A New Approach to Music Information Retrieval using Dynamic Neuronal Networks

L.E. Gomez[1], J.H. Sossa[1], R. Barron[1], J.F. Jimenez[1],

[1] Centro de Investigación en Computación-IPN, Unidad Profesional Adolfo-López Mateos, Av. Juan de Dios Bátiz s/n and M. Othón de Mendizábal, Zacatenco, México, DF. 07738, Mexico
sgomezb08@sagitario.cic.ipn.mx, hsossa@cic.ipn.mx, rbarron@cic.ipn.mx, jfvielma@cio.mx

Abstract. The majority of work in music information retrieval (MIR) has been focused on symbolic representations of music. However, most of the digitally available music is in the form of raw audio signals. Although various attempts at monophonic and polyphonic transcription have been made, none has been successful and general enough to work with real world signals. So far, many researchers have been done to develop efficient music retrieval systems. In this paper, we develop a novel music retrieval system based on dynamic neural networks, which are trained with the signal melody, and not with traditional descriptors.

Keywords: Music Information Retrieval; Dynamic Neuronal Networks; musical descriptors.

1 Introduction

With the explosive expansion of digital music and audio contents, efficient retrieval of such data is getting more and more attention, especially in large-scale multimedia database applications. In the past, music information retrieval was based on textual metadata such as title, composer, singer or lyric. However, these various metadata-based schemes for music retrieval have suffered from many problems including extensive human labor, incomplete knowledge and personal bias.

Compared with traditional keyword-based music retrieval, content-based music retrieval provides more flexibility and expressiveness. Content-based music retrieval is usually based on a set of extracted music features such as pitch, duration, and rhythm.

In some works, such as [1][2], only pitch contour is used to represent melody. Music melody is transformed to a stream of U, D, R, which stands for a note is higher than, lower than, or equal to the previous note, respectively. But it simplifies the melody so much that it cannot discriminate music very well, especially when the music database is large.

In order to represent the melody more accurately and discriminatively, new feature sets have been proposed. In [3], pitch interval and rhythm are considered as well as

Advances in Computer Science and Engineering.
Research Computing Science 45, 2010, pp. 41-51

Received 30/04/10
Final version 19/05/10
Accepted 14/05/10

pitch contour. In [4], relative interval slope is used in music information retrieval. And [5] introduces four basic segment types (A,B,C,D) to model music contour.

When rhythm and pitch interval is considered, more complex similarity measure and matching algorithm should be used. [5] uses two-dimensional augmented suffix tree to search the desired song, rather than approximate string matching algorithm used in [1][2]. In [6], a new distance metrics between query and songs is proposed. But its computation is very time-consuming because it need adjust many parameters step by step to find the minimum distance.

Neural networks are characterized by dynamic dependence of events in past moments. Within the neural networks are dynamic networks are inherently dynamic, such as networks Hopfield, Jordan and Elman [1]. On the other hand, there are networks multilayer, which are static in nature but, achieve a dynamic behavior reinforced their own inputs samples of their previous outings.

In this paper, we propose a novel music retrieval system based on the use of dynamic neural networks, training with these melodies and using their synaptic weights as descriptors for the recovery of the melody.

The rest of this paper is organized as follows. In Section 2, we present an overview of ongoing research for analyzing music features and constructing MIR systems. In Sections 3, we describe our music retrieval system using dynamic neural networks. In Section 4, we report on some of the experimental results. Section 5 concludes this paper and describes our future directions.

2 Related work

In this section, we review some of typical techniques and systems for music information retrieval. As we know, music can be represented in two different ways. One is based on musical scores such as MIDI and Humdrum [7]. The other is based on acoustic signals which are sampled at a certain frequency and compressed to save space. Wave (.wav) and MPEG Layer-3 (.mp3) are examples of this representation.

2.1 Symbolic analysis

Many research efforts to solve the music similarity problem have used symbolic representation such as MIDI, musical scores, note lists and so on. Based on this, pitch tracking finds a ‘‘melody contour’’ for a piece of music. Next, a string matching technique can be used to compare the transcriptions of songs [1],[8],[9],[10],[11].

String matching has been widely used in music retrieval because melodies are represented using a string sequence of notes. To consider human input errors, dynamic programming can be applied to the string matching; however, this method tends to be rather slow. An inexact model matching approach [12] was proposed based on a quantified inexact signature-matching theory to find an approximate model to users’ query requirements. It can enhance the reusability of a model repository and make it possible to use and manage a model repository conveniently and flexibly. Zhuge tried to apply this theory to a problem-oriented model repository system PROMBS [13].

