Chapter 7

Neural Network Configurations Analysis for Multilevel Speech Pattern Recognition System with Mixture of Experts

Washington Luis Santos Silva,
Priscila Lima Rocha and
Allan Kardec Duailibe Barros Filho

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.72575

Abstract

This chapter proposes to analyze two configurations of neural networks to compose the expert set in the development of a multilevel speech signal pattern recognition system of 30 commands in the Brazilian Portuguese language. Then, multilayer perceptron (MLP) and learning vector quantization (LVQ) networks have their performances verified during the training, validation and test stages in the speech signal recognition, whose patterns are given by two-dimensional time matrices, result from mel-cepstral coefficients coding by the discrete cosine transform (DCT). In order to avoid the pattern separability problem, the patterns are modified by a nonlinear transformation to a high-dimensional space through a suitable set of Gaussian radial base functions (GRBF). The performance of MLP and LVQ experts is improved and configurations are trained with few examples of each modified pattern. Several combinations were performed for the neural network topologies and algorithms previously established to determine the network structures with the best hit and generalization results.

Keywords: automatic speech recognition, neural network, DCT models, multilayer perceptron, learning vector quantization, Gaussian radial basis function, mixture of experts, multiclass task

1. Introduction

The human ability to recognize patterns involves the sophisticated neural and cognitive systems that, from the accumulation of experience on a given environment, can extract the relevant characteristics that shape a given situation and store that information for using when there is a need. This ability makes the decision-making process much faster. Thus, many

researchers work to understand the biological pattern recognition mechanism of human for the development of computational algorithms for learning machines that are increasingly robust for use in practical applications [1, 2].

Pattern recognition is a scientific area that aims to classify patterns, also called instances or examples, according to their characteristics that form a multidimensional space (space of characteristics) in distinct sets, which are called classes or labels or categories so that an action can subsequently be better performed according to each category. Since pattern examples are needed to obtain the distinct sets, the pattern recognition process involves a statistical analysis to obtain the models, as well as the insertion or not of the expert knowledge in the application domain, which can characterize a supervised or unsupervised classification, respectively.

The task of speech signals recognition is challenging, since the signals obtained in the speech production process are highly variable, due to the great amount of attributes of the human speech, besides the specific characteristics involved in speech, such as environment noise and the properties of each language. The development of systems based on speech signal pattern recognition is one of the practical applications in using pattern classification. Indeed, speech is the most natural and expressive mode in human communication, and thus methodologies for analysis and recognition of speech signal have been developed and influenced by the knowledge of how this task is solved by humans [1, 2].

Currently, speech recognition system applications cover a wide area of domains such as dictation tools in text editors, automatic answering services in telephone exchanges, hands-free car-based systems, people with motor disabilities, mobile interface via speech, ticket reservation applications in airlines, security systems by speech identification, and so on. Then, the pattern recognition task involves different steps and its efficient execution guarantees greater accuracy. The development stages required of a pattern recognition system are as follows [3–5]: data acquisition, preprocessing and extraction of the most relevant characteristics; data representation and definition of the classifier for decision-making.

The techniques of digital signal processing and digital signal coding are the tools that support the representation of the patterns. Advances in digital speech processing methodologies allow the maximum use of speech signal attributes for use in the speaker or speech recognition, depending on the application [6, 7]. In addition to the need of good attributes extraction that represents the patterns to be recognized, it is also important that they are coded in a reduced number of parameters. Indeed, the more information you add to the system, the greater the probability of good results. However, this relationship must be taken with caution because this increase in data expands configuration complexity and computational cost of the system. For this reason, appropriate digital signal coding techniques contribute significantly to determine the equilibrium between number of parameters and computational cost [8].

After speech signal coding process and obtainment of representative patterns, the recognition task can be performed efficiently using algorithms of patterns identification, according to the third step mentioned. These algorithms (also called classifiers) develop models that generalize each category or class belonging to the system from patterns set (called training set). The classification algorithm is responsible for establishing relationship between patterns and their respective categories. Then, in testing stage, the classifier can determine to which category the

new pattern belongs. A crucial point for classifiers is to determine the decision boundaries between each class, that is, to specify the model that allows the identification of new data. It becomes more complex as the number of classes increases. However, many of classification methodologies were developed based on solving the problem of two classes, because of the dichotomy algorithms (called binary classifiers). In reality, it shows that classification problems require solution for more than two classes (multiclass) [9, 10].

The use of only one compact structure classifier to solve multiclass task can increase computational cost and generalization capacity of the classifier. Overcoming this problem and from the principle of *divide and conquer*, the ensemble method aims to fragment the characteristics space so that a set of simpler topology classifiers learn the specificities of each subspace. Finally, the classification result is given by individual results or by choice from result of one of the classifiers topology, according to a certain rule hence, the result of the multiclass task is obtained from simpler classifiers [11, 12].

Among patterns identification algorithms that can be used in the approach of ensembles, the neural networks configure as high potential classifiers. Neural networks are intelligent computational algorithms that simulate the behavior of biological neurons. It results in a robust system with low rate of recognition errors. The robustness provided in the classification task is a result of the inherent adaptive characteristic of neural networks, allowing them to be able to learn complex patterns and trends present in the set of data available for identification, changing rapidly to modifications in the environment in which is inserted [13–15]. The neural networks have several configurations for solution of the most problems and among such configurations with the best results in solving pattern classification problems are multilayer perceptron (MLP) and the learning vector quantization (LVQ) [16, 17].

