Implementation of Artificial Neural Network and Multilevel of Discrete Wavelet Transform for Voice Recognition

Bandhit Suksiri and Masahiro Fukumoto

Abstract This paper presents an implementation of simple Artificial Neural Network model and multilevel of Discrete Wavelet Transform as feature extractions, which is achieved to increase the high recognition rates up to 95 % instead of Short-time Fourier Transform in the conversation background noises at noises up to 65 dB. The performance evaluation has been demonstrated in terms of correct recognition rate, maximum noise power of interfering sounds, hit rates, false alarm rates and miss rates. The proposed method offers a potential alternative to intelligence voice recognition system in speech analysis-synthesis and recognition applications.

Keywords Discrete wavelet transform · Voice recognition · Artificial neural network · Feature extractions

1 Introduction

During the past 65 years, voice recognition is being extensively implemented for the classification of sound types. The variety of voice recognition techniques have been developed to increase the efficiency of recognitive accuracy, statistical pattern recognition, signal processing and recognition rates as shown in [1].

According to a lot of research, a number of algorithms have been proposed and suggested as potential solutions to recognize human's speech, i.e., the simply probability distribution fitting methods such as, Structural Maximum A Posteriori,

B. Suksiri (✉)
Graduate School of Engineering, Kochi University of Technology (KUT),
Kami City, Kochi 782-8502, Japan
e-mail: 187001v@gs.kochi-tech.ac.jp

M. Fukumoto (✉)
School of Information, Kochi University of Technology (KUT),
Kami City, Kochi 782-8502, Japan
e-mail: fukumoto.masahiro@kochi-tech.ac.jp

© Springer International Publishing Switzerland 2016

R. Lee (ed.), *Computer and Information Science*,
Studies in Computational Intelligence 656, DOI 10.1007/978-3-319-40171-3_2

Parallel Model Composition and Maximum Likelihood Linear Regression. However, the issue of sequential voice input had been being still unsolved.

Ferguson et al. has proposed Hidden Markov Model (HMM) in order to solve an issue of sequential voice input. HMM was employed double stochastic process using an embedded stochastic function in order to determine the value of the hidden states as shown in [1]. High recognition rates design was essentially required state of the art of architecture in HMM using Gaussian Mixture Model (GMM) as shown in [2, 3]. GMM has been traditionally utilized voice models for voice recognition using two feature extractions, a power logarithm of FFT spectrum in order to create Log-power spectrum feature vectors and Mel-Scale Filter Bank Inverse FFT Dimension Reduction in order to created Mel Frequency Cepstral Coefficient feature vectors. GMM offered high voice recognition rate from 60 to 95 % in a static environment by comparison with other machine learning model such as Support Vector Machine and Dual Penalized Logistic Regression Machine as shown in [2]. Nonetheless, large amounts of computational resource are required in GMM.

Pitch-Cluster-Maps (PCMs) model was proposed by Yoko et al. [4] in order to replace the complex training sets with Binarized Frequency Spectrum resulted from simple codebook sets using Short-time Fourier Transform [5–7]. Vector Quantization Approach method was employed lead suitable Real-Time computation than GMM. Nonetheless, PCMs offered voice recognition rate up to 60 % for 6 sound sources environment under low frequency resolution.

This paper aim to propose, the alternative voice recognition utilize Artificial Neural Network and Multilevel of Discrete Wavelet Transform with 3 main advantages. First, Discrete Wavelet Transform has resolved the low frequency prediction issue in order to increases low frequency prediction. Second, the normal conversation background noise issue resolves in the proposed voice recognition. Last, the proposed voice recognition has been improved recognition rates up to 95 % by comparison with other model.

2 Proposed Voice Recognition

The overview of proposed voice recognitions consisted of feature extraction, feature normalization, machine learning as ANN and decision model which summarized in Fig. 1.

