IMPACT OF THE GLOTTAL SIGNAL ON THE PREDICTION OF SPEECH

Danijela D. Protić
General Staff of the Serbian Army, Department
of Telecommunications and Information Technology (J-6),
Centre for Applied Mathematics and Electronics, Belgrade,
e-mail: adanijela@ptt.rs

DOI: 10.5937/vojtehg63-6357
FIELD: Telecommunications
ARTICLE TYPE: Original Scientific Paper
ARTICLE LANGUAGE: English

Summary:

In this paper, several linear and nonlinear techniques for speech processing based on AR, ARX, ARMAX, WLS and FNN models are proposed. The impact of the glottal wave to modelling is also shown in details. GD, BPA and LM approximations are used for model training and optimization. A comparative experimental analysis of five considered models is done based on the prediction of a speech signal. The results on training and testing are presented through learning and training errors for all models given.

Key words: *Linear models, Prediction, Glottal signal, Feed-forward neural network, Speech.*

Introduction

When speech occurs, the air from the lungs propagates along the trachea and the vocal tract to the lips, where it is radiated into the environment. Vibrations of the vocal cords change the airflow that passes through the glottal and vocal tract, where the shapes of the nasal cavity, the tongue, teeth and the lips determine the output wave, i.e. speech. Speech is classified to unvoiced and voiced, depending on the nature of the excitation (Burrows, 1996). For unvoiced speech, the vocal cords are wide apart and the air passes freely through the glottal tract where a noise-like, low-power signal arises. The excitation is due to turbulence generated by the airflow passing through a narrow constriction and tends to be random in nature. For voiced speech, the excitation of the vocal tract originates at the glottis. When the vocal cords are close together, the air pressure causes them to vibrate and thus forms a strong signal, i.e. vowel. This vibration is periodical, and its frequency (or pitch), is controled by the tension in the vocal cords.

The most popular technique for speech processing is Linear Prediction (LP). LP uses a source-filter arrangement to model the system which assumes that the source is located at the glottis and that the linear filter can be used to model the frequency properties of the vocal tract. The main disadvantage of LP is that the source and the vocal tract filter are not decoupled in the analysis and the LP filter thus combines the effects of the source and vocal tract. Other approaches interpret the voiced speech signal following Auto Regressive (AR) model, Auto Regressive with eXogenous input (ARX) model, and Auto Regressive Moving Average with eXogenous input models. AR model parameters are estimated for time series using the variants of Linear-Squares (LS) method that minimizes the summed squares of errors which are assumed to be normally distributed. For multivariate data, ARX is used. A current output depends on previous outputs, previous and delayed inputs as well as a white noise disturbance value. A generalization of the ARX model, ARMAX, also includes the output error.

It is usually assumed that the response data is of equal quality and, therefore, has the constant variance. If this assumption is violated, the Weighted Least Squares (WLS) algorithm can be used to improve the fitting process by including the additional scale factors (weights). The weights determine how much each response value influences the final parameter estimate.

When the input/output dynamics of a system contains a nonlinear component, a common linear modelling procedure has to resort to changing to the nonlinear dynamic modelling. The most used nonlinear models for the prediction of speech are multilayered networks generally called the Multi Layer Perceptrons (MLPs). They allow non-linear mappings by a learning procedure that consists of adjusting synaptic weights which are fully connected and arranged in layers (Sainath et al, 2011), (Pamučar, Đorović, 2012), (Milićević, Župac, 2012). MLPs have become very popular in solving various problems such as regression, classification, time series processing, identification and control of dynamical systems (Haykin, 1994), (Narendra, Parthasaranthy, 1990). An MLP is a feed-forward neural network (FNN) with one or more hidden layers between the input layer and the output layer. Feed-forward means that data flows in one direction from the input layer to the output one. For the FNNs having differentiable activation functions, there exists a computationally efficient method, called the Back-Propagation Algorithm (BPA), used for finding the derivatives of an error function with respect of the network weights. Typically, the BPA uses the gradient descent (GD) training algorithm. The network weights are moved along the negative of the gradient to find a minimum of the error function (Silva et al., 2008), (Wu et al., 2011). However, the GD is relatively slow and the network solution may become trapped in one of the local minima instead of the global minimum. For these reasons, there are some other procedures such as the Levenberg-Marquardt (LM) algorithm, available to use in order to improve the standard BPA. It gives efficient solutions of convergence and better optimization than the GD (Riecke et al., 2009), (Shahin, Pitt, 2012). The LM combines advantages of the GD method (that is, minimization along the direction of the gradient) with the Newton method (that is, using a quadratic model to speed up the process of finding the minimum of a function) (Levenberg, 1944), (Marquardt, 1963).

This paper presents the impact of the glottal signal to the prediction of speech that is based on five different models. The AR model parameters are estimated by a training procedure based on the LS method. The goal is to prove that a high order model can improve modelling even though the glottal signal is not used for the prediction. Additionally, the WLS is used to demonstrate the influence of weights to the LS. Furthermore, in order to obtain the vocal tract transfer function and the glottal source parameters, the ARX model is estimated. In this way, the influence of the glottal signal on the evaluation of the model should decrease the error. For the ARMAX model, a sample of the output error is used to improve prediction. However, it does not influence modelling when a vowel is used for model estimation. Finally, the FNN with one hidden layer and the tangent hyperbolic activation functions for all neurons are used for non-linear modelling. The LM algorithm is applied for model evaluation. The results show that the mapping function gives better results for the FNN model than for all other models. The minimum training error is the estimation criterion for the model training. Finally, the models are tested and the test errors are used to compare the quality of prediction.

The article is organized as follows. Second section presents the optimal linear and nonlinear models. Linear prediction, the WLS, the influence of the glottal wave and the FNN learning are shown in details. The LS and the weighted LS are presented. The GD and the BPA are shown in more detail. The principles of the LM method are shown. The results are given in Section three. Finally, the paper ends with some concluding remarks.

Linear and nonlinear parametric models

Although they are two mutually separated and independent processes, the speech analysis and the speech synthesis are often implemented simultaneously. The analytical process determines the characteristics of excitation, the glottis and the vocal tract. The synthesis generates signals that can be used for speech or speaker recognition, to simulate or reject the side effects, etc. The analysis involves the phonetic features of the spoken content but the level of the estimated error is high, and the assessment methodology encompasses a wide range of models with a high degree of freedom. In the synthesis, the excitation signal can be a pulse or noise, or may be generated by the Linear Prediction Coder (LPC), which is applied in order to ensure a high quality of speech, assuming that the speech sample is a linear combination of the previous samples. The LPC is carried out as follows: 1) the new model parameters are estimated, 2) the Mean Squares Error (MSE) is calculated to re-perform the synthesis, and 3) acceptable results are obtained by all-pole models, as it will be explained in details later in the paper. Thereafter, the spectrum of the excitation, and the transfer function of the vocal tract are simulated. The main advantages of this technique are the automatic analysis of the original signal and the accuracy of the estimate. Still, there are discontinuities in all-pole modelling because models do not take into account the characteristics of nasals, plosives and fricatives, which enter zeroes to transfer functions.

For that reason, linear and nonlinear modelling is presented here. The influence of the model order, the glottal signal and the disturbance factors to the speech prediction are shown.