2. Theoretical fundamentals

2.1. Multiclass learning

Bayes' statistic decision theory or Bayes' decision theory is the classic fundament for mathematically defining the task of pattern recognition. This approach expresses the problem solution in probabilistic terms. Classifiers projected from Bayes' decision theory constitute optimum classifiers, in which new classification approaches can take them as a reference for comparison of results. The classification rule based on Bayes' theory can be better understood when it is analyzed to make a decision between two classes. This definition can be generalized for multiclass task solution. Then, it is possible for calculating a posteriori probability of class γ_i to occurring when input vector $\boldsymbol{x}$ is presented through Bayes' formula given by Eq. (1):

$$P(\gamma_i \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid \gamma_i)P(\gamma_i)}{p(\boldsymbol{x})} \tag{1}$$

where, γ_i is ith class defined in problem; $P(\gamma_i \mid \boldsymbol{x})$ is *a posteriori* probability of γ_i class; $p(\boldsymbol{x} \mid \gamma_i)$ is the joint distribution of pattern $\boldsymbol{x}$ into γ_i class; $P(\gamma_i)$ is the *priori* probability of γ_i class; $p(\boldsymbol{x})$ is the probability density function. Considering the classification in more than two classes, that is, when

the objective is to discriminate the feature vector $\boldsymbol{x}$ in one of C classes in set $\zeta = \{\gamma_1, \ldots, \gamma_C\}$, the conditional probability of each class is obtained by the Bayes' formula (2):

$$P(\gamma \mid \boldsymbol{x}) = \left\{P(\gamma_1 \mid \boldsymbol{x}), \ldots, P(\gamma_C \mid \boldsymbol{x})\right\} \tag{2}$$

Then, according to general Bayes decision rule, the vector $\boldsymbol{x}$ is allocated to the most probability class, given by (3):

$$\hat{\gamma} = \underset{\gamma_i \in \zeta}{argmax}\ \hat{P}(\gamma_i \mid \boldsymbol{x}) \tag{3}$$

Despite Bayesian mathematical formalism, there is a great difficulty in practical applications due to estimation of the quantities on right side of Eq. (1). This difficulty increases when the number of estimates in a multiclass problem must be defined simultaneously with high accuracy, since the boundaries among different classes may not be well defined. Thus, new methodologies are proposed to obtain more robust results in multiclass tasks [7, 18].

2.2. Speech recognition systems

Speech recognition systems extract significant characteristics of the speech signal to obtain a pattern that represents this signal and classify it into a class target space defined in recognition project. A class is a group of patterns that have similar characteristics. The purpose of speech recognition allows that these systems are divided into three ways: speaker recognition, language identification and word recognition. Speaker recognition systems are those whose focus is the recognition of the speaker who pronounced a certain word or sentence among different individuals. For identification of language, the purpose of the recognition system is to determine in which language that word or sentence is pronounced. Finally, the word recognition has the interest in identify which word or sentence was pronounced. It has the division into two different forms when the objective of the speech recognition system is to distinguish the spoken word or sentence: speaker-dependent word recognition and speaker-independent word recognition. The first one, it has the trained system to identify the word that was spoken by a specific individual. In the second case, the system identifies word or sentence spoken by people different from those used during the training because it is not important who spoke the word. Besides question of speaker dependence or not, word recognition can be accomplished through isolated words or continuous speech. In first case, it is necessary to have an interval between each word. This is done to have a clear distinction from start to finish of the word, avoiding effect of the coarticulation that causes change in the way of pronouncing the sounds. For continuous speech case, the speaker pronounces on natural way, and consequently, it is difficult to distinguish the beginning and end of the word, causing word concatenation. Continuous speech recognition is more complex because there is no pause between one word and another, generating a single sound. Systems that work with this form of recognition are based on smaller units of the word, such as syllables, phonemes, diphones, triphones, and so on [19, 20].

2.3. Radial basis function

Radial basis functions are important tools in modeling of classification and prediction tasks. They comprise a particular class of functions that have a monotonically increase or decrease

http://dx.doi.org/10.5772/intechopen.72575

response with distance from the origin or a central point, such that $\Phi(\boldsymbol{x}) = \Phi(\|\boldsymbol{x}\|)$ or $\Phi(\boldsymbol{x}, \boldsymbol{\mu}) = \Phi(\|\boldsymbol{x} - \boldsymbol{\mu}\|)$, respectively. In general, the norm used in radial basis functions is the Euclidean distance, but other distance functions may be used. Mathematically, a function $\Phi: \mathbb{R}^s \rightarrow \mathbb{R}$ is said radial if there is a univariate function, $\varphi: [0, \infty) \rightarrow \mathbb{R}$ such that (4):

$$\Phi(\boldsymbol{x}) = \varphi(r) \tag{4}$$

where $r = \|\boldsymbol{x} - \boldsymbol{\mu}\|$ and $\|\cdot\|$ is some norm in $\mathbb{R}^s$; Euclidean norm is usually used. Gaussian radial basis function is the most function used among radial functions, computed as (5):

$$\varphi(r) = e^{-\frac{r^2}{2\sigma^2}} \tag{5}$$

This function is defined by $\boldsymbol{c}$ parameter that defines Gaussian center and σ^2 represents the variance, which characterizes the base widening of the curve and indicates how dispersed a vector $\boldsymbol{x}$ in analysis is in relation to center $\boldsymbol{\mu}$. These parameters may be obtained from the data that belong to the problem to be modeled. Radial basis functions are used to make nonlinear mapping between two feature spaces. Thus, in pattern classification problems, for example, given a set χ of N patterns, $\boldsymbol{x} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ of m_0-dimension, where each one of these vectors is assigned to one of two classes, χ_1 and χ_2, if these patterns cannot be linearly separated in the original dimensional space, a set of radial basis functions can be used to map in a space that allows this separation. Then, each pattern $\boldsymbol{x}$ of the set χ is defined as new vector, where each element is represented by response of radial basis function set $\{\varphi_i(\boldsymbol{x}) \mid i = 1, 2, \ldots, m_1\}$, applied to vector $\boldsymbol{x}$ as (6):

$$\Phi(\boldsymbol{x}) = \left[\varphi_1(\boldsymbol{x}), \varphi_2(\boldsymbol{x}), \ldots, \varphi_{m_1}(\boldsymbol{x})\right]^T \tag{6}$$

The vector $\Phi(\boldsymbol{x})$ maps the vectors from m_0-dimensional input space into a new m_1-dimensional space. For the classification of complex patterns, the increase in radial basis function number creates a space of high dimensionality that increases the probability of data linear separation in this new space, making classification problem simpler. This property is supported by Cover's separability theorem that demonstrate how pattern classification problem in a high-dimensional space is more probable to be linearly separable than in low-dimensional space [21].

