



A parallel neural network approach to prediction of Parkinson's Disease

Freddie Åström^{a,*}, Rasit Koker^{b,c,1}

^a Computer Vision Laboratory, Department of Electrical Engineering, Linköping University, SE-58183 Linköping, Sweden

^b Engineering Faculty Esentepe Kampus, Computer Engineering Department, Sakarya University, 54187 Sakarya, Turkey

^c Faculty of Engineering and Natural Sciences, Department of Computer Engineering, International University of Sarajevo, 71000 Sarajevo, Bosnia and Herzegovina

ARTICLE INFO

Keywords:

Parallel neural networks
Parkinson's Disease
Decision support system

ABSTRACT

Recently the neural network based diagnosis of medical diseases has taken a great deal of attention. In this paper a parallel feed-forward neural network structure is used in the prediction of Parkinson's Disease. The main idea of this paper is using more than a unique neural network to reduce the possibility of decision with error. The output of each neural network is evaluated by using a rule-based system for the final decision. Another important point in this paper is that during the training process, unlearned data of each neural network is collected and used in the training set of the next neural network. The designed parallel network system significantly increased the robustness of the prediction. A set of nine parallel neural networks yielded an improvement of 8.4% on the prediction of Parkinson's Disease compared to a single unique network. Furthermore, it is demonstrated that the designed system, to some extent, deals with the problems of imbalanced data sets.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

The cause of Parkinson's Disease (PD) is unknown, however research has shown that a degradation of the dopaminergic neurons affect the dopamine production to decline (Mahlon & Jorge, 2009). Dopamine is used by the body to control movement, hence the less dopamine that is in circulation the more difficult the person has to control the movements and may experience tremors and numbness in extremities. As a direct cause of reduced control of motor-neurons in the central nervous system, the ability of articulating vocal phonetics is reduced. In this case the symptom (the inability to articulate words) is related to the presence of Parkinson's Disease and is described as Dysphonia, a reduced functionality of the vocal cords. One of the immediate effects of vocal Dysphonia is that the voice is experienced as more course by fellow listeners (Ropper & Samuels, 2009).

The features used in the prediction of Parkinson's Disease in this study have been obtained from vocal records of people. The field of speech processing and development of speech recognition systems have received considerable attention during the last decades. Separation of voice and background noise are important issues. With the emerge of portable phones and studio recording microphones analysing methods involving traditional digital signal processing approaches such as hidden Markov models, Kalman filter, short-time frequency analysis and wavelet analysis have been success-

fully used for both speech enhancement and speech recognition applications (Esposito & Marinaro, 2005; Gales, 1995; Hussain, Chetouani, Squartini, Bastari, & Piazza, 2007; McDonough et al., 2007; Silva & Joaquim, 2008; Skowronski & Harris, 2006; Sroka & Braid, 2005; Yan et al., 2008).

Previous related studies conducted on vocal recordings from Parkinson's Disease are scarce. The data set used in this study was collected by Little, McSharry, Hunter, and Ramig (2009) who used Support Vector Machine in order to distinguish between subjects of people who have normal vocal signs and subjects suffering from PD. They achieve a classification accuracy of 91.4% but they do not report single class true positive rates. This is noteworthy because of the unequal sick to healthy ratio data class distribution of the collected data. Das (2010) has made a comparative study of multiple classification algorithms on the same data set used in this study with regard to neural networks, DMNeural analysis, regression analysis and decision trees with the presented results of classification accuracy of 92.9%, 84.3%, 88.6% and 84.3%, respectively. In their paper the analysis was carried out on data exploration software. Another study published by Lee, Rhee, Kim, and Zhang (2009) on the imbalanced problem in biomedical data, uses a sampling scheme in collaboration with a Naïve Bayes classifier to deal with the imbalanced problem. The sampling pattern is to start with a small portion of the data to train the classifier, and then successively to increase the number of training samples regardless of the initial class distribution. The method shows promising results with positive predictive rates of 66.2% for normal subjects and 90.0% for subjects with PD. Neural networks have matured immensely since the first attempts of modelling the modern network architecture was introduced to the machine learning community half a century

* Corresponding author. Tel.: +46 13 28 2460; fax: +46 13 13 8526.

