Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 141–146

ISBN 978-83-60810-22-4

A Bezier Curve Approximation of the Speech Signal in the Classification Process of Laryngopathies

Jarosław Szkoła, Krzysztof Pancerz
Institute of Biomedical Informatics
University of Information Technology and Management
Rzeszów, Poland
Email: jszkola@wsiz.rzeszow.pl, kpancerz@wsiz.rzeszow.pl

Jan Warchoł
Department of Biophysics
Medical University of Lublin
Lublin, Poland
Email: jan.warchol@am.lublin.pl

***Abstract*—The research concerns a computer-based clinical decision support for laryngopathies. The classification process is based on a speech signal analysis in the time domain using recurrent neural networks. In our experiments, we use the modified Elman-Jordan neural network. In the preprocessing step, an original signal is approximated using Bezier curves and next the neural network is trained. Bezier curve approximation reduces the amount of data to be learned as well as removes a noise from the original signal.**

***Index Terms*—computer-based clinical decision support; recurrent neural networks; laryngopathies; Bezier curves; approximation**

I. Introduction

COMPUTER-BASED clinical decision support (CDS) is defined as the use of a computer to bring relevant knowledge to bear on the health care and well being of a patient [1]. Our research concerns designing methods for CDS in a non-invasive diagnosis of selected larynx diseases. Two diseases are taken into consideration: Reinke's edema and laryngeal polyp. In general, the diagnosis is based on an intelligent analysis of selected parameters of a patient's speech signal (phonation). The proposed approach is non-invasive. Comparing it to direct methods shows that it has several advantages. It is convenient for a patient because a measurement instrument is located outside the voice organ. This enables free articulation. Moreover, different physiological and psychological patient factors impede making a diagnosis using direct methods.

The majority of methods proposed to date are based only on the statistical analysis of the speech spectrum (e.g. [2]) as well as the wavelet analysis. In our research, we are going to propose a hybrid approach, which is additionally based on a signal analysis in the time domain. Preliminary observations of signal samples for patients from a control group and patients with a confirmed pathology clearly indicate deformations of standard articulation in precise time intervals. In our previous papers (see [3], [4], and [5]), we have taken into consideration the usage of recurrent neural networks (RNNs), especially, the Elman and Jordan networks [6], [7] also known as "simple recurrent networks." RNNs can be used for pattern recognition in time series data due to their ability of memorizing some information from the past. The Elman networks (ENs) are a classical representative of RNNs. To improve learning ability of ENs we have modified and combined them with the Jordan networks. Such networks manifest a faster and more exact achievement of the target pattern. Moreover, for the time domain analysis, RNNs have the capability of extracting the phoneme articulation pattern for a given patient (articulation is an individual patient feature) and the capability of assessment of its replication in the whole examined signal.

In contrast to approaches shown in [3], [4], and [5], in the approach presented in this paper, we do not learn the neural network using samples of a speech signal directly. Now, we introduce a preprocessing step. In this step an original signal is approximated using Bezier curves and next the neural network is trained. Bezier curve approximation reduces the amount of data to be learned (by one order of magnitude), as well as, removes a noise from the original signal.

Our paper is organized as follows. After introduction, we shortly describe the medical background related to larynx diseases (Section II). In Section III, we show the procedure used to find approximation of a speech signal using Bezier's curves. Section IV describes the use of the modified Elman-Jordan neural network in finding disturbances in a speech signal. In Section V, we present results obtained by experiments done on real-life data. Some conclusions and final remarks are given in Section VI.

II. Medical Background

A model of speech generation is based on the "source-filter" combination. The source is larynx stimulation, i.e., passive vibration of the vocal folds as a result of an increased subglottis pressure. Such a phenomenon of making speech sonorous in the glottis space is called phonation. The filter is the remaining articulators of the speech canal creating resonance spaces. A signal of larynx stimulation is shaped and modulated in these spaces. A final product of this process is called speech.

Pathological changes appearing in the glottis space entail a bigger or smaller impairment of the phonation functions of the larynx. The subject matter of presented research concerns diseases, which appear on the vocal folds, i.e., they have a direct influence on phonation [8].

We are interested in two diseases: Reinke's edema (*Oedema Reinke*) and laryngeal polyp (*Polypus laryngis*).