2.4. Neural networks

Artificial neural networks (ANNs) are systems whose computational structure is based on how the human brain processes information from environment. Also known as a connectionist model or distributed parallel processing, the ANNs arose after presentation of the simplified neuron by McCulloch and Pitts in 1943 [14, 15]. ANNs constitute distributed parallel systems composed of simple processing units (neurons) that calculate certain mathematical (usually nonlinear) functions. These units are arranged in one or more layers and interconnected by a large number of connections, usually unidirectional. In most models, these connections are associated with weights, which store the knowledge represented in the model and weighted the information received from inputs to each neuron in the network [22]. Among attractions for the use of ANNs in problem solutions, the main ones are their ability to learn through

examples presented to them and to generalize the information learned. Other characteristics that further enhance its use are: possibility of considering the nonlinear behavior of the physical phenomena responsible for generating the input data, requirement for few statistical knowledge about the environment which the network is inserted and knowledge represented by ANN structure itself and by its activation state [23].

2.5. Ensemble methods

Based on *divide and conquer* principle widely used in engineering, the ensemble method partitions a problem in subspaces, where each subspace is designated for simple expert algorithm to learn the characteristics of each partition. This way, the individual response of each expert contributes to final response of the problem, reducing the learning algorithm complexity. Also called multiclass system, this approach uses classifiers with simpler topologies and few adjustable parameters than if a single classify structure was used to solve the same task. Another advantage presented by this method is the decrease in the training time, since the training time of a large topological structure will be probably greater than the training time of several experts in parallel. The simplicity in expert structure also avoids super adjustment of data because when it has a large number of free parameters to be adjusted in relation to training set size, the risk of over fitting increases. The most common architecture of ensemble method has a classifiers set that learns the training data characteristics and they represent classifier base. Several learning algorithms, such as neural networks, can form this base. Normally, the base is formed by only one type of classifier, keeping the ensemble structure homogeneous; despite that, other methodologies may adopt different classifiers to form the base, that is, the ensemble become heterogeneous. There are three variations of ensemble approach and expert mixture is mostly used in neural networks area. Expert mix strategy uses simple sets of parametric model that learns task subspaces and the definition of decision rules that provide a general solution. In pattern classification tasks, a new sample may be classified by ensemble method in two ways: (1) it combines classifier outputs, according to certain procedure to obtain the final response in classification stage and (2) only the response of one classifier is taken as the final response, according to some selection criterion [24, 25].

3. Analysis methodology

In face of the theoretical fundamentals described, it is presented a multilevel classification approach using radial basis functions and artificial neural networks expert set for multiclass task solution represented by locutions in the Brazilian Portuguese language from following commands: “zero” (zero), “um” (one), “dois” (two), “três” (three), “quatro” (four), “cinco” (five), “seis” (six), “sete” (seven), “oito” (eight), “nove” (nine), “abaixo” (below), “abrir” (open), “acima” (up), “aumentar” (increase), “desligar” (turn off), “diminuir” (decrease), “direita” (right), “esquerda” (left), “fechar” (close), “finalizar” (finish), “iniciar” (start), “ligar” (turn on), “máximo” (maximum), médio” (medium), “mínimo” (minimum), “para trás” (back), “para frente” (forward), “parar” (stop), “repousar” (rest) and “salvar” (save).

Speech signal coding, which uses the mel-cepstral coefficients and the discrete cosine transform (DCT), provides patterns consisting of a reduced number of parameters obtained in speech signal preprocessing step by generating two-dimensional time matrices of order 2, 3 and 4. These matrices reproduce the global and local speech signal variations in time as well as the spectral envelope. Then, the original feature space is formed from two-dimensional time matrices transformed to column vectors, maintaining the alignment time of the extracted mel-cepstral coefficients.

A set of 30 Gaussian radial basis functions were modeled properly to transforming the primary feature space in a new high dimensional nonlinear space in order to increase the probability of linear separation of categories. This strategy makes easier the classification process, according to Cover's theorem. Gaussian radial basis functions were modeled by centroid and variance parameters extracted from training patterns that compose the different classes. Afterwards, each pattern obtained through DCT two-dimensional time matrix was mapped into 30-dimensional space by 30 Gaussian radial basis functions properly parameterized. Because the Gaussian radial basis functions are parameterized with center and variance characteristics of each class, in this space of high dimensionality, it is expected that there will be adequate clustering of these patterns. Therefore, vectors of 30 elements form the training set applied during classifier learning process, where each element represents the RBF's outputs when pattern from two-dimensional DCT time matrix is applied.

Once the training set is finalized, the design and definition recognizer is carried out through performance analysis of two neural network configurations widely used in the literature: multilayer perceptron (MLP) and learning vector quantization (LVQ). The proposed multilevel classifier uses a set of 15 neural networks and each of them is expert in each predefined partition of the mapped feature space by the RBFs. This division of feature space reduces the topological complexity of MLP and LVQ configurations, training time and generalization capacity.