E-mail addresses: freddie.astrom@liu.se (F. Åström), rkoker@ius.edu.ba (R. Koker).

¹ Tel.: +387 957210; fax: +387 33 957 105.

ago. The massive parallel computational structure of neural networks is what has contributed to its success in predictive tasks. It has been shown that the approach of using parallel networks is successful with respect to increasing the predictive accuracy of neural networks in robotics (Koker, 2005) and in speech recognition (Lee, 1997). In the case of the speech recognition application, Lee (1997) attempts to forward propagate unlearned data to a neighbouring neural network and achieves an increase for the classification accuracy of at most 6.7% compared to a single traditional multi-layer neural network approach.

This work presents a parallel network application with a rule-based decision system in order to further increase the predictive accuracy of Parkinson's Disease prediction based on vocal recordings. In the rule-based decision part the outputs of the neural networks are evaluated based on the majority outcome. Additionally, unlearned data of each neural network is used in the training of the following neural network to increase the predictive accuracy. For the proposed system it is shown with a case study of Parkinson's Disease that some of the difficulties with imbalanced data sets are resolved. The type of network used is the standard feed-forward backpropagation neural network, since they have proven useful in biomedical classification tasks (Mazurowskia et al., 2008). The performance of the trained neural networks is evaluated according to the true positive, true negative and accuracy rate of the prediction task. Furthermore the area under the receiver operating characteristic curve and the mean squared error are used as statistical measurements to compare the success of the different models.

The paper is organised as follows, firstly the data set of Parkinson's Disease is defined. Secondly, a short description of neural networks and the used training algorithm is presented in Section 3. Thirdly, the proposed algorithm and its configurations are illustrated in Section 4. Lastly, results are shown in Section 5 followed by a conclusion.

2. Parkinson's Disease data set

The data used in this study is a voice recording originally done at University of Oxford by Max Little (Asuncion & Newman, 2009; Little et al., 2009; Little, McSharry, Roberts, Costello, & Moroz, 2007). In the same studies a detailed presentation is made on the specificities of the recording equipment as well as in what environment the experiment was conducted. The recording consists of 195 entries collected from 31 people whom 23 are suffering from Parkinson's Disease. From the 195 entries, 147 are of Parkinson's Disease and the remainder 48 are of normal character. The attributes used in the prediction task in this study are MDVP:Jitter (Abs), Jitter:DDP, MDVP:APQ, Shimmer:DDA, NHR, HNR, RPDE, DFA, D2 and PPE. These attributes are computed from the vocal recording and they describe changes in fundamental frequency, amplitude variations, noise to tonal components and other nonlinear voice measurements. For more information on the data set and feature extraction process refer to references (Asuncion & Newman, 2009; Little et al., 2007; Little et al., 2009).

3. Neural networks

A simple topology of the neural network can be seen in Fig. 1. The backpropagation neural network structure was first constructed in the late 1980, since then it has become a very recognisable technique within the machine learning community (Haykin, 1999).

The backpropagation neural network was originally designed to depend on the so called delta rule which is also known as the steepest descendant training algorithm. The standard backpropa-

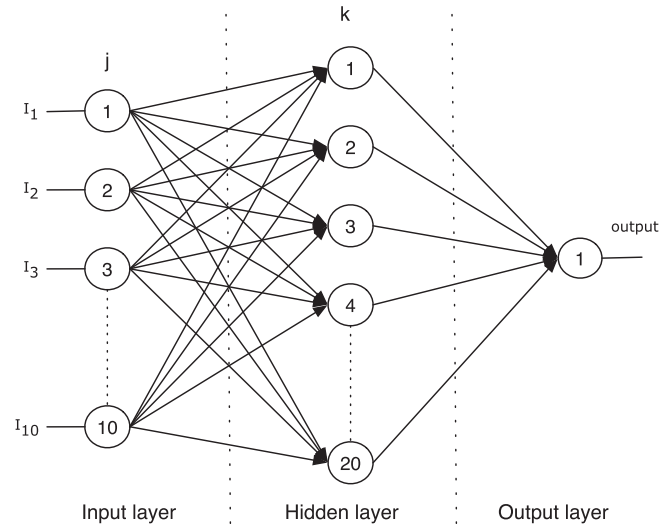


Fig. 1. Topology of a neural network with three layers. The input neurons 1–10 correspond to those input attributes as seen in Fig. 2.

