosen signal processing methods suitable to analyze and valuate pathological speech. This valuation is necessary during solving many medical diagnostics problems and when planning therapy and rehabilitation of certain types of diseases. All presented examples are used in clinical practice in the area of dentistry, dental surgery, otolaryngology and most of all, in phoniatrics and speech correction.

DOI: 10.12693/APhysPolA.123.995
PACS: 43.70.+I, 84.35.+i

1. Introduction

The issue of computer-based analysis and automatic evaluation of pathological speech has been taken up in many publications by authors of this paper from different points of view. There was emphasis on specificity of this subject in the area of preliminary (non-standard) methods of processing considered audio signals. They were studied in the range of extraction of given attributes which are basis of evaluation process and classification of particular patients. Finally, there was also research in the range of diagnostic and prognostic decisions methods concerning specific diseases and specific treatment methods.

Speech acoustics provides wide array of evaluation methods of speech signals quality, enabling its multidimensional analysis including visualization of results and characteristics showing how it changes during articulation [1, 2]. However, direct analysis of such characteristics is very complicated and requires experience, especially in case of pathological speech analysis. It leads to development and constant updates to methods of process control, analysis process, and speech signal recognition, and research results are shown in many papers, including this publication. Pathological speech analysis often leads to classification and determination of analyzed signals deformation type which may be in direct relationship with anatomical reasons and conditioning of considered disease. Pathological speech signal classification may make diagnosis easier by pointing most probable reason of pathological deformation of the speech signal. Such classification can also help with choosing optimal therapy and determining of rehabilitation advices.

Processing, analysis, classification, and recognition methods of speech signal have been seemingly known for many years, because it is possible to easily find many items in the literature which refer to the above-mentioned terms and present results of both fundamental researches and many application publications. Unfortunately, classical methods mentioned earlier generally skip the problem of speech signal evaluation, particularly in the context of supporting diagnostic, therapy optimization and rehabilitation monitoring. It has to be mentioned that for psycho-sociological reasons, in considered task there is a need for searching such audio signal evaluation methods which to a maximum degree fit the process of natural speech perception. It is crucial that evaluation and analysis results obtained by objective methods (using techniques of digital acoustic signal processing and cybernetic techniques of pattern recognition) maximally fit subjective evaluations made by humans who have to communicate with person suffering disability of pathological deformation of speech signal. By applying the criteria of cybernetics and psychoacoustics, it will be possible to use evaluations provided by technical system during planning therapy and rehabilitation for persons suffering from pathological speech. Artificial intelligence methods and learning systems are particularly interesting in this case. They are naturally suited for modeling (greatly simplified, of course) of chosen fragments of nerve system ability to adapt and learn of which is commonly known [3].

2. Problem definition and research material

Very often the process of recognition and evaluation of acoustic signal of pathological speech recognition consists in arranging results, obtained by research of acoustic images, to one of classes usually unknown earlier. It is worth to notice this particular detail and emphasize its meaning: during recognition of speech to e.g. control particular machines or devices by voice, collection of recognizable picture classes is given from the start. Similarly, in tasks related to speaker recognition, it is often the case of pre-determined collection of patterns. On the contrary, in recognizing deformation classes of pathological speech signal there are not any constant indices of deformed speech image classes. Additionally, the number of indicated classes cannot be pre-determined. The only

thing that can be said about any of classes is that their acoustic images are somewhat similar. In the considered case, recognition is connected with detection and specification of a feature that assigns given individual images into right groups number and essential characteristics of which are not known from the start. It is, which has to be emphasized again, a task apparently different from audio images recognition in the context of semantical identification of statement or in the context of recognition of speaking person.

The most common case of voice channel pathology (especially in children) is the cleft palate (med. palatschisis) direct effect of which is permanent and uncontrolled acoustic feedback between oral and nose channel revealed by forced nasalization of initially non-nose speech sounds. Cleft palate, both primarily and secondary, is a hard, unsightly malformation, most commonly encountered in face's fracture. An example of cleft palate is shown in Fig. 1.

Fig. 1. Example of cleft palate [4].