Linear prediction

Linear prediction (LP) determines the value of the n^{th} sample of the signal $y(n)$ that is based on the all-pole model. It is well known that the assumption of linearity does not exactly match the characteristics of speech. Nevertheless, a high-quality LP model has advantages over complex non-linear models, such as a simple sturcture and a minimal prediction error. The AR model that is the most commonly applied in the LP is given with the formula (1)

$$y(n)+a_1 y(n-1)+\ldots+a_{n_a} y(n-n_a)=e(n) \tag{1}$$

whereas $y(n)$ is an input, a_i (i=1…n_a) are the model parameters, and $e(n)$ is an error. If the extra input, in this case of the glottal signal, is also processed, the AR model expands to the ARX model. See (2).

$$\begin{aligned} y(n)+a_1 y(n-1)+\ldots+a_{n_a} y(n-n_a)&=\ldots \\ \ldots=b_1 u(n-1)+\ldots+b_{n_b} u(n-n_b)+e(n)& \end{aligned} \tag{2}$$

bb_i (i=1…n_b) are the eXogenos parameters. The generalization of the model, known as the ARMAX, includes the error propagation. See (3).

$$\begin{aligned} & y(n)+a_1 y(n-1)+\ldots+a_{n_a} y(n-n_a)=\ldots \\ & \ldots=b_1 u(n-1)+\ldots+b_{n_b} u(n-n_b)+\ldots \\ & \ldots+e(n)+c_1 e(n-1)+\ldots+c_{n_c}(n-n_c) \end{aligned} \qquad (3)$$

c_i (i=1…n_c) are the MA parameters, which are neglected if vowels are processed, because the disturbances, if at all present in vowels, are insignificant compared to the signal.

In this experiment, training was carried out by the BPA parameter changing. The optimal step size was reached by the GD method (Haykin, 1994), (Svarer 1995), given with the formula

$$E \approx E_0 + \left(\frac{\partial E}{\partial \mathbf{u}}\right)^T \delta\mathbf{u} + \frac{1}{2}\delta\mathbf{u}^T \mathbf{H} \delta\mathbf{u} \qquad (4)$$

where E is the error, E_0 is its approximation, **u** is a parameter vector, $\delta\mathbf{u}$ is the parameter deviation, and **H** is the Hessian symetric matrix of the second derivates of *E.*

$$\mathbf{u} = [u_1, u_2, \ldots, u_n]^T$$

$$\frac{\partial E}{\partial \mathbf{u}} = \left[\frac{\partial E}{\partial u_1}, \frac{\partial E}{\partial u_2}, \ldots, \frac{\partial E}{\partial u_n}\right]^T$$

$$\mathbf{H} = \frac{\partial^2 E}{\partial \mathbf{u}^2} = \begin{bmatrix} \frac{\partial^2 E}{\partial u_1^2} & \frac{\partial^2 E}{\partial u_1 \partial u_2} & \cdots & \frac{\partial^2 E}{\partial u_1 \partial u_n} \\ \frac{\partial^2 E}{\partial u_2 \partial u_1} & \frac{\partial^2 E}{\partial u_2^2} & \cdots & \frac{\partial^2 E}{\partial u_2 \partial u_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 E}{\partial u_n \partial u_1} & \frac{\partial^2 E}{\partial u_n \partial u_2} & \cdots & \frac{\partial^2 E}{\partial u_n^2} \end{bmatrix}$$

The parameter estimates are obtained in the following way

$$\delta\mathbf{u} = \mathbf{u}^* - \mathbf{u} = -\mathbf{H}^{-1}\frac{\partial E}{\partial \mathbf{u}} = 0 \qquad (5)$$

$$\mathbf{u}^{*} = \mathbf{u} - \mathbf{H}^{-1}\frac{\partial E}{\partial \mathbf{u}} \tag{6}$$

where $\mathbf{u}^{*}$ is the estimated parameter vector. A problem that raises is finding $\mathbf{H}^{-1}$. The number of calculations of the *n* dimensional Hessian matrix inverse is n^3 that is computer demanding, so $\mathbf{H}^{-1}$ has to be approximated. One of robust but very simple methods for matrix approximation is known as the Levenberg-Marquart algorithm (Le Cun et al., 1989), (Svarer, 1995), which is presented later in the paper.

Weighted Least-Squares

The WLS is an recursive algorithm with slowly decreasing weights, which is found to have a self-convergence property, i.e., it almost certainly converges to a certain random vector, irrespective of the control low design (Childers et al., 1995). This universal convergence result combined with a method of random regularization can easily be applied to construct a self-convergent and uniformly controllable estimated model and thus enable making a general framework for adaptive control (Guo, 1996). The WLS is an efficient method that makes good use of small data sets, having the ability to provide different types of easily interpretable statistical intervals for estimation, prediction, calibration and optimization. Given a sequence of the stochastic observation vector θ є R^n, let us consider the scalar process y_t generated according to the following time-varying equation

$$y_{t+1} = \theta^T \varphi_t + \omega_{t+1}$$

The scalar ω_t is a disturbance term, and φ_t є R^n is a stochastic sequence of unknown parameter vectors (regressors). The LS fitting technique is the most commonly applied way to estimate the parameter φ_t by minimizing the sum of the squares of the residuals. The estimate is the minimizer of the following criterion

$$J_t(\theta) = \frac{1}{2}\sum_{i=0}^{t}(y_{i+1} - \theta^T\varphi_i)^2 = \frac{1}{2}\sum_{i=1}^{t} e(t)^2$$

When the squares of the residuals are used, outlying points can have a disproportionate effect on the fit. The WLS reflects the behavior of the random errors in the model by incorporating extra nonnegative constants, or weights, associated with each data point, into the fitting criterion. Optimizing the criterion to find the parameter estimates allows the weights to determine the contribution of each observation to the final parameter estimates. The WLS error function *e*(*i*) given with the formula (7)

$$J_t = \frac{1}{2}\sum_{i=0}^{t} \alpha_i e(i)^2 \tag{7}$$

whereas 0<α_i≤1 is the weighting sequence, a so-called forgetting factor, which allows different measurements of interest. The forgetting factor is introduced to discount old data in favour of fresh information. The selection of the value for *αᵢ* is a user's choice as have been discussed by Ljung and Soderstrom (1983), Goodwin and Sin (1984), and (Campi, 1994). The size of the weight indicates the precision of the information contained in the associated observation. The forgetting factor α_i usually takes the exponential form $\alpha_i = \lambda^{t-i}$, 0<λ<1. Writing the criterion with an exponential forgetting factor

$$J_t = \frac{1}{2}\sum_{i=0}^{t} \lambda^{t-i} e(i)^2$$

Assuming that the non-stationary signal consists of stationary segments (λ<1, λ≈1), the forgetting factor is:

$$\lambda^t = e^{t\ln(\lambda)} = e^{t\ln(1+\lambda-1)} \approx e^{-t(1-\lambda)},$$

$$\lambda^t =^{-t/\tau},\ \tau = 1/(1-\lambda) \tag{8}$$

where τ is the effective memory of the algorithm, i.e. the memory length.