gation algorithm consists of a forward pass of the training sample data through an input layer, H hidden layers and then an output layer. In order to update the weight vector w_{ki} for all neurons i in layer k , a backward pass of the sample is made, hence the name “backpropagation neural network” comes from it. When all training samples have made one forward and one backward pass, it is said that the training data has been presented to the network in one epoch. The parameter n represents the number of conducted epochs. Considering a network of K layers, the output from layer k in the forward pass will be (disregarding the constant bias term)

$$\zeta_k(n) = \sum_{j=0}^m \omega_{kj}(n) \gamma_j(n) \quad (1)$$

where ω is the current weight vector, m is the number of neurons in layer k and γ is the output vector from the previous layer defined as

$$\gamma_j(n) = \alpha_j(\zeta_j(n)) \quad (2)$$

The forward pass for layer k is then finished and its output will be determined by an activation function α_k of any type. Commonly, the log sigmoidal, piecewise linear function or a hyperbolic tangent sigmoidal are used as activation functions. By defining the error in the forward pass in the output layer as the difference between the predicted value γ and the desired value d as the squared error

$$\epsilon_k(n) = \frac{1}{2} \sum_{j=0}^m \epsilon_j^2(n) = \frac{1}{2} \sum_{j=0}^m [d_j(n) - \gamma_j(n)]^2 \quad (3)$$

and differentiating ϵ with respect to ϵ_j , it can be shown by the use of partial derivatives that the delta rule can be expressed as (Haykin, 1999),

$$\Delta \omega_{kj}(n) = -\eta \frac{\partial \epsilon(n)}{\partial \omega_{kj}(n)} \quad (4)$$

where η defines the so called learning rate.

3.1. Levenberg–Marquardt training algorithm

In this study the Levenberg–Marquardt (LM) training algorithm was used. By modifying the steepest descendant rule the it can be shown that the LM training algorithm can be obtained. The LM training algorithm is convergence faster and is a more accurate error minimising algorithm compared to the steepest descendant

rule. Viewing the problem of finding an alternative expression of minimising the error rate, let's consider to approximate the error function ϵ around value of n and the correction δ with a Taylor expansion (Lourakis, 2008),

$$\epsilon(n + \delta) = \epsilon(n) + J(n) \cdot \delta \quad (5)$$

where $J(n)$ is the Jacobian matrix. It can then be shown that the Levenberg–Marquardt delta update rule is defined as the solution to an optimisation problem for which it can be proven has the following solution (Hagan & Menhaj, 1994)

$$\delta(n) = J^T(n)[J^T(n)J(n) + \mu I]^{-1}\epsilon(n) \quad (6)$$

The free parameter μ defines the robustness of the algorithm. This is realised by considering the value of the Jacobians approach to an extreme point on the error surface, thus the parameter μ adjusts for the Jacobians approach to zero (Fan & Pan, 2009). If the value of μ is defined as large then the LM delta rule will be approximately the steepest descendant, however if μ is defined as small then the algorithm can experience a decrease of robustness (Hagan & Menhaj, 1994).

4. Implementation of proposed parallel neural network structure

The proposed system consists of two steps; the first step is a set of parallel feed-forward neural networks with the Levenberg–Marquardt backpropagation training algorithm and the second step is a rule-based system. The rule-based system implements a simple voting schedule deciding the most probable outcome. All the proposed training techniques were implemented with the help of Matlab's Neural Network Toolbox. A descriptive block scheme can be seen in Fig. 2. The inputs seen in Fig. 2 were derived by Little et al. (Little et al., 2007; Little et al., 2009).

4.1. The parallel network block

The number of parallel neural networks should be an odd number. In this case it is easier to decide the prediction result by considering the majority output. The process of training the system is determined as follows; each network is presented with the full training set and as training progresses the first network propagates the data samples which it could not accurately predict to the second network. The second network is now trained with the full training data set and the additional data samples propagated from the first network; and so on. Algorithm 1 illustrates this.

