A Vibroacoustic Model of Selected Human Larynx Diseases

Zbigniew Witold Engel

Central Institute for Labour Protection – National Research Institute, Poland
Department of Mechanics and Vibroacoustics, AGH University of Science and Technology, Kraków, Poland

Maciej Kłaczyński
Wiesław Wszołek

Department of Mechanics and Vibroacoustics, AGH University of Science and Technology, Kraków, Poland

With the present development of digital registration and methods for processing speech it is possible to make effective objective acoustic diagnostics for medical purposes. These methods are useful as all pathologies and diseases of the human vocal tract influence the quality of a patient's speech signal. Diagnostics of the voice organ can be defined as an unambiguous recognition of the current condition of a specific voice source. Such recognition is based on an evaluation of essential acoustic parameters of the speech signal. This requires creating a vibroacoustic model of selected deformations of Polish speech in relation to specific human larynx diseases. An analysis of speech and parameter mapping in 29-dimensional space is reviewed in this study. Speech parameters were extracted in time, frequency and cepstral (quefrency) domains resulting in diagrams that qualified symptoms and conditions of selected human larynx diseases. The paper presents graphically selected human larynx diseases.

speech analysis   pathological speech   surgical treatment

1. INTRODUCTION

Larynx diseases, which are the main cause of dysphonia, are increasingly observed in laryngologic and phoniatric clinical practice. Lack of voice hygiene, which is related to the progress of civilisation, is considered the main factor leading to these diseases: an increasing number of people use their voice organ, the larynx, in their professional work. Harmful factors, such as nicotine, alcohol, unhealthy diet, allergies and chemical substances, are important causes of speech pathologies [1, 2]. In Poland, diseases of the voice organ constitute over 20% of all cases of occupational diseases. In addition, it is estimated that half of the number of people complaining of dysphonia suffer from hypertrophy changes on their vocal folds, such as laryngeal polypus, Reinke's oedema, vocal cord nodules, laryngeal papillomatosis, coccus and contact ulcer. Nearly 60% of all malignant diseases localised in the area of the head and neck are laryngeal cancers.

An assessment of pathological speech is necessary in medical diagnostics as well as in planning therapy and rehabilitation. This assessment should indicate the degree of signal deformation. Development of a quantitative measure (a scalar measure would be best) of the

difference between pathological and normal voice, which can be considered a standard, is the purpose of this study.

The condition of the larynx is established after detailed laryngological and glottographic examinations. However, such examinations are invasive, time consuming for specialists, troublesome for patients and expensive. That is why attention was directed towards professional methods of conversion, analysis, classification and detection of acoustic signals. An acoustic signal, in addition to semantics and individual voice features, contains information on the health of the voice tract and of the organs co-operating in speech production. Changes in the voice signal can be estimated subjectively by a physician. However, for precise and objective diagnostics of speech impairments and supervision of a rehabilitation process, it is much better to employ professional objective methods based upon conversion and analysis of an acoustic signal.

Articulation of speech and its pathological deformation were the object of the study; they are partially presented here. This includes tools and techniques used to detect deformations in the signal vocalised by a sick person and a comparison with correct speech produced by a healthy person [3]. The most important and, at the same time, the most difficult element of investigations—prior to practical utilisation of human speech as a source of useful medical diagnostic and prognostic information—is to select and describe those signal parameters that are as independent as possible from speech content and from individual properties of the assessed voice. In addition, the parameters of the speech signal must be maximally sensitive to speech deformations, even slight, corresponding to the structure and functioning of parts of the vocal tract generating the speech signal (both sound and noise), which is directly related to the structure of the voice tract used in the articulation process.

Development of approximate vibroacoustic models to represent selected larynx diseases requires mapping, visualisation and quantitative assessment of the degree to which speech signals are deformed. Those models consist of vectors of acoustic parameters of a deformed speech signal presented graphically.

2. MATERIAL

This study set out to assess the suitability of selected methods for mapping speech signals in diagnostics of pathological speech. Methods were tested on pathological and correct speech samples constituting a matched reference system. Analysis of pathological speech samples was part of a general investigation on pathological speech. The study was conducted in co-operation with the Chair and Clinic of Otolaryngology at Collegium Medicum, Jagiellonian University (CM UJ), in Cracow, Poland. The applied test was arranged and proposed by a group of otolaryngologists and phoniatrist experts. It had been used in earlier research [4, 5, 6, 7, 8].

A recording of sound pressure time waveforms, $p(t)$, of a speech signal was made in the anechoic chamber of the Department of Mechanics and Vibroacoustics, AGH University of Science and Technology, and in a room for hearing examinations and audiometric measurements of the Chair and Clinic of Otolaryngology, CM UJ. The recording system consisted of a professional digital magnetic recorder (PDR 1000; HHB Communications, the UK) for recordings in the frequency range from 20 to 20 000 Hz and a dynamic range exceeding 80 dB.

The patients had disease changes in their vocal folds, glottis and larynx. The database of deformed speech signals comprised recordings from 60 patients treated in the Clinic. Data were collected at three stages of the patients' medical treatment:

- the first recording—before surgery, ~7–14 days before treatment;
- the second recording—during early check-up, ~14–30 days after surgery; and
- the third recording—during late check-up, ~90 days after surgery.

In addition, voices of 36 persons were recorded before hospital treatment only, which corresponded to the first recording. During the same period normal voices of 128 persons, both

male and female, with no pathology of voice, were recorded as a reference for standard Polish.

3. METHOD

An unambiguous recognition and precise evaluation of the proposed set of parameters considered essential for representing pathological speech samples was difficult as phonetic data taken from the subjects differed from one another also with respect to aspects not undergoing an analysis (e.g., different speaking rates). Similarly reference speech material of healthy people with correct (standard) articulation varied. Our attention was focused on preliminary transformation of convert speech waveforms into a set of parameters whose values constituted a basis for a diagnostic description of the patient's disease. To construct models of speech deformation, it is necessary for a developed set of parameters describing the recorded phonetic samples to be arranged in an appropriate structure, a vector of features. In the model, various parameters of the acoustic signal and parametric space constituting sound patterns are assumed to be features or symptoms for detection of a given glottal disease.

Analysis of speech signals in the time domain was performed with the zero crossing analysis (ZCA), in the frequency domain with the short-term Fourier transform (STFT) and with the use of a mel scale, mel-frequency cepstral coefficients (MFCC). The ZCA, preceded with appropriate low-pass filtering of the recorded signal, estimated the course of the fundamental laryngeal tone of vocal fold vibration. This tone varies from cycle to cycle. To describe laryngeal tone fluctuations, in addition to an average value of the fundamental frequency, the mapping uses jitter (J) and shimmer coefficients (S). An average fundamental frequency of laryngeal tone is given in Equation 1:

$$\bar{f}_0 = \frac{\sum_{i=1}^{N} f_i}{N}, \qquad (1)$$

where f_i—frequency in the ith cycle.