2.1 Feature Extraction

The proposed voice recognition utilized the feature extraction as the pre-processing methods in order to transform the voices or signals to the time-frequency represented data. Three pre-processing methods were implemented for voices feature extraction consisted of Short-time Fourier Transform (STFT) and Discrete Wavelet

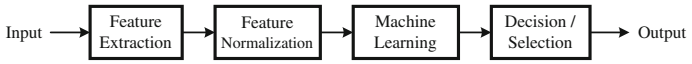

Fig. 1 The proposed voice recognition overview

Transform (DWT) [10]. In general case, Continuous Wavelet Transform (CWT) can be expressed as

$$X_{\Psi}(a,b) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} x(t) \cdot \overline{\Psi}\left(\frac{t-b}{a}\right) dt \tag{1}$$

where $\overline{\Psi}(t)$ is the conjugate of Wavelet function, a is input scales which represented as frequency variable, b is input time variable, $x(t)$ is the continuous signal to be transformed and $X_{\Psi}(a,b)$ is the CWT of a complex function represented the magnitude of the continuous signal over time and frequency based on specified Wavelet function.

In particular, DWT transformation decomposes the signal into mutually orthogonal set of wavelets, which is the main difference from the CWT [10] or its implementation for the discrete time series. DWT provides sufficient information in both time and frequency with a significant reduction in the computation time than CWT [10]. DWT can be constructed from convolution of the signal with the impulse response of the filter expressed as

$$\phi[n] = \sum_{i=-\infty}^{\infty} a_i \cdot \phi[Sn - i] \tag{2}$$

where $\phi[n]$ is the dilation reference equation as discrete signals from input to output states, S is a scaling factor to be assign value to 2, n is time index and a_i consists of two scaling functions obtained from each Wavelet function know as Quadrature Mirror Filter.

DWT equation can be represented as a binary hierarchical tree of LPF and HPF, in other words, it can be defined as Filter Banks as shown in Fig. 2. In Filter Banks analysis, lengths of discrete signals are reduced by halved per level. The effect of shifting and scaling process from (1) to (2) produces a time-scale representation as shown in Fig. 3. The graphs show the signal amplitudes in both of time and frequency domain using STFT for left-hand graph and CWT for right-hand graph. The vertical axis is represented frequency band and horizontal axis is represented time domain. It can be seen from a comparison with STFT and CWT, Wavelet Transform offers a superior temporal resolution of time resolution at high frequency components and scale resolution at low frequency components [10], which usually give a voice signal and its main characteristics or identity.

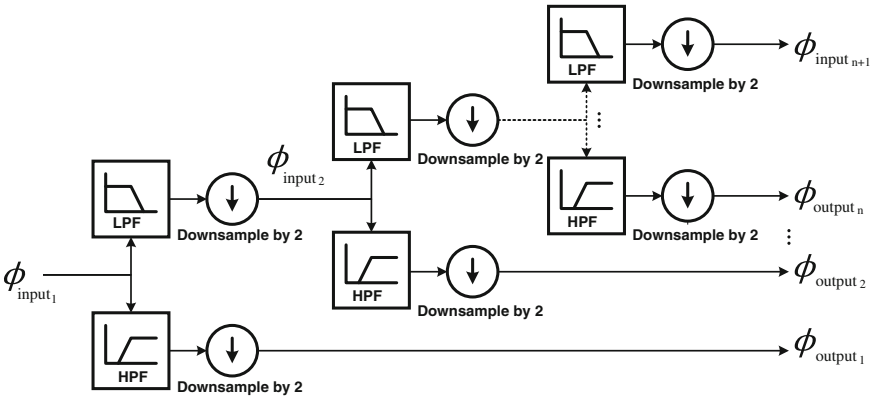

Fig. 2 DWT Filter Bank representation

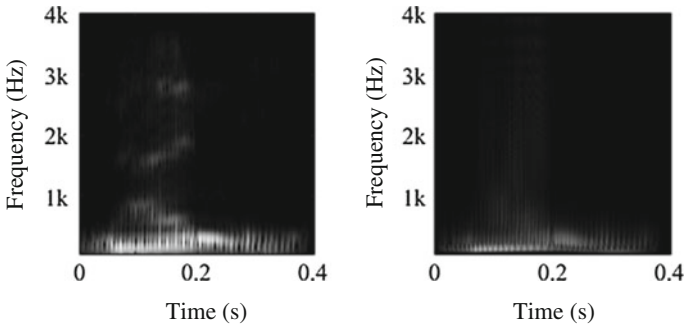

Fig. 3 The comparison of STFT (*left*) and CWT (*right*)

2.2 *Feature Normalization*

In order to increases the speed convergences of the machine learning algorithm, Feature Normalization method is utilized with simplest form is given as follow

$$\tilde{x} = \frac{x - \min(x)}{\max(x) - \min(x)} - \bar{x} \tag{3}$$

where $\tilde{x}$ is the normalized vector and x is original vector determine from feature extraction and $\bar{x}$ is its offset of signals from zero. Feature Normalization offers the range of original vector to scale the range between 0 and 1.