Algorithm 1. Pseudo code illustrating the parallel network block

```

for  $i = 1$  to number of networks do
  randomise (traindata, target);
  net( $i$ ) = train the network (data, target);
  simvec = simulate the training data (net, data, target);
  {Find performance parameters for training data.}
  if simvec  $\neq$  target then
    {if there are unlearned samples, extend the data set with them.}
    extend (traindata, target);
  end if
end for

```

In the training stage of the system only the parallel block is used. In order to train the system, a single network is trained until the training data set is fully recognised. The rule-based system comes into place when testing the system because there is no need to vote on the training set since each separate network should recognise the complete training set.

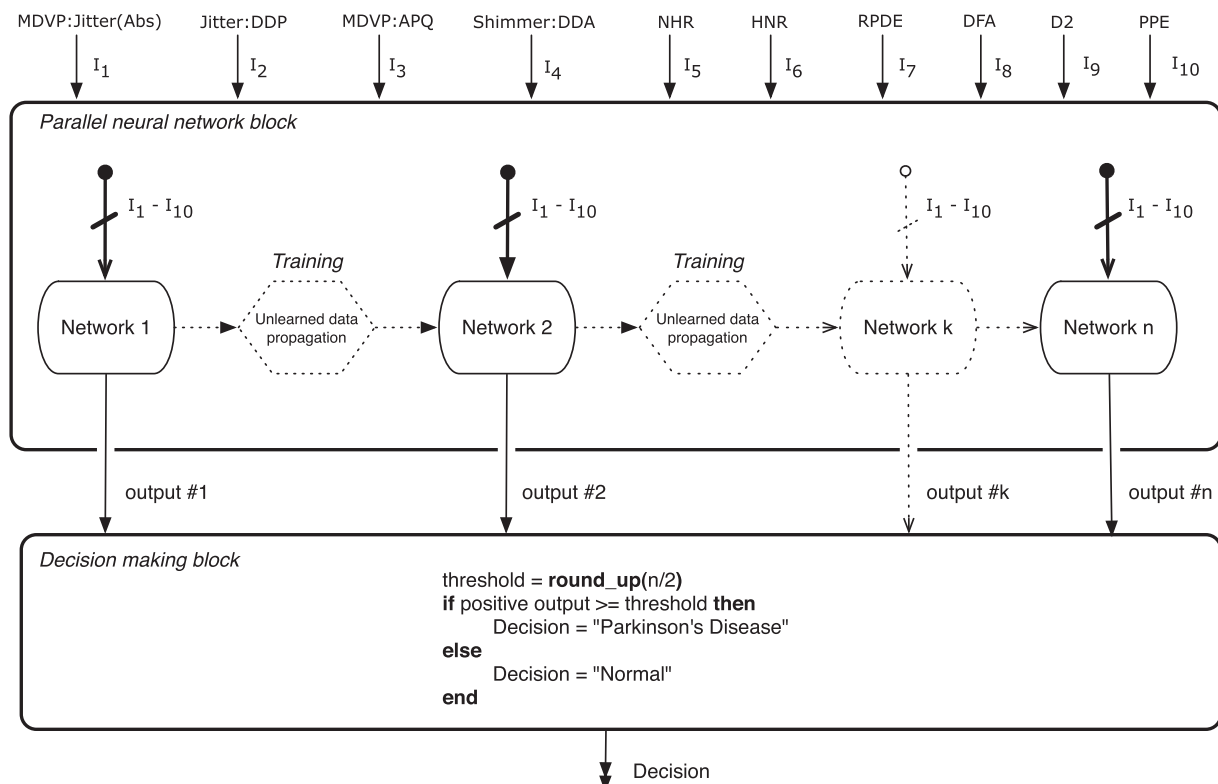


Fig. 2. A block diagram describing the overall system layout.

4.2. Rule-based system

The rule-based system is based on a voting decision scheme. The majority number of networks that are in favour for a class prediction will be the prediction of the particular sample. The purpose of designing a rule-based system is to increase the robustness of prediction. Algorithm 2 illustrates this:

Algorithm 2. Pseudo code illustrating the rule-based block

```

threshold = ceil (num of networks/2);
for i = 1 to num of test data do
  if sum of sim (i)  $\forall$  networks  $\geq$  threshold then
    recognised (i) = true;
  else
    recognised (i) = false;
  end if
end for
{Find performance parameters for test data.}

```

4.3. Training and evaluation

The PD data set was randomised and then divided into two partitions by random selection. Training of the network block is done with 60% of the data whereas the remaining 40% are used for evaluating the predictive accuracy of the system. The class distribution of the two sets can be seen in Table 1.

