A COMPARATIVE ANALYSIS OF SERBIAN PHONEMES: LINEAR AND NON-LINEAR MODELS

Danijela D. Protić
General Staff of the Serbian Army, Department
of Telecommunications and Information Technology (J-6),
Centre for Applied Mathematics and Electronics, Belgrade

DOI: 10.5937/vojtehg62-5170

FIELD: Telecommunications
ARTICLE TYPE: Original Scientific Paper
ARTICLE LANGUAGE: English

Summary:

This paper presents the results of a comparative analysis of Serbian phonemes. The characteristics of vowels are quasi-periodicity and clearly visible formants. Non-vowels are short-term quasi-periodical signals having a low power excitation signal. For the purpose of this work, speech production systems were modelled with linear AR models and the corresponding non-linear models, based feed-forward neural networks with one hidden-layer. Sum squared error minimization as well as the back-propagation algorithm were used to train models. The selection of the optimal model was based on two stopping criteria: the normalized mean squares test error and the final prediction error. The Levenberg-Marquart method was used for the Hessian matrix calculation. The Optimal Brain Surgeon method was used for pruning. The generalization properties, based on the time-domain and signal spectra of outputs at hidden-layer neurons, are presented.

Key words: *AR model; neural networks; speech.*

Introduction

For several years now, neural network (NN) models have enjoyed wide popularity, being applied to problems of regression, classification, computational science, computer vision, data processing and time series analysis (Haykin, 1994). They have been also successfully used for the identification and the control of dynamical systems, mapping

e-mail: adanijela@ptt.rs

the input-output representation of an unknown system and, possibly, its control law (Narendra, Parthasaranthy, 1990). Perhaps the most popular to date artificial neural networks (ANN) in speech recognition is the multilayer perceptron (MLP) which organizes non-linear hidden units into layers and has full weight connectivity between adjacent layers (Sainath et al., 2011). In training, these weights are initialized with small random values, which are adjusted to obtain the desired task by a learning procedure (Pamučar, Đorović, 2012), (Milićević, Župac, 2012). Many training algorithms are based on the gradient descent (GD) or the back-propagation algorithm (BPA) which is one of the most broadly used learning methods (Silva et al., 2008), (Wu et al., 2011), with input data and the target (predicted output). It uses an objective function *E* (error/cost/loss function) in order to assess the deviation of the predicted output values from the observed data. Problems concerned with MPLs relate to the random weight initialization and the objective function that is non-convex, which can stick training in poor local minimum. The pre-training allows much better initial weights, and resolves the first problem addressing with MLP estimate (Sainath et al., 2011). However, feed-forward neural networks (FNNs) prove to be very successful for solving both these problems, based largely on the use of the BPA and improved learning procedures, which include better optimization, new types of activation functions, and more appropriate ways to process speech. This also stands for acoustic modelling in speech recognition, sub-word and world level modelling (Mikolov et al., 2012), large vocabulary speech recognition, coding and classification of speech (Collobert, 2008), segmentation and word boundaries (Riecke et al., 2009), (Shahin, Pitt, 2012), as well as perception of boundaries in acoustic and speech signals (Mesbahi et al., 2011). According to Bojanić and Delić (2009), FNNs can also model the impact of emotion to the variation of speech characteristics on the level of fundamental frequencies of phonation (pitch), segmentation (changes in articulation quality), and intra-segmental level (general voice quality, whose acoustic correlates are glottal pulse shape and distribution of its spectral energy).

Serbian language belongs to a small group of tonal languages. For these languages, many successful identification techniques based on FNNs and character-level language models are commonly used. Unlike those with a striking accent, in which a syllable may simply be stressed or not, and where minimal pairs of words differ by changes in voice pitch during the pronunciation, in the Serbian language a different accent may indicate a difference in morphological categories (Sečujski, Pekar, 2014). From 1999 to 2010, scientists from Serbia were engaged in the AlfaNum project in order to resolve some problems related to Serbian speech such as phoneme-based continuous speech recognition, text-to-speech synthesis, lack of databases, etc. The results were speech databases and morphological dictionaries of the Serbian language (Delić, 2000), as well

as numerous published papers and books related to the speech technologies (Delić et al., 2010), (Pekar et al., 2010), machine learning (Kupusinac, Sečujski, 2009), speech synthesis (Sečujski et al., 2002), and speech recognition (Pekar et al., 2000). During the same period, Marković et al. (1999) analyzed Serbian vowels (a, e, i, o, u) and spoken digits (0, 1, 2,... 9) and detected the abrupt changes in speech signals. They have presented the results obtained from natural speech with natural and mixed excitation frames, and based on robust recursive and non-recursive approaches and the non-linear Modified Generalised Likelihood Ratio (MGLR) algorithm for the identification of non-stationarity of speech. In 2002, Arsenijević and Milosavljević explored non-linear models for consonant processing. In 2003, they also presented the MGLR algorithm based on FNNs. As it turned out, labial and dental consonants were significant for the articulation of voice and understanding of speech that was essential for synthesized speech, and primarily related to its intelligibility and naturalness. Protić and Milosavljević (2005) have presented the results on the generalization properties for various classes of linear and non-linear models. They have also analyzed the variations of test errors caused by the selection of models and modelling mode conditions (Protić, Milosavljević, 2006). In their research, they used the acoustic model and Gaussian noise to evaluate the impact of noise on speech recognition, recognition of phonemes of one or more speakers, comprehension, and performance evaluation.

This paper presents a comparative analysis of Serbian vowels (a, e, i, o, u) and non-vowels, the voiceless and sound sonant and consonants (labial, dental, anterior and posterior palatal). Men and women pronounced phonemes, in the context of words or isolated ones. The AR model parameters as well as the specific structure of FNNs were determined during the training, which was based on the BPA. The Levenberg-Marquart (LM) method was used to calculate the Hessian matrix and the Optimal Brain Surgeon (OBS) was enforced to prune the network parameters (Jing, 2012). The stopping criteria were reaching minima of normalized sum squared test errors ($NSSE_{TEST}$) and final prediction errors (FPEs). A novel method for multidimensional scaling, based on distance measure was developed for generalization properties testing. The results of the spectral analysis were also presented. Speech signals were represented by their spectro-temporal distribution of accoustic energy, the spectrograms. Finaly, $NSSE_{TEST}$ i FPEs were compared.

The paper is organized as follows. The following chapter deals with models for speech signal prediction. The third chapter describes speech signals. The methodology and the results are presented in Chapter four and the last chapter is the conclusion of the paper. The appendix consists of MATLAB algorithms for processing techniques.

Models

If voice is only one signal for speech prediction, linear Auto-Regressive (AR) two-pole model is usable to minimize the prediction error, and to model a speech production system. If the glottal signal is also available, the AR model with eXtra input (ARX) can be used. In addition, a Moving Average (MA) error correction model may also be taken into account, although it enters some instability in the learning processes and the instability of a model is possible if the error value is high. However, a fully connected FNN gives the best results, because it may prune parameters one by one, up to the partially connected structure, which gives the error minimum (Ljung, 1987). Linear models are very suitable for the purpose of speech signal processing when the structural simplicity of the model is an alternative to the training time or the minimal processing error. Nonlinear models are more complex but also more accurate than linear ones and, consequently, they accurately approximate transfer functions to a higher degree.

Linear AR model

Linear AR two-pole models for approximately (2**n*+1)*500Hz, n = 0, 1, ... poles are sutable for speech system modeling. For the purpose of this research, a 10-pole AR was used, which will be presented later in the paper. The following expression determines the AR model

AR:

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

$y(n)$ is a speech signal sample, a_i (i=1…n_a) are the AR parameters, $e(n)$ is an error that contaminates the speech signal with white (temporary independent) or coloured (temporary dependent) noise (Park, Choi, 2008).

Non-linear model

FNNs with one hidden-layer are mathematically expressed in a form

$$y_i\left(\mathbf{w},\mathbf{W}\right)=F_i\left(\sum_{j=1}^{q} W_{ij} f_j\left(\sum_{l=1}^{m} w_{jl} z_l + w_{j0}\right)+W_{f0}\right) \tag{2}$$

where y_i is output, z_i is input, $\mathbf{w}$ and $\mathbf{W}$ are sinaptic weight matrices, f_j i F_i are the activation functions of the hidden layer and the output layer, respectively. q and m represent the number of elements in the hidden

layer and the input layer, respectively. In many fundamental network models, the activation functions are of a sigmoid or logistic type, but for the networks used here, the activation function is tangent-hyperbolic (tanh).

$$\tanh(i) = \frac{1 - e^{-21}}{1 + e^{-2i}}$$

Speech signals

The speech production system consists of the lungs, the vocal cords, and the vocal tract. The lungs are the source of airflow and pressure, the vocal cords open and close periodically to produce voiced speech thus converting the airflow from the lungs to voice (glottal flow), and the vocal tract consists of a set of cavities above the vocal cords. It is an acoustic filter. At the output of this filter, the sound radiates to the surroundings through the lips and the nostrils. The main characteristic of vowels is the stationarity over the long-term. This feature allows the estimation of models having the minimum of estimation error. The excitation signal is quasi-periodic and of high power because the airflow from the lungs encounters a small diameter of aperture of the vocal cords. For non-vowels, stationarity is shorter, and the model evaluation is difficult. The excitation signal is noise or a mix of noise and it is of less strength because the opening between the vocal cords is high. It is well known that the analysis of the vibrating vocal cords during phonation presents a challenge because the larynx is not easily accessible. However, a non-invasive method such as electroglottography (EGG) is widely used to determine the glottal signal, and the resulting electroglottogram gives useful information for modelling. The excitation of non-vowels is the same as that of vowels, but the vocal cords do not vibrate. Figure 1 presents speech and the EGG signal. Figure 2 presents 4000 samples of the vowel ‘a’, and the corresponding training and testing sets used for the purpose of this research. Figure 3 presents the consonant 's' and the vowel a.