A. Reinke's Edema

Reinke's edema appears often bilaterally, and usually asymmetrically, on the vocal folds. It is created by transudation in a slotted epithelial space of folds devoid of lymphatic vessels and glands, called the Reinke's space. In the pathogenesis of disease, a big role is played by irritation of the laryngeal mucosa by different factors like smoking, excessive vocal effort, inhalatory toxins or allergens. The main symptoms are the following: hoarseness resulting from disturbance of vocal fold vibration or, in the case of large edemas, inspiratory dyspnea. In the case of Reinke's edemas, conservative therapy is not applied. They are microsurgically removed by decortication with saving the vocal muscle.

B. Laryngeal Polyp

Laryngeal polyp is a benign tumor arising as a result of gentle hyperplasia of fibrous tissue in mucous membrane of the vocal folds. In the pathogenesis, a big role is played by factors causing chronic larynx inflammation and irritation of the mucous membranes of the vocal folds: smoking, excessive vocal effort, reflux, etc. The main symptoms are the following: hoarseness, aphonia, cough, tickling in the larynx. In case of very big polyps, dyspnea may appear. However, not big polyps may be confused with vocal tumors especially when there is a factor of the load of the patient voice. The polyp may be pedunculated or may be placed on the wide base. If it is necessary, polyps are microsurgically removed with saving a free edge of vocal fold and vocal muscle.

III. Phoneme Approximation using the Bezier Curves

In order to approximate a speech signal (phoneme), we propose to use 4-point Bezier curves. A Bezier curve is a parametric curve very popular in different applications of computer graphics and related fields. A shape of the 4-point Bezier curve is determined by four control points $P_0 = [P_0^x, P_0^y]$, $P_1 = [P_1^x, P_1^y]$, $P_2 = [P_2^x, P_2^y]$, and $P_3 = [P_3^x, P_3^y]$. The curve interpolates points P_0 and P_3 and approximates points P_1 and P_2. Two parametric equations determine the shape of the curve:

$$x(t) = (1-t)^3 P_0^x + 3(1-t)^2 t P_1^x + 3(1-t)t^2 P_2^x + t^3 P_3^x,$$

$$y(t) = (1-t)^3 P_0^y + 3(1-t)^2 t P_1^y + 3(1-t)t^2 P_2^y + t^3 P_3^y,$$

where $t \in [0, 1]$. Application of the Bezier curves has the following advantages:

- a signal is encoded using a smaller number of values than a number of samples (by one order of magnitude), this can accelerate a learning process of neural networks,
- some noise from the original signal can be removed,
- due to transformations, signals with different magnitudes and gradients can be compared.

The main disadvantage is the complicated process of finding approximation of a signal in the form of a family of Bezier curves.

In this section, we propose an iterative algorithm for finding a family of Bezier curves best approximating a given signal curve. The algorithm inches forward (from the beginning) along the approximated signal curve (corresponding to a given phoneme) trying to find the best Bezier curve approximating the part of the signal curve. At each stage of searching, if a better curve cannot be found, the current Bezier curve is recorded for a covered part of the curve and new searching for the remaining part is started. In this algorithm, a number of Bezier curves approximating a given signal curve is not given in advance. The presented algorithm can be called the stepping algorithm. Algorithm 1 shows formally our procedure. An error ε between the Bezier curve and the signal curve in a given interval of samples (from $start$ to end) is calculated separately for coordinates x and y according to the following formulas:

$$\varepsilon^x = \sum_{i=start}^{end} |t_B^i - t_F^i|,$$

$$\varepsilon^y = \sum_{i=start}^{end} |v_B^i - v_F^i|,$$

where $\{t_B^i, v_B^i\}_i$, $\{t_F^i, v_F^i\}_i$ are sequences of samples of the Bezier curve B and the signal curve F, respectively.

The presented algorithm has a polynomial time complexity.

Algorithm 1 uses Algorithm 2 for finding the best moving of a given control point. This algorithm checks an error between the Bezier curve and the original curve. If the error ε is worsened (with respect to the previous one ε_p) after moving a control point, then moving is canceled and its direction for the next step is changed to opposite. A number of searching steps is limited (in our experiments to 100).