Each individual network consists of three layers, where the hidden layer has 20 neurons. The training algorithm for the networks is the Levenberg–Marquardt training algorithm configured by the following parameters; initially μ is set to 10^{-3} and as training progresses, the parameter is decreased or increased with 10^{-1} and 5 respectively. Training stops if $\mu \geq 10^{10}$. The LM training algorithm is used because of its fast convergence times and excellent generalisation ability. Activation function in the hidden and the output layer is of hyperbolic tangent sigmoidal type.

Performance parameters are computed according to a confusion matrix. From the matrix, the true positive rate (TPR) is the rate of positive (subjects with PD) outcomes that are predicted as positive. Whereas, the true negative rate is the rate of which negative (normal subjects) outcomes are predicted as negative. The accuracy (ACC) of the prediction is the rate of which both positive and negative outcomes are correctly predicted. The values are calculated as follows

$$TPR = \frac{TP}{TP + FN} \quad (7)$$

$$TNR = \frac{TN}{TN + FP} \quad (8)$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (9)$$

where TP are true positive, TN are true negative, FP are false positive and FN are false negative outcomes. Also, the mean square error (MSE) is computed to show the performance of the proposed system.

$$MSE = \frac{1}{N} \sum_{j=0}^N [d_j(n) - \gamma_j(n)]^2 \quad (10)$$

with the same meaning of the variables as in Section 3 and N is the number of samples. Finally, the area (AUC) under the receiver operating curve (ROC) is used as a method of comparing the accuracy of the system's robustness.

The system is tested with 1, 3, 5, 7 and 9 parallel networks. The result is depicted from an average run of 30 trials, with the motivation that after 30 randomly initialised network weights and randomised training data the variability of the system parameters is significantly large.

In order to test the system, two experiments are conducted:

- *Experiment 1*, the system is tested without forward propagation of the unlearned data samples in the neural network training stage.
- *Experiment 2*, the system is tested with forward propagation of the unlearned data samples in the neural network training stage.

5. Results and discussion

To demonstrate the increased robustness of the system and to justify forward propagation of unlearned data samples, two experiments are conducted. The first experiment is a control test, where the system is evaluated without forward propagation of the unlearned data samples. Results from the first experiment can be seen in Table 2. The second experiment is the system evaluated with forward propagation of unlearned data samples, the result of the second experiment is depicted in Table 3.

It has been shown in this study that parallel neural networks in combination with a rule-based system increase performance of true recognition rates in an imbalanced data set. In both conducted experiments all measurement parameters are improved compared to single network predictions. From the two experiments it is evident that the parallel network system with forward propagation of unlearned data samples increases the robustness and decrease the variability as seen in the system which does not have this feature. Despite the advantages of having an accurate system prediction, the training time and complexity of the parallel network algorithm do increase as the number of parallel networks increases.

The data set is imbalanced with regard to the class distribution. Out of 195 samples, 75.4% are of Parkinson's Disease type and the remainder are of normal character. It implies that the baseline prediction is 75.4% and any prediction accuracy less than the baseline is not relevant. A common problem with an imbalanced data set is high single class true recognition rates of the majority class. The problem is that if the positive class has three times more samples than the negative class then a 100% true recognition rate of the positive class yield that the baseline prediction of 75% is achieved with the worst case of zero recognition rate of the negative class. These problems are significant to address in order to achieve a robust and general prediction systems. Viewing the results as seen in Tables 2 and 3 it was noticed that the true positive rate and especially the true negative rate was significantly increased as the number of parallel networks are increased. The best result can be seen with nine parallel networks with forward propagation of unlearned data enabled. This clearly shows that the proposed prediction algorithm can improve single class predictions. In the literature false positive rates up to 25%–30% of the positive class have been reported (Lee et al., 2009). It has been demonstrated in this study that a true positive rate up to 90.5% and 93.5% of each class can be achieved by using nine parallel networks. This is a

Table 1
Description of data set divided into training and test set.