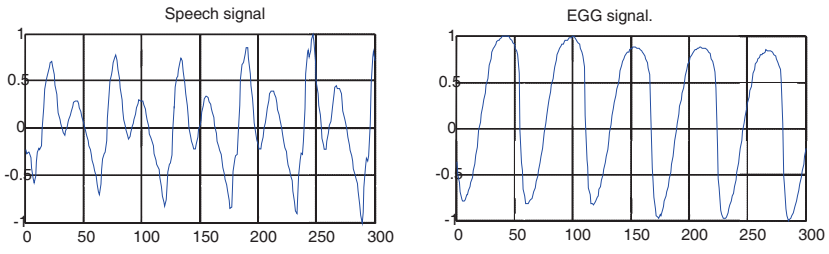
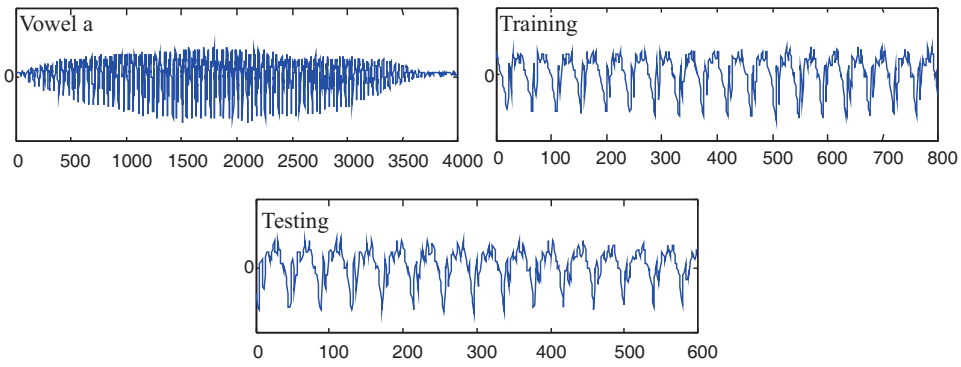


Figure 1 – Speech and the EGG signal
Slika 1 – Govorni i EGG signal



F*igure 2* – Vowel a, training and test set
Slika 2 – Vokal a, trening i test skup

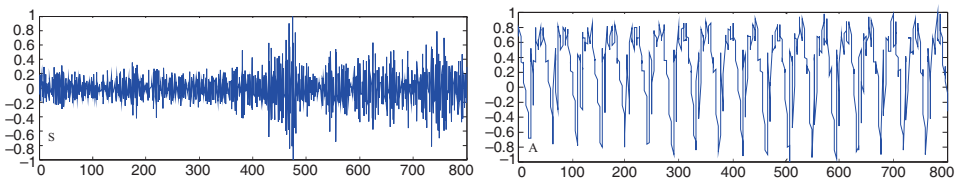


Figure 3 – Consonant s and vowel a
Slika 3 – Konsonant s i vokal a

For the testing purpose, 10 men and women pronounced all the vowels. The recorded analog signals were afterwards sampled with 8kHz and 10kHz frequencies. For the purpose of this paper, the results were given for the signals sampled with frequency f_s=8kHz. The signals consisted of 4,000 samples. Each signal was divided into two equal parts. The training sets (800 samples) were chosen from the first 2,000 samples while the testing sets (600 samples) were parts of the other 2,000 samples. The resulting sets were normalized by the *dscale* function (A.1), to have a zero mean and a variance equal to one (Haykin, 1994). This pre-processing removes offset, variance and correlation of the input data. For testing the non-vowels, analog signals were sampled with f_s=22050Hz. The phonemes were also pronounced in the context of words, or out of it, isolated. The Serbian phonemes were sorted in the following way (1) voiced sonant (j, l, lj, m, n, nj and r), (2) voiced consonants (f, c, s, t, č, š, h, k, b, p), (3) unvoiced sonant (v), and (4) unvoiced consonants (d, đ, dž, z, ž, and g).

Model learning

Training

For the training sets, FNN and AR models were trained. Training was carried out by changing the parameters based on the BPA. The LM approximation of the Hessian matrix was used (Svarer, 1995), (Le Cun et al., 1989). The optimal step size of the error changing was approximated by a Taylor series (Haykin, 1994), (Svarer, 1995). See (3).

$$E = E_0 + \left(\frac{\partial E}{\partial \mathbf{u}}\right)^T \delta\mathbf{u} + \frac{1}{2}\delta\mathbf{u}^T \frac{\partial^2 E}{\partial \mathbf{u}^2}\delta\mathbf{u} + \ldots \tag{3}$$

The Gauss-Newton approximation of an error is given with (4)

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

E is an error fuction approximation, E_0 is its value in the point of approximation, **u** is the parameter vector, w_{jk} and W_{ij} are sinaptic weights, $\delta\mathbf{u}$ is the parameter deviation of **u**, and **H** is the Hessian matirx.

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

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

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

The error minimum and the estimated parameters are given with the formulae

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

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

The number of the calculations for the Hessian matrix inverse is computer demanding (for a matrix of the dimension *nxn,* the numer of calculations is ~ n^3). The LM algorithm accelerates the process of matrix estimation, as it is described by the following expressions

$$u_i^* = u_i - \frac{\frac{\partial E}{\partial u_i}}{\frac{\partial E}{\partial u_i^2}} \qquad u_i^* = u_i - \frac{\left(\frac{\partial E_{train}}{\partial u_i} + \frac{2\alpha_{u_i}}{p} u_i\right)}{\frac{\partial E_{train}}{\partial u_i^2} + \frac{2\alpha_{u_i}}{p} u_i}$$

u_i is the i^{th} parameter estimation. For the square error function

$$\frac{\partial^2 E_{train}}{\partial W_{ij}^2} = \frac{2}{p}\sum_{\mu}\left(\left(V_j^{\mu}\right)^2 - \left(\zeta_i^{\mu} - O_i^{\mu}\right)\frac{\partial V_j^{\mu}}{\partial W_{ij}}\right) = \frac{2}{p}\sum_{\mu}\left(V_j^{\mu}\right)^2 \tag{5}$$

$$\frac{\partial^2 E_{train}}{\partial w_{jk}^2} = \frac{2}{p}\sum_{i\mu}\left(\left(W_{ij}\frac{\partial V_j^{\mu}}{\partial h_j^{\mu}}\xi_k^{\mu}\right)^2 - \left(\zeta_i^{\mu} - O_i^{\mu}\right)W_{ij}\frac{\partial^2 V_j^{\mu}}{\partial\left(h_j^{\mu}\right)^2}\left(\xi_k^{\mu}\right)^2\right) \tag{6}$$

$$\frac{\partial^2 E_{train}}{\partial w_{jk}^2} \approx \frac{2}{p}\sum_{i\mu}\left(W_{ij}\frac{\partial V_j^{\mu}}{\partial h_j^{\mu}}\xi_k^{\mu}\right)^2 \tag{7}$$

This approximation implies the non-corelation of the input ξ_k^{μ}, and the error $(\zeta_i^{\mu}-O_i^{\mu})$ that may include the non-modeled dynamics, which can be reduced by increasing the order of the model and may represent the measurement noise in the output data (Jing, 2012). It also ensures the correct direction of error estimation (Silva et al., 2008). FNN training stops at reaching the minimum of training error (E_{train}) or after 500 parameter changes. FNN training lasts from a few minutes to half an hour, which depends on the complexity of its structure. The initial values of parameters are random. For the FNN training, *nnarx* (A.2) and *marq* (A.3) are used. AR-10 is also trained by *nnarx*. The output y_i is predicted based on its *p* previous values (8)

$$y_i = \sum_{k=1}^{p} a_k y_{i-k} \qquad 1 \le k \le p \tag{8}$$

The formant characteristics of vowels and the distribution of formant frequencies determine the parameters of the model. The AR model is

stable, simple and, considering computer recourses, not very demanding so it can model spectral envelope to make the spectrum of residuals flat if p is sufficiently large. The prediction optimization is based on the minimum squared error (MSE) criterion; its partial derivates by parameters must equate zero.

$$E = \sum_i e_i^2 = \sum_n \left(y_i + \sum_{k=1}^{p} a_k y_{i-k} \right)^2$$

At the frequency domain, the modelled signal spectrum tends to the original signal as p increases. It becomes computer demanding and takes a long time, but the results are more accurate. A criterion for the optimization is the threshold criterion

$$1 - \frac{V_{p-1}}{V_p} < \delta$$

$V_{p\text{-}1}$ and V_p are the normalized prediction errors for p-1 and p, and δ is the threshold. Tipically, the number of coefficients is 8, 10, or 12. These models shift lower formants by adding biases to formants, which have high-energy value. It creates problems in analyses of male speech, considering that the basic frequencies of those signals are much lower than the basic frequencies of signals spoken by women or children.

