The Past, Present, and Future of Neural Networks for Signal Processing

Anniversary Guest Editor’s Message

In this issue, the Signal Processing Society’s Neural Networks for Signal Processing Technical Committee (NNSP-TC) provides the third article in our 50th Anniversary series. Compared to the Audio and Electroacoustics Technical Committee that contributed the last article in this series (published in the September 1997 issue), the NNSP-TC is relatively young (founded in 1990). Since its inception, its popularity has been evidenced by the increasing number of submissions to ICASSP in the neural-networks area, the well-attended annual Neural Networks Signal Processing workshops, and a number of special journal issues organized by the Technical Committee. Furthermore, the Technical Committee is special in that its impact has gone beyond the SPS’s boundary, because neural networks really form a core technology of many other societies of IEEE. The IEEE Neural Networks Council represents the technology’s interests across a dozen IEEE societies, with strong participation from members of the Neural Networks for Signal Processing Technical Committee in Council conferences and publications.

In this article, members of the NNSP-TC review, with great insight, the fundamentals of neural networks and report recent progress. Topics covered include dynamic modeling, model-based neural networks, statistical learning, eigen-structure-based processing, active learning, and generalization capability. Current and potential applications of neural networks are also described in detail. Those applications include optical character recognition, speech recognition and synthesis, automobile and aircraft control, image analysis and neural vision, and several medical applications.

Essentially, neural networks have become a very effective tool in signal processing, particularly in various recognition tasks. Many of you may recall that, during the infancy of the development of neural-networks technology, one thing that excited people’s interest was its analogy to biological systems. Even though there is still a lot to be understood about the learning processes of human neural systems, artificial neural networks have, without a doubt, provided solutions to many problems in the signal-processing area. As you read this article, you will understand how that happened.

Enjoy!

Tsuhan Chen
Guest Editor
The main attributes of neural processing are its nonlinear and adaptive learning capability, which enables machines to recognize possible variations of a same object or pattern and/or to identify unknown functions and mappings based on a finite set of training data, which can be noisy with missing information. Based on this training by example property with strong support of statistical and optimization theories, neural networks are becoming one of the most powerful and appealing nonlinear and adaptive data analysis tools for a variety of signal processing applications. The Neural Networks for Signal Processing Technical Committee (NNSP TC) was established in 1990 to accommodate an exponentially increasing number of paper submissions to the *Transactions on Signal Processing* (T-SP) and the International Conferences on Acoustics, Speech and Signal Processing (ICASSP) in the IEEE Signal Processing Society (SPS).

In the past seven years, in addition to helping reviewing and organize sessions related to the NNSP for ICASSP, the NNSP TC also consistently organized the well-received annual international workshops—the IEEE Workshops on Neural Networks for Signal Processing (NNSP). These workshops (NNSP'91 in Princeton, New Jersey, USA; NNSP'92 in Helsingor, Denmark; NNSP'93 in Linthicum Heights, Maryland, USA; NNSP'94 in Ermiion, Greece; NNSP'95 in Cambridge, Massachusetts, USA; NNSP'96 in Kyoto, Japan; NNSP'97 in Amelia Island Plantation, Florida, USA) are designed to serve as a regular forum for researchers from universities and industry who are interested in interdisciplinary research on neural networks for signal-processing applications. In the present scope, the workshop encompasses up-to-date results in several key areas, including learning theory, neural models, speech processing, signal processing, image processing, communications, pattern recognition, and system implementation. The associated Conference Proceedings is crafted to be an archival reference in the rapidly growing field of NNSP. For more information about the past and future workshops, please visit the official Web page of the NNSP TC at http://pierce.ee.washington.edu/~nns.

Thanks to the significant effort of Professor Yu-Hsin Hu, the Chief-Guest-Editor, along with five other TC members, the NNSP TC also successfully organized a special issue on NNSP for the T-SP. The special issue was approved during ICASSP 1996 in Atlanta. An unexpectedly large number of manuscripts were received (101) covering all scopes of NNSP from all over the world. There were 28 papers chosen to appear in the NNSP special issue, which was published in November 1997.

Our TC members have also been appointed by the SPS President to serve as AdCom members in the IEEE Neural Network Council (NCN). Several TC members have helped organize the International Conference on Neural Networks (ICNN), the annual NNC conference, and special issues for the *IEEE Transactions on Neural Networks*. Some other members also participate in organizing activities for the International Neural Network Society (INNS).