There are also researches for symbolic MIR based on the ideas from traditional text IR. Using traditional IR techniques such as probabilistic modeling is described in [14] and using approximate string matching in [15]. Some work addressed other IR issues such as ranking and relevance. Hoashi [16] used relevance feedback for music retrieval based on the tree-structured vector quantization method (TreeQ) developed by Foote. The TreeQ method trains a vector quantizer instead of modeling the sound data directly.

2.2 Acoustic signal analysis

There are many techniques to extract pitch contour, pitch interval, and duration from a voice humming query. In general, methods for detecting pitches can be divided roughly into two categories: time-domain based and frequency-domain based.

In the time-domain, ZCR (zero crossing rate) and ACF (auto correlation function) are two popular methods. The basic idea is that ZCR gives information about the spectral content waveform cross zero per unit time [17]. In recent works, ZCR appeared in a different form such as VZCR (variance of ZCR) or SZCR (smoothing ZCR) [18]. On the contrary, ACF is based on the cross correlation function. While a cross correlation function measures the similarity between two waveforms along the time interval, ACF can compare one waveform with itself.

In the frequency-domain, FFT (fast Fourier transformation) is one of the most popular methods. This method is based on the property that every waveform can be divided into simple sine waves. But, a low spectrum rate for longer window may increase the frequency resolution while decreasing the time resolution. Another problem is that the frequency bins of the standard FFT are linearly spaced, while musical pitches are better mapped on a logarithmic scale. So, Forberg [19] used an alternative frequency transformation such as constant Q transform spectrums which are computed from tracked parts.

In recent works for the automatic transcription, they used probabilistic machine learning techniques such as HMM (hidden Markov model) and NN (neural network) to identify salient audio features and reduce the dimensionality of feature space. Ryynanen and Klapuri [20] proposed a singing transcription system based on the

HMM-based notes event modeling. The system performed note segmentation and labeling and also applied multiple-F0 estimation method [21] for calculating the fundamental frequency.

2.3 Recent MIR systems

For decades, many researchers have developed content based MIR (Music Information Retrieval) systems based on both acoustic and symbolic representations [1],[8],[22],[11].

Ghias [1] developed a QBH system that is capable of processing acoustic input in order to extract appropriate query information. However, this system used only three types of contour information to represent melodies. The MELDEX system [8] was designed to retrieve melodies from a database using a microphone. It first transformed

acoustic query melodies into music notations, and then searched the database for tunes containing the hummed (or similar) pattern. This web-based system provided several match modes including approximate matching for interval, contour, and rhythm.

MelodyHound [22], originally known as the ‘‘TuneServer’’, also used only three types of contour information to represent melodies. They recognized the tune based on error-resistant encoding. Also, they used the direction of the melody only, ignoring the interval size or rhythm. The C-BRAHMS [23] project developed nine different algorithms known as P1, P2, P3, MonoPoly, IntervalMatching, PolyCheck, Splitting, ShiftOrAnd, and LCTS for dealing with polyphonic music.

Suzuki [24] proposed a MIR system that uses both lyrics and melody information in the singing voice. They used a finite state automaton (FSA) as a lyric recognizer to check the grammar and developed an algorithm for verifying a hypothesis output by a lyric recognizer. Melody information is extracted from an input song using several pieces information of hypothesis such as song names, recognized text, recognition score, and time alignment information.

Many other researchers have studied quality of service (QoS)-guaranteed multimedia systems over unpredictable delay networks by monitoring network conditions such as available bandwidth. McCann [25] developed an audio delivery system called Kendra that used adaptability with a distributed caching mechanism to improve data availability and delivery performance over the Internet. Huang [26] presented the PARK approach for multimedia presentations over a best-effort network in order to achieve reliable transmission of continuous media such as audio or video.

3 Dynamic Neuronal Network applied to MIR

Dynamic neural network is the extension of static neural network via the consideration of time. The proposed dynamic models are developed based on static MLFN. In general, dynamics can be expressed by using a tapped-delay line, external dynamics and internal dynamics [27]. Tapped-delay line approach uses a sequence of delay to express dynamics and forms time-delay neural network [28],[29]. External dynamics approach uses the historical information of output itself to show dynamics and forms autoregressive type neural network [30],[31].