The WLS parameter estimation can easily be constructed so that the corresponding estimated model is almost surely self-convergent and controllable. Using weights that are inversely proportional to the variance yields the most precise parameter estimates possible (Ljung, Soderstrom, 1983), (Guo, 1996). Consider the following ARMAX model

$$\begin{aligned}
&\mathbf{A}(z)y_t = \mathbf{B}(z)u_t + \mathbf{C}(z)w_t, && t \geq 0\\
&\mathbf{A}(z) = 1 + a_1 z^1 + \cdots + a_p z^p, && p \geq 0\\
&\mathbf{B}(z) = b_1 z^1 + \cdots + b_q z^q, && q \geq 1\\
&\mathbf{C}(z) = 1 + c_1 z^1 + \cdots + c_r z^r, && r \geq 0
\end{aligned}$$

where y_t, u_t, and w_t are the system output, input, and noise sequence, respectively, and **A**(*z*), **B**(*z*), and **C**(*z*) are polynomials in the backward-shift operator *z* with unknown coefficients and known upper bounds *p*, *q*, and *r*, for orders. To describe the WLS algorithm for estimating an unknow parameter wector

$$\boldsymbol{\theta} = [-a_1 \cdots - a_p b_1 \cdots b_q c_1 \cdots c_r]^T$$

the recursive algorithm is applied. It has the following form

$$
\begin{aligned}
\theta_{t+1} &= \theta_t + L_t(y_{t+1} - \theta_t^T) \\
L_t &= \frac{P_t \varphi_t}{\alpha_t^{-1} + \varphi_t^T P_t \varphi_t} \\
P_{t+1} &= P_t - \frac{P_t \varphi_t \varphi_t^T P_t}{\alpha_t^{-1} + \varphi_t^T P_t \varphi_t} \\
\varphi_t &= [y_t \cdots y_{t-p+1} u_t \cdots u_{t-p+1} \omega_t \cdots \omega_{t-r+1}] \\
\omega_t &= y_t - \theta_t^T \varphi_{t-1}, \ t \geq 0
\end{aligned}
$$

where α_t is the weighting sequence, and the initial values θ_0 and $P_0=\alpha\mathbf{I}$, ($0 < \alpha < 1$) are chosen arbitrary. Various versions of this algorithm are studied by many authors (Lee et al., 1981), (Ljung and Soderstrom 1983), (Campi, 1994), (Guo, 1996), (Macchi,1986), (Widrow et al., 1976), (Kovačević et al., 2000), (Jing 2012). Their work aims at studying the performances of algorithms in a stochastic framework. The following questions motivate almost all the papers pertaining to the performance analysis of adaptive identification algorithms: a) Is the algorithm able to keep estimation error bounded? b) What does the estimation error depend on and in what way?

The impact of the glottal signal on modeling

The noninvasive methodology for recording the signal that flows through the glottis, before it modulates into speech, is known as the Electro-GlottoGraphy (EGG). The method examines the vibration of the vocal cords by measuring the impedance through the throat of the subject. Electrodes are placed outside, on the larynx. When the vocal cords are closed together, electricity passes through the person's neck, and the impedance is low, while the opened vocal cords make that extremely difficult, and the impedance is high. The change of impedance indicates a change of the glottal flow.

According to Fant (1960), the speech wave is the response of the vocal tract filter system to the sound sources. This rule is know as the source-filter theory of speech production. For vowels, the source of sound is the regular vibration of the vocal cords, and the filter is a vocal tract tube between the larynx and the lips. Regular vibrations of the vocal cords result in the periodic excitation source, which is always in the larynx, usually in the glottis. A period is the duration of one glottal cycle (opening and closing phase). The waveform of the sound is complex, i.e. its wave-shape depends on the relationship between various frequencies that it contains. In the source-filter theory, the frequencies (formants) are responses of the vocal tract filter. Literature suggests that at least a pair of poles is needed for each formant representation (10-16 poles), which is expected in the frequency range, and another pair of poles for the impact of the glottal flow (Kovačević et al., 2000). The glottal-flow velocity can be thought of as a low-pass filter filtering of an impulse stream (Gutierrez-Osuma, 2011). Vowel '*a*', and a corresponding glottal signal (*egg*), for a female subject during normal phonation, are presented in Fig. 1. The sampling frequency for the signals is f_s=10kHz. Each sample is 0.1ms apart. Therefore, n=300 samples of signals is equivalent to the time period of 30ms.

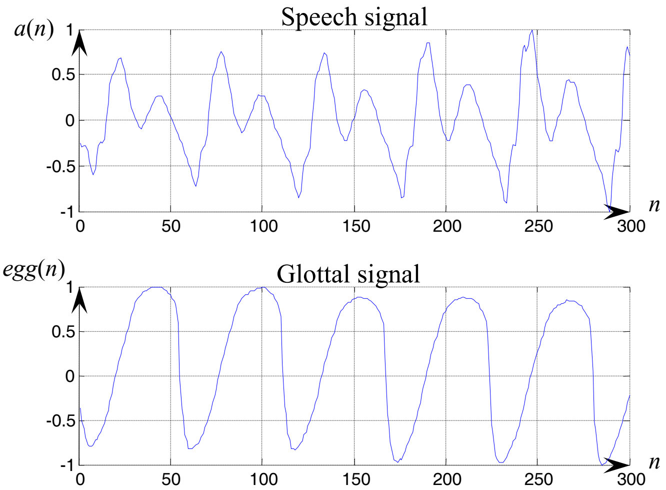

Figure 1 – Upper panel: time-domain speech signal (vowel '*a*').
Lower panel: glottal flow waveform of the vowel '*a*' (*egg*)
Slika 1 – Gore: govorni signal (vokal '*a*'). Dole: oblik glotalnog talasa za vokal '*a*' (*egg*)

Direct observation of the glottal behaviour is rather difficult, which implies the development of computational procedures for the estimation of the glottal source directly from the speech signal. Some of the most known and used models are Rosenberg (Degottex, 2010), Liljencrants-Fant (de Oliviera Dias, 2012), Klatt (Klatt, Klatt, 1990) and Strube model (Kovačević et al., 2000) that is given with the formula (9).

$$u_g(t)=\begin{cases}\sin^2\dfrac{\pi t}{2T_0}, & 0\le t\le T_s\\ \cos\dfrac{\pi(t-T_0)}{2T_n}, & T_s\le t\le T_{og},\quad T_{og}=T_s+T_n\\ 0, & T_{og}\le t\le T_0,\quad T_0=T_{og}+T_{cg}\end{cases}\tag{9}$$

where $u_g(t)$ is the glottal flow, T_0 is the fundamental frequency period, T_{cg} and T_{og} are the periods of the open and the close phase of the glottal wave, respectively, T_s and T_n indicate the slow growth phase (T_s) and the phase of fast decrease (T_n), which make the phase of the open glottis (T_{og}). But in this paper, the influence of the glottal signal obtained by the EGG on the prediction of the corresponding speech signal is examined. The polynomial model of the glottal flow is used as an exogenous part of the ARX and ARMAX models. Also, it is used to improve the training of the FNN.