The jitter coefficient (J) represents average departure of the fundamental frequency from frequency $\bar{f}_0$ given in Equation 1 in consecutive cycles. It is given in Equation 2:

$$J = \frac{\sqrt{\frac{1}{N-1} \sum_{i=1}^{N-1} (f_i - f_{i-1})^2}}{\frac{1}{N} \sum_{i=1}^{N-1} f_i} \bullet 100\%. \qquad (2)$$

The shimmer coefficient (S) determines variations of the fundamental tone amplitude from the average amplitude in consecutive cycles:

$$S = \frac{\sqrt{\frac{1}{2N-1} \sum_{i=1}^{2N-1} (A_i - A_{i-1})^2}}{\frac{1}{2N} \sum_{i=1}^{N-1} A_i} \bullet 100\%, \qquad (3)$$

where the average fundamental tone amplitude is given by

$$\bar{A}_0 = \frac{1}{N} \sum_{i=1}^{N} A_i. \qquad (4)$$

Signal conversion from the time domain into the frequency domain with STFT [9, 10] makes it possible to represent variations of the speech signal in time in the form of a dynamic time-frequency spectrum $G_n(t, f)$. Dynamic spectra obtained in this way were often used directly as vectors of essential features in the analysis and estimation of pathologic speech signals [5, 6, 11, 12]. In particular, the preliminary study showed the necessity to reveal the reason for the incorrect time–frequency structure of voice signals generated with various pathologies. In order to standardise the measurement and to assure mutual comparability of the results, the same system of signal conversion and processing was used. The amplitude resolution quantum, $\Delta s = 0.2$ dB, was assumed; the signal was discretised uniformly within the frequency range from $f_l = 125$ Hz to $f_h = 1\,200$ Hz, using frequency step $\Delta f = 125$ Hz at every $\Delta t = 8$ ms. Having determined the spectrum of a speech signal $G_n(t, f)$ it was possible to estimate parameters describing its envelope. This was done with spectrum moments, where the moment of the mth order in the jth time instant, t_j, for the nth recording is given by [11]

$$M_n(m) = \sum_{i=f_1}^{i=f_h} |G_n(t_j, f_i)| [f_i]^m, \qquad (5)$$

where $G_n(t_j, f)$—frequency spectrum in the jth time instant, t_j, f_i—mid-band frequency of the ith band selected for frequency analysis, f_1, f_h—lower and higher limiting frequencies for the frequency band in which the spectrum moment was determined on the discrete frequency scale.

Formants, the next parameters describing the shape of the spectrum, were defined as local maxima in the amplitude frequency spectrum envelope. Frequencies at which they occur are called formant frequencies (F_1, F_2, F_4, ... , etc.). The general form defining formants is as follows [11]:

$$F_{ij} \equiv f_{ij} \Leftrightarrow \left[\left(\frac{\partial G_n(t_j, f)}{\partial f} = 0 \right) \wedge \right.$$
$$\left. \left(\frac{\partial^2 G_n(t_j, f)}{\partial f^2} < 0 \right) \wedge \left(f_1 \le F_{ij} \le f_h \right) \right], \qquad (6)$$

where i—formant number $i = 1, 2, 3, 4, ...$, etc., at the jth time instant, t_j.

In addition to the speech parameters just described, the so-called coefficients of relative power were estimated in the time–frequency domain. These parameters determine the ratio of signal power in the selected frequency band $f\,0$ <f_1, f_h> to wide-band signal power. It is proposed that the band selection will follow the structure of formants for Polish-language vowels, using set vowels recorded from the patients. The power coefficient of the 1st, 2nd and 3rd power signal is described with Equation 7:

$$W_m = \frac{\displaystyle\sum_{t=t_b}^{t_e} \sum_{f=f_1}^{f_h} G_n(t, f)}{\displaystyle\sum_{t=t_b}^{t_e} \sum_{f=125}^{f_e=12\,000} G_n(t, f)}, \qquad (7)$$

where t_b, t_e—beginning and end of speech signal recording, f_1, f_h—corner frequencies for selected frequency band.

Features that make it possible to recognise sound patterns in pathological speech should be derived from the amplitude–frequency spectrum of the acoustic signal [6, 13, 14, 15]. An experienced physician–diagnostician is able to recognise deformations in the patient's vocal tract already during a short conversation during a check-up visit. This prompted these authors to map the speech signal in the frequency domain using the mel scale [7, 13, 16, 17, 18]. Moreover, the mel-scaled cepstral analysis is based on the response of human hearing [9, 16, 17, 18]. The following procedures of digital signal processing were used in calculating MSCC:

- Hamming's time-weighting window (length of $N = 384$ samples);
- the 384-point digital Fourier transform (DFT) calculated every 8 ms (signal sampling frequency of 48 000 Hz);
- rescaling the frequency scale into a mel scale, according to Equation 8 [9, 18]:

$$\text{mel}(f) = 2\,595 \bullet \log(1 + \frac{f}{700}), \qquad (8)$$

- mel filtering, i.e., spectral conversion using triangular band pass filters corresponding to a mel scale (by summing weighted spectral lines). The number of filters in the set is $N = 12$;
- calculation of cepstral coefficients in mel-scale bands, MFCC, as a discrete conversion of logarithm cosinuses of the parameters of filter data, according to Equation 9:

$$C_n = \sqrt{\frac{2}{N}} \sum_{i=1}^{N} \log(s_i) \bullet \cos\left[\frac{\Pi \cdot n}{N} (i - \frac{1}{2}) \right], \quad (9)$$

where C_n—nth cepstral coefficient, s_i—ith coefficient obtained from signal conversion with a set of filters, N—the number of filters in the set, $N = 12$.