3.1 Multi-layer Feed-forward Neural Network (MLFN)

The most common network structure the MLFN which is a parallel distributed processing network. Parallel distributed processing network is formed by many basic units called neurons. Information processing takes place through the interactions of a large number of neurons can solve difficult tasks.

This is the idea of learning tasks via different angles via neurons and the interactions between neurons. It is a good option for model selection. Figure 1 shows the scheme network of MLFN.

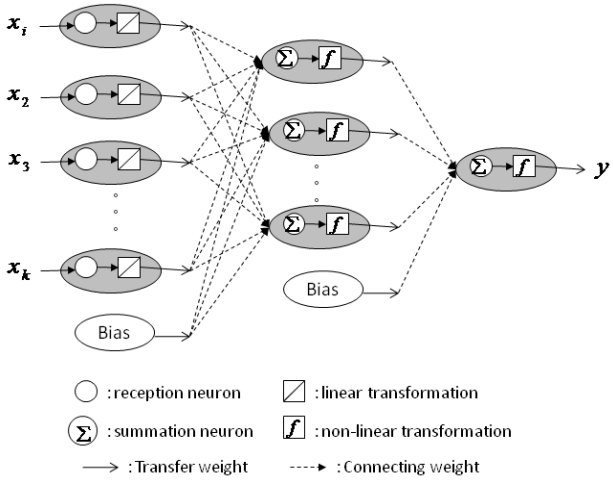

Fig. 1. MLFN Network Structure.

3.2 Time Delay Neural Network

Time delay neural network (TDNN) comes under dynamic neural networks, which are designed to explicitly include time relationships in the input-output mappings. Time-lagged feedforward networks (TLFNs) are a special type of dynamic networks that integrate linear filter structures inside a feedforward neural network to extend the non-linear mapping capabilities of the network with a representation of time [32]. Thus, in TLFN the time representation is brought inside the learning machine. The advantage of this technique is that the learning machine can use filtering information while the disadvantage is that the learning becomes complex since the time information is also coded in. TDNN is one of the specific cases of TLFN where a tapped delay line is given in the input followed by a multilayer perceptron (MLP) as shown in the block diagram in Fig. 2. Current input (at time t) and D delayed inputs (at time $t-1, t-2, ..., t-D$) can be seen by the TDNN. The TDNN can be trained by using gradient descent back propagation. The ordered training patterns must be provided during training process [33].

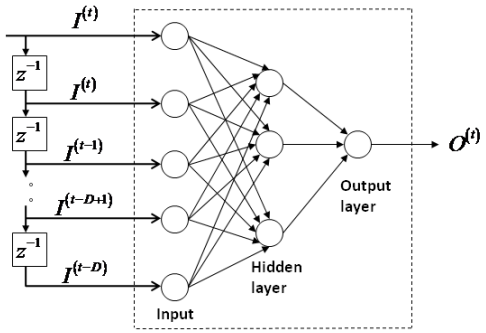

Fig. 2. Structure of Time Delay Neural Network.

3.3 Proposed method

It used WAV files, each file is trained in a dynamic neural network (TDNN), it shows a diagram in Figure 3. At the end of the training is obtained the weight matrix (WNN), as this is used as a descriptor of melody trained.

This method is novel because it works on the time domain, not you the frequency domain which gives a digital signature, such as: 1) Music features: pitch, duration, and rhythm. 2) Traditional descriptors: pitch contour, zero crossing rate, cross correlation, FFT, and others.

It is not necessary obtain the digital signature or features of the melody, because it is used in full. This reduces the level of a-priori knowledge of the melody by the user.

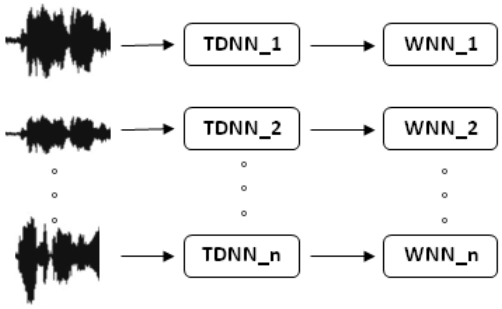

Fig. 3. Structure training of the TDNN with melodies.

The recovery of melodies is performed query with a segment of a melody, this segment is processed in the TDNN and stepdaughter and the descriptors of the melodies, get the error recovery the melody, and finally with the argument minimum, you get the index gives melody that was query, it shows a diagram in Figure 4.