2.3 *Artificial Neural Network Model*

Artificial Neuron Network (ANN) [8] is an adaptive system that changes structure based on external and internal information that flows through the network. ANN is considered nonlinear statistical data modeling tools where the complex relationships between inputs and outputs are modeled or patterns are found. Therefore, the proposed voice recognition utilized ANN in order to recognize a characteristics or identity of human speech.

The novel network topology name the nth-order All-features-connecting topology is represented by H_n as illustrated in Fig. 4 where x_f is an input vector in each frequency band which is calculated from feature extraction model and y is a class probability vector which is calculated by ANN. H_n model utilizes the network of A, B and C-class in order to construct simple network topology with those network were shown in Table 1. The four main conditions of the novel network topology are defined. First, numbers of layers are defined from order of H_n where $n > 0$. Second, numbers of input networks are related to number of input time index, i.e., size of

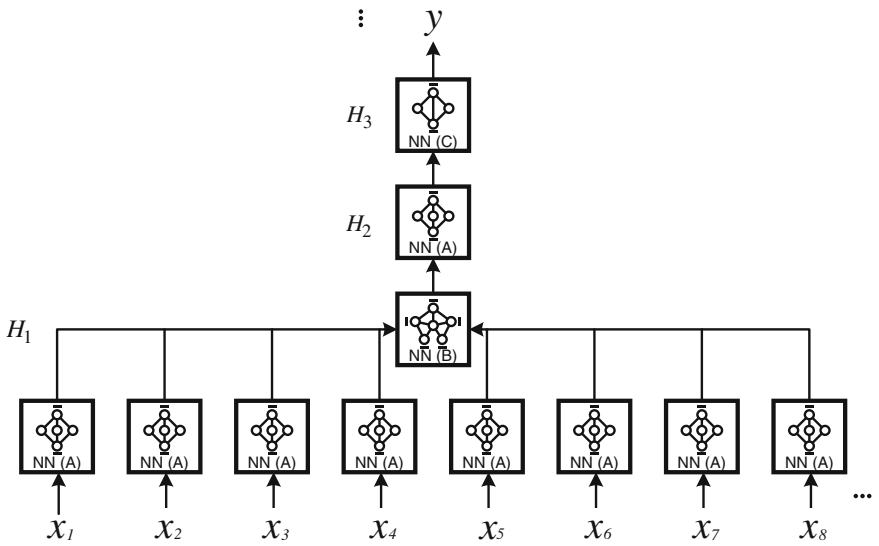

Fig. 4 All-features-connecting topology

Table 1 The Proposed neuron network architectures

Class name	Control input	Control output	Transfer function
A-class	Single	Single	Log-sigmoid
B-class	Multiple	Single	Log-sigmoid
C-class	Single	Single	Softmax

scales vector for CWT and number of levels for DWT. Third, input networks are required the connection of all single outputs to the first block of network series. The input, middle and output networks are A, B and C-class network, respectively, which give as a last condition.

In order to train specified ANN, Scaled Conjugate Gradient Backpropagation [9] supervised learning rule is employed. Additionally, specified ANN utilizes pre-learning rule using Autoassociator [11, 12] to initiate weights approximation of final solution lead to accelerate the convergence of the error Backpropagation learning algorithm and reduce dimension from wavelet packet series.

2.4 Decision Model

The output of ANN is represented as vector of class possibility value base on feature set. The decision model is expressed as maximum of class possibility value

$$c = \underset{i \in \aleph}{\operatorname{argmax}}(y_i) \tag{4}$$

where c is maximum possibility class value and y_i is element of y where $y = (y_1, y_2, \ldots, y_n)^T$ in each class number i which is calculated from ANN.