Pathology forming after surgical procedure manifests itself with less elastic tissue forming speech organs, and in many cases also deformed from the physiological norm. Rhinolalia, typical for cleft palate, is caused mostly by increased distance from back wall of throat and palate [5]. These deformed elements of speech organs are not capable of shaping sound as subtly as healthy organs, so they cannot ensure right correlation between voice articulated by child and something that forms a sensual experience in the ear of person who receives this sound.

Examination was carried out on 60 children divided into four groups by kind of defect operated by surgeon because of preliminary and secondary cleft palate in the Clinic of Plastic Surgery (Medical University of Gdańsk). Children chosen for the research were aged 4–7. Additionally, control group of pre-school children was examined.

Medical evaluation of each child was conducted by a team of multi-specialization jaw and face orthopedists of the Institute of Stomatology (Medical University of Gdańsk). They have performed evaluation of teeth contact defect and the effect of given disorder on child's correctness of speech based on examination of facial features, diagnostic models, and functional examination of chewing organ. Then, children were examined by a laryngologist-phoniatrist who, apart from evaluation of aesthetical-morphological changes, conducted also subjective examination of speech. Objective examinations of speech signal were made in the Department of Mechanics and Vibroacoustics at the AGH University of Science and Technology in Kraków cooperating with the Broadcasting Station of Polish Radio in Gdańsk where phonetic research material was registered by Nagra IV SJ measurement recorder in studio conditions.

Because of looseness of the sample and children becoming tired quickly, acoustic sample size was reduced to a necessary minimum. Each child (both in all four examined groups and in the control group) was instructed to pronounce the same sentence:

Rudy pies goni szybko kota. (Polish)

Red dog chases cat quickly. (English)

This sentence has been chosen considering acoustic needs, because it contains mean spectrum of /s/ character for correct speech (control group) and for pathological speech (groups of children operated because of cleft palate), as well as all phonemes, where deformations caused by the cleft palate were expected. It also meets psychological needs (sentence refers to situation that is easy to understand and inspires child's imagination, so repeating it – often many times – is less tiring to the patient). Figures 2 and 3 present examples of the averaged spectral sounds /s/ for correct speech (control group) and for pathological speech, respectively.

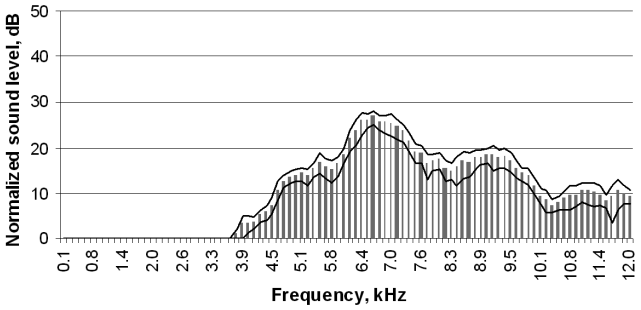

Fig. 2. Example of mean phoneme spectrum /s/ — standard, children.

As one can see from Fig. 2, the entire spectrum of the phoneme model /s/ is contained in the frequency range from 3.5 kHz to 12 kHz. The spectrum of the phoneme /s/ in speech pathology is shifted towards lower frequencies with a decrease of level of the phoneme-band model.

By observing change of signal in particular ranges of frequency, it is possible to determine if a sound is correctly formed by speech organs and whether these frequencies are mixed together, or, generally speaking, blurry. In the first picture with correct speech (Fig. 2) many subtle details can be seen indicating the precision of articulation, and in the following pictures with pathological speech (Fig. 3), in the range of the same sound events in pathological speech, blurred time and frequency

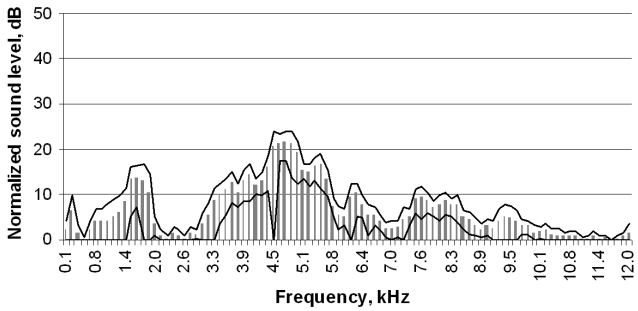

Fig. 3. Example of mean phoneme spectrum /*s*/ — cleft palate preliminary and secondary, absolute, both sided.

subtle spectrum structures can be clearly seen [6].