Pruning

The parameters of trained FNNs were OBS pruned (Haykin, 1994), (Norgaard, 2001). The full Hessian matrix is calculated iteratively (Svarer, 1995). The error change is given by the formula

$$\delta E \approx \frac{1}{2} \delta \mathbf{u}^T \mathbf{H} \delta \mathbf{u}$$

$\delta \mathbf{u}$ is a parameter change. The pruning of the parameter u_m to zero requires that

$$\delta u_m + u_m = 0$$

which corresponds to

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

$\mathbf{e}_m$ is the unit vector, and is of the same dimension as $\delta\mathbf{u}$. The goal of this methodology is to prune the parameter u_m, which would cause the minimum increase in the error E. This gives LaGrange's equality

$$L_a = \frac{1}{2}\delta\mathbf{u}^T\mathbf{H}\delta\mathbf{u} + \lambda\left(\mathbf{e}_m^T\delta\mathbf{u} + u_m\right)$$

λ is a LaGrange multiplier. If

$$\frac{\partial L_a}{\partial(\delta\mathbf{u})} = \delta\mathbf{u}^T\mathbf{H} + \lambda\mathbf{e}_m^T = 0$$

the Hessian matrix is a positive definite, and it is possible to find its inverse as follows

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

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

following

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

The main criterion for stopping the pruning algorithm is to achieve the error minimum. The method that determines the balance between too many and too few parameters takes into account the number of parameters, the size of the training set, the Hessian matrix size, the correlation of input data, etc. It is based on the available data, which enables adjusting the model parameters to the optimum. There are various algorithms for model optimization. One presented here stops pruning when the generalization error (E_{gen}), the smallest error determined based on the independent set of data having the same distribution as the training set, reaches the minimum. The method requires large training and testing sets, but it is widely used, and gives good results. Akaike (1969) developed a method for the approximation of Taylor's series expansion of learning (E_{learn}) and generalization (E_{gen}) errors (Ljung, 1987), (Larsen, 1993), (Hansen, Rasmusen, 1994), (Kashyap, 1980). The errors are shown in Figure 4 and given by expressions (9) - (10).

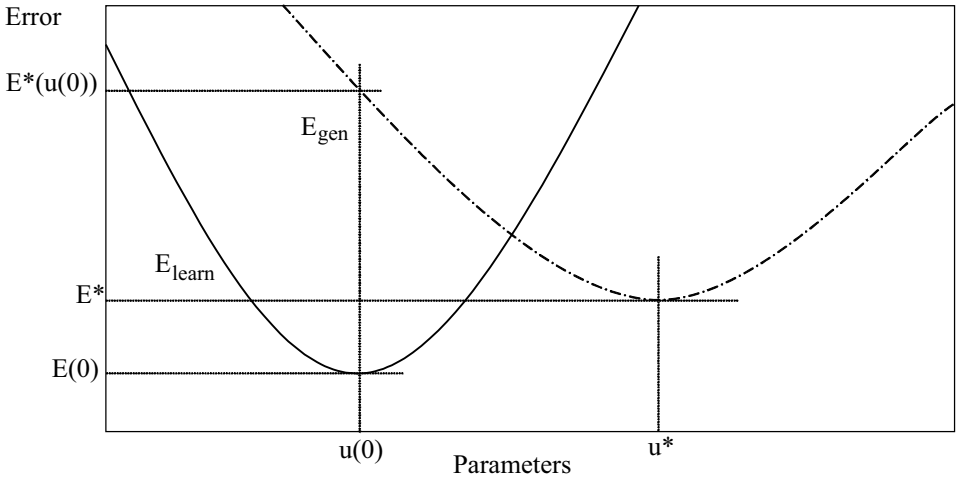


Figure 4 – Approximation of E_{learn} *and* E_{gend}
Slika 4 – Aproksimacija E_{learn} i E_{gend}

$$E_{learn} = E_0 + \left(\frac{\partial E}{\partial \mathbf{u}}\right)^T \delta\mathbf{u} + \frac{1}{2}\delta\mathbf{u}^T \mathbf{H}\delta\mathbf{u} + o\left(\left\|\delta\mathbf{u}\right\|^3\right) \tag{9}$$

$$E_{gen} = E^* + \left(\frac{\partial E^*}{\partial \mathbf{u}}\right)^T \delta\mathbf{u}^* + \frac{1}{2}\delta\mathbf{u}^{*T} \mathbf{H}^*\delta\mathbf{u}^* + o\left(\left\|\delta\mathbf{u}^*\right\|^3\right) \tag{10}$$

$\delta\mathbf{u}^*$ is the vector of the parameter changes on a minimum of E_{gen}, $\partial E/\partial\mathbf{u}$ and $\partial E^*/\partial\mathbf{u}$ are the first derivates of the given functions, respectively. $\mathbf{H}$ and $\mathbf{H}^*$ are corresponding Hessian matrices, $o(||...||)^3$ is a part of the Taylor series which equals zero. The first derivates of $\mathbf{u}_0$ i $\mathbf{u}^*$, are also equal to zero. E_{learn} is

$$E_{learn} = E_0 + \frac{1}{2}\delta\mathbf{u}^T \mathbf{H}\delta\mathbf{u}$$

$$E_{gen} = E^* + \frac{1}{2}\delta\mathbf{u}^{*T} \mathbf{H}^*\delta\mathbf{u}^*$$

E_{test} is equal to $E_{gen}(\mathbf{u}_0)$. The problems that arise here are unknown values of E^*, $\mathbf{u}^*$ and $\mathbf{H}^*$. The following assumptions imply that the difference

between the values of $\mathbf{u}_0$ and $\mathbf{u}^*$ are small, and the FNN is well-trained. It is also assumed that the second derivate of E_{gen} can be equated with the second derivate of the training error, so E_{gen} is given with the formula

$$E_{gen} \approx E^* + \frac{1}{2}\delta\mathbf{u}^{*T}\mathbf{H}^*\delta\mathbf{u}^*$$

Akaike's estimate of FPE provides the way to estimate E_{gen} from the given FNN structure, if the numer of parameters is known (Akaike, 1969). If the unknown value of noise variance can be removed from E^* then

$$\hat{E}_{gen}(\mathbf{u}_0) \approx \frac{\left(1+\frac{N_M}{p}\right)}{\left(1-\frac{N_M}{p}\right)}\hat{E}_{learn}(\mathbf{u}_0)$$

N_M is the dimenson of the parameter vector, and p is the size of the training set. E_{gen} can be computed when it is necessary to compare different structures of neural networks, if the training sets are the same (Haykin, 1994), (Svarer, 1995), (Akaike, 1969). It follows that

$$K_{FPE}E_{learn} = \frac{\left(1+\frac{N_M}{p}\right)}{\left(1-\frac{N_M}{p}\right)}E_{learn}, \qquad N_M < p$$

K_{FPE} is the FPE coefficient. As the number of parameters increase E_{learn} decreases to zero. In addition, K_{FPE} increases from one to ∞ when the ratio N_M/p changes from zero to one. Also, the parameter change always leads to the point of the E_{gen}, because it exists. Furthermore, the minimum value of E_{FPE} exists within the limits determined by the number of parameters, and pruning stops when E_{FPE} reaches its minimum. It should be noted that Akaike's criterion showed some inconsistency related to the determination of AR model orders when there is a Gaussian noise and if $N_M << p$. The lower limit of this relation is 0.156 (Kashyap, 1980).

To compare AR and NNAR models for non-vowels, the FPE gain is introduced (10)

$$G_{FPE} = 10\log\frac{E}{FPE} \tag{11}$$

E is the normalized sum of errors (NSE) and *N* is the size of the training set

$$E = \frac{1}{N}\sum_{i=1}^{N} y(i)^2$$

$$FPE = \frac{N+d}{N-d}\left(\frac{1}{N}\sum_{i=1}^{N}\left(y(i)-\hat{y}(i)\right)^2\right) \quad (12)$$

y(i) is *i*th speech sample, $\hat{y}(i)$ is its estimated value and *d* is the number of model parameters.

Validation

Validation is performed for all the vowels and all speakers on the independent training and test sets. The results presented here are given for the structures that are selected to give the minimum error values for one of the criteria

1) minimum $NSSE_{TEST}$ and
2) minimum FPE.

The results are presented in the following chapter.

Results

Vowels

For the training sets of vowels, the linear 10-pole AR model (AR-10) as well as FNNs with 10 inputs, 1 output, and 3, 5, 7, 9 or 11 neurons in the hidden layer are trained. The obtained structures are OBS pruned, to a maximum of 20 iterations retraining at each rejection of parameters. MATLAB function *nnprune* (A.4) is used. The results are NSSE for the training set ($NSSE_{TRAIN}$), the test set ($NSSE_{TEST}$) and the FPE for each parameter. Pruning of the structures presented in this paper lasted from a half an hour up to four hours. Validation is carried out by the MATLAB function *nnvalid* (A.5). Along with the errors of the FNNs, the NSSE for AR-10 ($NSSE_{AR}$) is also calculated. The process of training and testing the AR-10 model lasted to a maximum of 10s. Table 1 shows the minimum error values FNN and AR-10 for all the vowels and all the speakers.

Table 1 – Minimum values of $NSSE_{TEST}$, FPE and $NSSE_{AR}$
Tabela 1 – Minimalne vrednosti $NSSE_{TEST}$, FPE i $NSSE_{AR}$

		$NSSE_{TEST}$	NN	FPE	NN	$NSSE_{AR}$
VOWEL a	a2	0,0038	10-7-1	0,0013	10-11-1	0,0605
	a3	0,0089	10-13-1	0,0030	10-13-1	0,0152
	a6	0,0479	10-11-1	0,0025	10-13-1	0,1057
	a7	0,0067	10-5-1	0,0010	10-13-1	0,0109
	a8	0,0383	10-9-1	0,0012	10-13-1	0,0817
VOWEL e	e2	0,0099	10-13-1	0,0023	10-13-1	0,0509
	e3	0,0067	10-7-1	5,12e-04	10-13-1	0,0147
	e6	0,0083	10-7-1	0,0011	10-13-1	0,0416
	e7	0,0065	10-9-1	0,0016	10-13-1	0,0147
	e8	0,0026	10-13-1	0,0040	10-13-1	0,0840
VOWEL i	i2	0,0045	10-9-1	4,17e-04	10-13-1	0,0129
	i3	0,0044	10-3-1	4,34e-04	10-13-1	0,0132
	i6	0,0021	10-13-1	2,44e-04	10-13-1	0,0135
	i7	0,0022	10-9-1	6,99e-04	10-13-1	0,0024
	i8	0,0047	10-5-1	7,10e-04	10-13-1	0,0162
VOWEL o	o2	0,0015	10-13-1	1,85e-04	10-13-1	0,0061
	o3	5,10e-04	10-13-1	9,40e-05	10-11-1	0,0015
	o6	6,51e-04	10-13-1	1,14e-04	10-13-1	0,0054
	o7	9,33e-04	10-13-1	1,41e-05	10-11-1	0,0010
	o8	0,0027	10-5-1	3,74e-04	10-7-1	0,0095
VOWEL u	u2	9,88e-05	10-11-1	1,98e-04	10-13-1	1,34e-04
	u3	2,10e-04	10-9-1	5,07e-05	10-11-1	5,13e-04
	u6	3,91e-04	10-7-1	1,15e-04	10-13-1	5,51e-04
	u7	3,41e-04	10-7-1	1,13e-04	10-11-1	4,06e-04
	u8	1,96e-04	10-7-1	8,34e-05	10-13-1	1,80e-04

Non-vowels (voiceless sonant, voiceless consonant, sonar sonant, and sonar consonant)

Table 2 shows the results of the analysis of non-vowels pronounced by men and women. The FPE gains for AR (G_{FPE} AR) and NNAR (G_{FPE} NNAR) are determined.