	Training set	Test set	Total
Normal	21	27	48
Parkinson's Disease	96	51	147
Total	117	78	195

Table 2
Experiment 1: Performance measurements from an average of 30 trails runs for Parkinson's Disease. The system is trained *without* forward propagation of unlearned training samples.

Number of networks	ACC	TPR	TNR	AUC	MSE
1	84.10 ± 10.80	82.10 ± 11.00	62.90 ± 39.00	80.70 ± 19.50	0.1732 ± 0.1143
3	89.10 ± 2.60	89.00 ± 3.40	90.20 ± 5.40	96.00 ± 1.20	0.0867 ± 0.0184
5	89.20 ± 2.60	89.10 ± 3.00	90.30 ± 5.40	96.10 ± 1.60	0.0823 ± 0.0199
7	89.90 ± 2.90	89.40 ± 4.20	92.20 ± 4.30	96.70 ± 0.70	0.0777 ± 0.0177
9	89.70 ± 2.90	89.00 ± 3.30	91.50 ± 3.30	96.60 ± 1.10	0.0770 ± 0.0138

Table 3
Experiment 2: Performance measurements from an average of 30 trails runs for Parkinson's Disease. The system is trained *with* forward propagation of unlearned training samples.

Number of networks	ACC	TPR	TNR	AUC	MSE
1	84.10 ± 9.70	85.70 ± 9.90	72.00 ± 31.1	85.00 ± 18.10	0.1472 ± 0.1016
3	87.90 ± 5.00	87.70 ± 5.50	86.70 ± 5.50	95.80 ± 1.10	0.0967 ± 0.0273
5	89.50 ± 2.60	89.10 ± 3.70	91.60 ± 4.70	96.60 ± 0.90	0.0826 ± 0.0156
7	90.20 ± 2.00	89.20 ± 2.60	93.10 ± 4.40	96.70 ± 0.80	0.0773 ± 0.0131
9	91.20 ± 1.60	90.50 ± 2.10	93.00 ± 3.30	96.80 ± 0.70	0.0703 ± 0.0082

significant improvement compared to previously demonstrated results.

Furthermore, it can be seen from the mean squared error (MSE) in Tables 2 and 3, that after a certain number of parallel networks, the overall accuracy of the prediction does not improve. For the case of forward propagation of unlearned data, this threshold is after seven networks. But for the case of no forward propagation of unlearned data there is no significant improvement after five parallel networks. This is particularly evident when considering the low standard deviation of each performance measurement. Here the trade-off between a more accurate predictive system and the system's complexity is well demonstrated. Considering the area (AUC) under the receiver operating curve it can be seen that the area has drastically increased when utilising three parallel networks compared to one network in both conducted experiments. The higher value of AUC implies that there were more samples (from both the positive and negative class) that were predicted correctly.

6. Conclusion

The proposed prediction system is based on using parallel neural networks and evaluating the outputs to find the best prediction result. It is known that in parallel systems the reliability increases. In the same way it is evidently observed that the performance of the prediction has been increased in this paper compared to the use of a unique network. Using unlearned data in the next neural network also gave profound impact on the robustness of the system. It has also been shown that after a certain number of parallel networks, the accuracy of the prediction does not improve anymore.

References

- Asuncion, A., & Newman, D. (2009). UCI machine learning repository, university of california, irvine, school of information and computer science, <<http://www.ics.uci.edu/~mllearn/MLRepository.html>>.
- Das, R. (2010). A comparison of multiple classification methods for diagnosis of Parkinson disease. *Expert Systems with Applications*, 37, 1568–1572.
- Esposito, A., & Marinaro, M. (2005). *Nonlinear speech modeling and applications. chapter Some Notes on Nonlinearities of Speech. Lecture Notes in Computer Science* (Vol. 3445). Berlin/ Heidelberg: Springer, pp. 1–14.
- Fan, J., & Pan, J. (2009). A note on the Levenberg–Marquardt parameter. *Applied Mathematics and Computation*, 207, 351–359.