Figure 3 shows that the existing low-frequency (up to about 3 kHz) spectrum bands may indicate potential defects after surgery.

Vowel sounds, as the results of formant structures, usually depict only deformations of nose channel geometry. Deformations of consonants phonation show potential defects or degree of articular defect remaining after surgical treatment. Deformation of consonant sounds allows trying to interpret the degree of functional damage, because in the range of consistent trembling sound, the process of articulation is a dynamic process. If noisy, it is correctness of consistency forming, where noise and turbulence appear. In Fig. 4a, a spectrogram of word "szybko" ("quick") is shown as pronounced by a child speaking correctly [7]. Fig. 4b shows the same word pronounced by another child with right-sided cleft palate after surgical treatment. Spectrograms (dynamic spectra) show how during the process of creating a quote its spectral composition was changing over time [6].

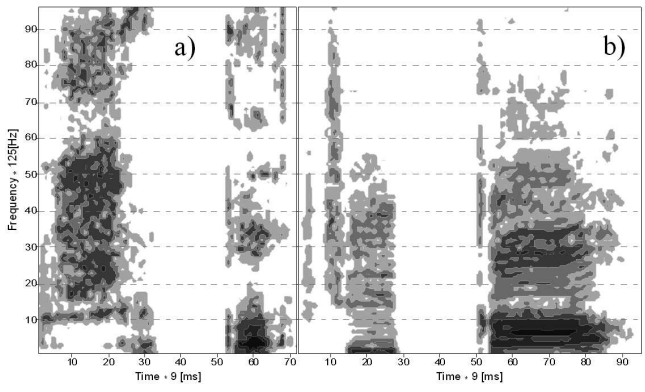

Fig. 4. Word "*szybko*": (a) standard; (b) cleft palate.

3. Artificial intelligence methods

Lack of specific means allowing fully algorithmic projection of speech signal into a numerical value, being the measure of degree of its psycho-acoustic acceptability, leads to search for such a model form of signal processing and analysis in which the process of creating suitable rules and new methods of signal transformations could be made mostly automatic. Particularly interesting are neural networks which are naturally suited for modeling (greatly simplified) chosen fragments of nerve system and which ability to adapt and learn is commonly known [3, 8–10].

In the described research, neural networks were used to analyze, evaluate, and recognize pathological speech signal. It can be considered, to some extent, as a cybernetics approach to modeling fragments of nerve part of real biological hearing analyzer. Very often the process of recognition of pathological speech acoustic signal deformations consists in arranging acoustic pictures obtained via research to one of classes, usually unknown earlier.

On the other hand, in case of recognition of pathological speech signal deformation classes there are no constant indices of deformed speech image classes. Additionally, it cannot be predetermined how many classes can be distinguished. Each class can be only characterized as having some similarity in terms of acoustic images.

Quality estimation of deformation degree of pathological speech often requires presenting it as properly designed sound image, instead of hearing it. By hearing it is possible to obtain information if speech is correct or deformed. However, it is not possible to evaluate it quantitatively. If speech is transformed to image by proper visualization form, a human can easily compare images and determine the state of deformation [9].

4. Results analysis

Object can be described by parameters that reflect its properties. Making assumption that objects (their states) are the study subjects which are sources of acoustic signals in measurable bandwidth, it can be stated that reflection of object's state is a measurable acoustic signal. This signal, as some reflection of object's state, usually contains more information than is actually needed [2]. In this paper, the problem is approached in context of evaluation of speech deformation degree, considering different reasons of pathological speech and leading to evaluation of the degree of deformation. Many scientific researches, including the one conducted by present authors, have led to conclusion that the most important (and the most difficult) element of research preceding using speech as source of medically useful diagnostic and prognostic information was finding and describing signal parameters which were as independent from both context and each person's voice attributes as possible. Additionally, the desired signal attributes have to be maximally sensitive to its even tiny deformations in the aspect connected with build and functionality of speech signal generators (larynx and choke being the source of noisy speech elements) and with structure of speech route used in articulation. Most parameters come from frequency-magnitude characteristics of the signal. Following this approach, the

structure of discussed vector of attributes, which is an input to analyze and identification algorithms, can be expressed as [6]

$$\langle f_1, f_2, f_3, ..., f_n \rangle = X_1, \tag{1}$$

where f_i is the mean magnitude of i-th frequency band of dynamic spectrum*, and n is the number of spectrum lines used in further analysis. This number depends on the highest frequency of the analyzed signal. In the present authors' work it is assumed to be $n = 96$.