Table 2 – G_{FPE} for isolated phonemes
Tabela 2 – G_{FPE} za izolovane foneme

	G_{FPE} [dB] AR		G_{FPE} [dB] NNAR		G_{FPE}NNAR – G_{FPE}AR [dB]	
	Women	Men	Women	Men	Women	Men
B	34,6519	30,167	38,5847	34,3293	3,9328	4,1623
C	2,2075	2,9722	5,6637	6,559	3,4562	3,5868
Ć	11,5334	7,0614	15,5047	10,9525	3,9713	3,8911
Č	8,8546	10,4918	13,1923	14,0135	4,3377	3,5217
D	31,3806	25,9894	37,1691	30,9179	5,7885	4,9285
Đ	29,4399	14,9543	34,0653	19,4483	4,6254	4,494
DŽ	13,6488	12,789	17,5563	16,8545	3,9075	4,0655
F	3,5165	11,365	6,7754	15,1115	3,2589	3,7465
G	28,1664	19,1006	35,7757	24,087	7,6093	4,9864
H	11,064	9,7711	14,2458	12,8508	3,1818	3,0797
J	26,3422	21,7075	30,025	24,9229	3,6828	3,2154
K	19,1537	10,6407	22,8868	14,1928	3,7331	3,5521
L	29,5687	31,0459	32,8997	35,0852	3,331	4,0393
LJ	31,5999	28,8233	35,2027	32,4762	3,6028	3,6529
M	33,4249	36,0233	37,3764	39,9362	3,9515	3,9129
N	33,3766	36,6123	37,4727	40,2027	4,0961	3,5904
NJ	33,6963	34,1175	38,1354	38,2625	4,4391	4,145
P	11,548	34,1175	14,9509	38,4441	3,4029	4,3266
R	24,8785	28,8893	28,4164	33,3224	3,5379	4,4331
S	2,8275	4,9219	6,081	8,1239	3,2535	3,202
Š	10,9265	11,2751	14,222	15,1396	3,2955	3,8645
T	8,351	7,3746	12,3908	11,1201	4,0398	3,7455
V	23,8889	25,068	27,2998	28,3466	3,4109	3,2786
Z	10,1461	14,3512	13,7734	18,4116	3,6273	4,0604
Ž	12,441	12,4569	15,9663	15,808	3,5253	3,3511

The average G_{FPE} for the NNAR model is approximately 4dB higher than G_{FPE} for the AR model, indicating better properties of NNAR as compared to the same order AR model. Figure 5 shows G_{FPE} for the phonemes that were pronounced out of the context of words (isolated).

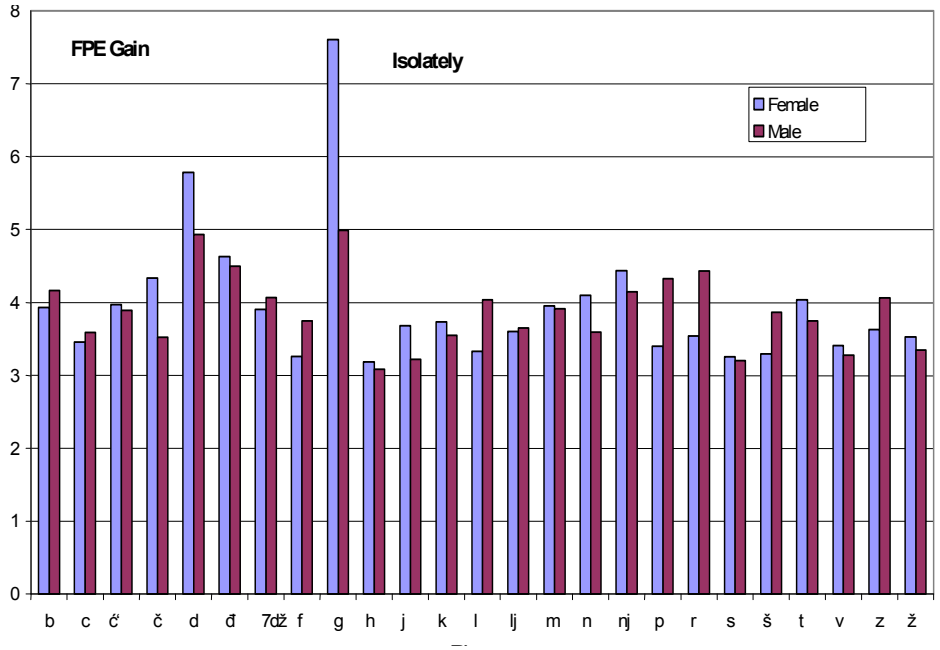


Figure 5 – FPE Gain for isolately pronounced phonemes
Slika 5 – FPE pojačanje za izolovano izgovorene foneme

From Table 2, the following grouping of phonemes can be noticed: sonar sonant (j, l, lj, m, n, nj, and r), which always have a high value of G_{FPE}, and sonar consonant (f, c, s, t, č, š, h, and k), which do not have it, excluding "b" and "p". The voiceless sonant "v" has a high value of G_{FPE}, while voiceless consonants (d, đ, dž, z, ž, and g) mostly have an average G_{FPE}, depending on the model or the gender of a speaker.

Table 3 and Figure 6 show the results of the analysis for non-vowels that were spoken in the context of words.

Table 3 – G_{FPE} for phonemes pronounced in the context of words
Tabela 3 – G_{FPE} za foneme izgovoren u kontekstu reči

	G_{FPE} [dB] AR		G_{FPE} [dB] NNAR		GNNAR - GAR	
Content	Women	Men	Women	Men	Women	Men
B	32,6346	33,0303	36,0158	36,1753	3,3812	3,145
C	2,6291	1,7716	5,7897	5,2355	3,1606	3,4639
Ć	12,2942	9,0452	16,2301	12,6387	3,9359	3,5935
Č	9,442	10,5968	13,6114	14,7664	4,1694	4,1696
D	33,9978	30,7685	37,8104	35,2031	3,8126	4,4346
Đ	14,3192	13,5345	17,7639	16,9531	3,4447	3,4186

	G_{FPE} [dB] AR		G_{FPE} [dB] NNAR		GNNAR - GAR	
Content	Women	Men	Women	Men	Women	Men
DŽ	14,9507	0	17,5563	0	2,6056	0
F	0	0	0	0	0	0
G	33,1103	27,2414	36,7543	30,8057	3,644	3,5643
H	10,365	7,9697	13,5085	11,018	3,1435	3,0483
J	21,1887	27,3908	24,6641	30,9087	3,4754	3,5179
K	11,6569	14,5669	15,9135	18,2566	4,2566	3,6897
L	25,7282	30,6369	30,3882	34,2097	4,66	3,5728
LJ	21,0459	33,7669	26,0215	37,3985	4,9756	3,6316
M	33,4249	31,5708	37,0044	34,8602	3,5795	3,2894
N	26,9488	33,0333	30,6018	36,9415	3,653	3,9082
NJ	27,4645	30,1661	32,5246	35,1878	5,0601	5,0217
P	6,1535	11,4462	11,3976	14,9627	5,2441	3,5165
R	26,3231	25,6694	30,0101	30,2615	3,687	4,5921
S	1,9138	3,5753	5,0039	7,017	3,0901	3,4417
Š	13,9972	11,201	17,5547	14,9197	3,5575	3,7187
T	11,3133	4,3862	15,359	7,5557	4,0457	3,1695
V	29,2793	33,852	32,6862	37,532	3,4069	3,68
Z	13,8724	23,0607	17,2613	27,119	3,3889	4,0583
Ž	0	0	0	0	0	0

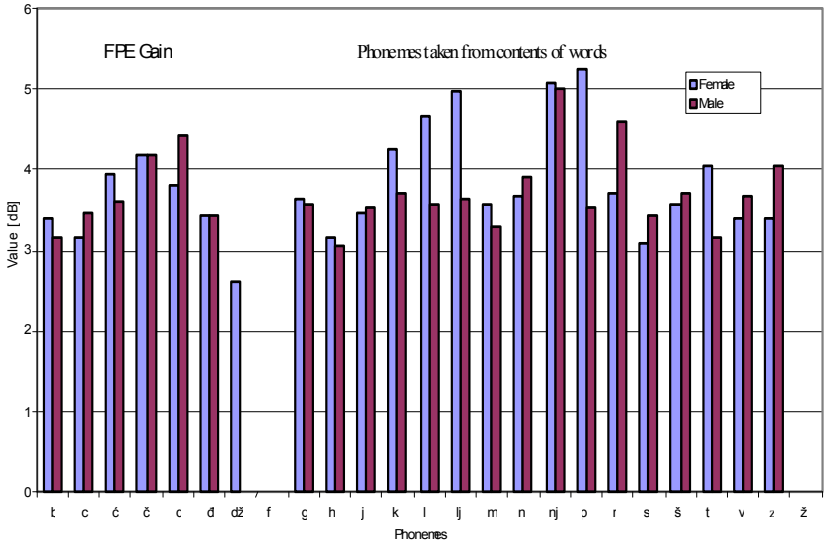


Figure 6 – FPE gain for non-vowels pronounced as parts of words
Slika 6 – FPE pojačanje za nevokale izgovorene u delovima reči

Generalization properties

To test the generalization of the given models, the parameters of which were estimated based on the training set of one speaker, the testing was carried out on the sets of other speakers. The minima of $NSSE_{TEST}$ and FPE, as well as the matrices of the mean($NSSE_{TEST}$) were calculated for all vowels. Figure 7 shows the mean($NSSE_{TEST}$) for the vowel 'a' and the FNN structure 10-3-1. The arrows mark points of error jumps, which are evident for 5-8, 13-18, and 21-25 parameters remained in FNN after pruning.