Feed-forward neural network learning

Each nonlinear system can be modeled by the dynamic parameter function

$$g(y_t, \delta_t, t) = \varepsilon(t)$$

where $\delta_t^{\mathrm{T}}=[-y_{-1},\ldots,-y_{t-n}]$ is the vector of n samples of the sequence y, $\varepsilon(t)$ is an error, and g is the parametric function known in advance (Svarer, 1995), (Arsenijević, Milosavljević, 2000). It is shown that the FNN with three layers (input, hidden and output), and sigmoidal-type nonlinearity can approximate any nonlinear function and generate any complex decision region needed for clasification and recognition tasks (Azimi-Sadjadi, Liou, 1992), if the choice of inputs, the dimensionality of weight space and the transition of learning are properly suited. For the given inputs and weights, the output of the FNN is given with the following expression

$$y_i(\mathbf{w},\mathbf{W}) = F_i\left(\sum_{j=1}^{q} W_{ij} f_j\left(\sum_{l=1}^{m} w_{ij} z_l + w_{j0}\right) + W_{f0}\right)$$

y_i is the output, **w** and **W** are the synaptic weight matrices, f_j and F_i are the activation functions of the hidden and output layer, respectively, while q and m represent the number of nodes in the network (Arsenijević, 2001).

The problem of the neural network learning can be seen as a function optimization problem. Let us consider the FNN with differentiable activation functions of both input variables and weights. Each unit computes a weighted sum of its inputs

$$a_j = \sum_i w_{ji} z_i$$

where z_i is the activation which sends a connection to the unit j, and w_{ji} is the weight associated with the connection. The summation is transformed by a nonlinear function $g(\ldots)$ to give the activation z_j of the unit j in the form $z_i=g(a_i)$ The error function, which is a sum of all paterns in the training set, is defined on each pattern separately

$$E = \sum_b E^n$$

where $E^n = E^n(y_{1,\,\ldots,}\, y_c)$. The goal is to evaluate derivatives of the error E^n with respect to the weights

$$\frac{\partial E^n}{\partial w_{ij}} = \frac{\partial E^n}{\partial a_j}\frac{\partial a_j}{\partial w_{ij}} \tag{10}$$

where

$$\delta_j = \frac{\partial E^n}{\partial a_j}, \quad \frac{\partial a_j}{\partial w_{ji}} = z_i \tag{11}$$

which gives

$$\frac{\partial E^n}{\partial w_{ij}} = \delta_j z_i$$

For the output units, the error δ_k is given with the equation

$$\delta_k = \frac{\partial E^n}{\partial a_k} = g'(a_k)\frac{\partial E^n}{\partial y_k}$$

where $g'(a)$ substitutes $\partial E^n/\partial y$, while for the hidden units

$$\delta_j = \frac{\partial E^n}{\partial a_j} = \sum_k \frac{\partial E^n}{\partial a_k}\frac{\partial a_k}{\partial a_j}$$

that gives the back-propagation formula:

$$\delta_j = g'(a_j)\sum_k w_{kj}\delta_k$$

δ's can be evaluated backward since δ's from the outputs are known.

The BPA can also be applied for the calculation of other derivatives. Let us consider the evaluation of the *Jacobian* matrix, whose elements are given by the derivatives of the network outputs y_k with respect to the network inputs x_i

$$J_{ki} = \frac{\partial y_k}{\partial x_i} = \sum_k \frac{\partial y_k}{\partial a_j}\frac{\partial a_j}{\partial x_i} = \sum_j w_{ji}\frac{\partial y_k}{\partial a_j} = \sum_j w_{ji}\sum_l \frac{\partial y_k}{\partial a_l}\frac{\partial a_l}{\partial a_j}\cdots$$

$$\cdots = \sum_j w_{ji} g'(a_j)\sum_l w_{lj}\frac{\partial y_k}{\partial a_l} \tag{12}$$

To evaluate second order derivatives, let us consider the following error derivatives:

$$\frac{\partial^2 E}{\partial w_{ij}\partial w_{lk}} = \sum_n \frac{\partial y^n}{\partial w_{ji}}\frac{\partial y^n}{\partial w_{lk}} + \sum_n \left(y^n - t^n\right)\frac{\partial^2 y^n}{\partial w_{ji}\partial w_{lk}} \tag{13}$$

are the elements of the *Hessian* matrix. If the network outputs y^n are very closely to the target values t^n, then the second term in (13) can be neglected, which gives an LM formula:

$$\frac{\partial^2 E}{\partial w_{ij}\partial w_{lk}} = \sum_n \frac{\partial y^n}{\partial w_{ji}}\frac{\partial y^n}{\partial w_{lk}}$$

The LM algorithm provides a numerical solution to the problem of minimizing a (generally nonlinear) function, over a space of the parameters for the function (weights) (Kashyap, 1980), (Ljung, 1987), (Larsen, 1993), (Hansen, Rasmusen, 1994), (Fahlman, 1988). See (5)-(6). The LM basically consists of solving the equation

$$(\mathbf{H} + \lambda\mathbf{I})\boldsymbol{\delta} = \mathbf{J}^T\mathbf{E}$$

where *λ* is the Levenberg's damping factor adjusted at each iteration guiding the optimization process, and **δ** is the weight update vector that shows how much the network weights should be changed to achieve a better solution. If the reduction of **E** is rapid, a smaller value of *λ* brings the algorithm closer to the Gauss-Newton algorithm, whereas if the iteration gives insufficient reduction in the residual, *λ* can be increased, giving a step closer to the GD direction.

The problem of parameter adjustment (see 13) has been solved by Hassibi and Stork (1993). They have used the outer product approximation to develop a computationally efficient procedure for approximating the inverse of Hessian:

$$\mathbf{H}_N = \sum_{n=1}^{N} \mathbf{g}^n \left(\mathbf{g}^n\right)^T$$

where *N* is the number of the parameters in the data set, and the vector **g** is the gradient of the error function. The sequential procedure for building up the Hessian is obtained by separating the contribution from the data point *N*+1 to give:

$$\mathbf{H}_{N+1} = \mathbf{H}_N + \mathbf{g}^{N+1}\left(\mathbf{g}^{N+1}\right)^T$$

In order to evaluate the inverse Hessian, let us consider the matrix identity:

$$(\mathbf{A} + \mathbf{BC})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}\left(\mathbf{I} + \mathbf{CA}^{-1}\mathbf{B}\right)^{-1}\mathbf{CA}^{-1}$$

where **I** is the identity matrix. If **A=H**$_n$, **B=g**$^{N+1}$, **C= (g**$^{N+1}$)$^{\mathrm{T}}$

$$\mathbf{H}_{N+1}^{-1} = \mathbf{H}_{N}^{-1} - \frac{\mathbf{H}_{N}^{-1}\mathbf{g}^{N+1}\left(\mathbf{g}^{N+1}\right)^{T}\mathbf{H}_{N}^{-1}}{1+\left(\mathbf{g}^{N+1}\right)^{T}\mathbf{H}_{N}^{-1}\mathbf{g}^{N+1}} \tag{14}$$

The initial matrix **H**$_0$ is chosen to be α**I**, where α is small quantity, so that the algorithm actually finds the inverse of **H**+α**I**.

The updating parameter procedure is carried out in the following way:

Step 1: propagate the input signal through the FNN in the forward direction to obtain actual outputs for each training signal, at each layer.

Step 2: generate the output signal at the output of each layer for each node. At the output layer, this error is simply formed by comparing the actual outputs with the desired signal. For other layers, the error is propagated backward through those layers with updated weights until the errors at the outputs of the lower layer with weights to be updated are generated.