The performance analysis of the multilevel classifier is carried out in two phases: training, validation and individual test process of the experts and final test process. In this first procedure, predetermined topological elements and training algorithms for MLP and LVQ network configurations are combined to define the best characteristics of the 15 experts, where each one of them is responsible for learning the specificities of two classes. Thus, it is possible to verify the behavior of the MLP and LVQ networks and to select the expert topologies that presented the greatest global validation hit. These selected experts are tested individually to check the level of generalization for the classes they were assigned. Because of this, the level of accuracy is determined for each expert and these levels are part of the rules defined in the final classification stage. So, the expert topologies that obtained the highest accuracy are selected for the final test step. This step consists of the definition of rules for selection of the expert that will provide the final solution of the classification.

A new pattern generated by the DCT two-dimensional time matrix, different from those used in the training step, are used as inputs to the 30 Gaussian radial basis functions parameterized with the characteristics of each problem class. In addition to mapping the DCT pattern to a high dimensionality space, the outputs of each RBF provide a measure of input pattern probability belonging to a given class. The RBF outputs provide a preclassification rule in the multilevel recognition system, and their responses direct the appropriate expert to complete the classification. In order

to ensure that the preclassification stage by the RBF selects the correct expert, a second selection rule is adopted. The final classification result given by the neural network chosen is compared to the accuracy result of the same class verified in the individual test step. The LVQ neural network performance study as expert in this work provides an alternative approach to the classifier, since the MLP configuration is the most executed neural network in pattern recognition problems.

3.1. Speech signal preprocessing

The locutions used in this work were recorded at sampling frequency f_a = 22,050 Hz, with 16-bit resolution. The speech signal preprocessing step was carried out through samples obtained from three different voice banks. After that, characteristics of each class were extracted to constitute feature space. Signal preprocessing step consists of segmenting and windowing speech signal from database. For this proposed work, it was defined windowing of the segments through Hamming function to speech signal preprocessing algorithm. The overlap between the windows was 50%. The window size on samples was calculated by multiplying the window duration T_ω = 20 ms by the sampling frequency f_a.

3.2. Extraction of the mel-cepstral coefficients from speech signal

The coefficients are attributes extracted from the speech signal. These coefficients have vocal tract characteristics that are important information for speech recognition. In addition, its formulation makes analog to perception of sounds by humans. Then, a filter bank spaced in the mel scale was developed to obtain the mel-cepstral coefficients from speech signal samples. This filter bank covers the range of 0–4600 Hz. The bank is distributed in 20 filters, in which up to limit frequency for uniform segmentation, given by Fu = 1 kHz, filters are distributed in 10 uniform intervals. The mel-cepstral coefficients were obtained using the energy calculated for each frequency band, according to Eq. (7):

$$mfcc[k] = \sum_{i=1}^{N_F} E[i] \cos\left[\frac{i(k-0.5)\pi}{N_F}\right] \tag{7}$$

where $k = 1, 2, \cdots, K$ is the number of mel-cepstral coefficients, N_F is the number of filters used and $E[i]$ is the energy log output of the ith band.

3.3. Generation of DCT two-dimensional time matrix

After obtaining the mel-cepstral coefficients from speech signal, the coding was performed through discrete cosine transform (DCT), which allows synthesizing the long-term variations of the spectral envelope of the speech signal [26]. The result of this coding was the generation of a DCT two-dimensional time matrix that was obtained according to Eq. (8):

$$C_k(n, T) = \frac{1}{N} \sum_{t=1}^{T} mfcc_k(t) \cos\left[\frac{(2t-1)n\pi}{2T}\right] \tag{8}$$

where k, which varies from $1 \le k \le K$, is the kth component line of the ith segment of the matrix; K is the number of mel-cepstral coefficients; n, which varies from $1 \le n \le N$, is the nth column. n is the order of the DCT matrix; T is the number of observation vectors of the mel-cepstral coefficients in the time axis; $mfcck(t)$ represents mel-cepstral coefficients. Each locution of **D**

http://dx.doi.org/10.5772/intechopen.72575

digit has a DCT two-dimensional time matrix C_{kn}^{jm}, where $j = 1, 2, 3, \ldots, 30$ represents the class of commands to be recognized and $m = 1, 2, 3, \ldots, 20$ represents the example taken for each command. Each two-dimensional time matrix was transformed into column vector called C_N^{jm}, that preserve the time alignment of the mel-cepstral coefficients and they are given by (9):

$$C_N^{jm} = \left[c_{11}^{jm}, c_{12}^{jm}, \ldots, c_{1n}^{jm}, c_{21}^{jm}, c_{22}^{jm}, \ldots, c_{2n}^{jm}, \ldots, c_{kn}^{jm}\right]', \quad j = 1, 2, \ldots, 30 \mid m = 1, 2, \ldots, 20 \tag{9}$$

The vectors C_N^{jm} were used to form original training set or original feature space. So, DCT two-dimensional time matrices C_{kn}^{jm} of order $n = 2$, 3 and 4 were generated in order to compare the multilevel classifier performance when the number of parameters that compose primary speech patterns is increased. Thus, as a result, patterns represented by C_N^{jm} were obtained, where $N = k \times n = 4$, 9, 16, respectively.

3.4. Structuring of the multilevel speech recognition system with mixture of expert neural networks

After speech signal coding to generate command patterns used by recognizer, parameters of the Gaussian radial basis function set and topology design of expert neural networks were started. Radial basis functions and neural networks integrate the multilevel speech recognition system. The parameters required to model each RBFs are obtained from patterns generated by the DCT two-dimensional time matrix. These RBFs are responsible for the change of the feature space and for the preclassification stage of the multilevel system. The design of the expert neural networks set is carried out through simulations and based on results obtained in other similar pattern classification works. Two neural network configurations, MLP and LVQ are analyzed to constitute the experts in the proposed system. The choice for analyzing these two configurations in this chapter is justified because they are neural networks with great applicability and good results in pattern recognition field [14, 15, 22, 23]. According to presented methodology of this chapter, MLP and LVQ networks were analyzed for their performance in the pattern classification through two distinct steps. The analysis procedures of each step for integration between Gaussian RBFs and MLP and LVQ experts were carried out using the patterns from DCT two-dimensional time matrices of order 2, 3 and 4, and it was observed the multilevel system behavior under study. A block diagram of training step is shown in **Figure 1**.