Figures 1a and 1b show sample time variation of 12 cepstral coefficients along the mel scale, determined for a prolonged phoneme /i/ (as in, e.g., /pit/) recorded before and after surgery.
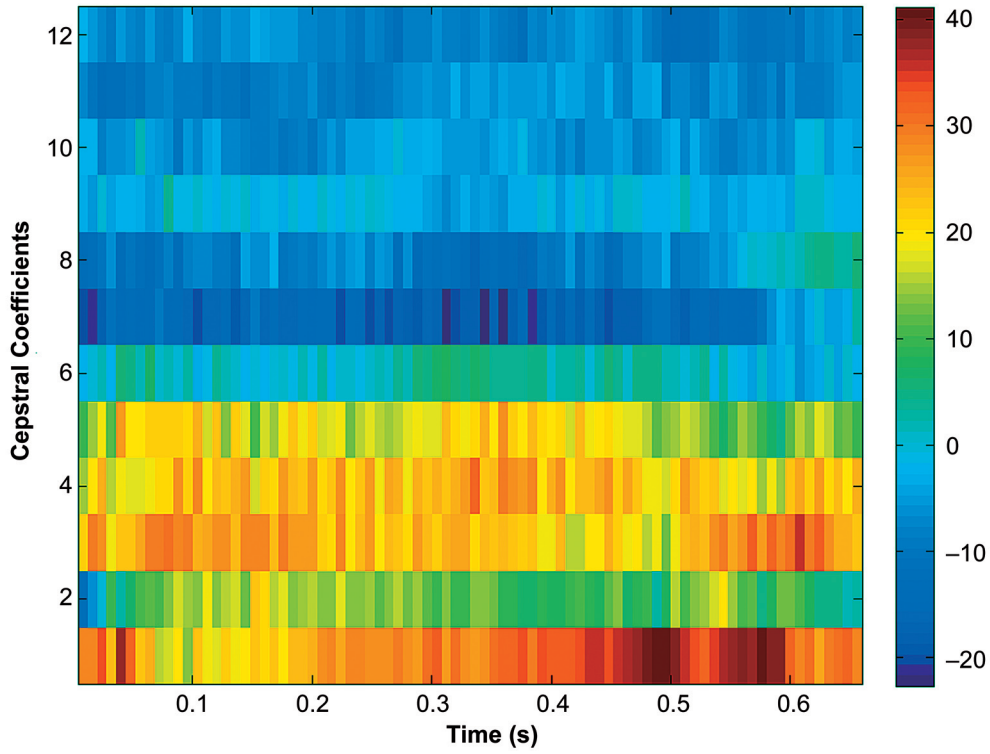
Figure 1a. Mel-frequency cepstral coefficients (MFCC)—deformed male speech (laryngean polypus); vowel /i/ (e.g., /pit/) with prolonged phonation before surgery.
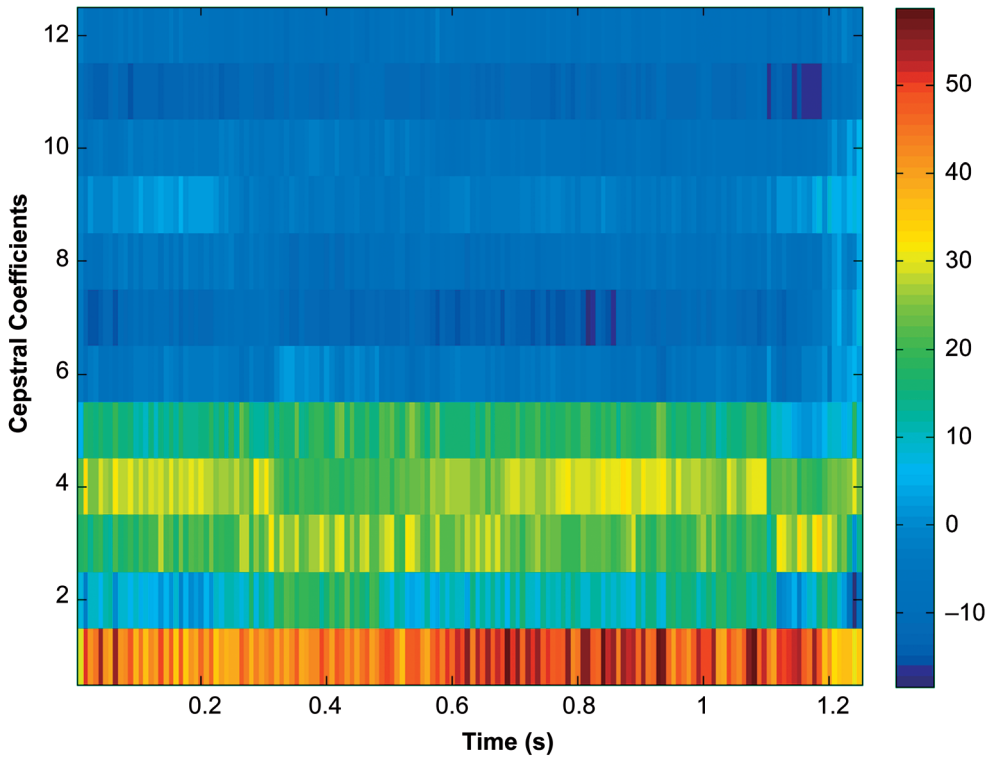

Figure 1b. Mel-frequency cepstral coefficients (MFCC)—deformed male speech (laryngean polypus), vowel /i/ (e.g., /pit/) with prolonged phonation after surgery.

4. RESULTS

Collected research material analysed with the ZCA, STFT and MFCC techniques made it possible to create a 29-element feature vector:

$$X_n = <M_0, M_1, M_2, F_1, F_2, F_3, F_4, AF_1, AF_2,$$
$$AF_3, AF_4, WM_1, WM_2, WM_3, C_1, C_2, C_3, C_4, (10)$$
$$C_5, C_6, C_7, C_8, C_9, C_{10}, C_{11}, C_{12}, \overline{f_0}, J, S >.$$

Figures 2a and 2b provide visualisation of the feature vector of the deformed speech recorded from a male patient diagnosed with laryngeal