## NNSP Now and in the Future

### Present State-of-the-Art in NNSP

Engineering in general and signal processing in particular are still exploring linear models such as Gaussianity assumptions and stationarities, although the world is nonlinear, non-Gaussian, and nonstationary. Neural networks are nonlinear adaptive systems that have the potential to push the technology barrier beyond conventional approaches. Growing numbers of neural-network solutions to real-world signal-processing problems have been reported. Some recent examples in engineering include:

- **Optical-Character Recognition**: The LeNet5 developed by AT&T [1] got the best results in the NIST OCR Special Data Bases 1 and 3 (binary images of handwritten digits), achieving an accuracy of 99.1%. LeNet5 is included in an NCR product to automatically read bank checks. These neural network topologies are specialized multilayer perceptrons (MLPs) with convolution layers

---

### List of Current NNSP TC Members

<table>
<thead>
<tr>
<th>Member Name</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tulay Adali, Univ. of Maryland</td>
<td>Baltimore County</td>
</tr>
<tr>
<td>Charles Bachmann, Naval Research Laboratory</td>
<td></td>
</tr>
<tr>
<td>Andrew Back, Laboratory for Artificial Brain Systems</td>
<td>RIKEN</td>
</tr>
<tr>
<td>Andrzej Cichocki, Laboratory for Artificial Brain Systems</td>
<td>RIIKN</td>
</tr>
<tr>
<td>Anthony G. Constantinides, Imperial College</td>
<td></td>
</tr>
<tr>
<td>Bert de Vries, David Sarnoff Research Center</td>
<td></td>
</tr>
<tr>
<td>Kevin R. Farrell, T-NetIX, Inc.</td>
<td></td>
</tr>
<tr>
<td>Horacio Franco, SRI International</td>
<td></td>
</tr>
<tr>
<td>Hsin Chia Fu, National Chiao-Tung University</td>
<td></td>
</tr>
<tr>
<td>Ling Guan, University of Sydney</td>
<td></td>
</tr>
<tr>
<td>Lars Kai Hansen, Technical University of Denmark</td>
<td></td>
</tr>
<tr>
<td>Jenq-Neng Hwang, University of Washington</td>
<td>TC Chair</td>
</tr>
<tr>
<td>Shigiru Katagiri, ATR (TC Secretary)</td>
<td></td>
</tr>
<tr>
<td>Jan Larsen, Technical University of Denmark</td>
<td></td>
</tr>
<tr>
<td>Fa-Long Luo, Universitat Erlangen-Nuernberg</td>
<td></td>
</tr>
<tr>
<td>Elias S. Manolakos, Northeastern University</td>
<td></td>
</tr>
<tr>
<td>Nelson H. Morgan, International Computer Science Institute</td>
<td></td>
</tr>
<tr>
<td>Mahesan Niranjan, Cambridge University</td>
<td></td>
</tr>
<tr>
<td>Ananth Sankar, SRI International</td>
<td></td>
</tr>
<tr>
<td>Volker Tresp, Siemens AG, Central Research</td>
<td></td>
</tr>
<tr>
<td>Marc M. Van Hulle, Laboratorium voor Neurofysiologie of Belgium</td>
<td></td>
</tr>
<tr>
<td>Elizabeth J. Wilson, Raytheon Company</td>
<td></td>
</tr>
</tbody>
</table>
and shared weights. A neural-network-based character classifier is also used to provide robust recognition of hand-printed English text in new models of Apple Computer's Newton MessagePad [2].

**Automobile Control:** Engine idle and misfiring controls are being developed by Ford Corporation [3] using a recurrent neural network trained with multistreaming, which is an adaptation of the Decoupled Extended Kalman Filter training. Ford researchers believe that recurrent networks are the most promising technology to help meet the new stringent emission levels of the Clean Air Act.

**Self-Organizing Feature Maps:** The self-organizing map (SOM) converts complex, nonlinear statistical relationships between high-dimensional data into simple geometric relationships on a low-dimensional display. It thereby compresses information while preserving the most important topological and metric relationships of the primary data elements. The SOMs have been applied to hundreds of practical systems [4].

**Aircraft Control:** Neural controllers for aircraft are being developed by several groups [5] both extending optimal control ideas to nonlinear systems as well as new ideas based on dynamic programming (adaptive critics). The LoFLYTE hypersonic waverider is a joint project of NASA and the US Air Force to design experimental aircraft reaching Mach 5 speeds. Neural networks (adaptive critics) are being developed for the aircraft control by Accurate Automation Inc.