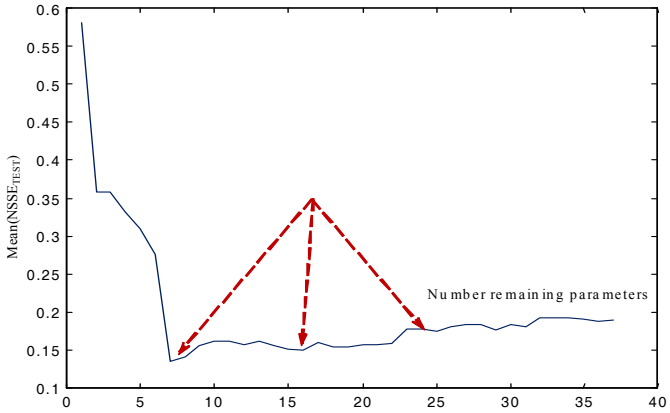


Figure 7 – Mean($NSSE_{TEST}$)
Slika 7 – Srednja vrednost $NSSE_{TEST}$

To measure the variability, a new distance measure, based on FNN is defined. Two models based on the speech signals are given with formulae (13), (14)

$$y_{i,t} = g\left(y_{i,t-1}, \delta_{t-1}\left(y_i\right)\right) + \varepsilon_{i,t}, \quad i = 1,2 \tag{13}$$

$$y_{i,t} = g_{W_i}\left(y_{i,t-1}, \delta_{t-1}\left(y_i\right)\right) + \varepsilon_{i,t} \quad i = 1,2 \tag{14}$$

For the parameters that are estimated based on the corresponding training sets the variable

$$C\left(W_{Y_i}, Y\right) = \frac{1}{2N}\sum_{i=1}^{N}\left(y_i - g_{W_{Y_i}}\left(y_{t-1}, \delta_{t-1}(y)\right)\right)^2 \tag{15}$$

presents the MSE of the residuals of the signal Y for g_{WYi} (a transfer function of FNN trained with Y_i), giving the expression for the distance between two signals Y_1 and Y_2

$$D(Y_1, Y_2) = \frac{1}{2}\left[\log\frac{C(W_{Y_1}, Y_2)}{C(W_{Y_1}, Y_1)} + \log\frac{C(W_{Y_2}, Y_1)}{C(W_{Y_2}, Y_2)}\right] \quad (16)$$

Originally developed for regression problems, the MSE function is obtained by the maximum likelihood principle assuming the independence and Gaussianity of the target data (Bishop, 1995) (Silva et al., 2008). However, although most classical approaches in speech processing are based on linear techniques, which rely on the source-filter model, these linear approaches cannot capture the complex dynamic of speech. It has been shown that the Gaussian linear prediction analysis cannot be used to extract all dynamical structures of real time speech series (Khanagha et al., 2012), (Little et al., 2006). However, in this particular case, the difference between a speech signal sample and its predicted value is temporary independent, so *e*(*n*) in the model (1) is assumed to be a zero mean white Gaussian process (Marković et al., 1999) and the prediction error ε is given with the formula

$$\varepsilon = \hat{e}_k = y_k + \sum_{i=1}^{p} \hat{a}_i \, y_{k-i} \quad (17)$$

In their work, Stanimirović and Ćirović (2008) describe an adaptive algorithm for the adaptive classification of speech and pause, and describe noise and the residuals, which are Gaussian, in this case this states for both Y_1 and Y_2 (Park, Choi, 2008). Table 4 shows the distances $D_{NSSETEST}$, D_{FPE} and $D_{AR\text{-}10}$.

Table 4 – $D_{NSSETEST}$, D_{FPE} i D_{AR-10}
Tabela 4 – $D_{NSSETEST}$, D_{FPE} i D_{AR-10}

	$D_{NSSETEST}$					D_{FPE}					D_{AR-10}				
	a2	a3	a6	a7	a8	a2	a3	a6	a7	a8	a2	a3	a6	a7	a8
a2	0	2,58	1,77	1,43	1,24	0	2,29	1,71	1,54	1,01	0	1,26	0,89	0,64	0,58
a3	2,58	0	1,67	2,32	1,85	2,29	0	1,60	2,76	1,75	1,26	0	0,74	1,40	1,37
a6	1,77	1,67	0	2,66	1,51	1,71	1,60	0	2,56	1,44	0,89	0,74	0	1,56	1,17
a7	1,43	2,32	2,66	0	2,19	1,54	2,76	2,56	0	2,11	0,64	1,40	1,56	0	1,28
a8	1,24	1,85	1,51	2,19	0	1,01	1,75	1,44	2,11	0	0,58	1,37	1,17	1,28	0
	e2	e3	e6	e7	e8	e2	e3	e6	e7	e8	e2	e3	e6	e7	e8
e2	0	3,92	2,82	1,93	1,17	0	3,90	1,95	0,47	0,86	0	2,42	2,02	0,59	1,15
e3	3,92	0	1,46	2,24	3,67	3,90	0	2,46	2,60	3,77	2,42	0	0,36	1,29	1,93
e6	2,82	1,46	0	1,96	0,05	1,95	2,46	0	2,18	1,80	2,02	0,36	0	0,97	1,41
e7	1,93	2,24	1,96	0	0,32	0,47	2,60	2,18	0	1,25	0,59	1,29	0,97	0	1,17
e8	1,17	3,67	0,05	0,32	0	0,86	3,77	1,80	1,25	0	1,15	1,93	1,41	1,17	0

	i2	i3	i6	i7	i8	i2	i3	i6	i7	i8	i2	i3	i6	i7	i8
i2	0	1,16	2,20	0,88	1,37	0	1,16	2,20	0,88	1,37	0	0,00	2,63	0,91	1,20
i3	1,16	0	3,04	1,36	2,29	1,16	0	3,04	1,36	2,29	0,00	0	2,59	0,86	1,21
i6	2,20	3,04	0	2,17	2,30	2,20	3,04	0	2,17	2,30	2,63	2,59	0	1,09	1,76
i7	0,88	1,36	2,17	0	2,00	0,88	1,36	2,17	0	2,00	0,91	0,86	1,09	0	1,27
i8	1,37	2,29	2,30	2,00	0	1,37	2,29	2,30	2,00	0	1,20	1,21	1,76	1,27	0
	o2	o3	o6	o7	o8	o2	o3	o6	o7	o8	o2	o3	o6	o7	o8
o2	0	3,44	3,08	1,16	1,74	0	2,83	2,29	2,98	2,00	0	1,10	1,16	0,56	0,54
o3	3,44	0	3,45	1,70	4,00	2,83	0	3,56	4,27	3,42	1,10	0	0,68	1,09	1,48
o6	3,08	3,45	0	2,05	3,53	2,29	3,56	0	4,81	3,33	1,16	0,68	0	1,57	1,71
o7	1,16	1,70	2,05	0	1,68	2,98	4,27	4,81	0	4,33	0,56	1,09	1,57	0	0,98
o8	1,74	4,00	3,53	1,68	0	2,00	3,42	3,33	4,33	0	0,54	1,48	1,71	0,98	0
	u2	u3	u6	u7	u8	u2	u3	u6	u7	u8	u2	u3	u6	u7	u8
u2	0	5,10	1,16	0,23	1,00	0	1,78	1,33	1,18	2,33	0	0,68	0,99	0,18	0,79
u3	5,10	0	4,18	4,13	1,70	1,78	0	0,56	1,22	2,48	0,68	0	0,45	0,62	1,28
u6	1,16	4,18	0	0,63	1,10	1,33	0,56	0	0,25	1,21	0,99	0,45	0	0,34	0,80
u7	0,23	4,13	0,63	0	0,03	1,18	1,22	0,25	0	1,13	0,18	0,62	0,34	0	0,34
u8	1,00	1,70	1,10	0,03	0	2,33	2,48	1,21	1,13	0	0,79	1,28	0,80	0,34	0

For the purpose of this work, the signals within the FNN structure were also analysed. The training and pruning of FNN (10-3-1 structure), were based on the joined training set formed in the following way: the signal sets of vowels a3, a4, a6, a7, and a8 were 'glued' to the following one. The testing was done with the corresponding test set. The total length of the joined training set was 4,000 samples and the total length of the joined test set was 3,000 samples. The validation was performed over the independent vowel, in this paricular case it was the vowel 'a' that was pronounced by the second speaker (a2). Error jumps occured after 22, 13, and 5 parameters remained after pruning. The $NSSE_{TRAIN}$, $NSSE_{TEST}$, and FPE are shown in Figure 8. The graph also shows the $NSSE_{AR}$ for AR-5, AR-10, and AR-15 models.