The RBF's modeling and MLP and LVQ neural network training were carried out with speech signal patterns from locutions obtained by EPUSP (Polytechnic School of the University of São Paulo), INATEL (National Institute of Telecommunications) and IFMA (Federal Institute of Maranhão) banks. The training set $\Omega_{NL}^{Tr} = \{C_N^{11}, C_N^{12}, \ldots, C_N^{jm}\}$ is composed of 600 locution with 20 examples of each command to be recognized (m = 20), where $N = \{4, 9, 16\}$ represents the parameter number of patterns; L is total number of locutions and Tr indicates that set is training. The used training set is balanced type, that is, all classes have the same quantitative of examples, which avoid bias of the classifier. The training set was partitioned into the estimation subset Ω_N^E, which contains 80% of the training patterns, and into the validation set Ω_N^V, representing the remaining 20% of the training set $\Omega_{N600}^{Tr} = \{\Omega_N^E \cup \Omega_N^V\}$. The test phase is done using set Ω_{NL}^T consisting of 40 speakers, where 20 speakers are male (Ω_{NL}^{TM}) and 20 speakers are female (Ω_{NL}^{TF}). All speakers belong to IFMA voice bank, but they are speakers who did not participate in pronunciations for training set. Then, test set available to verify the generalization of the multilevel recognition system has 6000 samples in total $(\Omega_{N6000}^T = \{\Omega_{N3000}^{TM} \cup \Omega_{N3000}^{TF}\})$.

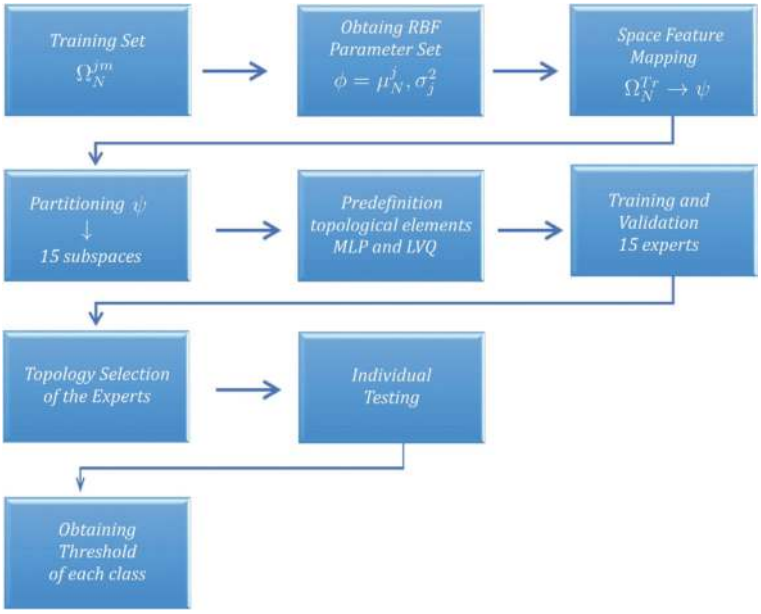

Figure 1. Block diagram of training step.

3.4.1. Parameterization Gaussian radial basis functions

The multilevel speech recognition system with mixture of experts use a set of 30 Gaussian radial basis functions that have two purposes in proposed system: the first of them, in training step, is to mapping the patterns C_N^{Jm} into a new high-dimensional nonlinear space to making easier the separability of the patterns. The second goal in testing step is to providing a preclassification rule for speech signal sample, in addition to mapping this sample into high-dimensional space. The number of chosen Gaussian radial basis functions is related to number of problem classes. Thus, the centroid parameters $\boldsymbol{\mu}_j$ and variances σ_j^2 of each class j were determined through training set Ω_{NL}^{Tr}. A suitable method for this purpose, called *k-means* [24], was used to obtain the 30 RBF centroids, whose purpose is interactively position the k-Gaussian centers in regions where the input patterns will tend to cluster. The training set Ω_{NL}^{Tr} was applied to *k-means* algorithm, where k was defined as 30, as shown in **Figure 2**.

The variance σ_j^2 was determined by criterion of the average quadratic distance. The variance σ_j^2 is expressed as:

$$\sigma_j^2 = \frac{1}{m^{(j)}} \sum_{x^{(m)} \in \Omega_j} \sum_{i=1}^{n_1} \left(x_i^{(n)} - c_j \right)^2 \tag{10}$$

Therefore, at the end of these procedures, all vectors C_N^{Jm} from training set Ω_{NL}^{Tr} are mapped into a nonlinear 30-dimensional space by radial basis functions set $\Phi(\boldsymbol{x}) = \{\varphi_1, \varphi_2, \ldots, \varphi_{30}\}$ duly modeled with characteristics of the classes.

http://dx.doi.org/10.5772/intechopen.72575

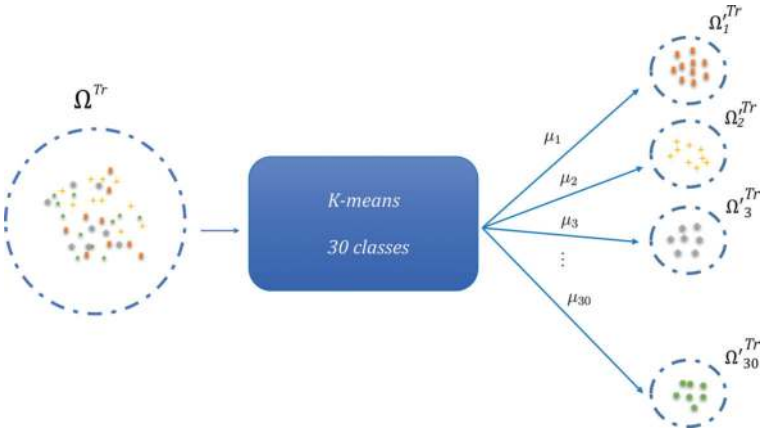

Figure 2. Schematization of the *k-means* algorithm.