cancer. Correspondingly, Figures 3a and 3b present feature vectors of deformed speech of a female patient diagnosed with chronic laryngitis. Reference values of feature vectors for normal speech are marked as shadow areas in Figures 2a, 2b, 3a and 3b. Reference vectors are the averages over the whole control group with subdivisions for the gender and the spoken phoneme. Individual feature vectors of pathological speech signals are shown with solid lines. Such a display is convenient for diagnostic purposes: it is possible to make an instant comparative visual analysis of features specific for a disease.
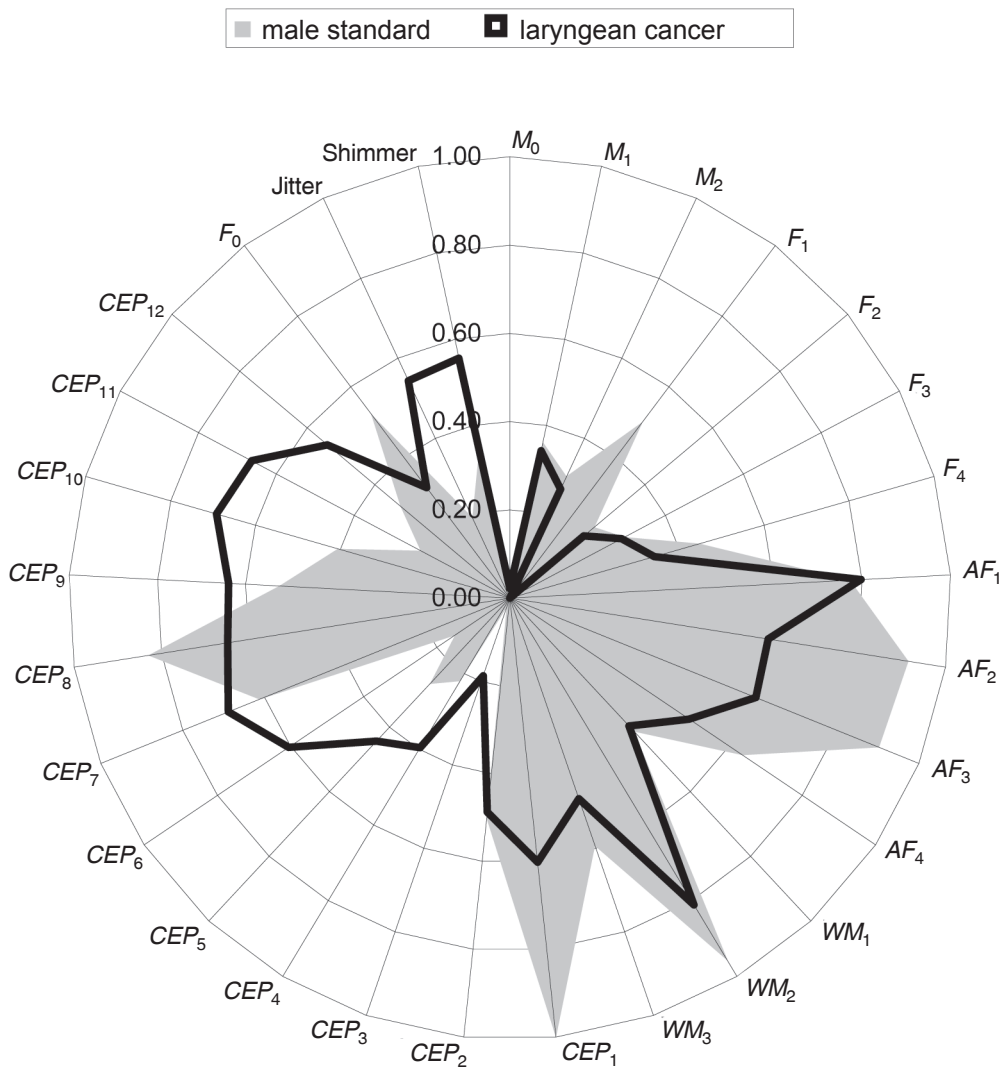

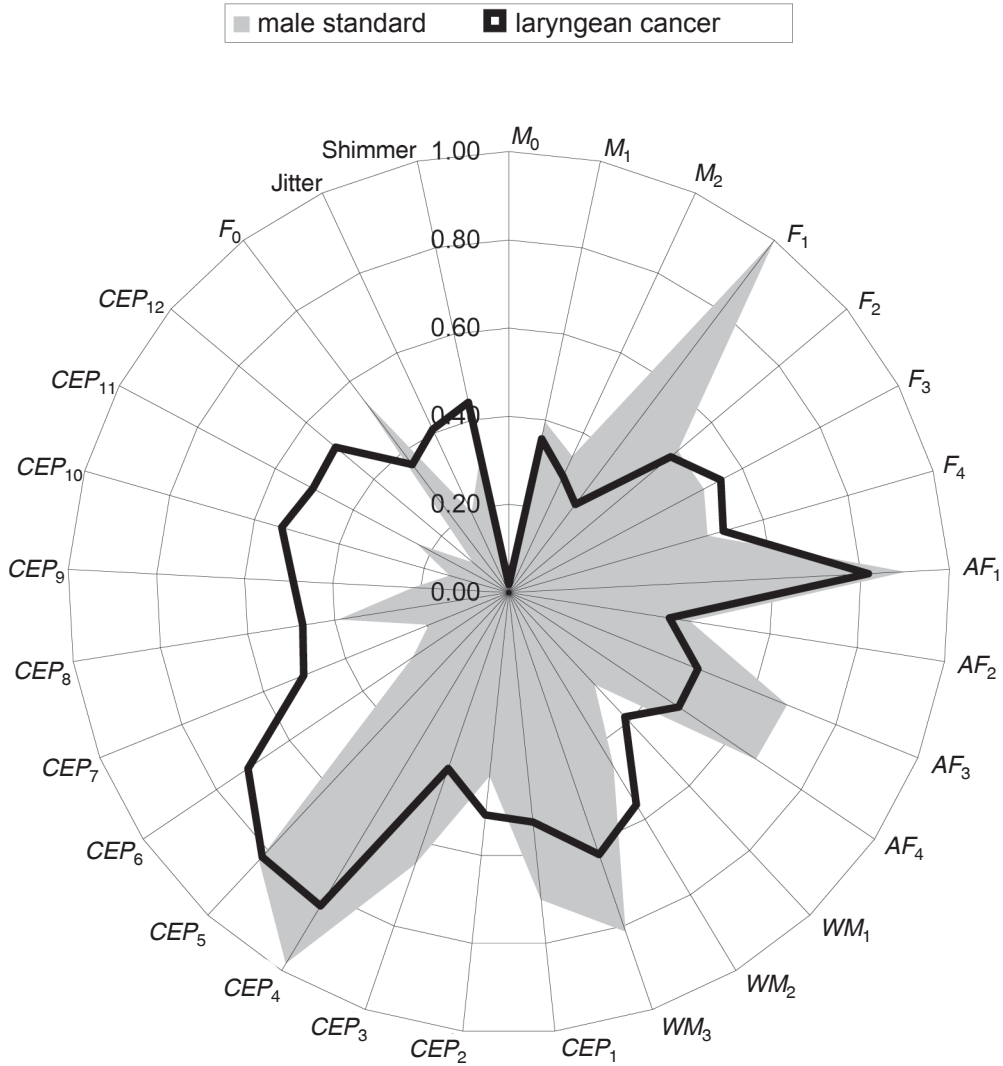**Figure 2a. Graphic interpretation of a feature vector for deformed male speech (laryngean cancer). Vowel /a/ (e.g., /pat/) with prolonged phonation.** *Notes.* M_0–M_2, F_0–F_4, AF_1–AF_4, WM_1–WM_3, CEP_1–CEP_{12}—co-ordinates of the feature vector.

Figure 2b. Graphic interpretation of a feature vector for deformed male speech (laryngean cancer). Vowel /a/ (e.g., /pat/) with prolonged phonation. *Notes.* M_0–M_2, F_0–F_4, AF_1–AF_4, WM_1–WM_3, CEP_1–CEP_{12}—co-ordinates of the feature vector.