**Chaotic Dynamic Reconstruction:** Chaotic dynamic reconstruction has been applied to model complex natural phenomena generated by deterministic nonlinear phenomena such as sea clutter [6]. The universal approximation properties of neural networks (either MLPs or radial-basis functions) coupled with regularization have been shown to identify the nonlinear system that produced the time series [7]. The section titled "Neural Networks for Dynamic Modeling," written by Jose C. Principe, offers a clear perspective of creating a nonlinear model for the time series using only the available samples of the time series.

**Cocktail-Party Problem:** Separation of blindly mixed signals (such as voices in a party—the cocktail-party effect), where neither knowledge of the signals nor of the mixing process is available, is an important problem in inverse modeling and communications (channel equalization, teleconferencing, and hands-free telephony). It has been shown that the unique identification of the mixing-system parameters requires knowledge beyond second-order statistics. Neural networks have been shown to be naturally applicable to this problem due to their nonlinear activation functions and learning-by-example capabilities. Bell and Sejnowski [8], among many others, exemplified such a solution recently, and much of the work in the area is being propelled by the use of nonlinear systems.

**Natural Speech Synthesizer:** A text-to-speech synthesizer including five cooperating neural networks, each specializing in a particular area of human natural language ability, represents a significant advance in naturalness and adaptability to different voice qualities over previous rule-based and concatenative approaches [9].

**Neural Vision System:** By integrating a model-based network as a low-level vision subsystem and a hierarchically structured neural array for higher-level analysis and recognition tasks, the neural vision system has been used for a number of complex real-world image-processing applications such as computer-aided diagnosis of breast cancers in digital mammograms; biomedical vision to analyze neurological disorders; and identifying and classifying underwater mines through 3-D sonar image processing and visualization. For more details, please refer to the section titled "Model-Based Neural Networks for Image Processing," contributed by Ling Guan and Sun-Yuan Kung.

**3-D Heart Contour Delineation:** Neural networks have also been used in representing 3-D objects, e.g., a neural-network architecture is used to represent 3-D endocardial (inner) and epicardial (outer) heart contours and quantitatively estimate the motion of left ventricles of human hearts from ultrasound images acquired using ultrasonic echocardiography. The absolute error measured compares favorably with the human interobserver variability reported for analyzing distances. With this unique representation, it is also possible to systematically and effectively measure the amount of 3-D heart motion [10].

The great advantage of MLPs for classification is that they carve the input space through the use of saturating nonlinearities in the (neural) processing elements (PEs) using directly the information contained in the data and through a well-established learning algorithm (back-propagation) [11]. Multilayer perceptrons are universal approximators [12], and an important theorem by Barron [13] shows that MLPs are very efficient for function approximation in high-dimensional spaces. Unlike polynomial approximators, where the rate of the convergence of the error decreases with the dimension of the input space, the rate of convergence of MLPs is independent of the input space dimensionality. This may explain the reason why MLPs perform better than other statistical procedures for large-dimensional problems.

Back-propagation is in itself a refinement of gradient-descent learning since it enables the computation of the sensitivities in layered networks with a computational complexity of $O(N^2)$, where $N$ is the number of processing elements (neurons). This savings is due to the clever use of the topology to help compute the sensitivities. The importance of the back-propagation algorithm in adaptive systems is comparable to that of the fast Fourier transform (FFT) algorithm in digital signal processing.

Scientists and engineers have already conquered reasonably well the problems and applications of static nonlinear mappings (e.g., system identification and pattern recognition). Neural networks were very important in this arena since they have injected new blood in applied
statistics. The section written by Jenq-Neng Hwang titled “A Unified Perspective of Statistical Learning Networks,” brings close the missing links between the model-free statistical regression/classification with the nonparametric neural-network learning.

Besides the least-squares methods for linear systems, the eigen-decomposition (ED) and singular-value decomposition (SVD) methods have been the most powerful tools in most fields of signal processing. Despite many proposals of recursive and adaptive algorithms to perform ED and SVD, a good compromise involving performance, implementation, and computational complexity has not been reached yet. As an alternative, neural-network approaches have recently attracted much attention as reviewed in the section titled “Neural Networks for Eigen-Structure Based Signal Processing,” written by Fa-Long Luo and Rolf Unbehauen.

The other big challenge is how to go about extracting more information from the data beyond second-order statistics by incorporating more domain knowledge, instead of purely data-driven training. As described in the section titled “From Pattern Classification to Active Learning,” written by Yu-Hen Hu, one such technique is known as importance sampling. Based on one’s knowledge of the underlying system, the importance-sampling method samples the data from a twisted distribution so that the variance of the parameter estimate can be made significantly smaller than that obtained using random sampling. The trick here is to find this twisted distribution. In machine-learning literature, another variance-reduction technique, called active learning, has received much attention recently. In active learning, the learning algorithm dictates which point to sample. The active learning does so by selecting the most informative sample where the most knowledge can be gained.