3.4.2. Expert neural network design

The distribution of each 30 classes among the defined experts is shown in **Table 1**, both for LVQ and MLP configuration.

It is necessary to specify the best structure for characteristics learning of each class of training set Ω^{Tr}_{NL}. Thus, for the 15 expert neural networks, both MLP and LVQ configurations, the topological elements and training algorithms were combined during the training step. Next, it is shown how the LVQ and MLP configurations were specified in the training step [27].

3.4.2.1. LVQ experts

For the structure of the LVQ neural network, it was necessary to define the η learning rate and the *n* number of neurons of the competitive layer. The defined values in η set are often used in the specialized literature [17, 18, 26] and the *n* set was specified considering that the number of neurons in hidden layer should be greater than the number of inputs and greater than the number of neural network outputs. Because the vectors C^{lm}_N, where $N = \{4, 9, 16\}$ are mapped into a 30-dimensional space, the input of 15 LVQ experts is a set with 30 source nodes. The number of classes that integrate each specified subset gives the output number of each expert. Due to recognition problem of this work has partitioned the Ω^{Tr}_{NL} set into 15 subsets, the output of each expert should have two neurons, that is, one neuron for each class. It was defined a neuron set represented by multiple numbers of neural network inputs, starting with 60 neurons as the smallest number of neurons in hidden layer. The increase of neurons in hidden layer until neuron maximum value of *n* set allows to observe the network behavior in relation to increase the number of neurons in hidden layer.

It is summarized in **Table 2**, the topology elements and training algorithm for the simulations of the LVQ expert neural networks.

Classes	Experts	Classes	Experts	Classes	Experts
'ZERO'	1	'ABAIXO'	6	'INICIAR'	11
'UM'		'ABRIR'		'LIGAR'	
'DOIS'	2	'ACIMA'	7	'MAXIMO'	12
'TRÊS'		'AUMENTAR'		'MÉDIO'	
'QUATRO'	3	'DESLIGAR'	8	'MINIMO'	13
'CINCO'		'DIMINNUIR'		'PARA TRÁS'	
'SEIS'	4	'DIREITA'	9	'PARA FRENTE'	14
'SETE'		'ESQUERDA'		'PARAR'	
'OITO'	5	'FECHAR'	10	'REPOUSAR'	15
'NOVE'		'FINALIZAR'		'SALVAR'	

Table 1. Division of classes among experts.

Elements	Symbol
No. of neurons	n = {60, 90, 120, 150}
Learning rate	η = 0.01
No. of epoch	Epoch = 1000
Training algorithm	LVQ-1

Table 2. LVQ neural network elements.

3.4.2.2. MLP experts

The MLP neural network structure is defined by some variable elements that, properly chosen, allow a good performance of the neural network in solution of the proposed problem. It is presented in **Table 3**, these variable elements that are combined in some simulations to define the best topology.

In addition to defining the network topological elements, four different training algorithms were used in MLP network. This way, it can be verified the algorithm that presents better results to pattern set presented to network. The chosen training algorithms were: gradient descendent (GD); gradient descendent with momentum (GDM); resilient propagation (RP); Levenberg-Marquardt (LM). The simulated number of hidden layers was defined by fact that, for pattern classification problems, the use of up to two layers is sufficient for this application. The η set and the n set were defined according to same criteria of LVQ configuration. For simulations involving MLP networks of two hidden layers, it was defined that second hidden layer presents 30 neurons. This value was specified because it is a smaller number than all those belonging to n set and greater than number of neural network outputs. This value is fixed for all combinations with n set. The used activation function in all neurons is the hyperbolic tangent function. For each combination of training algorithm "versus" number of layers "*versus*" number of neurons "*versus*" learning rate were carried out in 100 training algorithms. Each of them used different initializations of the weights, made over a random uniform distribution between the values [−0.01, 0.01]. This interval of random initialization of weights is justified by the fact that it is smaller than the range of values that comprise the parameters of the training set patterns,

Elements	Symbol	Typical range
No. of hidden layers	Θ	1 e 2
No. of hidden neurons 1° layer	n	60, 90, 120, 150
No. of hidden neurons 2° layer	n_1	30
Learning rate	η	0.01, 0.1 ,0.5, 0.9
Momentum constant	α	0.8

Table 3. Variable elements of the multilayer perceptron.

avoiding the saturation of activation function that prevent the convergence of the neural network [28]. So, it was possible to observe the neural network behavior in relation to training time and generalization capacity, since an adequate set of initial weights allows reduction in training time and high probability of reaching the global minimum of function error. Moreover, this set can significantly improve performance in generalization. Simulated topologies are trained using $\Omega_{4L}^{Tr}, \Omega_{9L}^{Tr}, \Omega_{16L}^{Tr}$ sets and this way, it is verified MLP network response to parameter number increment of the speech signal patterns presented in the original feature space.

4. Experimental results

4.1. Training and validation of LVQ experts

It is shown in **Figure 3(a)** and **(b)**, respectively, the global hit result (in percentage) of the commands in training and validation in relation to the n neuron set simulated to original training set Ω_{NL}^{Tr} with $N = 16$. It was observed that, by using the patterns C_{16}^{jm}, the mean of global training hit increased over the experiments using patterns with four and nine parameters, reaching 97.5%. The result of global validation hit mean for this experiment was 91.45%.