Data shown in Figures 2a, 2b, 3a and 3b refer to diseases related to partial impairment of vocal folds. Significant differences in comparison with the control group, useful for diagnosis of the disease, can be seen in values of the following parameters: F_1, AF_2, AF_3, AF_4, WM_3, CEP_1, CEP_3, CEP_6, CEP_7, CEP_9, CEP_{10}, F_0, jitter (J) and shimmer (S). Jitter, shimmer and F_0 are directly related to vibrations of vocal folds. Further differences are seen in cepstral coefficients of the lower order (CEP_1 and CEP_3),

which provide additional information on the disease. Higher-order cepstral coefficients (CEP_6, CEP_8, CEP_9, CEP_{10}) are related to changes in geometry, mass and form of the vocal tract itself.

Power coefficient WM_3 illustrates changes in the acoustic signal spectrum in the frequency band from 2 125 to 4 375 Hz. Results of the analysis reveals that spectrum variation in this frequency range significantly corresponds to changes in the larynx condition and symptoms qualifying the disease. The analysis of formant
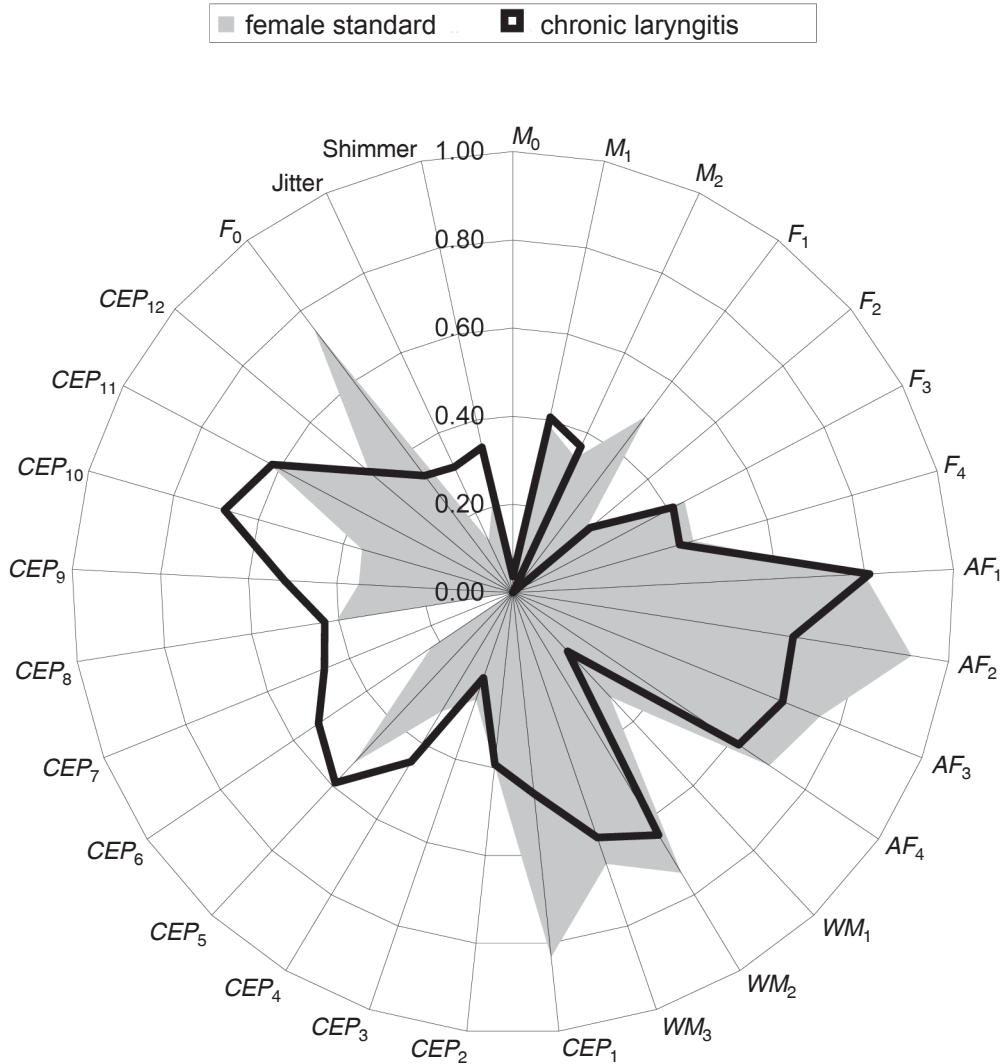
Figure 3a. Graphic interpretation of a feature vector for deformed female speech (chronic laryngitis). Vowel /e/ (e.g., /test/) with prolonged phonation. *Notes.* M_0–M_2, F_0–F_4, AF_1–AF_4, WM_1–WM_3, CEP_1–CEP_{12}—co-ordinates of the feature vector.

amplitudes—indicating energy changes in the extreme values of spectrum envelope—is also a valuable diagnostic factor.

Criteria for an objective assessment of speech signals are based on a "distance" in the space of features. Corresponding calculations were performed for simple Hamming and standardised Euclidean metrics. They make it possible to objectively order certain measures, which subjectively correspond to aurally perceived differences among the patient's voice and an average for normal voice. Those metrics also open a possibility to treat the problem quantitatively,

including the size of a scatter of an individual diagram, variability within a standard group, as well as variability within voices deformed by the same kind of disease. Most importantly, such approach makes it possible to confront acoustic data of persons with a certain type of deformation with a reference group of normal voices and, thus, to measure quantitatively the degree of voice pathology [5, 7].