For each structure, the spectra of signals at the outputs of neurons in the hidden layer were also analysed. Figure 9 shows the spectra of the validation signal and the signals at the outputs of neurons in the hidden layer for 5, 13, and 22 parameters remained after pruning the 10-3-1 FNN. The spectra were calculated by Burg's method. It is evident that the spectra of the outputs of the hidden-layer group around the formant frequency of the validation signal. The signal range from one neuron shows strong grouping around one and less around other formant frequencies. It should be noted that 2nd neuron was rejected when 5 parameters remained. The FNNs whose total numer of parameters in the pruning exceeds 25 show overfitting in the assessment of the validation signal while those FNNs with 5 parameters remaining make good assessment, which indicated the existence of a non-linear structure with a minimal number of parameters.

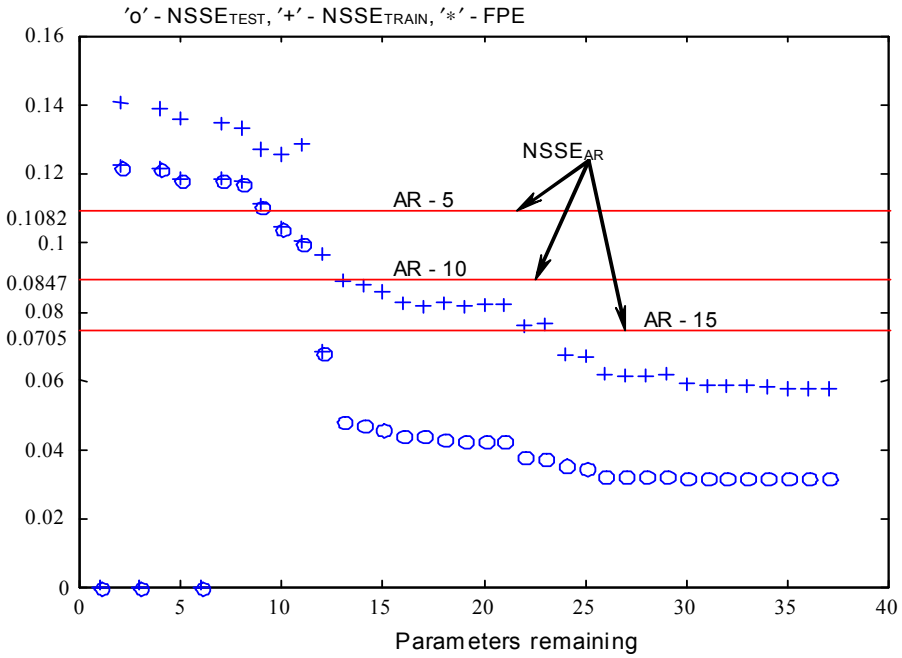


Figure 8 – $NSSE_{TEST}$, $NSSE_{TRAIN}$, FPE, $NSSE_{AR}$
Slika 8 – FPE pojačanje za nevokale izgovorene u delovima reči

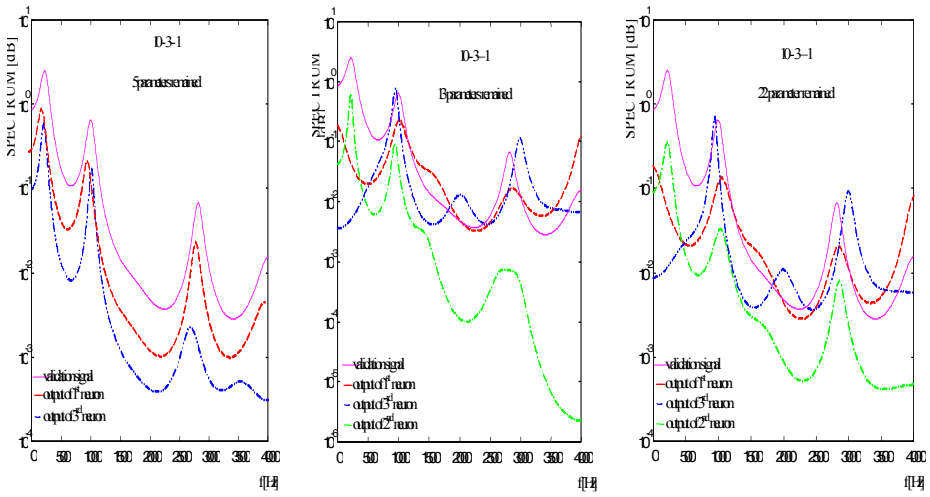


Figure 9 – Signal spectra
Slika 9 – Spektar signala

In addition to the above, the cross-correlations up to the shift 30 of the given signals were also determined. The cross-correlation of two signals x_1 and x_2 is given by the following expression

$$xcorr_{x,y}(k) = \sum_{n=-\infty}^{\infty} x(n)y(n-k) \tag{15}$$

The results of the cross-correlation analysis are shown in Figure 10.

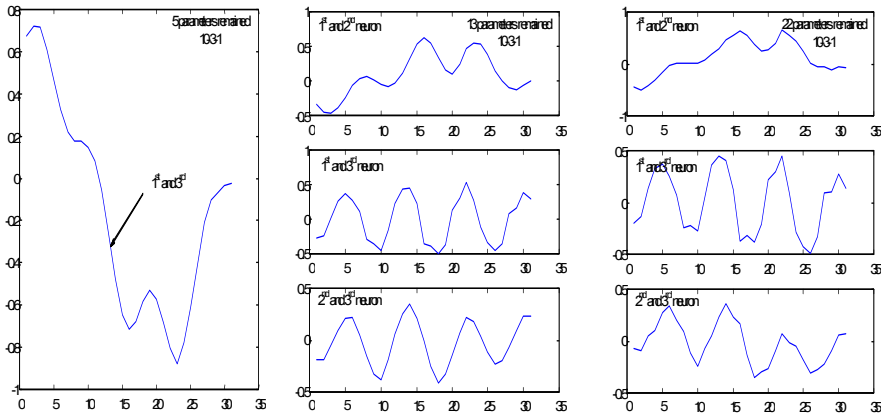


Figure 10 – Crosscorrelations
Slika 10 – Kroskorelacije

Moreover, as shown in Figure 11, the cumulative sum of the absolute values of the cross-correlations was given for the clarity of results.

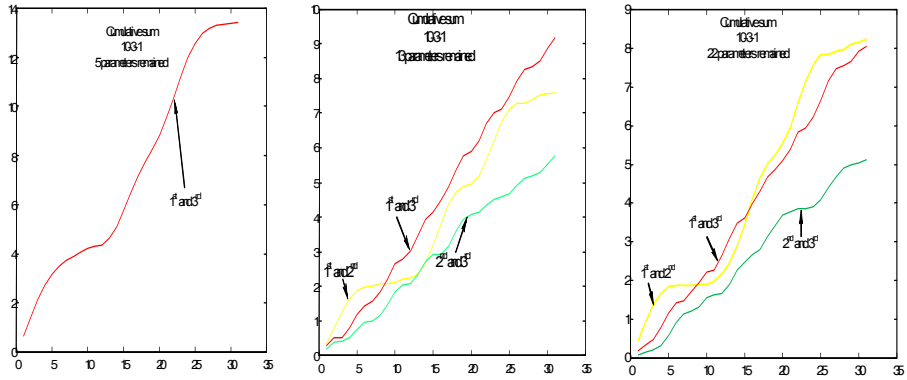


Figure 11 – Cumulative sum of the absolute values of cross-correlation signals
Slika 11 – Kumulativne sume apsolutnih vrednosti kroskorelacionih signala

The average cross-correlation is shown in Table 5. The ratio of the first and the fifth element of cumulative sums (of the absolute values of cross-corelations) whith the number of signal shifts (1 and 5), give average values for a given magnitude. These values are in the range from 0.117- 0.387 for 13 and 22 parameters remained, which represents weak stochastic dependence. The value for five parameters is 0.636, which indicates a medium stochastic dependence.

Table 5 – Cumulative sum of cross-correlations at the outputs of hidden-layer neurons
Tabela 5 – Kumulativne sume kroskorelacija na izlazima neurona skrivenog sloja

	1st and 2nd neuron		1st and 3rd neuron		2nd and 3rd neuron	
Parameters remaining	1st element	av(xcorr) n_{max}=5	1st element	av(xcorr) n_{max}=5	1st element	av(xcorr) n_{max}=5
5	-	-	0,6714	0,63618	-	-
13	0,3408	0,37804	0,2690	0,23198	0,1836	0,14376
22	0,4391	0,36744	0,1975	0,23222	0,0696	0,11702

Conclusion

This paper presents a comparative analysis of Serbian phonemes (vowels and non-vowels). The FNN and AR-10 models are trained and tested. The characteristics of vowels are long-term quasi-periodicity and power spectrum with clearly visible formants. Non-vowels are characterised by short quasi-periodicity and a low power excitation signal. The methodology of generalization enabled a choice of network architectures with improved properties, based on pruning and significant reduction of model parameters. Limited architectures are characterized by a minimal number of parameters within the given margins of errors. In order to review the discriminatory properties of the selected models, a new method for multi-dimensional scaling based on the measurement of distance is developed. The analysis of discrimination loss suggests that the FNNs have a much higher discrimination power, which makes them usable in a wide class of speech recognition usage. The spectral analysis shows a good corellation of the signals at the outputs of hidden-layer neurons and the input signal. The time-domain analysis indicates a week statistical dependence of these signals for the low ranks of cross-correlation (up to the fifth order). The analyses indicate a slight advantage of $NSSE_{TEST}$ compared to FPE criteria. If training sets are short, the FPE is an acceptable criterion. The results indicate that the proposed FNN model, as well as a choice of architecture with the best generalization properties, provides high accuracy and an internally distributed structure that correspond to the natural time-frequency contents of input signals, as well as high discrimination properties for the same number of parameters, as compared to the traditional linear model.

Appendix

A.1 DSCALE

[X,Xscale]=dscale(X) scales data to zero mean and variance 1.

INPUTS:

X: Data matrix (dimension is # of data vectors in matrix * # of data points)

OUTPUTS:

X: Scaled data matrix

Xscale: Matrix containing sample mean (column 1) and standard deviation (column 2) for each data vector in X.