Step 3: compute the matrices for updating weights.

Step 4: determine the state of the particular node. If the input to this node is within the ramp region, then proceed; otherwise, there is no need for weight updating and then examine the next node.

Step 5: Update the weight vector using the recursion, and repeat steps 4 and 5 for the next node until all the weight vectors in this layer are updated.

These steps are performed for all the layers several times for a given training set until the error converges to within an acceptable range. After the network updating is finished, the pruning of parameters is carried out in the following way

$$\delta u_m + u_m = 0$$

$$\mathbf{e}_m^T \delta\mathbf{u} + u_m = 0$$

where u_m is the m-th parameter, $\mathbf{e}_m$ is the unit vector of the same dimensions as $\delta\mathbf{u}$. The objective of this methodology is to prune the parameter u_m that would cause minimum increase of an error in the following way

$$\delta\mathbf{u} = -\lambda\mathbf{H}^{-1}\mathbf{e}_m$$

$$\lambda = \frac{u_m}{\mathbf{e}_m^T\mathbf{H}^{-1}\mathbf{e}_m}$$

$$\delta\mathbf{u} = -\frac{u_m}{\mathbf{e}_m^T\mathbf{H}^{-1}\mathbf{e}_m}\mathbf{H}^{-1}\mathbf{e}_m$$

The Hessian matrix inverse is used to identify the least significant weights (Silva et al., 2008).

Results

A comparative analysis of five different models, estimated on the basis of speech and glottal signals, provides the understanding of the impact of the glottal signal on the estimation of the model parameters. The evaluation criterion is the minimum training error, which is presented graphically and in the percentages for all the models. The parameters are estimated on 300 samples of a vowel '*a*' and the corresponding glottal signal, pronounced by a female speaker, during normal phonation. The sampling frequency is 10 kHz. The training set length captures about 5 glottal cycles (not quite). For the evaluation of the algorithms, the MATLAB functions are applied. The hyperbolic tangent (tanh) function is the activation function for all neurons, because it is the rational function of exponential, i.e. the first and second derivatives of tanh always exist (Wall, 1948.). Since the output of tanh was limited to approximately [-1, 1] for all inputs within [-1, 1], speech and glottal signals were also normalized to the same limits.

The model orders were as follows:

AR:	n_a=25
ARX:	n_a=14, n_b=4
ARMAX:	n_a=14, n_b=4, n_c=1
WLS:	n_a=25, α=0,95.

where n_a corresponds to the speech, n_b to the glottal signal, n_c to the output error, and α to the initial weight. The FNN with one hidden layer is trained using a training set that consists of 14 samples of speech and 4 samples of the glottal signal for the prediction of one speech sample. A hidden layer contains three neurons and the output layer a single neuron, i.e. the network structure is 18-3-1. Prior to the training, the weights were initialized to small random numbers. The LM training is used to progressively reduce the total network training error.

Fig. 2 shows the training sets of a vowel '*a*' and the egg signal as well as the corresponding training errors for the WLS (u_{hatt}), FNN ($e_{NNARMAX}$), AR (e_{AR}), ARX (e_{ARX}) and ARMAX (e_{ARMAX}) models, respectively.

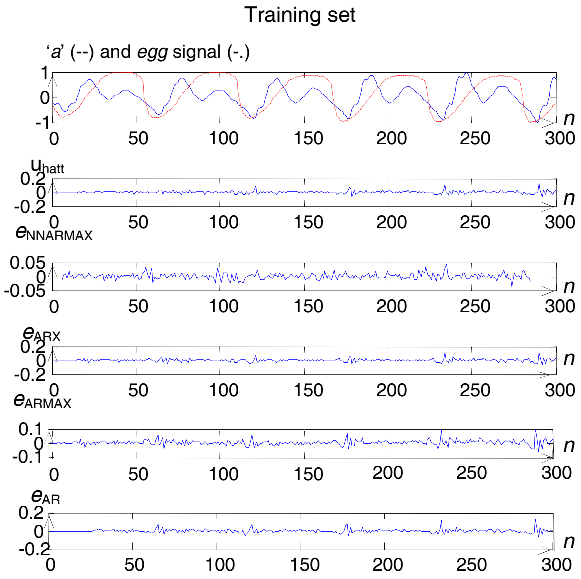

Figure 2 – Vowel '*a*' and the glottal e_{gg} signal (1), u_{hatt} (2), e_{NNARMAX} (3), e_{ARX} (4), e_{ARMAX} (5) and e_{AR} (6) – training set
Slika 2 – Vokal '*a*' i glottalni e_{gg} signal (1), u_{hatt} (2), e_{NNARMAX} (3), e_{ARX} (4), e_{ARMAX} (5) i e_{AR} (6) – obučavajuci skup

The speech and glottal signal sets contain 600 samples. The sets are divided into two equal parts. The training set (300 samples) is composed of the first 300 samples while the test set consists of the following 300 samples.

For all models, the errors indicate the opening and closing of the vocal cords. As expected, the training of the FNN gives the lowest error value. The input/output mapping function shows that the model does find the minimum training error. Also, the e_{ARMAX} shows a large impact of the glottal signal on the model evaluation, which is not the case for the e_{AR}, e_{ARX} and u_{hatt}. As expected, the error of the FNN model is the lowest in comparison to other errors.

After training, the models are tested. The results, which are presented in Fig. 3, show that all the test errors are higher than the corresponding training errors. The u_{hatt} is about four times higher, the e_{ARMAX} increase is about three times, and the e_{AR} and the e_{ARX} are doubled. The FNN shows the increase of the error a little less than four times; however, this value is also significantly lower than the values of the test errors of other models.

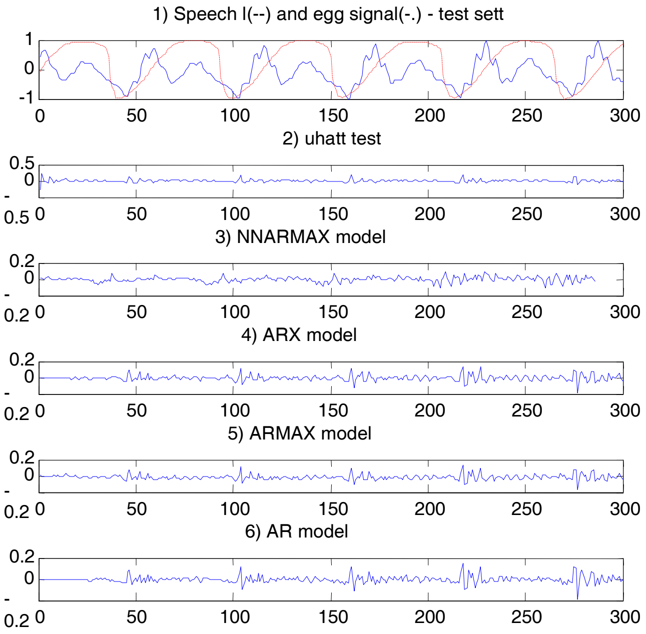

Figure 3 – Vowel '*a*' and the glottal e_{gg} signal (1), u_{hatt} (2), e_{NNARMAX} (3), e_{ARX} (4), e_{ARMAX} (5) and e_{AR} (6) - test set

Slika 3 – Vokal '*a*' i glottalni e_{gg} signal (1), u_{hatt} (2), e_{NNARMAX} (3), e_{ARX} (4), e_{ARMAX} (5) i e_{AR} (6) - test skup

Table 1 summarizes the results of the minimum and maximum training and the test errors for the given models.