Examples of distance metrics between pathological and standard speech are presented in Figures 4a and 4b; they illustrate consecutive stages of patients' medical examination
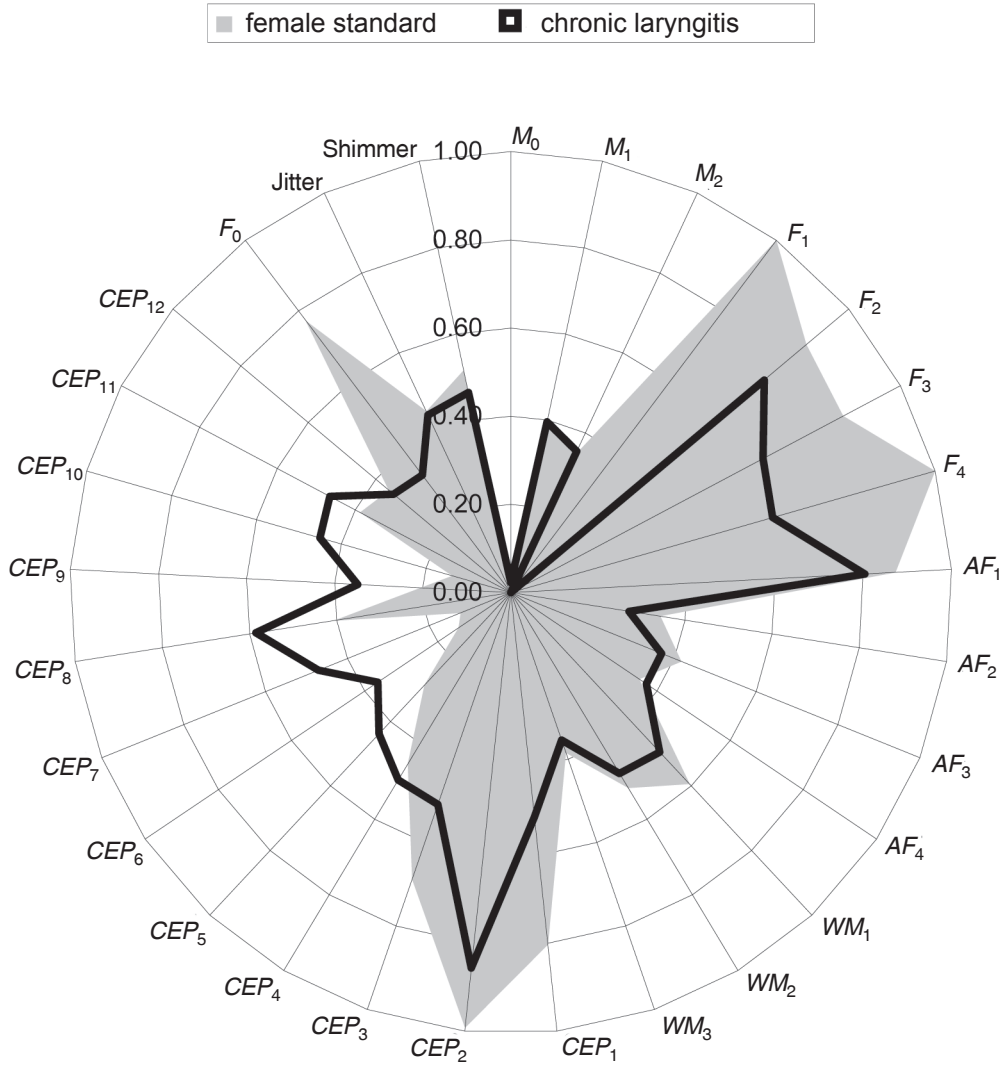
Figure 3b. Graphic interpretation of a feature vector for deformed female speech (chronic laryngitis). Vowel /u/ (e.g., /puk/) with prolonged phonation. *Notes.* M_0–M_2, F_0–F_4, AF_1–AF_4, WM_1–WM_3, CEP_1–CEP_{12}—co-ordinates of the feature vector.

and treatment. On the one hand the metrics contribute, by means of providing distance from the standards, to objectivisation of certain measures, which are subjectively perceptible by hearing. On the other hand those metrics make approaching the problem quantitatively possible. Generally, with this approach acoustic data of persons with proper articulation can be compared with data of persons with a determined kind of pathological articulation. Thus, we can "measure" quantitatively different kinds of pathology. High distance metric values represent deep speech signal deformation, whereas close to zero values

mean that examined signals can be considered the same (near zero deformation).