4.2. Individual test of the LVQ experts

In view of these results of training and validation, the topologies for each expert that presented global validation hit greater than 80% were tested. Besides the criterion of the value of the global validation hit for the application of the tests, the choice of a simple topology with the acceptable validation error is also necessary. Consequently, through the training and validation results, the LVQ expert neural networks with 60 neurons in the competitive layer were chosen for the individual test step. The individual test step has the objective of verifying the expert networks generalization capacity for classes that they were trained. From the results achieved in this step, a classification threshold for outputs of each expert was defined. The information of classification threshold is part of the decision rules of the multilevel speech recognition system with mixture expert neural networks. The established criteria for choice of the best topology were applied for each experiment carried out in the training step. The test sets Ω_{N3000}^{TM} and Ω_{N3000}^{TF} with $N = \{4, 9, 16\}$ were applied to the topologies in the three experiments performed. The individual classification results tests applied to topologies that presented global validation hit above 80% and lower topological complexity (60 neurons) for the training sets using the original patterns C_4^{jm}, C_9^{jm} and C_{16}^{jm} are shown in **Table 4**, where it is observed the results of global hit for each expert,

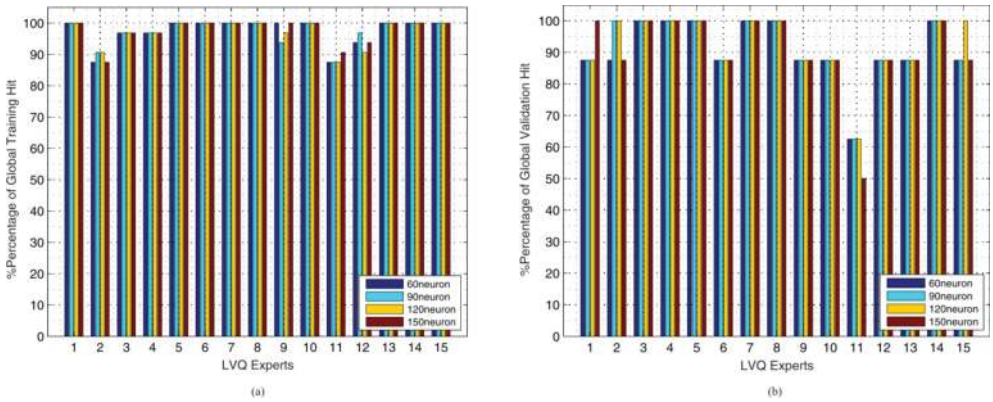

Figure 3. LVQ experts C_{16}^{jm}: result of global training and validation hit.

Experts	C_4^{jm}			C_9^{jm}			C_{16}^{jm}		
	% Global Hit	% Hit Output Class 1	% Hit Output Class 2	% Global Hit	% Hit Output Class 1	% Hit Output Class 2	% Global Hit	% Hit Output Class 1	% Hit Output Class 2
1	82.1	85.3	78.9	95.8	93.2	98.4	90.3	85.3	95.3
2	79.5	88.9	70.0	91.6	93.7	89.5	86.1	94.7	77.4
3	99.2	98.9	99.5	99.5	100	98.9	100	100	100
4	91.1	83.2	98.9	94.5	94.2	94.7	94.2	93.2	95.3
5	91.8	91.6	92.1	95.5	95.3	95.8	97.1	94.2	100
6	88.9	85.8	92.1	87.4	93.2	81.6	82.4	87.9	76.88
7	93.2	88.9	97.4	94.2	90	98.4	99.5	100	98.9
8	97.1	96.3	97.9	97.1	95.3	98.9	99.7	99.5	100
9	77.9	70.0	85.8	72.4	71.6	73.2	93.9	96.3	91.6
10	82.9	75.3	90.5	85.8	93.7	77.9	92.4	92.1	92.6
11	75.0	90.0	60.0	73.4	86.8	60	72.1	65.8	78.4
12	82.1	76.3	87.9	73.7	54.2	93.2	73.4	69.5	77.4
13	99.2	98.4	100	99.5	98.9	100	99.7	99.5	100
14	97.1	98.9	95.3	97.6	97.9	97.4	95.8	96.8	94.7
15	75.8	72.1	79.5	67.9	60.5	75.3	71.3	62.1	80.5

Table 4. Individual test of the expert LVQ with 60 neurons.

as well as individual results of their classes. It is assumed that the values % Hit Output Class 1 and % Hit Output Class 2 constitute the classification threshold for each command

4.3. Training and validation of MLP experts

At the end of all simulations that combine topological elements and training algorithms and number of hidden layers, it can be observed the behavior of the proposed topologies and

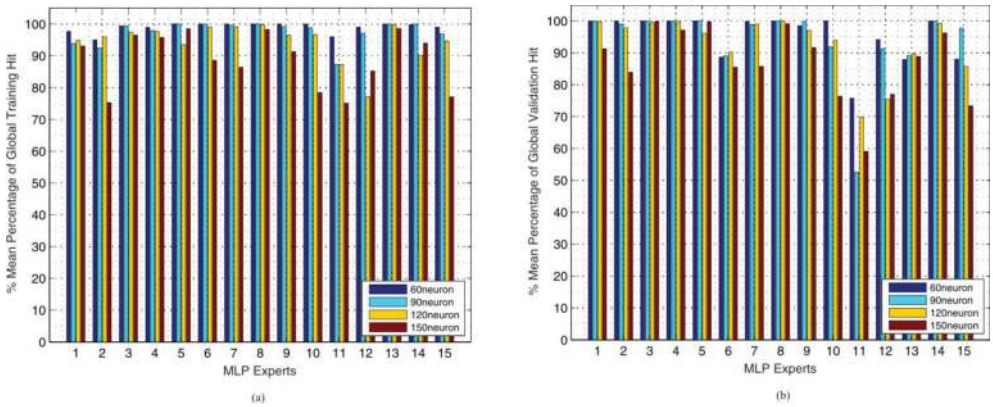

Figure 4. MLP expert C_{16}^{jm}: result of mean training and validation global hit - 1 hidden layer RP algorithm.