One of key observations of pathology effects in the area of speech signal articulation leads to statement that in pathological speech, the signal energy distribution in particular spectral band is completely different than in the correct speech pattern. This feature is highlighted by coefficients of relative power WS proposed by authors as a new coordinate of the attribute vector. This remark led to using in the research a vector of acoustic attributes with following structure:

$$\langle M_0, M_1, M_2, WS_F, WS_1, WS_2, WS_3 \rangle = X_2, \tag{2}$$

where WS_F is the relative power coefficient determining the ratio of acoustic power in reference bandwidth of phoneme (set from statistical research on non-deformed speech) to the signal acoustic power in whole analyzed signal band of pathological speech, and WS_k is the relative power coefficient determining the ratio of acoustic signal in i-th band ($i = 1, 2, 3$) to acoustic power in the whole analyzed band (matching dedicated bands borders is one of main research problems). General form of the signal power coefficient W_k can be expressed as

$$W_k = \frac{\sum_{j=1}^{m} \sum_{i=p}^{i=K} G(t_j, f_i)}{\sum_{j=1}^{m} \sum_{i=f_d}^{i=f_g} G(t_j, f_i)}. \tag{3}$$

General form of the spectral momentum of m-th grade in j-th moment in time can be described as

$$M_m(j) = \sum_{i=f_d}^{i=f_g} |G(t_j, f_i)| \, [f_i]^m \,, \tag{4}$$

where $G(t_j, f_i)$ is the dynamic spectrum (time-frequency), f_i is the middle frequency of *i-th* band, t_j is the j-th time interval that momentary spectrum is computed for, and f_d, f_g are the bottom and the top band frequencies for which spectral moment is computed, respectively.

In further research parameters groups described by both (1) and (2) are used. Parameterized momentary spectrums that compose analyzed quotes are used as input signals to artificial neural networks of different kinds.

For processing the feature vectors describing the signals of speech (pathological speech for research purposes and correct speech for purposes connected with its analysis and recognition) into a form which is optimal for interpretation by a human (physician), Kohonen type neural networks (known also as Self Organizing Maps) have been applied, which are commonly recognized as a highly useful tool for presentation of multidimensional phenomena in the form of plots and one- or, more often, two-dimensional drawings.

By regarding an arbitrary statement as a sequence of momentary spectra presenting the state of the speech organs in consecutive stages of the articulation process, a sequence of "winning" neurons is obtained from the Kohonen network. By connecting these neurons during the signal visualization, a characteristic trajectory is obtained for every statement, clearly distinct for signal obtained from a person with correct articulation and considerably different for particular forms of speech pathology. Visual evaluation of the trajectory shape can provide a convenient basis for evaluation of the degree of the speech signal pathology by a physician, and in the long run this form of the signal presentation (as a specific "acoustic image") can be the basis for the automated diagnosis and automated attempts to evaluate the signal quality.

Images of pathological phones presented against the background of summarized images of the reference phones (obtained by overlapping). The image of the pathological phone is plotted in black, while the images of the reference phones are plotted in green.

The results of visualization of the speech signal samples, obtained by means of Kohonen type neural networks and shown in Fig. 5 provide a possibility of considerable reduction of the amount of information being presented, while preserving the proper discrimination between samples of correct and pathological speech.