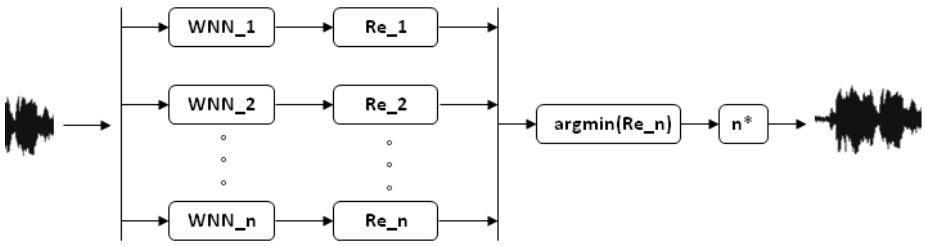

Fig. 4. Structure retrieval of melodies using descriptors of the TDNN.

4 Experimental results

It used 16-bit WAV files (stereo mode), each file is trained in a dynamic neural network, and these networks have a maximum of 100 iterations, and 10 neurons in the hidden layer. At the end of the training is obtained the weight matrix, as this is used as a descriptor of melody trained.

Tests were with different numbers of neurons in the hidden layer, as well as different numbers of iterations, the results of this test are shown in Tables 1,2 and Figures 5,6,7,8. The compares the errors rate training and recovery.

Table 1. Table of error rate training and recovery, with different numbers of neurons

	Training			**Recovery**		
N. neurons	Minimum	Average	Maximum	Minimum	Average	Maximum
5	2.99E-04	2.48E-03	6.12E-03	8.45E-03	**1.41E-02**	**2.21E-02**
6	**2.93E-04**	**2.48E-03**	**5.61E-03**	**5.19E-03**	1.88E-02	5.05E-02
7	5.59E-04	3.67E-03	6.46E-03	6.09E-03	1.95E-02	4.83E-02
8	2.94E-04	3.52E-03	6.43E-03	1.01E-02	1.69E-02	3.00E-02

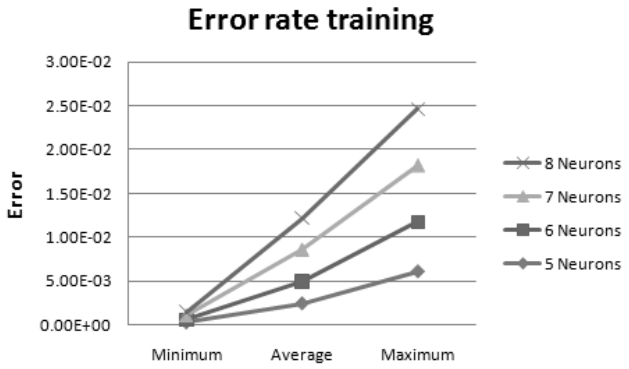

Fig. 5. Graphic of error rate training, with different number of neurons.

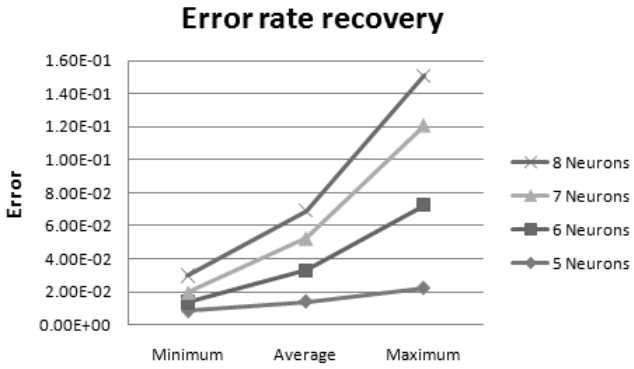

Fig. 6. Graphic of error rate recovery, with different number of neurons.

Table 2. Table of error rate training and recovery, with different numbers of iterations

	Training			**Recovery**		
N. iterations	Minimum	Average	Maximum	Minimum	Average	Maximum
10	2.07E-03	9.67E-03	2.64E-02	9.96E-03	3.00E-02	6.37E-02
25	1.15E-03	5.65E-03	1.09E-02	**5.49E-03**	2.14E-02	5.68E-02
50	**2.99E-04**	**2.48E-03**	**6.12E-03**	8.45E-03	**1.41E-02**	**2.21E-02**
75	3.55E-04	5.29E-03	1.05E-02	7.94E-03	2.11E-02	5.08E-02

Fig. 7. Graphic of error rate training, with different number of iterations.