- Gales, M.J.F. (1995). *Model-Based Techniques for Noise Robust Speech Recognition*, Ph.D. thesis Gonville and Caius College.
- Hagan, M. T., & Menhaj, M. B. (1994). Training feedforward networks with the Marquardt algorithm. *IEEE Transactions on Neural Networks*, 5, 989–993.
- Haykin, S. (1999). *Neural networks – A comprehensive foundation* (2nd ed.). Prentice Hall, Inc.
- Hussain, A., Chetouani, M., Squartini, S., Bastari, A., & Piazza, F. (2007). *Progress in nonlinear speech processing. chapter Nonlinear Speech Enhancement: An Overview*, Springer Berlin/ Heidelberg, Vol. 4391, pp. 217–248.
- Koker, R. (2005). Reliability-based approach to the inverse kinematics solution of robots using Elman's networks. *Engineering Applications of Artificial Intelligence*, 18, 685–693.
- Lee, B. J. (1997). Parallel neural networks for speech recognition. *International Conference on Neural Networks*, 4, pp. 2093–2097.
- Lee, M. S., Rhee, J. -K., Kim, B. -H., & Zhang, B. -T. (2009). AESNB: Active example selection with naïve bayes classifier or learning from imbalanced biomedical data. In *2009 Ninth IEEE International Conference on Bioinformatics and Bioengineering*, (pp. 15–21).
- Little, M. A., McSharry, P. E., Hunter, E. J., & Ramig, L. O. (2009). Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. *IEEE Transactions on Biomedical Engineering*, 56, 1015–1022.
- Little, M. A., McSharry, P. E., Roberts, S. J., Costello, D. A., & Moroz, I. M. (2007). Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. *BioMedical Engineering OnLine*, 2007, 6.
- Lourakis, M.I.A. (2008). *A Brief Description of the Levenberg–Marquardt Algorithm Implemented by Levmar*. Technical Report Institute of Computer Science Vassilika Vouton, P.O. Box 1385, GR 711 10, Heraklion, Crete, Greece.
- Mahlon, D., & Jorge, J. (2009). *Harrison's principles of internal medicine*, <<http://www.accessmedicine.com/content.aspx?aid=2905868>>, chapter 366. Parkinson's Disease and Other Extraparasympathetic Movement Disorders. <<http://www.accessmedicine.com/content.aspx?aid=2905868>>: The McGraw-Hill Companies, Inc. (17th ed.).
- Mazurowskia, M. A., Habasa, P. A., Zuradaa, J. M., Lob, J. Y., Bakerb, J. A., & Tourassib, G. D. (2008). Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Advances in Neural Networks Research: IJCNN '07*, In *2007 International Joint Conference on Neural Networks IJCNN '07*, 21, pp. 427–436.
- McDonough, J., Kumatani, K., Gehrig, T., Stoimenov, E., Mayer, U., Schacht, S., Wölfel, M., & Klakow, D. (2007). *Machine learning for multimodal interaction. chapter To Separate Speech, A System For Recognizing Simultaneous Speech*. (pp. 283 – 294). Springer-Verlag Berlin Heidelberg volume 4892/2008.
- Ropper, A.H., & Samuels, M.A. (2009). *Adams and victor's principles of neurology. chapter 23. Disorders of Speech and Language*. <<http://www.accessmedicine.com/content.aspx?aid=3633872>>: The McGraw-Hill Companies, Inc. (9th ed.).
- Silva, L. A. D., & Joaquim, M. B. (2008). Noise reduction in biomedical speech signal processing based on time and frequency Kalman filtering combined with spectral subtraction. *Computers and Electrical Engineering*, 34, 154–164.
- Skowronski, M. D., & Harris, J. G. (2006). Applied principles of clear and lombard speech for automated intelligibility enhancement in noisy environments. *Speech Communication*, 48, 549–558.
- Sroka, J. J., & Braid, L. D. (2005). Human and machine consonant recognition. *Speech Communication*, 45, 401–423.
- Yan, Q., Vaseghi, S., Zavarehei, E., Milner, B., Darch, J., White, P., et al. (2008). Kalman tracking of linear predictor and harmonic noise models for noisy speech enhancement. *Computer Speech and Language*, 22, 69–83.