A.2 NNARX

Determine a nonlinear ARX model of a dynamic system by training a two-layer neural network with the Marquardt method. The function can handle multi-input systems (MISO).

[W1,W2,critvec,iteration,lambda]=nnarx(NetDef,NN,W1,W2,trparms,Y,U)

INPUTS:

U: Input signal (= control signal) (left out in the nnarma case)

dim(U) = [(inputs) * (# of data)]

Y: Output signal. dim(Y) = [1 * # of data]

NN: NN=[na nb nk].

na = # of past outputs used for determining prediction

nb = # of past inputs used for determining prediction

nk = time delay (usually 1)

For multi-input systems nb and nk contain as many columns as there are inputs.

W1,W2: Input-to-hidden-layer and hidden-to-output layer weights. If they are passed as [] they are initialized automatically

trparms : Contains parameters associated with the training (see MARQ), if trparms=[] it is reset to trparms = [500 0 1 0]. For time series (NNAR models), NN=na only.

See the function MARQ for an explanation of the remaining input arguments as well as of the returned variables.

A.3 MARQ

Train a two layer neural network with the Levenberg-Marquardt method. If desired, it is possible to use regularization by weight decay. Also pruned (ie. not fully connected) networks can be trained. Given a set of corresponding input-output pairs and an initial network

[W1,W2,critvec,iteration,lambda]=marq(NetDef,W1,W2,PHI,Y,trparms)

trains the network with the Levenberg-Marquardt method. The activation functions can be either linear or tanh. The network architecture is defined by the matrix 'NetDef' which has two rows. The first row specifies the hidden-layer and the second row specifies the output layer.

E.g.: NetDef = ['LHHHH'; 'LL---'] (L = Linear, H = tanh)

Notice that the bias is included as the last column in the weight matrices.

INPUT:
NetDef: Network definition
W1: Input-to-hidden-layer weights. The matrix dimension is dim(W1) = [(# of hidden units) * (inputs + 1)] (the 1 is due to the bias)
W2: hidden-to-output layer weights, dim(W2) = [(outputs) * (# of hidden units + 1)]
PHI: Input vector. dim(PHI) = [(inputs) * (# of data)]
Y : Output data. dim(Y) = [(outputs) * (# of data)]
trparms : Vector containing parameters associated with the training
trparms = [max_iter stop_crit lambda D]
max_iter : max # of iterations.
stop_crit : Stop training if criterion is below this value
lambda: Initial Levenberg-Marquardt parameter
D: Row vector containing the weight decay parameters. If D has one element, a scalar weight decay will be used. If D has two elements, the first element will be used as weight decay for the hidden-to-output layer while the second one will be used for the input-to hidden-layer weights. For individual weight decays, D must contain as many elements as there are weights in the network.
Default values are (obtained if left out): trparms = [500 0 1 0]

OUTPUT:
W1, W2 : Weight matrices after training
critvec: Vector containing the criterion evaluated at each iteration
iteration: # of iterations
lambda: The final value of lambda. Relevant only if retraining is desired

A.4 NNPRUNE

This function applies the Optimal Brain Surgeon (OBS) strategy for pruning neural network models of dynamic systems. That is networks trained by NNARX, NNOE, NNARMAX1, NNARMAX2, or their recursive counterparts.

```
[theta_data,NSSEvec,FPEvec,NSSEtestvec,deff,pvec]=...
nnprune(method,NetDef,W1,W2,U,Y,NN,trparms,prparms,U2,Y2,skip,Chat)
```

INPUT:
method: The function applied for generating the model. For example method='nnarx' or method='nnoe' NetDef, W1, W2, U, Y,trparms: See for example the function MARQ
U2,Y2: Test data. This can be used for pointing out the optimal network architecture is achieved. Pass two []'s if a test set is not available.
skip (optional): See for example NNOE or NNARMAX1/2. If passed as [] it is set to 0.
Chat (optional): See NNARMAX1
prparms: Parameters associated with the pruning session
prparms = [iter RePercent]
iter: Max. number of retraining iterations
RePercent : Prune 'RePercent' percent of the remaining weights (0 = prune one at a time)
if passed as [], prparms=[50 0] will be used.

OUTPUT:
theta_data: Matrix containing the parameter vectors saved after each weight elimination round.
NSSEvec: Vector containing the training error (SSE/2N) after each weight elimination.
FPEvec: Contains the FPE estimate of the average generalization error
NSSEtestvec : Contains the normalized SSE evaluated on the test set
deff: Contains the "effective" number of weights
pvec: Index to the above vectors

A.5 NNVALID

Validate a neural network input-output model of a dynamic system.
I.e., a network model which has been generated by NNARX, NNRARX, NNARMAX1+2, NNRARMX1+2, or NNOE. The following plots are produced:

- Observed output together with predicted output
- Prediction error
- Auto-correlation function of prediction error and cross-correlation between the prediction error and input
- A histogram showing the distribution of the prediction errors
- Coefficients of extracted linear models

Network generated by NNARX (or NNRARX):
[Yhat,NSSE] = nnvalid('nnarx',NetDef,NN,W1,W2,Y,U)
Network generated by NNARMAX1 (or NNRARMAX1):
[Yhat,NSSE] = nnvalid('nnarmax1',NetDef,NN,W1,W2,C,Y,U)
Network generated by NNARMAX2 (or NNRARMX2):
[Yhat,NSSE] = nnvalid('nnarmax2',NetDef,NN,W1,W2,Y,U)
Network generated by NNOE:
[Yhat,NSSE] = nnvalid('nnoe',NetDef,NN,W1,W2,Y,U)
Network generated by NNARXM:
[Yhat,NSSE] = nnvalid('nnarxm',NetDef,NN,W1,W2,Gamma,Y,U)
NB: For time-series, U is left out!

References

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

Arsenijević, D., Milosavljević. M., 2002, *Analysis of Neural Network Models in Serbian Speech Consonants*, Electronic Review, Faculty of Electrical Engineering, Banja Luka.

Bishop, C., 1995, *Neural networks for pattern recognition.* Oxford University Press.

Bojanić, M., Delić, V., 2009, Automatic Emotion Recognition in Speech: Possibility and Significance. *Electronics*, Vol.13, No.2, pp.35-40.

Collobert, R., Weston, J., 2008, A unified architecture for natural language processing: deep neural networks with multitask learning. In *Proceedings of the 25th International conference on machine learning*, pp.160-167. New York, NY, USA.

Delić, V., 2000, Speech Databases in Serbian Language Recorded with the AlfaNum Project. DOGS conference, pp.29-32, September 21st-22nd, 2000, Novi Sad.

Delić, V., Sečujski, M., Jakovljević, N., Janev, M., Obradović, R., Pekar, D., 2010, *Speech Technologies for Serbian and Kindered South Slavic Languages.* Chapter 9 in the Shabtai, N. ed book Advances in Speech Recognition, pp.141-165.

Hansen, L.K., Rasmusen, C.E., 1994, Pruning from adaptive regularization. *Neural Computation* 6(6), pp.1223-1232.

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

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.

Khanagha, V., Yahia, H., Daoudi, K., 2011, Reconstruction of Speech Signals from Their Unpredictable Points Manifold, *Nonlinear Speech Processing*, 2011 7015, pp.1-7, Available at

http://hal.inria.fr/docs/00/64/71/97/PDF/KHANAGHA_Reconstruction_of_speech_from_UPM.pdf, Retrieved on January 22, 2014.

Kupusinac, A., Sečujski, M., 2009, Part of Speech Tagging Based on Combining Markov Model and Machine Learning. *Speech and Language.* November 13th-14th, 2009, Belgrade.

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

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

Little, M., McSharry, P.E., Moroz, I., Roberts, S., 2006, Testing the assumptions of linear prediction analysis in normal vowels. *Journal of the Acoustic Society of America*, 119, pp.549-558.

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

Marković, M., Milosavljević, M., Kovačević, M., Veinović, M., 1999, Robust AR Speech Analysis Based on MGLR Algorithm and Quadratic Clasifier with Sliding Training Set. In *Proceedings of IMACS/IEEECSCC'99*, pp.2401-2408.

Mesbahi, L., Jouvet, D., Bonneau, A., Fohr D., Illina, I. Laprie, Y., 2011, Reliability of non-native speech automatic segmentation for prosodic feedback. In *SlaTE, 2011*, Venice, Italy.

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

Mikolov, T., Sutskever, I., Deodoras, A., Le, H.S., Kombrink, S., Cernocky, J., 2012, Subword language modelling with neural networks. Unpublished.

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

Norgaard, M., 2001, *Neural Network Based System Identification Toolbox, Version 1.2*, Technical University of Denmark, Department of Automation Department of Mathematical Modelling, Technical Report 97-E-851.

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.

Park, S., Choi, S., 2008, A constrained sequential EM algorithm for speech enhancement, *Neural Networks* 21, pp.1401-1409.

Pekar, D., Obradović, R., Delić, V., Krčo, S., Šenk, V., 2002, Connected Words Recognition. DOGS conference, September 21st-22nd, 2002, pp.21-24, Novi Sad.

Pekar, D., Mišković, D., Knežević, D., Vujnović Sedlar, N., Sečujski, M., Delić, V., 2010, Chapter 7 in the Shabtai, N. ed book Advances in Speech Recognition, pp.105-122.

Protić, D., Milosavljević, M., 2005, Generalizaciona svojstva različitih klasa linearnih i nelinearnih modela govornog signala, *Festival informatičkih dostignuća INFOFEST*, *Festivalski katalog*, pp.247-258, Budva.

Protić, D., Milosavljević, M., 2006, NNARX Model of Speech Signal Generating System: Test Error Subject to Modeling Mode Selection, *Conference MIEL*, *IEEE Catalog*, May 2006, pp.685-688, Belgrade.