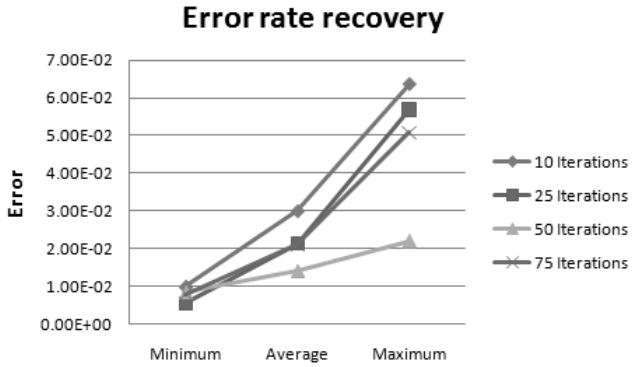

Fig. 8. Graphic of error rate recovery, with different number of iterations.

The system works effectively and efficiently, with a query with a segment less than 1 percent the total melody, the melody is recovered.

For example, if you train the TDNN with a melody the 132.221 frames, and recovers with only 350 frames, which is 0.264 percent of the total of the melody, with these we can conclude that is recovering with less than 1 percent of the melody.

4 Conclusions

Content-based music retrieval is a very promising method for large music library, yet it is a very challenging task. We discussed various features of music contents for content-based retrieval.

In this paper, we have presented a preliminary approach to Music Information Retrieval. The goal of this study was to explore a new line research within the field of MIR. Not all people are experts in models auditory perception, so we have chosen a TDNN network type that is capable of solving the problem from the samples without using any traditional descriptor or digital signature. Neither has made any preprocessing before the music files, these apply changes a melody can distort or

highlight the information contained therein. Therefore, unlike other techniques, MIR, we have original melody introduced directly into the net try find this as a suitable height relations present in the spectrum signal.

The inputs to TDNN network as a series of amplitudes obtained from stereo melody. The output of network is the encoding of a musical descriptor in a matrix of weights.

The system works effectively and efficiently, as a query with a segment less than 1 percent the total melody, the melody is recovered.

It can be concluded that the system retrieval using dynamic neural network is a success, achieving very faithful to the melody identification, in any case, the results of this study open many lines promising for further research on MIR by dynamic neural networks.

Acknowledgements. We wish to thank the Centro de Investigación en Computación of the I.P.N. by the support to accomplish this project. L.E. Gomez and J.F. Jimenez thanks CONACYT by the scholarship received to complete his doctoral studies. R. Barron thanks the SIP-IPN under grant 20100379 for the support. H. Sossa thanks the SIP-IPN under grant 20091421 for the support. H. Sossa also thanks CINVESTAV-GDL for the support to do a sabbatical stay from December 1, 2009 to May 31, 2010. Authors thank the European Union, the European Commission and CONACYT for the economical support. This paper has been prepared by economical support of the European Commission under grant FONCICYT 93829. The content of this paper is an exclusive responsibility of the CIC-IPN and it cannot be considered that it reflects the position of the European Union. Finally, authors thank the reviewers for their comments for the improvement of this paper.

References

[1] Ghias, A.: Query By Humming-Musical Information Retrieval in an Audio Database. Proc.s of ACM Multimedia 95, pp231-236, 1995.

[2] Blackburn, S., De Roure, D.: A Tool for Content Based Navigation of Music, Proc. ACM Multimedia 98, pp 361-368, 1998.

[3] McNab, R.J.: Towards the Digital Music Library: Tune Retrieval from Acoustic Input. Proc. of Digital Libraries, pp 11-18, 1996.

[4] Lemstrom, K., Laine, P., Perttu, S.: Using Relative Interval Slope in Music Information. Retrieval. In Proc. of International Computer Music Conference 1999 (ICMC '99), pp. 317-320, 1999.

[5] Chen, A.L.P., Chang, M., Chen, J.: Query by Music Segments: An Efficient Approach for Song Retrieval. In Proc. of IEEE International Conference on Multimedia and Expo., 2000.

[6] Francu, C. Nevill-Manning, C.G.: Distance Metrics and Indexing Strategies for a Digital Library of Popular Music. In Proc. of IEEE International Conference on Multimedia and Expo. 2000.

[7] Kornstadt, A.: Themefinder: A web-based melodic search tool. In: Computing in Musicology 11. MIT Press., 1998.

[8] McNab, R.J. et al.:. The New Zealand digital library melody index. Digital Libraries Magazine., 1997.