Feature vectors represent standard speech and deformed signals in 29-dimensional space. This multidimensional structure is difficult to interpret visually. Therefore, it is important to develop an understandable and perceptible graphical illustration of such set of data. This problem is sometimes solved by applying Sammon's representation [19, 20]. Sammon's representation locates multi-variant data on a plane [21]. Transformation is based upon selection of two-dimensional vectors created in a way which
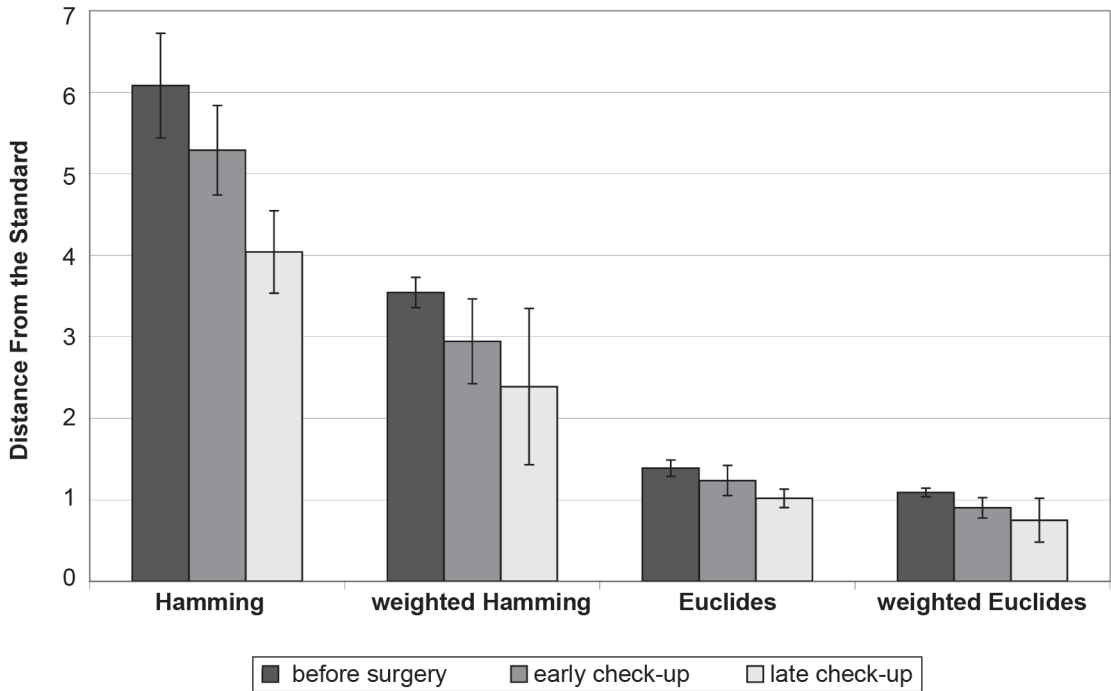
Figure 4a. Average distance of a deformed speech signal from the standard for vowels of prolonged phonation—/a/, /e/, /i/, /u/ (e.g., /pat/, /test/, /pit/, /puk/)—male speech (Reinke's oedema).
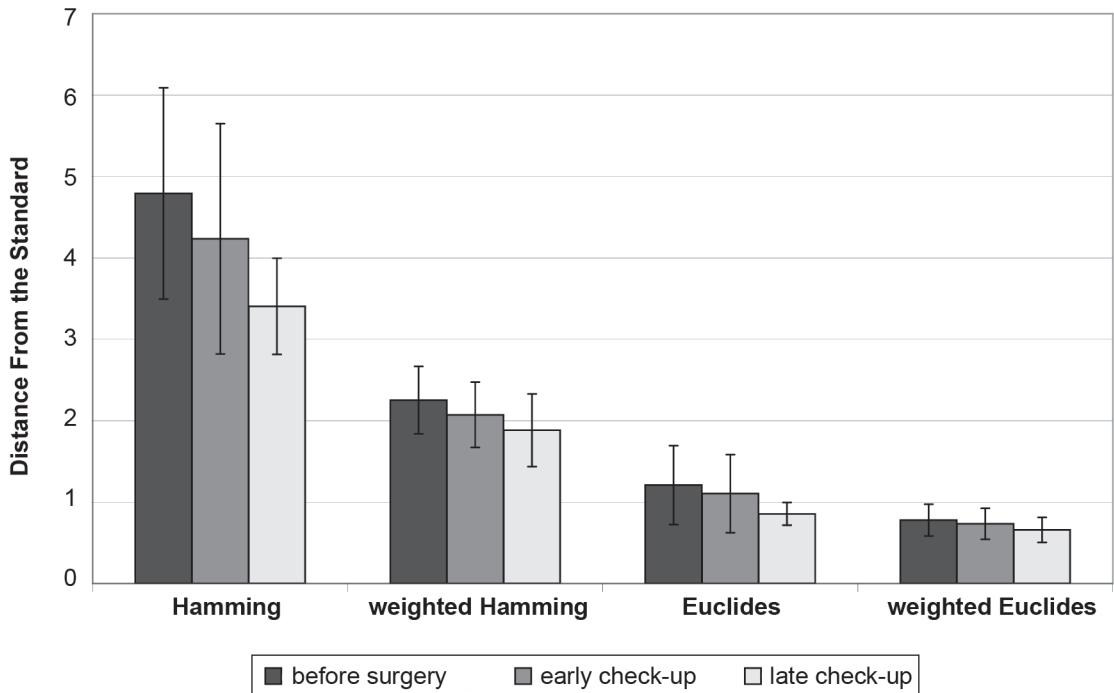
Figure 4b. Average distance of a deformed speech signal from the standard for vowels of a prolonged phonation—/a/, /e/, /i/, /u/ (e.g., /pat/, /test/, /pit/, /puk/)—male speech (laryngeal polypus).
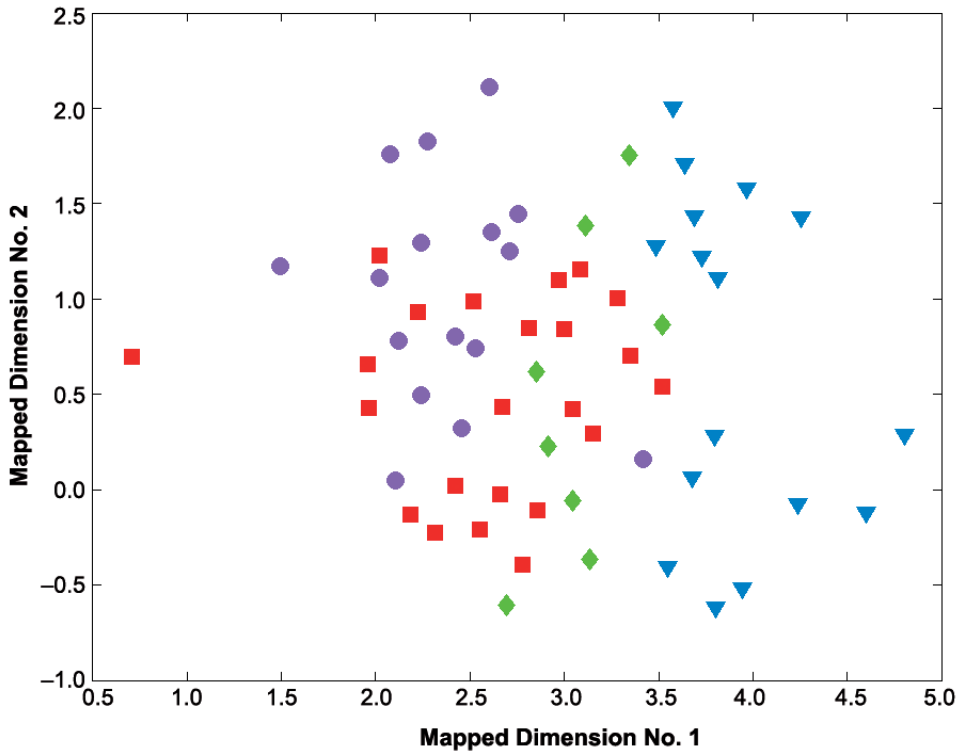
Figure 5a. Visualisation of 29-dimensional feature vectors with Sammon's representation—male speech. *Notes.* ▼—normal speech, ◆—chronic laryngitis, ■—laryngeal polypus, ●—laryngeal cancer.

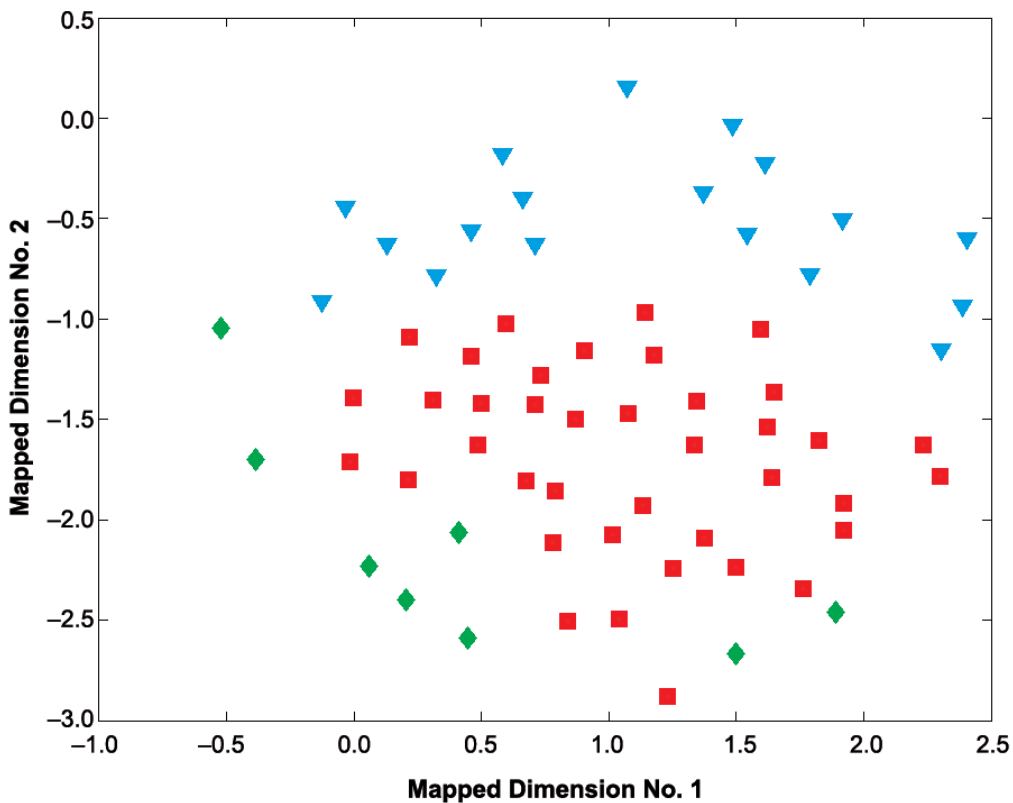**Figure 5b. Visualisation of 29-dimensional feature vectors with Sammon's representation—female speech.** *Notes*. ▼—normal speech, ◆—chronic laryngitis, ■—laryngeal polypus, ●—laryngeal cancer.

minimises the error function of the distance between the considered objects.