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, 2010 IEEE Workshop*, 11-15 December 2011, pp.30-35, Waikoloa, HI.

Sečujski, M., Pekar, D., 2014, Evaluacija različitih aspekata kvaliteta sintetizovanog govora. Available at http://www.savez-slijepih.hr/hr/kategorija/evaluacija-razlicitih-aspekata-kvaliteta-sintetizovanog-govora-452/. Retrieved on February 16, 2014.

Shahin, A.J., Pitt, M.A., 2012, Alpha activity making world boundaries mediates speech segmentation, *European Journal of Neuroscience*, Vol.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.

Stanimirović, Lj., Ćirović, Z., 2008, Digitalna obrada govornog signala, Retrieved from www.viser.edu.rs/download/uploads/2371.pdf Accessed January 24, 2013.

Svarer, C., 1995, *Neural Networks for Signal Processing*, Technical University of Denmark.

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.

UPOREDNA ANALIZA FONEMA SRPSKOG JEZIKA: LINEARNI I NELINEARNI MODELI

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

Sažetak

U radu je prikazana analiza karakteristika vokala i nevokala srpskog jezika. Vokale karakteriše kvaziperiodičnost i spektar snage signala sa dobro uočljivim formantima. Nevokale karakteriše kratkotrajna kvaziperiodičnost i mala snaga pobudnog signala. Vokali i nevokali modelovani su linearnim AR modelima i odgovarajućim nelinearnim modelima koji su generisani kao feed-forward neuronska mreža sa jednim skrivenim slojem. U procesu modelovanja korišćena je minimizacija srednje kvadratne greške sa propagacijom unazad, a kriterijum izbora optimalnog modela jeste zaustavljanje obučavanja, kada normalizovana srednja kvadratna test greška ili finalna greška predikcije dostignu minimalnu vrednost. LM metod korišćen je za proračun inverzne Hessianove matrice, a za pruning je upotrebljen Optimal Brain Surgeon. Prikazana su generalizaciona svojstva signala u vremenskom i frekvencijskom domenu, a kroskorelacionom analizom utvrđen je odnos signala na izlazima neurona skrivenog sloja.

Uvod

Unazad nekoliko godina NN su primenjivane u procesima obrade podataka, pa samim tim i govornog signala. Značajan napredak u ovoj oblasti kreće se u pravcu ubrzanja konvergencije algoritama obučavanja. Pored izbora strukture NN, izbor prenosnih funkcija takođe je veoma bitan. Nadzirano obučavanje sa ulaznim podacima i predefinisanim izlazom zahtevaju korišćenje funkcije gubitaka ili greške za utvrđivanje odstupanja očekivane, prediktovane vrednosti od tačnih vrednosti podataka. Od mnogo primenjenih algoritama u radu je korišćen BPA, koji je istovremeno i najrasprostranjeniji algoritam obučavanja u ovoj oblasti. Analizirani su vokali i nevokali koje su izgovarali i muškarci i žene, u kontekstu reči ili izolovano. BPA je korišćen uz standardni gradijentni metod, koji je prilagođen LM metodom. U radu je korišćen OBS za pruning. Kriterijum zaustavljanja pruninga su minimizacija $NSSE_{TEST}$ i FPE.

Prikazane su vrednosti dobijenih grešaka za vokale i nevokale, pojačanja FPE, kao i rezultati kroskorelacione analize signala na izlazima neurona skrivenog sloja FNN.

Modeli

*Ukoliko je u obradi govora dostupan samo govorni signal koriste se AR modeli sa dva pola na približno (2n+1)*500Hz, n = 0, 1,... Ukoliko je na raspolaganju i signal sa glotisa koriste se ARX linearni modeli*

sa dodatnim ulazom. Uz to, pokretna srednja vrednost greške koristi se u ARMA(X) modelima, kada je dostupna korekcija greške. Međutim, tada postoji problem nestabilnosti u procesu obučavanja ukoliko je vrednost greške velika, što može dovesti do nestabilnosti modela. Zbog toga se u modelovanju koristi nelinearna FNN na koju je moguće primeniti pruning, odnosno proces odbacivanja viška parametara u odnosu na potpuno povezanu strukturu, tako da ukupna greška obučavanja ne prelazi dozvoljenu vrednost. Kriterijum zaustavljanja pruninga je dostizanje minimuma $NSSE_{TEST}$, $NSSE_{TRAIN}$ ili FPE. Nelinearni modeli su, u opštem slučaju, tačniji, ali proces njihovog obučavanja traje duže.

Obučavanje modela

FNN i AR modeli su obučavani trening skupovima. Obučavanje je izvedeno promenom parametara po BPA. Korišćena je LM aproksimacija za proračun Hessianove matrice. Optimalni korak promene greške aproksimiran je Taylor-ovim nizom. Aproksimacija drugog reda ukazuje na nekorelisanost ulaza sa dobijenom greškom, što omogućuje ispravan smer korekcije greške. Korišćene su MATLAB-ove metode nnarx i marq. Treniran je i AR-10 čiji je red jednak broju ulaza u FNN (10), odnosno procenjeni izlaz dobijen je na osnovu 10 prethodnih vrednosti datog signala. Inicijalna vrednost parametara je slučajna. Formantne karakteristike vokala su takve da njihov broj i raspored određuju parametre modela. AR model je stabilan, jednostavan i računarski malo zahtevan. Predikcija je bazirana na MSE kriterijumu. Za FNN korišćen je OBS pruning. Za promene greške računa se puna Hessian-ova matrica. Akaike-ova FPE omogućuje da se proceni generalizaciona greška za datu FNN, kada je poznat broj parametara. Da bi bilo moguće uporediti AR i NNAR modele uvedeno je pojačanje FPE, tj. odnos MSE za AR model i FPE za FNN, a validacija je izvedena za sve vokale i sve govornike. Isti proces izveden je i za govornike i nevokale koji su izgovarani u kontekstu reči ili van njih.

Signali govora

Vokalno-nazalni trakt je deo sistema za proizvođenje govora, čija se prenosna funkcija može aproksimirati akustičkim filtrom. Vazduh, pobuda iz pluća, prolazi kroz vokalno-nazalni trakt i, u zavisnosti od toga da li glasne žice vibriraju ili ne, formira se vokal ili nevokal. Zvuk koji se čuje kao govor nastaje zračenjem sa usana i iz nosa. Vokali su kvaziperiodični u dužem vremenskom periodu, pobuda je snažna, a glasne žice vibriraju. Kod ostalih fonema kvaziperiodičnost je zanemariva, pobuda je slab signal ili kombinacija takvog signala sa šumom.

Rezultati

Za obučavajuće skupove trenirani su AR-10 i FNN, strukture 10-3-1. Pruning je izveden OBS metodom sa maksimalno 20 iteracija retreninga po odbacivanju jednog parametra. Korišćen je algoritam

nnprune. Dobijene su NSSE za obučavajući i test skup, i FPE. U radu su prikazane strukture koje zaustavljaju pruning dostizanjem minimalnih vrednosti $NSSE_{TEST}$ i FPE. Izračunata je i NSSE za AR-10. Validacija je izvedena funkcijom nnvalid. Za nevokale računato je pojačanje FPE za žene i za muškarce. Uvedena je mera rastojanja dva signala (u spektralnom domenu) i poređeni su spektri snage signala na izlazima neurona skrivenog sloja. Takođe, izvedena je kroskorelaciona analiza i kumulativno sumiranje apsolutnih vrednosti kroskorelacionih signala za male distance.

Zaključak

U radu je analizirana klasa FNN, strukture sa 10 ulaza, promenljivim brojem neurona u skrivenom sloju i jednim izlazom, za predikciju govornog signala, tj. fonema srpskog jezika. Metodologija izbora arhitektura sa dobrim generalizacionim osobinama, zasnovana na pruningu, omogućila je znatno smanjenje broja parametara modela i veću tačnost, u odnosu na linearne AR modele. Granične arhitekture odlikuju se minimalnim brojem parametara u okviru zadate margine greške. Pri analizi vokala uočen je uticaj nevokalizovanih fonema koji su takođe prediktovani FNN i AR modelima. Radi sagledavanja diskriminacionih osobina izabranih klasa modela razvijena je metoda višedimenzionog skaliranja zasnovana na novoj meri rastojanja. Analiza gubitka diskriminatornosti ukazuje na činjenicu da FNN modeli za foneme u srpskom jeziku imaju znatno veću diskriminacionu snagu, što ih čini upotrebljivim u širokoj klasi prepoznavanja govornih elemenata. Spektralna analiza pokazuje da su izlazni signali neurona skrivenog sloja dobro korelisani sa dominantnim formantnim karakteristikama ulaznog signala. Vremenska karakteristika ukazuje na slabu statističku zavisnost ovih signala za niske redove kroskorelacione zavisnosti (do petog reda). Analize ukazuju na blagu prednost kriterijuma $NSSE_{TEST}$ u odnosu na FPE kriterijum, na nezavisnom signalu. U slučaju kratkih obučavajućih skupova FPE je prihvatljiv kriterijum.

Rezultati ukazuju na činjenicu da predložena klasa FNN modela srpskog jezika i izbor arhitektura sa najboljim generalizacionim svojstvima obezbeđuju modele visoke tačnosti sa internom distribuiranom strukturom koja odgovara prirodnom vremensko-frekvencijskom sadržaju ulaznih signala, i visokih su diskriminaconih svojstava za isti broj parametara u odnosu na tradicionalne linerane modele.

Ključne reči: *AR model, neuronske mreže, govor.*

Datum prijema članka/Paper received on: 18. 12. 2013.

Datum dostavljanja ispravki rukopisa/Manuscript corrections submitted on: 06. 03. 2014.

Datum konačnog prihvatanja članka za objavljivanje/Paper accepted for publishing on: 08. 03. 2014.