IV. Recurrent Neural Networks for Learning Phoneme Patterns

For each Bezier curve represented by four control points $P_0 = [P_0^x, P_0^y]$, $P_1 = [P_1^x, P_1^y]$, $P_2 = [P_2^x, P_2^y]$, and $P_3 = [P_3^x, P_3^y]$, we calculate three Euclidean distances:

$$D_1 = \sqrt{(P_1^x - P_0^x)^2 + (P_1^y - P_0^y)^2},$$

$$D_2 = \sqrt{(P_2^x - P_0^x)^2 + (P_2^y - P_0^y)^2},$$

$$D_3 = \sqrt{(P_3^x - P_0^x)^2 + (P_3^y - P_0^y)^2}.$$

Let $\mathcal{B}$ be a family of Bezier curves approximating a given phoneme. The phoneme is represented by sequences of distances calculated for each Bezier curve. Such sequences are used to learn the modified Elman-Jordan network. The modified Elman-Jordan network has been proposed in [4].

Algorithm 1: Algorithm for determining a family of Bezier curves approximating a phoneme.

Input : $F = [(t_0, v_0), (t_1, v_1), \ldots, (t_{n-1}, v_{n-1})]$ - a phoneme (a vector of speech signal samples), tau - error threshold, c_{max} - a maximal number of attempts of searching, $[s_1^x, s_1^y]$, $[s_2^x, s_2^y]$ - moving vectors, where $s_1^x, s_1^y, s_2^x, s_2^y \in [0, 1]$

Output: $\mathcal{B}$ - the family of 4-point Bezier curves approximating F.

$\mathcal{B} \leftarrow \emptyset$;
$start \leftarrow 0$; $end \leftarrow 1$;
$P_1 \leftarrow null$; $P_2 \leftarrow null$;
$\varepsilon_{p1} \leftarrow null$; $\varepsilon_{p2} \leftarrow null$;
$c \leftarrow 0$; $stop \leftarrow false$;
while $stop = false$ **do**
  $P_0 \leftarrow (t_{start}, v_{start})$;
  if $P_1 = null$ **then**
    $P_1 \leftarrow (t_{start}, v_{start})$;
  end
  if $P_2 = null$ **then**
    $P_2 \leftarrow (t_{start}, v_{start})$;
  end
  $P_3 \leftarrow (t_{end}, v_{end})$;
  Create a curve B on the basis of P_0, P_1, P_2, P_3;
  Calculate an error ε_1 between B and F in the interval $[t_{start}, t_{end}]$
  Move point P_1 using Algorithm 2 by vector $[s_1^x, s_1^y]$;
  Calculate an error ε_2 between B and F in the interval $[t_{start}, t_{end}]$;
  Move point P_2 using Algorithm 2 by vector $[s_2^x, s_2^y]$;
  if *($\varepsilon_{p1}^x < \tau$ and $\varepsilon_{p1}^y < \tau$) or ($\varepsilon_{p2}^x < \tau$ and $\varepsilon_{p2}^y < \tau$)* **then**
    if $1 + end < n$ **then**
      $end \leftarrow end + 1$;
    else
      $stop = true$;
    end
    $\varepsilon_{p1} \leftarrow null$; $\varepsilon_{p2} \leftarrow null$;
    $c \leftarrow 0$;
  else
    $c \leftarrow c + 1$;
  end
  if $c > c_{max}$ **then**
    if $1 + end < n$ **then**
      $\mathcal{B} \leftarrow \mathcal{B} \cup \{B\}$;
      $start \leftarrow end$;
      $end \leftarrow start + 1$;
      $P_1 \leftarrow null$; $P_2 \leftarrow null$;
      $\varepsilon_{p1} \leftarrow null$; $\varepsilon_{p2} \leftarrow null$;
      $c \leftarrow 0$;
    else
      $stop \leftarrow true$;
    end
  end
end
Return $\mathcal{B}$;

Algorithm 2: Algorithm for moving a control point.

Input : P - a control point to be moved, ε_p - a last error, ε - a current error, $[s^x, s^y]$ - a moving vector.