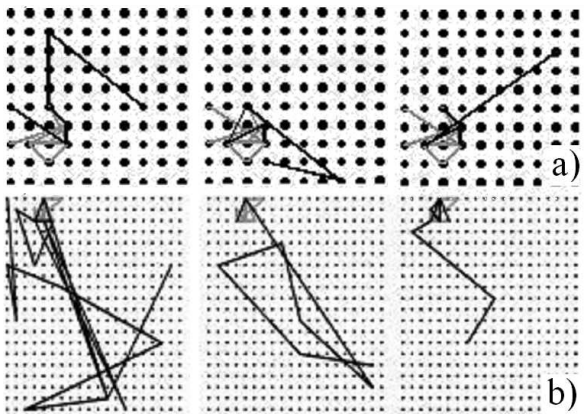

Fig. 5. (a) Sound /s/, cleft palate; (b) sound "sz" /S/, incomplete cleft palate. Black color — correct, grey color — pathological [6].

For tasks concerning the analysis and recognition of the pathological speech it is often more important to provide the respective person (e.g. physician) with necessary indications required for qualitative evaluation of the speech than to provide its perfect automated recognition. By

* Most often coordinates of this attribute vector are product of frequency analysis with constant band width which means $\Delta f_i =$ const for the whole analyzed band.

ear it is easy to receive an information that a particular speech is correct or that it is deformed, while an attempt to provide a quantitative evaluation usually fails. If the speech is transformed to a proper visualized form, then, by comparing such images, man is able to evaluate the degree of its deformation.

Figure 5 shows that for a variety of sounds (/s/ and /S/) and cleft palates, the drawn shapes have different global trajectory.

To recognize speech pathology induced cleft palate different in this paper uses neural networks. The network's task was to transform the parametrized speech signal samples into the values which can be interpreted as the evaluation of the speech signal deformation level.

In the described study, a triple-layer network is used of a *feedforward* organization type. The topography of the network included the input layer (96 or 7 neurons), one hidden layer (49 and 7 neurons), and the output layer 5 neurons. The problem strictly connected with the choice of the network is the decision about the size of (one or more) hidden layers [11]. The rules governing the choice of the number of elements in the hidden layers have been presented e.g. in [12]. The results for the selected simulations are presented in Figures 6, 7, and 8, where C_o denotes the cleft palate, primary and secondary total bilateral, C_{lp} — cleft palate total primary and secondary left- or right-hand-sided, C_a — total secondary cleft palate, I_e — incomplete secondary cleft palate.

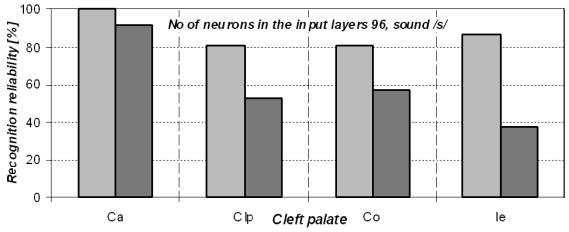

Fig. 6. Percent of recognition in class function of palate for "s" /s/ sound.

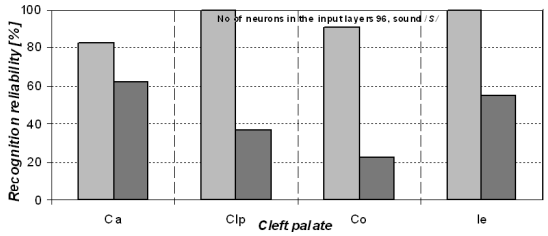

Fig. 7. Percent of recognition in class function of palate for "sz" /S/ sound.

Based on results presented in Figures 6, 7, and 8 it can be said with high probability that it is possible to classify pathological speech signal due to the type of cleft. Statistically, the highest percentage (92% for the sounds

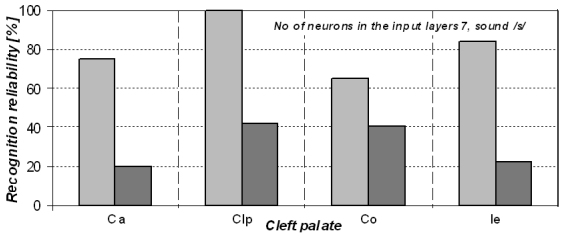

Fig. 8. Percent of recognition in class function of palate for "sz" /S/ sound.