3 Experiment Setup

The proposed voice recognition is implemented using MATLAB®. The recording devices utilizes Audio-Technica® AT-VD3 microphone and ROLAND® UA-101 Hi-speed USB audio capture. The samples selects 5 Japanese including 2 youth in both male and female speakers and 1 middle-age Japanese male speaker. In order to perform word classification, the speaker pronounces the reference words from International Phonetic Alphabet (IPA) [13] datasets which were described in Table 2. The features set assigns voice input to 8 kHz sampling frequency, 16 bits

Table 2 Features set for experimentation

Class	Words	IPA's	Class	Words	IPA's
1	パン	páɴ	10	雑用	zátˢuzi
2	番	báɴ	11	山	jamá
3	先ず	mázu	12	脈	mjakú
4	太陽	táijo	13	風	kaze
5	段々	daɴdaɴ	14	外套	gaito:
6	通知	tˢu: tˢi	15	医学	ígaku
7	何	náni	16	善意	zéɴi
8	蘭	ɽáɴ	17	鼻	hana
9	数字	su:si	18	わ	wa

data resolution and 8000 sample points. The numbers of features set are 450 elements obtains from reference words in dataset with 5 times repeated. In order to perform the performance evaluation, the experiments are selected 20 % for tests set and 80 % for features set.

4 Experimental Results

The performance evaluation was established in term of correct recognition rate which calculated from the summation of true positive and true negative rates in each class. Moreover, maximum noise power of interfering sounds with nonlinear logarithmic scale defines as follows

$$P_{\text{noise, dB}} = 10 \log_{10}\left(\frac{P_{\text{noise}}}{P_{\text{ref}}}\right) \tag{5}$$

where $P_{\text{noise,dB}}$ is noise power level in decibel (dB), P_{noise} is noise power level in watt and P_{ref} is reference power level in watt (W). The experimentation assign P_{ref} is 10^{-12} W as a reference for ambient noise level in order to map voice signal conditions over a spatial regime.

Three experiments were conducted with subject to word classification in order to examine appropriate values of Wavelet and ANN parameters. The first experiment proposed an examination of Wavelet function category and its order using set of static parameters shown in Table 3. The experimental results consisted of three Wavelet functions included Daubechies, Symlet and Coiflet Wavelet function with each order from 1 to 16. It is definitely seen from Table 4 that several Wavelet function achieved word classification with correct recognition rates greater than 80 % and noise power of interfering sounds greater than 50 dB. The Wavelet function was selected by two satisfied conditions, maximum values of noise power of interfering sound and correct recognition rate. Therefore, Daubechies 15 Wavelet function revealed the satisfied maximum values of noise power of interfering sound 65.5 dB and correct recognition rate 96.22 %.

Table 3 The first experimental configuration

Parameter name	Value
Subject	Word classification
Feature extraction method	Discrete Wavelet Transform (DWT)
Wavelet level	6
Wavelet function	*Variable parameter*
Network topology	3rd-order All-features-connecting topology (H_3)
Node size in each layer	{1000, 4000, 1000, 18}

Table 4 The first experimental results

Wavelet function						
Order	Daubechies (db)		Symlet (sym)		Coiflet (coif)	
	$P_{\text{noise, dB}}$	Recognition Rate (%)	$P_{\text{noise, dB}}$	Recognition rate (%)	$P_{\text{noise, dB}}$	Recognition rate (%)
1	24.50	90.44	*None*		64.50	94.89
2	60.38	93.11	0.00	88.44	62.50	93.33
3	63.63	94.00	63.75	94.22	63.50	96.00
4	64.25	92.44	55.50	94.00	42.50	92.00
5	55.13	94.67	63.25	94.67	*None*	
6	67.75	94.89	65.25	96.00	*None*	
7	56.00	93.78	61.00	94.67	*None*	
8	57.50	95.56	35.50	91.11	*None*	
9	34.50	92.44	67.50	94.22	*None*	
10	36.25	93.78	64.25	94.44	*None*	
11	59.00	94.00	0.00	84.00	*None*	
12	61.50	95.56	61.25	95.11	*None*	
13	67.50	95.11	67.75	95.78	*None*	
14	58.00	95.33	65.25	94.67	*None*	
15	**65.50**	**96.22**	27.75	91.33	*None*	
16	61.50	96.88	65.75	94.44	*None*	

However, cost functions of proposed voice recognition were obviously influenced by the effect of Wavelet function, Wavelet level and ANN network topology. Hence, the second experiment was designed to optimize Wavelet level and ANN network topology using set of static parameters as shown in Table 5. It is obviously seen from Table 6 that H_3 model with Wavelet level 4 to 8 achieved word classification with noise power of interfering sounds greater than 60 dB and correct recognition rates greater than 94 %. Hence, H_3 model with Wavelet level 6 was selected with two satisfied conditions criteria, minimizing computation and verify the validity inside the ROI in human speech frequency form 130 to 4 kHz. H_3 model with Wavelet level 6 was selected which gives the maximum values with correct recognition rate 94.67 % and noise power of interfering sound 61 dB.