Output: P - a moved control point

if $\varepsilon_p = null$ **then**
  $\varepsilon_p \leftarrow \varepsilon$;
end
if $\varepsilon^x \geq \varepsilon_p^x$ **then**
  $P^x \leftarrow P^x - s^x$;
  $s^x \leftarrow -s^x$;
end
if $\varepsilon^y \geq \varepsilon_p^y$ **then**
  $P^y \leftarrow P^y - s^y$;
  $s^y \leftarrow -s^y$;
end
if $\varepsilon^x < \varepsilon_p^x$ **then**
  $\varepsilon_p \leftarrow \varepsilon$;
end
if $\varepsilon^y < \varepsilon_p^y$ **then**
  $\varepsilon_p \leftarrow \varepsilon$;
end
$P^x \leftarrow P^x + s^x$;
$P^y \leftarrow P^y + s^y$;
Return P;

Moreover, some learning abilities were tested in [5]. The modified Elman-Jordan network consists of (see Figure 1):

- an input layer,
- a hidden layer,
- a context layer,
- an output layer,
- feedbacks for a hidden layer through the context layer, such feedbacks are used in the Elman networks,
- feedback between an output layer and a hidden layer through the context layer, such feedback is used in the Jordan networks,
- feedback for an output layer.

Output feedback accelerates a learning process and causes seamless modification of weights. Generally, the modified Elman-Jordan network needs a smaller number of epochs (sometimes by 50 per cent) for learning a given pattern (see [5]).

In the approach presented in this paper, we use similar procedure to that presented in [3], [4], and [5]. A difference is that we use as the input for the neural network sequences of distances calculated for Bezier curves approximating an original speech signal instead of samples of that signal. Articulation is an individual patient feature. Therefore, we cannot train a neural network on the independent patterns of phonation of individual vowels. For each patient, parameters of Bezier curves of a speech signal are used for both training and testing of a neural network. The procedure is as follows. We divide the speech signal of an examined patient into time windows corresponding to phonemes. The next step is

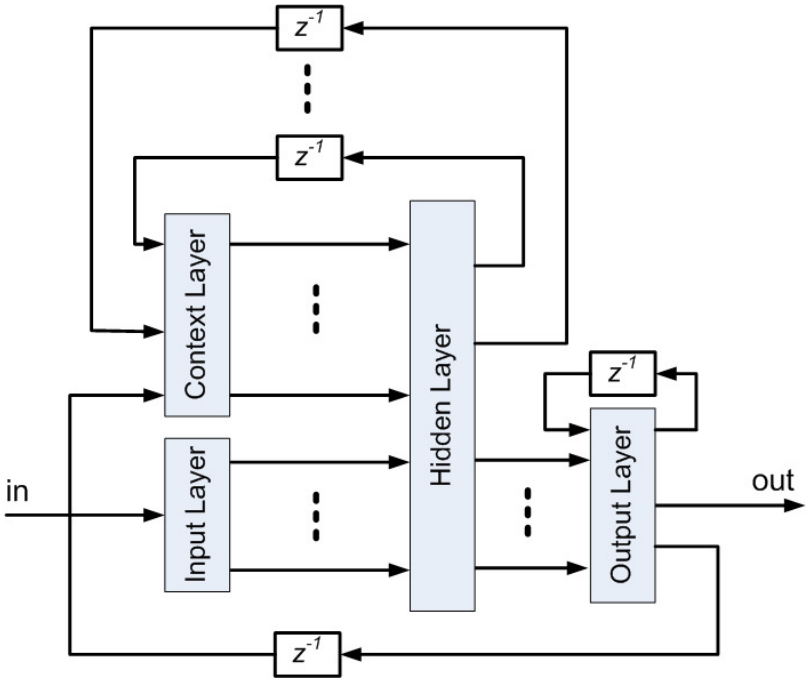

Fig. 1. A structure of the trained Elman-Jordan neural network.

Algorithm 3: Algorithm for calculating an average mean squared error corresponding to deformations in a speech signal.

Input : S - a speech signal of a given patient (a vector of samples), N - a neural network.
Output: $\overline{E}_N$ - an average mean squared error corresponding to deformations in S.