Sample results of Sammon's representation for the examined group of voices are presented in Figures 5a and 5b. Separation of the objects related to various diseases and control groups is clearly seen in these figures confirming that the selection of 29 parameters describing the speech signal is useful for diagnostic purposes. There are clear distinctions between areas occupied by objects representing normal speech (solid triangles) and areas occupied by objects representing the diagnosed diseases. Similar distinct differences are observed between particular diseases: chronic laryngitis (open diamonds), laryngeal polypus (open squares) and laryngeal cancer (open circles). It is remarkable that Sammon's representation seen in Figures 5a and 5b clearly separates such diseases as the most dangerous laryngeal cancer (open circles) from chronic laryngitis (open diamonds), which medically and personally may be considered as a much less serious problem.

5. CONCLUSIONS

Integrated acoustical analysis of deformed pathological speech including mapping, visualisation and quantitative assessment was discussed in this paper. This analysis, conducted among groups of patients, showed that speech pathology caused by various laryngeal diseases can be assessed using acoustical methods supported by distance metrics measuring the degree of signal deformation. Assessment executed before surgery and during recovery period corresponds to changes in the vocal tract, may be used for checking the decrease in speech deformation and may be a useful tool to control prosthetic restoration and rehabilitation.

The obtained parameters (co-ordinates of a feature vector), were useful for the development of a vibroacoustic model of the diseases of the human vocal tract. In this model, the feature vectors of the deformed speech signal are presented as graphs or tables. Information presented in such a way can be used by phoniatrists, laryngologists and speech therapists

as an additional objective tool for estimating the degree of speech deformation. Physicians can use visual representation of an acoustical analysis to assess changes in a speech signal at different stages of medical treatment.

The presented method has been already applied in control examinations of patients who underwent larynx oncological surgery. A broader range of applications of this method in medical treatment of various diseases of larynx, oral cavity and nasal sinuses will be possible in the future. The feature vector introduced in this study is suitable for advanced interpretation by artificial neurone networks dedicated to recognition of medical images.

REFERENCES

1. Jurkiewicz D, Dzaman K, Rapiejko P. Laryngeal cancer risk factors. Pol Merkur Lekarski. 2006;21(121):94–8. In Polish, with an abstract in English.

2. Modrzejewski M. Skuteczność chirurgicznego leczenia chorych na raka krtani piętra głośni. Kraków, Poland: Collegium Medicum, Jagiellonian University (CM UJ); 1998.

3. Wszołek W. Selected methods of pathological speech signal analysis. Arch Acoust. 2006;31(4):413–30.

4. Engel Z, Modrzejewski M, Wszołek W. Akustyczna ocena operacji krtani z wykorzystaniem parametrów tonu podstawowego. Zeszyty Naukowe AGH Mechanika. 1997;16(1):67–79.

5. Engel Z, Tadeusiewicz R, Tosińska-Ostrój H, Wszołek W. Analiza zmienności sygnału mowy jako metoda oceny wyników wybranej klasy zabiegów chirurgicznych. Zeszyty Naukowe AGH Mechanika. 1993;12(1):29–37.

6. Tadeusiewicz R, Izworski A, Wszołek W. Pathological speech evaluation using the artificial intelligence methods [abstract]. Med Biol Eng Comput. 1997;35:307.

7. Wszołek W, Kłaczyński M. Acoustic methods of voice estimation after surgical treatment of the vocal tract. Arch Acoust. 2005;30(4 Suppl):193–7.

8. Wszołek W, Kłaczyński M. Comparative study of the selected methods of laryngeal tone determination. Arch Acoust. 2006; 31(4):219–27.

9. Deller JR Jr, Proakis JG, Hansen JHL. Discrete-time processing of speech signals. New York, NY, USA: Macmillan; 1993.

10. Zieliński T. Cyfrowe przetwarzanie sygnałów. Warszawa, Poland: Wydawnictwo Komunikacji i Łączności; 2005.

11. Tadeusiewicz R. Sygnał mowy. Warszawa, Poland: Wydawnictwo Komunikacji i Łączności; 1988.

12. Wszołek W, Szaleniec J. Application of new acoustic parameters in pathological speech diagnosis In: 2nd International Conference From Scientific Computing to Computational Engineering. Book of abstracts. Patras, Greece: Laboratory of Fluid Mechanics and Energy (LFME), University of Patras; 2006. p. 21.

13. Grocholewski S. Statystyczne podstawy systemu ARM dla języka polskiego. Poznań, Poland: Wydawnictwo Politechniki Poznańskiej; 2001.

14. Gubrynowicz R. Komputerowe modelowanie artykulacji głosek języka polskiego. Warszawa, Poland: IPPT PAN; 2000.

15. Titze IR. Principles of voice production. Englewood Cliffs, NJ, USA: Prentice Hall; 1994.

16. Campbell JP Jr. Speaker recognition: a tutorial. Proceedings of the IEEE. 1997;85: 1437–62.

17. Rabiner LR. Fundamentals of speech recognition. Englewood Cliffs, NJ, USA: Prentice Hall; 1993.

18. Fant CGM. Acoustic description and classification of phonetic units. In: Fant G, editor. Speech sounds and features. Cambridge, MA, USA: MIT Press; 1973. p. 32–83. (Reprinted from Ericsson Technics. 1959;15(1).)

19. Cioch W. The objective selection of multi-dimensional technical state characteristics vectors. Machine Dynamics Problems. 2003;27(3):57–67.

20. Sobczak W, Malina W. Metody selekcji i redukcji informacji. Warszawa, Poland: WNT; 1985.

21. Sammon J. A nonlinear mapping for data structure analysis. IEEE Trans. Computers. 1969;18:401–9.

22. Dukiewicz L. Fonetyka. In: Wróbel H, editor. Gramatyka współczesnego języka polskiego (Fonetyka i fonologia). Kraków, Poland: Wydawnictwo Instytutu Języka Polskiego PAN; 1995. p. 9–103.

23. Stevens KN. Acoustics phonetics. Cambridge, MA, USA: MIT Press; 1998.