Table 5 The second experimental configuration

Parameter name	Value
Subject	Word classification
Feature extraction method	Discrete Wavelet Transform (DWT)
Wavelet level	*Variable parameter*
Wavelet function	Symlet 7 (sym7)
Network topology	*Variable parameter*
Node size in each layer	*Variable parameter*

Table 6 The second experimental results

Wavelet level	Network topology	Node size in each layer	$P_{\text{noise, dB}}$	Recognition rate (%)
1	H_1	{1000,18}	0.00	90.89
2	H_1	{1000,18}	33.00	93.56
1	H_2	{1000,1000,18}	0.00	90.44
2	H_2	{1000,1000,18}	29.75	93.33
3	H_2	{1000,1000,18}	39.75	94.00
4	H_2	{1000,1000,18}	48.50	94.67
1	H_3	{1000,4000,1000,18}	0.00	90.22
2	H_3	{1000,4000,1000,18}	28.25	92.22
3	H_3	{1000,4000,1000,18}	38.25	94.89
4	H_3	{1000,4000,1000,18}	52.25	94.44
5	H_3	{1000,4000,1000,18}	54.00	94.00
6	H_3	{1000,4000,1000,18}	**61.00**	**94.67**
7	H_3	{1000,4000,1000,18}	**60.00**	**95.33**
8	H_3	{1000,4000,1000,18}	**62.25**	**95.78**

Finally, the last experiment was designed to verify the hypothesis which Wavelet Transform feature extraction is suitable for the voice recognition application instead of STFT as shown in Tables 7 and 8. It is apparent seen that the correct recognition rates and noise power of interfering sounds in DWT achieved to increase high recognition rates than of STFT by reason of DWT theoretically employs multi-resolution [10] lead to offers the main characteristics or identity of voice at low frequency boundary which depends on Wavelet function and length of input signal.

Table 7 The third experimental configuration

Parameter name	Value
Subject	Word classification
Feature extraction method	*Variable parameter*
Wavelet level	6
Wavelet function	Symlet 7 (sym7)
STFT windows	Hamming
STFT time slot	1 ms
STFT frequency separation	8
Network topology	3rd-order All-features-connecting topology (H_3)
Node size in each layer	{1000,4000,1000,18}

Table 8 The third experimental results

Feature extraction method	$P_{\text{noise,dB}}$	Recognition rate (%)
Discrete Wavelet Transform (DWT)	61.00	94.67
Short-time Fourier Transform (STFT)	0.00	88.67

5 Discussions

The summaries of the optimized parameters with both of word and gender classification were described in Table 9. It can be seen that the proposed voice recognition with the optimized parameters offered high correct recognition rate and noise power were 96.22 % and 65.5 dB which sufficient for word classification. Moreover, the proposed voice recognition with the optimized parameters offered the correct recognition rate and noise power were 99.8 % and 72.25 dB which acceptable for gender classification.

The proposed voice recognition performance was established in term of the boundary of hit rate, false alarm and miss rate [8] with gender classification in order to compare with other models, i.e., simple sound database named Pitch-Cluster-Maps (PCMs). The performance of PCMs models established in term of Detection Error Tradeoff (DET) [4] curves with gender classification, in other words, it can be defines as upper and lower boundary both of false alarm rate and miss rate. The best performance of hit rate requires set of predicted data which approach to 100 % on true positive rate. In contrast, the best performance of false alarm and miss rate requires set of predicted data to approach on the false positive rate and false negative rate being closely equal to 0 and 0 %, respectively. Therefore, lower boundary of hit rate, upper boundary of false alarm rate and upper boundary of miss rate were important for performance evaluation. The proposed voice recognition performance was shown in Table 10.