Table 1

Minimum and maximum errors for the training and the test sets

Tabela 1 Minimalne i maksimalne greške za obučavajuci i test skup

	Training set		Test set	
Error	min	max	min	max
u_{hatt}	-0,0646	0,1285	-0,2646	0,2305
e_{ARMAX}	-0,0614	0,0996	-0,1777	0,1213
e_{AR}	-0,0737	0,1251	-0,1956	0,1411
e_{ARX}	-0,0738	0,1084	-0,1893	0,1308
e_{NNARMAX}	-0,0366	0,0440	-0,1161	0,0927

Results indicate the following:

Training error of the AR model (n_a=25), having twice of the required parameters than common AR models (n_a=10-16), is similar to the one of the ARX model (n_a=14, n_b=4).

u_{hatt} and e_{ARMAX} (n_a=14, n_b=4, n_c=1) are smaller than the errors of AR and ARX models, which shows the impact of the weights on the LS modelling, as well as the impact of the glottal signal and the output error on the prediction of speech.

Training error of three-layer FNN that has 18 inputs (14 inputs for speech and four inputs for glottal signal samples) provides almost half the training error than other models.

Test errors are 3-4 times higher than the training errors for each model, which is particularly noticeable for u_{hatt}.

Minima and maxima of the test errors for AR, ARX and ARMAX models differ in ~1%.

The results show that the WLS model has the greatest volatility in testing; the u_{hatt} is approximately two times higher than the errors of other linear models.

After testing, the FNN does not show changes in characteristics, although the test error is slightly higher.

Conclusion

This paper presents the impact of the glottal signal on the prediction of speech, which is based on linear and nonlinear models. AR, ARX, ARMAX models, the WLS algorithm and the FNN are used for the prediction. The training of the models is performed on a vowel ‘a’, pronounced by a female speaker, during normal phonation.

For the training, the BPA is used for fitting the model parameters. The parameter change is carried out by propagation along the negative of the gradient to find a minimum of the error function. The LM algorithm, which is used to speed up and ease calculation of the Hessian matrix, showed significant advantage over the GD algorithm. The LM combines the minimization along the negative direction of the gradient and the Newton method based on a quadratic model to speed up the process of finding the minimum of a function.

A comparative analysis of the training and the test errors shows that the high-order AR model as well as the WLS algorithm give higher errors if compared with ARX and ARMAX models for which the glottal signal influence prediction. The training errors show that the impact of the glottal signal is higher in the phase of the open glottis than in the phase of the closed glottis. The LP models also show robustness. The results indicate that the high-order AR model can be an adequate substitute for the ARX model if the glottal signal is not available for the prediction. The

WLS model improves prediction by including the weight parameter, while the ARMAX model shows significant reducing of the training errors, because of the glottal signal and the output error, which were used for training. The results also show the minimum error of the FNN model. The FNN with one hidden layer and tanh activation functions for all neurons showed that its input-output mapping gives the model that predicts the speech signal much more precisely than linear models.

According to the results, if the glottal signal is available for model training, the FNN should be used whenever possible, due to the precision of estimates, although the sensitivity of the model is increased and training time takes longer. However, if this is not the case, the high-order AR model can be a replacement for ARX and ARMAX models. The results of the WLS training show that, although the training gives satisfying results, the testing shows higher errors, so models based on WLS should not be used for this purpose.

References

Akaike, H., 1969, *Fitting Autoregressive Models for Prediction*, Ann. Ins. Stat. Mat.

Arsenijević, D., 2001, *Analiza neuronskih modela vokala srpskog jezika*, Magistarski rad, Elektrotehnički fakultet, Beograd.

Arsenijević, D., Milosavljević, M., 2000, O jednoj meri rastojanja govornih signala zasnovanoj na neuronskim modelima, *Zbornik radova DOGS*, Novi Sad.

Azimi-Sadjadi, M.R., Liou, R., 1992, Fast Learning Process of Multilayer Neural Networks Using Recursive Least Squares Method, *IEEE Transaction on Signal Processing*, Vol. 40, No. 2, pp.446-450.

Burrows, T.L., 1996, *Speech Processing with Linear and Neural Network Models*, PhD Thesis, Queens' College, Cambridge University, England.

Campi, M.C., 1994, Exponentially Weighted Least Squares Identification of Time-Varying Systems with White Disturbances, *IEEE Transactions on Signal Processing*, Vol. 42, No. 11, pp.2906-2914.

Childers, D.G., Principe, J.C. Thing, Z.T., 1995, Adaptive WRLS_VFF for Speech Analysis, *IEEE Transactions on Speech and Audio Processing*, pp.209-213.

Degottex, G., 2010, *Glottal source and vocal-tract separation. Estimation of glottal parameters, voice transformation and synthesis using glottal model.* PhD thesis, Universite Paris, France.

De Oliviera Dias, S., 2012, *Estimation of the glottal pulse from speech or singing voice, Master's Thesis*, School of Engineering of University of Porto.

Fahlman, S.E., 1988, Fast-learning variation on back propagation: An empirical study, pp.38-51., *Proceedings of the 188 Connectionist Model Summer Schools*, San Mateo, Pittsburgh, USA.

Fant, G.1960, *Acoustic Theory of Speech Production*. Mouton, The Hague

Guo, L., 1996, Self-Convergence of Weighted Least-Squares with Applications to the Stochastic Adaptive Control, *IEEE Transaction on Automatic Control*, Vol. 41, No. 1, pp. 79-89.

Gutierrez-Osuna, R., 2011, *Introduction to speech processing, CSE@TAMU*, Available at: http://research.cs.tamu.edu/prism/lectures/sp/l8.pdf

Hansen, L.K., Rasmusen, C.E., 1994, Pruning from adaptive regularization, *Neural Computation* vol. 6, no. 6, pp.1223-1232.

Hassibi, B., Stork, D.G., 1993, Second order derivatives for network pruning: optimal brain surgeon. In S.J. Hanson, J.D. Cowan, C.L. Giles (Eds.) *Advances in Neural Information Processing Systems*, Volume 5, pp.164-171.

Haykin, S., 1994, *Neural networks: A comprehensive foundation*, New York: Macmillan.

Jing, X., 2012, Robust adaptive learning of feed forward neural networks via LMI optimizations, *Neural Networks* 31, pp.33-45.

Kashyap, R.L., 1980, Inconsistency of the AIC Rule for Estimating the Order of AR Models, *IEEE Transaction on Automatic Control.* AC-25, pp.996-998.

Klatt, D., Klatt, L., 1990, Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of Acoustical Society of America* 87, pp.820-257.

Kovačević, B., Milosavljević, M., Veinović, M., Marković, M., 2000, *Robusna digitalna obrada govornog signala*. Akademska misao, Beograd.

Larsen, J., 1993, *Design of Neural Networks*, Ph.D. Thesis, Electronic Institute, DTH, Lyngby.