Doing well on unseen data may at first seem unattainable, but the ability to generalize in very complex environments is nevertheless one of the most striking properties of neural systems, and indeed one of the reasons that neural networks have been shown useful in practical time-series applications. However, the training with MSE and the MLP topology do not control directly the generalization ability, so the designer has to deal directly with this issue. A rigorous study of the “generalization” capability in neural networks for time-series prediction is given in the section titled “Generalization: The Hidden Agenda of Learning,” written by Jan Larsen and Lars Hansen.

Future Challenges of NNSP

Time processing is probably the greatest challenge with neural networks, although others exist, too. There has not been a genuine attempt to solve time-varying problems with nonlinear structures. Time is considered an extra dimension for the representation of information instead of being utilized directly for information processing.

In most signal-processing problems, data arrives sequentially and the underlying systems generating the data tend to be nonstationary. For these problems, one needs sequential (adaptive) training algorithms (such as the extended Kalman filter) as well as clever use of prior knowledge for model selection. The section titled “Step Size Selection for On-Line Training of Adaptive Systems,” written by Scott C. Douglas and Andrzej Cichocki, reviews learning techniques (especially the learning step-size parameters that control the degree of averaging) used in sequential training of adaptive systems, which continually adapts its internal states as new signal measurements become available.

Multilayer perceptrons are purely static and are incapable of processing time information. One way to extend MLPs to time processing is by creating a time window over the data to serve as memory of the past as first done in the time-delay neural network (TDNN) [14]. Alternatively, dynamic neural networks bring the memory inside the neural-network topology. There are basically two ways to create a dynamic neural network: either the PEs are extended with local memory structures or the network topology becomes recurrent. The first type is simpler to study than the second with respect to stability. Moreover, the memory structures are generally linear, which means that the concepts of filtering and linear modeling can be brought to bear in their study [15]. The gamma model [16] is a good example of the use of linear filtering concepts to study dynamic networks and shows that dynamic networks are predefined neural networks for the processing of time information. Recently, Sandberg [17] proved that MLPs extended with short-term memories are universal approximators for a large class of functional mappings. Sontag showed that, unlike linear systems, a nonlinear system provides a unique model for system identification [18]. These works establish the foundation to use dynamic neural networks for nonlinear system identification.

Recurrent connections across the topology provide very powerful mappings but bring two problems: networks are not guaranteed to be stable, and they cannot be trained with standard back-propagation. Clever but partial recurrent topologies have been proposed to avoid these problems. A simple extension to the existing feed-forward structure to deal with temporal sequence data is the partially recurrent network, also called simple recurrent networks (SRNs). An SRN has the connections that are mainly feed-forward but include a carefully chosen set of feedback. In most cases the feedback connections are fixed and not trainable. Due to recurrency, it remembers cues from the past and does not appreciably complicate the training procedure. The most widely used SRNs are Elman’s network [19] and Jordan’s network [20]. In general, researchers have to face some approximation deficiencies in using SRNs. Narendra extended the ARMA models with nonlinearities [21] and recently the same group provided observability and controllability conditions for the recurrent networks around equilibrium.
points. Model stability has to be studied using Lyapunov stability, but progress has been slow in this area with some results provided by Sonntag [22].

In terms of learning algorithms there are two basic approaches to train dynamic networks: real-time recurrent learning (RTRL) [23] and back-propagation through time (BPTT) [24]. They produce equivalent gradient updates (provided the initial conditions are identical), but they chain the computations very differently and result in different properties. RTRL is global in the topology and local in time, while BPTT is local in the topology and global in time (anticipatory). So if on-line learning is required, RTRL should be utilized (BPTT is computationally more efficient but has storage requirements that are dependent upon the trajectory length). The recent development of decoupled extended Kalman filter (DEKF) [25] learning provides a powerful alternative to the two above-mentioned procedures.

A hidden Markov model (HMM) is a doubly stochastic process with an underlying stochastic process that is not observable (i.e., hidden), but can only be observed through another set of stochastic processes that produce the sequence of observed symbols [26]. The trellis diagram realization of an HMM can be considered as a BPTT network expanding in time since its connections (transition probabilities) are carrying the information about the environment, and it consists of a multilayer network of simple units activated by the weighted sum of the unit activations at the previous iteration. In addition, the learning technique used in HMMs has a close algorithmic analogy with that used in the BPTT networks [27].