$W_{all} \leftarrow Div2Win(S)$;
$W_{sel} \leftarrow SelWin(W_{all})$;
for *each window* $w \in W_{sel}$ **do**
  $B \leftarrow Bezier(w)$;
  $Train(N, B)$;
  for *each window* $w^* \in W_{sel}$ **do**
    if $w^* \neq w$ **then**
      $B^* \leftarrow Bezier(w^*)$;
      $E[w^*] \leftarrow MSE(Test(N, B^*))$;
    end
  end
  $\overline{E}[w] \leftarrow Avg(E)$;
end
$\overline{E}_N \leftarrow Avg(\overline{E})$;
Return $\overline{E}_N$;

random selection of a number of time windows. This set of selected windows is used for determining some coefficient characterizing deformations in the speech signal. This coefficient is constituted by an error obtained during testing of the neural network. We propose to use the approach similar to the cross-validation strategy. For each time window we determine a family of Bezier curves approximating it (see Algorithms 1 and 2) and next calculate distances for each curve according to formulas shown at the beginning of this section. Bezier curve parameters of one time window are taken for training the neural network, whereas Bezier curve parameters of the remaining ones, for testing of the neural network. The network learns parameters of a selected time window. If parameters of the remaining windows are similar to the selected one in terms of the time patterns, then for such windows an error generated by the network in a testing stage is small. If significant replication disturbances in time appear for patients with the larynx disease, then an error generated by the network is greater. In this case, the time pattern is not preserved in the whole signal. Therefore, the error generated by the network reflects non-natural disturbances in the patient phonation. Our approach can be expressed formally as it is shown in Algorithm 3. In the algorithm we use the following functions (procedures):

- $Div2Win(S)$ - dividing the speech signal S into time windows corresponding to phonemes,
- $SelWin(W)$ - selecting randomly a number of time windows from the whole set W,
- $Bezier(w)$ - calculating a set of parameters of Bezier curves approximating w,
- $Train(N, B)$ - training a neural network N on a given set B of parameters of Bezier curves,
- $Test(N, B)$ - testing a neural network N on a given set B of parameters of Bezier curves,
- $MSE(E)$ - calculating a mean squared error for the

TABLE I
SELECTED RESULTS OF EXPERIMENTS FOR WOMEN FROM THE CONTROL GROUP OBTAINED USING THE MODIFIED ELMAN-JORDAN NETWORK.

ID	$\overline{E}_{EJN}$	$\overline{n}_{EJN}$
w_{CG1}	0.0106	104
w_{CG2}	0.0108	103
w_{CG3}	0.0120	110
w_{CG4}	0.0017	111
w_{CG5}	0.0055	99
w_{CG6}	0.0159	113
w_{CG7}	0.0064	131
w_{CG8}	0.0128	113
w_{CG9}	0.0128	105
w_{CG10}	0.0186	104

TABLE II
SELECTED RESULTS OF EXPERIMENTS FOR WOMEN WITH LARYNGEAL POLYP OBTAINED USING THE MODIFIED ELMAN-JORDAN NETWORK.

ID	$\overline{E}_{EJN}$	$\overline{n}_{EJN}$
w_{P1}	0.2184	98
w_{P2}	0.0429	71
w_{P3}	0.0139	87
w_{P4}	0.0201	120
w_{P5}	0.0155	132
w_{P6}	0.0375	80
w_{P7}	0.0148	210
w_{P8}	0.0184	94
w_{P9}	0.0229	88
w_{P10}	0.0462	109

absolute error vector E:

$$MSE(E) = \frac{1}{n}\sum_{i=1}^{n}(E_i)^2,$$

where n is a number of elements in the vector E, $E_i = y(x_i) - z(x_i)$ and $y(x_i)$ is the obtained output for x_i whereas $z(x_i)$ is the desired output for x_i.

- $Avg(E)$ - calculating an arithmetic average for the vector E of errors.

V. EXPERIMENTS

In the experiments, sound samples were analyzed. Samples were recorded for two groups of patients [2]. The first group included patients without disturbances of phonation. They were confirmed by phoniatrist opinion. The second group included patients of Otolaryngology Clinic of the Medical University of Lublin in Poland. They had clinically confirmed dysphonia as a result of Reinke's edema or laryngeal polyp. The information about diseases was received from patients' documentations. Each recording was preceded by a course of breathing exercises with an instruction about a way of articulation. The task of all examined patients was to utter separately Polish vowels: "A", "I", and "U" with extended articulation as long as possible, without intonation, and each on separate expiration. Samples were normalized to the interval $[0.0, 1.0]$ before providing them to the next block. After normalization, samples (as double numbers) were provided to the block calculating the Bezier curve parameters.