Ljung, L., 1987, *System Identification: Theory for the User*, Prentice Hall Inc.

Ljung, L., Soderstrom, T., 1983, *Theory and Practice of Recursive Identification*, Cambrige MA: MIT Press, p.36.

Le Cun, Y., Denker, J.S., Solla, S.A., 1989, Optimal Brain Damage, *Advances in Neural Information Processing Systems* 2, pp.598-605.

Levenberg, K., 1944, A Method for the Solution of Certain Problems in Least Squares, *Quart. Appl. Math.* Vol. 2, pp.164-168.

Marquardt, D., 1963, An Algorithm for Least-Squares Estimation of Nonlinear Parameters, *SLAMJ. Appl. Math.* Vol. 11., pp.431-441.

Milićević, M.R., Župac, Ž.G., 2012, Objektivni pristup određivanju težina krierijuma, *Vojnotehnički glasnik/Military technical courier*, Vol. 60, No. 1, pp.39-56.

Narendra, K.S., Parthasaranthy, K., 1990, *IEEE Transactions on Neural Networks,* 1, p.4.

Pamučar, S.D., Đorović, D.B., 2012, Optimizing models for production and inventory control using genetic algorithm, *Vojnotehnički glasnik/Military technical courier*, Vol. 60, No. 1, pp.14-38.

Riecke, L., Esposito, F., Bonte, M., Formisano, E., 2009, Hearing illusory sound in noise: the timing of sensory-perceptual transformations in auditory cortex, *Neuron* 64, pp.550-561.

Sainath, T.N., Kingsbury, B., Ramabhadran, B., Fousek, P. Novak, P., Mohamed, A., 2011, Making deep belief networks effective for large vocabulary continous speech recognition, In *Automatic Speech Recognition and Understanding,* pp.30-35., *2010 IEEE Workshop*, 11-15 December 2011, Waikoloa, HI.

Shahin, A.J., Pitt, M.A., 2012, Alpha activity making world boundaries mediates speech segmentation, *European Journal of Neuroscience*, 36, pp.3740-3748.

Silva, L., Marques de Sa, J., Alexandre, L.A., 2008, Data classification with multilayer perceptrons using a generalized error function, *Neural Networks* 21, pp.1302-1310.

Svarer, C., 1995, *Neural networks for signal processing*, Technical University of Denmark.

Wall, H.S., 1948, *Analytis Theory of Continued Fractions*, New York: Chelsea

Wu, W., Wang, J., Cheng, M., Li., Z., 2011, Convergence analysis of online gradient method for BP neural networks, *Neural Networks* 24, pp.91-98.

ВЛИЯНИЕ ГЛОТТАЛЬНОГО СИГНАЛА НА ПРЕДИКЦИЮ РЕЧИ

ОБЛАСТЬ: телекоммуникация
ВИД СТАТЬИ: оригинальная научная статья
ЯЗЫК СТАТЬИ: английский

Краткое содержание:

В настоящей работе рассматриваются несколько линейных и нелинейных способов распознавания речи, основанных на моделях: АР, АРХ, АРМАХ, и алгоритмах WЛС и ФНН. В работе подробно представлено влияние глоттального сигнала. Апроксимации ГД, БПА и ЛМ, применяемые при обучении и оптимизациии. Проведен сравнительный экспериментальный анализ пяти исследуемых моделей, основанных на предикции речевого сигнала. В работе приведены результаты обучения и тестирования, полученные на материале допущенных ошибок, при обучении применяемых моделей.

Ключевые слова: *линейные модели, предикция, глоттальный сигнал, нейронная сеть, речь.*

UTICAJ GLOTALNOG SIGNALA NA PREDIKCIJU GOVORA

OBLAST: telekomunikacije
VRSTA ČLANKA: originalni naučni članak
JEZIK ČLANKA: engleski

Rezime:

U radu je prikazano nekoliko linearnih i nelinearnih tehnika za obradu govora, koje su zasnovane na AR, ARX, ARMAX modelima, WLS algoritmu i FNN. Detaljno je opisan uticaj glotalnog signala. GD, BPA i LM aproksimacija korišćene su za obučavanje i optimizaciju. Izvedena je komparativna, eksperimentalna analiza pet razmatranih modela koja je zasnovana na predikciji govornog signala. Rezultati obučavanja i testiranja predstavljeni su pomoću grešaka dobijenih u fazi učenja i treninga za svaki od modela.

Uvod

Kad nastaje govor, vazduh iz pluća, preko trahee, ulazi u grlo i pobuđuje glasne žice, koje menjaju njegov protok, pa novonastali signal prolazi kroz glotalni i vokalni trakt, gde oblik usne i nosne šupljine, jezika i zuba formira signal govora. Ukoliko su glasne žice razdvojene, vazduh prolazi između njih i nastaje šumolik signal male snage, a ukoliko su sastavljene, potisak iz pluća ih tera da kvaziperiodično vibriraju formirajući snažan signal, tj. vokal.

Najpoznatija tehnika za obradu govora je linearna predikcija (LP), koja koristi source-filter sistem za modelovanje sistema, koji podrazumeva da je pobuda locirana na glotisu, dok se linearan filter koristi za modelovanje frekvencijskih karakteristika vokalnog trakta. Takođe, koriste se AR, ARX i ARMAX modeli, čiji se parametri procenjuju na osnovu odbiraka govora (AR), glotalnog signala (X) i uticaja greške (MA). Iako se uglavnom podrazumeva da su podaci odziva takvi da imaju istu varijansu, ukoliko ova pretpostavka nije tačna koristi se weighted Least Squares (WLS) tehnika, kojom se procenjena greška koriguje težinskim faktorima.

Kada je narušena ulazno-izlazna dinamika sistema, odnosno kada sistem sadrži nelinearne komponente, koriste se nelinearni modeli kao što je višeslojni perceptron (MLP), koji omogućuju modelovanje po proceduri obučavanja koja je zasnovana na podešavanju sinaptičkih težina koje su organizovane po slojevima i međusobno povezane. MLP je Feed-Forward neuronska mreža (FNN), što znači da se mapiranje izvodi u smeru od ulaza ka izlazu. Parametri mreže podešavaju se propagacijom greške unazad (BPA) po principu pada gradijenta (GD).Za ubrzavanje ove procedure koristi se Levenberg-Marquardt (LM) koji omogućuje smanjenje broja operacija u podešavanju parametara mreže, direktnom procenom Hessianove matrice. Trening i test greške za sve modele korišćene su radi poređenja dobijenih rezultata.