[9] Uitdenbogerd, A., Zobel, J.: Melodic matching techniques for large music databases. In: Proceedings of ACM Multimedia Conference. pp. 57–66., 1999.
[10] Hwang, E., Rho, S.: FMF(fast melody finder): A web-based music retrieval system. In: Lecture Notes in Computer Science, vol. 2771. Springer-Verlag, pp. 179–192., 2004.
[11] Hwang, E., Rho, S.: FMF: Query adaptive melody retrieval system. Journal of Systems and Software 79 (1), 43–56. 2006.
[12] Zhuge, H.: An inexact model matching approach and its applications. Journal of Systems and Software 67 (3), 201–212. 2003.
[13] Zhuge, H.: A problem-oriented and rule-based component repository. Journal of Systems and Software 50 (3), 201–208. 2000.
[14] Pickens, J.: A comparison of language modeling and probabilistic text information retrieval approaches to monophonic music retrieval. Proceedings of the 1st Annual International Symposium on Music Information Retrieval (ISMIR2000). 2000.
[15] Lemstrom, K., Wiggins, G.A., Meredith, D.: A threelayer approach for music retrieval in large databases. In: Second International Symposium on Music Information Retrieval. Bloomington, IN, USA. pp. 13–14. 2001.
[16] Hoashi, Matsumoto, Inoue.: Personalization of user profiles for content-based music retrieval based on relevance feedback. ACM Multimedia, pp. 110–119. 2003.
[17] Gerhard, David.: Pitch Extraction and Fundamental Frequency: History and Current Techniques. Technical Report TR-CS 2003-06. 2003.
[18] Huang, R., Hansen, J.H.L.: Advanced in unsupervised audio classification and segmentation for the broadcast news and NGSW Corpora. IEEE Trans. on Audio, Speech and Language Processing 14 (3), 907–919. 2006.
[19] Forberg, Johan.: Automatic conversion of sound to the MIDIformat. TMH-QPSR 1-2/1998. 1998.
[20] Ryynanen, Matti., Klapuri, Anssi.: Transcription of the singing melody in polyphonic music, ISMIR 2006. 2006.
[21] Klapuri, Anssi P.: A perceptually motivated multiple-f0 estimation method. 2005 IEEE workshop on applications of signal processing to audio and acoustics, 291–294. 2005.
[22] Typke, R., Prechelt, L.: An interface for melody input. ACM Transactions on Computer–Human Interaction., 133–149. 2001.
[23] Ukkonen, E., Lemstrom, K., Makinen, V.: Sweepline the music. Lecture Notes in Computer Science 2598, 330–342. 2003.
[24] Motoyuki Suzuki, et al.: Music information retrieval from a singing voice based on verification of recognized hypothesis. ISMIR 2006. 2006.
[25] McCann, J.A. et al.: Kendra: Adaptive Internet system. Journal of Systems and Software 55 (1), 3–17. 2000.
[26] Huang, C.M. et al.: Synchronization and flow adaptation schemes for reliable multiple-stream transmission in multimedia presentation. Journal of Systems and Software 56 (2), 133–151. 2001.
[27] Nelles, O.: Nonlinear system identification. Springer, Germany. 2001.
[28] Yun, S.Y., Namkoong, S., Rho, J.H., Shin, S.W. and Choi, J.U.: A performance evaluation of neural network models in traffic volume forecasting, Mathematic Computing Modelling, Vol. 27, No.9-11, 293-310. 1998.
[29] Lingras, P. and Mountford, P.: Time delay neural networks designed using genetic algorithms for short term inter-city traffic forecasting. In L. Monostori, J. Vancza and A. Moonis (eds.), IEA/AIE 2001. Springer, Berlin. 2001.
[30] Campolucci, P., Uncini, A., Piazza, F. and Rao, B.D.: On-line learning algorithms for locally recurrent neural networks, IEEE transactions on neural networks, Vol. 10, No. 2, 253-271. 1999.

[31] Tsai, T-H., Lee, C-K. and Wei, C-H.: Artificial Neural Networks Based Approach for Short-term Railway Passenger Demand Forecasting, Journal of the Eastern Asia Society for Transportation Studies, Vol. 4, 221-235. 2003.
[32] Haykin, S.: Neural Networks: A Comprehensive Foundation, 2nd Edition, Prentice Hall PTR. ISBN 0-13-273350-1, p. 837, 1998.
[33] Weibel, A., Hanazawa, T., Hinton, G., Shikano, K., Lang, K.: Phenomena Recognition Using Time-delay Neural Networks. In: IEEE Transactions on Acoustics, Speech, and Signal Processing 37, pp. 328-339, 1989.