In Tables I and II, we present results of experiments carried out using the modified Elman-Jordan network described in Section IV. Table I includes results for women from the control group as well as Table II includes results for women with laryngeal polyp. Both tables include results for women uttering vowel "A". We give consecutively the average mean squared error $\overline{E}_{EJN}$ and an average number $\overline{n}_{EJN}$ of epochs in the training process.

It is easy to see that the modified Elman-Jordan network trained by parameters of Bezier curves approximating the speech signal has some ability to distinct between normal and disease states. The distinction ability presented here would be comparable with abilities obtained if the neural networks were trained using the original speech signals (cf. [4], and [5]). In the approach presented in this paper, we significantly reduce the amount of data to be learned by the neural network. Such observations are very important for further research, especially in the context of a created computer tool for diagnosis of larynx diseases.

VI. CONCLUSIONS

In the paper, we have shown the classification process of laryngopathies based on a speech signal analysis in the time domain using recurrent neural networks. In our experiments, we have used the modified Elman-Jordan neural network presented in our earlier papers. In the procedure, we have introduced the preprocessing step. In this step, an original signal is approximated using Bezier curves and next the neural network is trained. Bezier curve approximation reduces the amount of data to be learned as well as removes a noise from the original signal. The quality of the proposed method in terms of differentiating normal and pathological categories is not entirely satisfactory, but it shows the direction of further research. In the future, we will concentrate on two directions. The first one is the optimization of the process of finding a family of Bezier curves approximating the speech signal. The second one is an improvement (tuning) of the proposed method for better differentiating between cases belonging to different categories.

ACKNOWLEDGMENT

This research has been supported by the grant No. N N516 423938 from the Polish Ministry of Science and Higher Education.

REFERENCES

[1] R. Greenes, *Clinical Decision Support. The Road Ahead*. Elsevier Inc., 2007.

[2] J. Warchoł, "Speech examination with correct and pathological phonation using the SVAN 912AE analyser (in Polish)," Ph.D. dissertation, Medical University of Lublin, 2006.

[3] J. Szkoła, K. Pancerz, and J. Warchoł, "Computer diagnosis of laryngopathies based on temporal pattern recognition in speech signal," *Bio-Algorithms and Med-Systems*, vol. 6, no. 12, pp. 75–80, 2010.

[4] ——, "Computer-based clinical decision support for laryngopathies using recurrent neural networks," in *Proc. of the 10th International Conference on Intelligent Systems Design and Applications (ISDA'2010)*, A. Hassanien, A. Abraham, F. Marcelloni, H. Hagras, M. Antonelli, and T.-P. Hong, Eds., Cairo, Egypt, 2010, pp. 627–632.

[5] ——, "Improving learning ability of recurrent neural networks: Experiments on speech signals of patients with laryngopathies," in *Proc. of the International Conference on Bio-inspired Systems and Signal Processing (BIOSIGNALS'2011)*, F. Babiloni, A. Fred, J. Filipe, and H. Gamboa, Eds., Rome, Italy, 2011, pp. 360–364.
[6] J. Elman, "Finding structure in time," *Cognitive Science*, vol. 14, pp. 179–211, 1990.
[7] M. Jordan, "Serial order: A parallel distributed processing approach," University of California, San Diego, Institute for Cognitive Science, Tech. Rep. 8604, 1986.
[8] A. Lalvani, *Current Diagnosis and Treatment in Otolaryngology - Head and Neck Surgery*. McGraw-Hill, 2008.
[9] R. Orlikoff, R. Baken, and D. Kraus, "Acoustic and physiologic characteristics of inspiratory phonation," *Journal of the Acoustical Society of America*, vol. 102, no. 3, pp. 1838–1845, 1997.
[10] J. Warchoł, J. Szkoła, and K. Pancerz, "Towards computer diagnosis of laryngopathies based on speech spectrum analysis: A preliminary approach," in *Proc. of the Third International Conference on Bio-inspired Systems and Signal Processing (BIOSIGNALS'2010)*, A. Fred, J. Filipe, and H. Gamboa, Eds., Valencia, Spain, 2010, pp. 464–467.
[11] W. Winholtz and I. Titze, "Suitability of minidisc (MD) recordings for voice perturbation analysis," *Journal of Voice*, vol. 12, no. 2, pp. 138–142, 1998.