Linearni i nelinearni parametarski modeli

Analiza i sinteza govornog signala često se izvode zajedno. Analitičkim procesom utvrđuju se karakteristike izvora signala, glotisa i vokalnog trakta. Sintezom se dobijaju signali koji mogu koristiti za prepoznavanje govora ili govornika, simulaciju ili otklanjanje pratećih, neželjenih efekata na sintetizovani signal. Analiza signala podrazumeva ili analizu fonetskih karakteristika ili analizu izgovorenog sadržaja, ali je nivo greške procene visok, a metodologija procene podrazumeva širok spektar modela sa velikim stepenom slobode. Kod analize signala uvek postoji problem nepoznavanja izvora pobudnog signala, glotalnog talasa i prenosne funkcije vokalnog trakta. Kod sinteze signala pobudni signal na ulazu u filter za sintezu može se podeliti na generator impulsa i generator šuma ili se može koristi pobudni signal dobijen LPC analizom govornog signala. Ova tehnika koristi se da bi bio obezbeđen visok kvalitet govora, uz pretpostavku da je odbirak govornog signala linearna kombinacija uzastopnih, prethodnih odbiraka. Formira se linearna kombinacija n prethodnih odbiraka, a optimizacija se vrši minimizacijom greške predikcije. Dobar LP model može biti jednostavan, a davati zadovoljavajuće rezultate i na taj način imati prednost nad složenim, nelinearnim modelima. Najčešće korišćen LP model kod predikcije govornog signala je AR model. Ukoliko je u procesuiranju govornog signala dostupan i glotalni signal, moguće je formirat ARX, a generalizacija ovog modela uključuje i propagaciju greške, pa se primenjuje ARMAX model.

Pored klasičnog LS modela koristi se Weighted Least Squares (WLS) algoritam kod kojeg težinski faktori utiču na poboljšanje greške predikcije. Na ovaj način, težinama se koriguje varijansa greške, čime se poboljšava procena parametara modela. WLS je efikasan metod koji je dobro koristiti na malom skupu podataka. U radu, WLS algoritam rešava probleme konvergencije i uniformnosti.

U radu je opisana neinvanzivna metoda za snimanje signala sa glotisa koja je poznata pod nazivom elektroglotografija (EGG). Osnova metode je ispitivanje vibracija glasnih žica, merenjem impendanse kroz vrat ispitanika. Elektrode se stavljaju spolja, na larings. Kada su glasne žice zatvorene struja iz elektroda može da prolazi kroz njih i impendansa je mala, dok je kod otvorenih glasnih žica impendansa viša. Promena impendanse ukazuje na promenu karakteristike glotisa.

Direktna opsetvacija ponašanja glotisa je teška, što je uticalo na pojavu različitih računarskih procedura koje estimiraju glotalnu pobudu na osnovu izmerenog govornog signala. Jedan od najpoznatijih modela – Štrubeov model prikazan je u tekstu. Međutim, u proceni navedenih modela, glotalni signal je bio dostupan, pa je ova relacija navedena zbog primera. U radu je glotalni signal koršćen kao X deo kod procenjenig ARX i ARMAX modela, kao i za obučavanje FNN.

Nelinearni sistemi mogu se modelovati dinamičkom, nelinearnom, parametarskom prenosnom funkcijom. Po literaturi, FNN sa jednim skrivenim slojem i sigmoidalnim prenosnim funkcijama može generisati rešenja kompleksnih problema kao što su klasigikacija, prepoznavanje oblika i slično, ukoliko je izbor težina, dimenzija i pravila obučavanja adekvatan. Problem kod obučavanja neuronske mreže može se posmatrati kao optimizaciona funkcija, pri kojoj težine moraju da budu diferencijabilne. Greška se računa za svaku težinu i sve slojeve ponaosob, a zatim se njihove vrednosti menjaju propagacijom unazad. Za minimizaciju greške predikcije koristi se LM algoritam. U osnovi, LM algoritam je numeričko rešenje problema nelinearne funkcije, po vektoru parametara. Algoritam koristi dumping faktor kojim se LM približava Gauss-Newtonovom (GN) algoritmu za veliki korak greške, odnosno GD za manje vrednosti greške. Vrednost Hessianove matrice računa se iterativno, kao i vrednost inverzne Hessianove matrice.

Podešavanje parametara izvodi se u pet koraka: propagacija ulaznog signala ka izlazu, generisanje izlaznog signala na osnovu strukture mreže, proračun težinskih matrica, određivanje stanja za svaki čvor ponaosob i podešavanje vektora težina unazad. Nakon što je UI mapiranje mreže završeno, može se koristit pruning, tehnika kojom se odbacuje višak parametara modela.

Rezultati

U eksperimentima je za obučavanje AR, ARX i ARMAX modela, WLS i FNN korišćeno 600 odbiraka ženskog fonema 'a'. Broj parametara primenjenih modela bio je: AR (n_a=25), ARX (n_a=14, n_b=4), ARMAX

(n_a=14, n_b=4, n_c=1). Visoki red AR modela primenjen je da se proveri da li ima potrebe za uvođenjem glotalnog signala kod linearnog modelovanja. Kod nelinearnih modela korišćeni su isti redovi modela kao i za linearne modele, a broj ulaznih podataka odgovarao je broju ulaza u linearne modele. Prikazane su greške obučavanja i testiranja, koje ukazuju na činjenicu da slične rezultate daju AR i ARX modeli, WLS i ARMAX modeli, dok je greška na FNN znatno manja od ostalih grešaka, što je posebno primetno kod test skupa.

Zaključak

Rad predstavlja uticaj glotalnog signala na predikciju govora koja je bazirana na linearnim i nelinearnim modelima. AR, ARX i ARMAX modeli, WLS algoritam i FNN korišćeni su u predikciji. Modeli su obučavani na vokalu 'a' koji je izgovorila žena tokom normalne fonacije. Za obučavanje BPA je korišćen za podešavanje parametara modela. Promena parametara izvedena je propagacijom po pravcu negativnog gradijenta, za minimizaciju funkcije greške. LM algoritam, koji je korišćen da ubrza i olakša izračunavanje Hesianove matrice, pokazao je značajne prednosti nad GD algoritmom. LM kombinuje minimizaciju po pravcu negativnog gradijenta i Njutnov metod.

Komparativna analiza koja je zasnovana na trening i test greškama pokazuje da AR model sa velikim brojem parametara i WLS algoritam, koji su bazirani isključivo na govoru, daju veću grešku ukoliko se uporede sa ARX i ARMAX modelima, kod kojh glotalni signal utiče na predikciju. Trening greške pokazuje da je uticaj glotalnog signala veći u fazi otvorenog glotisa. ARX modeli i WLS poboljšavaju predikciju i znatno redukuju grešku. Rezultati, takođe, ukazuju na veću tačnost, odnosno minimum greške za FNN. FNN sa jednim skrivenim slojem i tanh aktivacionim funkcijama svih neurona pokazuje da njeno ulazno-izlazno preslikavanje može preciznije da prediktuje govorni signal od svih drugih modela.

Na osnovu svega što je ranije izneseno, može se zaključiti da, ukoliko je glotalni signal dostupan, FNN treba koristiti kad god je to moguće, zbog preciznosti procena, iako je osetljivost modela povećana, a vreme obučavanja traje duže. Ipak, ukoliko to nije slučaj, AR modeli visokog reda mogu biti zamena za ARX ili ARMAX modele. Obučavanje WLS pokazuje malu trening grešku. Međutim, kod testiranja greška izuzetno raste, pa modele zasnovane na WLS ne bi trebalo koristiti u ove svrhe.

Ključne reči: *linearni modeli, predikcija, glotalni signal, feed-forward neuronska mreža, govor.*

Datum prijema članka/Paper received on: 25. 06. 2014.

Datum dostavljanja ispravki rukopisa/Manuscript corrections submitted on: 15. 08. 2014.

Datum konačnog prihvatanja članka za objavljivanje/Paper accepted for publishing on: 17. 08. 2014.