/s/ and 63% for the sounds /S/), the correct classification was achieved for C_a. Lighter bars on the graphs represent the training data and darker ones the test data. This relatively low percentage of correct diagnoses in the other classes (types of clefts) can be explained by the fact that the allocation of patients to the class in terms of healthcare does not necessarily coincide with the acoustic speech signal distortion. The effects of post-operative rehabilitation and improve speech signal meant that speech pathology was close to the reference speech signal.

5. Discussion

The usefulness of sets of features (parameters of the signal description) dedicated for pathological speech evaluation tasks revealed as a result of the present study confirms the opinion that dedicated features should be used in the considered task and not the widely used parameters and signal description elements applied in the analysis and recognition of the normal speech signal. The revealed advantage of the neural network technique in the solution of problems formulated here is an obvious consequence of the generally known usefulness of this tool in tasks characterized by great shape complexity for the areas of the particular considered classes in the feature space. This result indicates a possibility of application of the neural network technique as a favorable alternative for other techniques (like pattern recognition or statistical methods), however this was not the primary objective of the study. The crucial role of the proper preparation of speech signal parameters is worth stressing once more, because in the case when the problem cannot be parametrized in a way ensuring a proper resolution of areas belonging to particular classes in the feature space, the results obtained by any methods (particularly the pattern recognition methods) are highly unsatisfactory.

6. Conclusion

The results quoted in the article confirm the assumption that the technique of neural networks can be a useful tool in evaluation of the pathological speech.

Presented examples can be practically used in clinical applications in areas of stomatology, jaw surgery, otolaryngology, and first and foremost in phoniatry and speech treatment. However, when qualitative assessment

of the degree of distortion of the speech signal is required (for example, in order to follow rehabilitation) the use of Kohonen network (SOM class) is preferred. In fact, these networks do not guarantee direct determination of the measure of the signal deformation degree (for this purpose, MLP network can be used for example). However, due to very clear visualization of the speech signal deformation degree obtainable by use of these networks, – it is possible to considerably support the work of physician and acquiring much more adequate (compared to other techniques) predictions of compensation degree of pathological speech signal deformation by using surgical operations, providing rehabilitation, or by referring to prosthesis techniques. It is worth stressing that in spite of the basic nature of the study directed towards the analysis of properties of various elements in the feature space, our proposal seems to be very helpful in practical applications such as selection of dentures and orthodontic apparatuses or qualification of patients for specific types of surgical treatments in order to minimize voice deformation after the treatment.

Acknowledgments

This study was supported by AGH grant No. 11.11.120.612.

References

[1] R. Tadeusiewicz, *Speech Signal*, WKŁ, Warszawa 1988, (In Polish).

[2] Cz. Basztura, *Acoustical Sources, Signals and Images*, WKŁ, Warsaw 1988, (in Polish).

[3] R. Tadeusiewicz, A. Izworski, in: *ICONIP 2006*, Springer, Berlin 2006, p. 211.

[4] http://www.cleftadvocate.org/galleryk.html.

[5] W. Wszołek, M. Modrzejewski, R. Tadeusiewicz, T. Wszołek, in: *Proc. EMBEC'02 Vienna 2002*, Eds. H. Hutten, P. Krösl, 2002, Vienna 2002, p. 536.

[6] W. Wszołek, *Methods of cognitive categorization for analysis and classification of selected cases of pathological speech, Monograph 232*, AGH Publishing, Kraków 2011, (in Polish).

[7] R. Gubrynowicz, *J. Phonet.* **14**, 525 (1986).

[8] W. Wszołek, T. Wszołek, in: *Intelligent Information Processing and Web Mining*, Eds. A. Kłopotek, S. Wierzchoń, K. Trojanowski, Springer, Berlin 2004, p. 609.

[9] W. Wszołek, M. Kłaczyński, in: *Fundamentals of Biomedical Engineering*, Eds. R. Tadeusiewicz, P. Augustyniak, AGH Publishing, Kraków 2009, p. 339, (in Polish).

[10] S. Howard, *The Cleft Palate-Craniofacial Journal* **50**, 207 (2013).

[11] K. Fukushima, N. Wake, *IEEE Trans. on Neural Networks* **2**, (1991).

[12] A.I. Maren, C. Harston, R.M. Papm, *Handbook of Neural Computing Applications*, Academic Press, London 1990.