On the one hand, subject to male classification, PCMs offered miss probability range from 2 to 12 % and false alarm probability range from 1 to 10 %, in other words, PCMs offered the upper boundary of miss and false alarm rate were 12 % and 10 %, respectively. Likewise, PCMs offered miss probability range from 2 to 20 % and false alarm probability range from 1 to 20 % for subject of female classification, in other words, PCMs offered the both of upper boundary of miss and

Table 9 The parameter optimization results

Parameter name	Subject	
	Word classification	Gender classification
Feature extraction method	Discrete Wavelet Transform (DWT)	Discrete Wavelet Transform (DWT)
Wavelet level	6	6
Wavelet function	Daubechies 15 (db15)	Daubechies 15 (db15)
Network topology	3rd-order All-features-connecting topology (H_3)	3rd-order All-features-connecting topology (H_3)
Node size in each layer	{1000,4000,1000,18}	{1000,4000,1000,2}
$P_{noise,dB}$	65.50	72.25
Recognition rate (%)	96.22	99.80

Table 10 Performance evaluation

Gender	Lower boundary	Upper boundary	
	Hit rate (%)	False alarm rate (%)	Miss rate (%)
Male	99.63	2.78	0.37
Female	97.22	0.37	2.78

false alarm rate were 20 %, which is appropriate for the female speaker identification from several words utterance.

On the other hand, proposed voice recognition with male classification offered upper boundary of miss and false alarm rate were 0.37 % and 2.78 %, respectively. With the female subjects, proposed voice recognition with male classification offered upper boundary of miss and false alarm rate were 2.78 % and 0.37 %, respectively, which is reduce the miss and false alarm rates leads to increases accuracy both of male and female classification with sufficient for work under normal conversation background noises conditions. However, the accuracy of gender classification is decreased since speech phase shift occurred. The variations of feature sets are further required for training the proposed voice recognition in order to implement large scale of word classification.

6 Conclusions

This paper presented an alternative voice recognition using combination of Artificial Neural Network and Multilevel of Discrete Wavelet Transform. The experimental results proved Wavelet Transform was achieved to increases high recognition rates up to 95 % instead of Short-time Fourier Transform feature extractions at noises up to 65 dB as in normal conversation background noises. The performance evaluation was demonstrated in terms of correct recognition rate, maximum noise power of interfering sounds, hit rate, false alarm rate and miss rate. The proposed method offers a potential alternative to intelligence voice recognition system in speech analysis-synthesis and recognition applications.

References

1. Furui, S.: 50 years of progress in speech and speaker recognition. In: SPECOM2005, pp. 1–9, Patras, Greece (2005)
2. Matsui, T., Tanabe, K.: Comparative study of speaker identification methods: dPLRM, SVM and GMM. In: IEICE Transactions on Information and System, vol. E89–D, no.3 (2006)
3. Matsui, T., Furui, S.: Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs. In: Acoustics, Speech, and Signal Processing, ICASSP, vol. 92. IEEE, San Francisco (1992)

4. Sasaki, Y., et al.: Pitch-cluster-map based daily sound recognition for mobile robot audition. J. Robot. Mechatron. **22**(3) (2010)
5. Zenteno, E., Sotomayor, M.: Robust voice activity detection algorithm using spectrum estimation and dynamic thresholding. IEEE Latin-American Communications, LATINCOM09 (2009)
6. Patil, S.P., Gowdy, J.N.: Exploiting the baseband phase structure of the voiced speech for speech enhancement. In: Acoustics, Speech and Signal Processing, ICASSP2014. IEEE, Florence, Italy (2014)
7. Krawczyk, M., et al.: Phase-sensitive real-time capable speech enhancement under voiced-unvoiced uncertainty. In: Signal Processing Conference, EUSIPCO2013, IEEE, Marrakech, Morocco (2013)
8. Bishop, C.M.: Pattern Recognition and Machine Learning (Information Science and Statistics). Springer, New York (2006). ISBN 978-0-387-31073-2
9. Møller, M.F.: A scaled conjugate gradient algorithm for fast supervised learning. Neural Netw. **6**(4), 525–533 (1993)
10. Mallat, S.G.: A theory for multiresolution signal decomposition: the wavelet representation. IEEE Pattern Anal. Mach. Intell. **11**(7), 674–693 (1989)
11. Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H.: Greedy layer-wise training of deep networks. In: Advances in Neural Information Processing Systems (2007)
12. Bengio, Y.: Learning deep architectures for AI. Found. Trends Mach. Learn. **2**(1), 1–127 (2009)
13. Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet. Cambridge University Press (1999). ISBN-0521637511

http://www.springer.com/978-3-319-40170-6

Computer and Information Science
Lee, R. (Ed.)
2016, XIII, 181 p. 68 illus., 43 illus. in color., Hardcover
ISBN: 978-3-319-40170-6