Neural networks have not surpassed the power of HMMs for speech recognition, while it is accepted by the practitioners that HMMs do not map well to the speech-recognition problem. Ingenuity is needed to propose new ways of looking at time processing, and biology may be again a motivating factor since humans excel in the extraction of information in time signals. The section titled “Applications of Neural Networks to Speech Recognition” contributed by Nelson Morgan and Horacio Franco, specifically discusses the issues of integrating the NN techniques into HMMs. More specifically, in these speech-recognition systems, neural networks trained to classify the state types can be shown (under some simple assumptions) to estimate the posterior probability of state types given the acoustic observations. These posterior probabilities can be converted to observation probabilities (using Bayes’ Rule) and used within the classical HMM framework. Networks used in this way have been shown to be comparable in performance to larger and more complex probabilistic estimators for large-vocabulary, continuous-speech recognition.

Neural networks have recently been shown to outperform linear models in the task of time-series prediction [28], which opens up a lot of applications in engineering as well as in the financial and service industries. Dynamic neural networks have also been recently applied to the closely related problem of chaotic dynamic reconstruction also known as dynamic modeling. Dynamic modeling attempts to identify the nonlinear system that generates the observable time series. This method is very important in modeling natural phenomena, developing synthetic models for the data, and producing improved engineering systems.

Medical diagnostics offers a number of challenging problems in the area of neural networks for signal processing. Many technical challenges such as the principle methods of dealing with missing data, fusion of information from different sources of information, dealing with the large variability in the way experts treat information, and thorough validation of neural-network solutions are likely to be addressed in the near future. Mahesan Niranjan provides the section titled “Examples in Medical Applications,” which describes two studies and discusses lessons to be learned from them.

Multimedia technologies represent a new opportunity for research interactions among a variety of media such as speech, audio, image, video, text, and graphics. The technologies will profoundly change the way we access information, conduct business, communicate, educate, learn, and entertain [29, 30]. Future multimedia technologies will need to handle information with an increasing level of intelligence, i.e., automatic recognition and interpretation of multimodal signals. The key attributes of neural processing essential to intelligent multimedia processing were already discussed in a section in [30] earlier this year. In that article, it was argued that adaptive neural-network technology offers a promising and unified solution to a broad spectrum of multimedia applications. The reason why neural networks should be perceived as a core technology for intelligent multimedia processing hinges upon their adaptive learning capability. Plenty of evidence exists to show why neural networks serve as a core technology for several vital multimedia functionalities, including (1) efficient representations for audio/visual information; (2) detection and classification techniques; (3) fusion of multimodal signals; and (4) multimodal conversion and synchronization.

A strong early motivation for interest in neural networks is our interest in understanding computations in the biological systems. As evidenced by this article, interesting applications resulting from statistical signal processing, exploiting the nonlinear function-approximation capabilities of neural networks, seems to have dominated the area, particularly in the NNSP community. It is likely in the future that there will be greater movement toward understanding learning in biological systems.

References


Neural Networks for Dynamic Modeling
José C. Principe, University of Florida
One of the goals in science is to develop models that explain the data. The linear model has been extensively utilized for this purpose: first in regression for static data and later in optimum filtering for time series. While a lot has been said about artificial neural networks for classification, much less attention has been devoted to nonlinear regression and its time counterpart, which we call here dynamic modeling. The goal of dynamic modeling is to create a nonlinear model for the incoming time series, using only the available samples of the time series. The problem formulation is reminiscent of the use of prediction to obtain the best linear model from a time series, and in fact it is an extension of the prediction formalism for nonlinear systems.

The first step in dynamic modeling is to create an embedding of the time series using the time delay embedding theorem [1] or one of its extensions [2]. The purpose of the embedding is to create a space, called the reconstruction space, where the dynamics of the autonomous nonlinear system that created the time series can be reconstructed. Theorem specifies conditions that guarantee that there is a diffeomorphism between the reconstructed trajectory and the original dynamical system. The practical implication is that we can use the reconstructed space to estimate global dynamical invariants about the original system (such as correlation dimension and Lyapunov exponents). Here we will use the reconstructed trajectories to create a nonlinear autoregressive (NAR) model for the dynamics, since it has been shown that the prediction of the next point in the reconstruction space identifies the predicted part of the model that defines the dynamical system [3]. These are the theoretical foundations for dynamical modeling. Next we will address the details for practically achieving it.

Lapedes and Farber [4] were the first to show that a multilayer perceptron (MLP) with time-delayed inputs was able to predict a chaotic time series (synthetically generated). Therefore, in principle, an MLP can be used...