Experts	C_4^{jm}			C_9^{jm}			C_{16}^{jm}		
	% Global Hit	% Hit Output Class 1	% Hit Output Class 2	% Global Hit	% Hit Output Class 1	% Hit Output Class 2	% Global Hit	% Hit Output Class 1	% Hit Output Class 2
1	97.6	95.8	99.5	98.4	97.4	99.5	98.7	97.9	99.5
2	80.5	83.2	77.9	87.4	93.7	81.1	86.3	95.8	76.8
3	100	100	100	100	100	100	100	100	100
4	98.7	97.9	99.5	96.3	93.2	99.5	98.7	97.9	99.5
5	96.8	93.7	100	97.1	94.2	100	97.4	94.7	100
6	93.7	87.9	99.5	94.7	95.8	93.7	90.3	87.4	93.2
7	96.9	93.8	100	96.8	98.4	95.3	98.4	98.4	98.4
8	97.6	95.3	100	98.2	96.3	100	97.4	94.7	100
9	81.8	72.1	91.6	90.8	87.4	94.2	98.7	99.5	97.9
10	90	81.6	98.4	93.7	88.4	98.9	92.1	84.2	100
11	76.1	73.2	78.9	85.5	86.3	84.7	90.5	88.9	92.1
12	85.8	76.3	95.3	81.6	82.1	81.1	98.7	98.9	98.4
13	100	100	100	99.7	99.5	100	99.7	99.5	100
14	98.7	97.9	99.5	98.9	97.9	100	99.2	98.4	100
15	77.4	74.2	80	78.2	75.8	80.5	77.6	72.1	83.2

Table 5. Individual test of the expert MLP with 60 neurons.

define the best result. It was verified during the simulations that the GD, GDM and LM algorithms did not reach good results for the problem of pattern recognition with the proposed coding, showing global results of training and validation of less than 50%. In addition, the MLP networks trained with two hidden layers did not present significant results in relation to trained networks with one hidden layer, which does not justify the increase of complexity of the network structure. For these reasons, only the results presented by networks trained

with the algorithm RP with a hidden layer are presented. The average results of global training and validation hit for each expert are shown in **Figure 4(a)** and **(b)**, respectively. These results were achieved by topologies trained with RP algorithm, one hidden layer and 16 input parameters.

4.4. Individual test of MLP experts

The adopted criteria to application of the tests in LVQ topologies were the same as MLP topologies. The best results (in percentage) found in the tests performed for each expert, considering the networks trained with a hidden layer of 60 neurons by the algorithms RP using C_4^{jm}, C_9^{jm} and C_{16}^{jm} patterns are summarized in **Table 5**.

4.5. Final test of the multilevel speech recognition system with mixture of expert neural networks

At the end of the expert design stage, given by analysis of the LVQ and MLP configurations, and defined the classification threshold for each expert output, it was performed the integration between radial basis functions and MLP and LVQ topologies with the best classification results. The flowchart of final test is presented in **Figure 5**.

Patterns from particular class are initially classified through the responses given by RBFs. The RBF that has the highest probability value at its output direct at which expert those patterns should be applied. It is highlighted that obtained results in this step are the same when it has used both MLP and LVQ networks as experts, since the test patterns are the same and

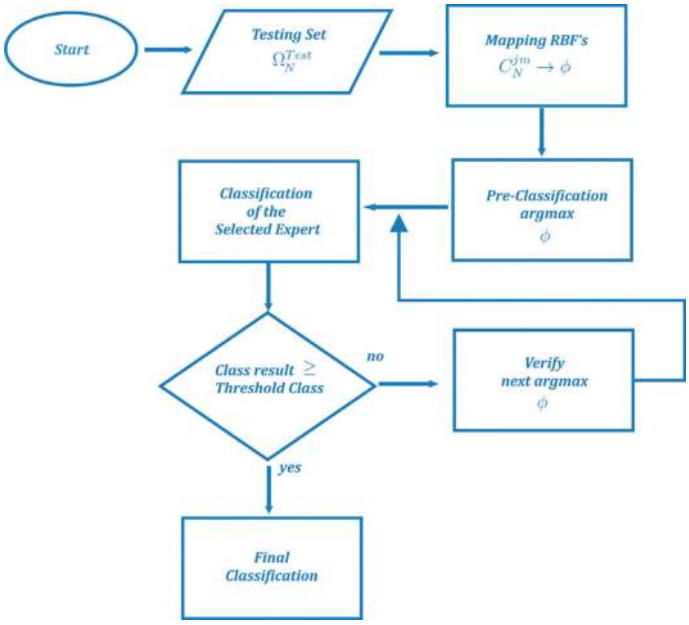

Figure 5. Final test flowchart of the multilevel speech recognition system.

Test class	RBF/%Max Pr/expert	Test class	RBF/%Max Pr/expert	Test class	RBF/%Max Pr/expert
'ZERO'	'ZERO'/80.5/1	'ABAIXO'	'ABAIXO'/57/6	'INICIAR'	'INICIAR'/71.5/11
'UM'	'UM'/54.5/1	'ABRIR'	'ABRIR'/78/6	'LIGAR'	'LIGAR'/48.5/11
'DOIS'	'DOIS'/87/2	'ACIMA'	'ACIMA'/61/7	'MAXIMO'	'MAXIMO'/93/12
'TRÊS'	'TRÊS'/58/2	'AUMENTAR'	AUMENTAR'/70.5/7	'MÉDIO'	'MAXIMO'/40.5/12
'QUATRO'	'QUATRO'/61/3	'DESLIGAR'	'DESLIGAR'/47/8	'MINIMO'	'MINIMO'/41.5/13
'CINCO'	'CINCO'/85.5/3	'DIMINNUIR'	'SETE'/67/4	'PARA TRÁS'	'PARA TRÁS'/35.5/13
'SEIS'	'SEIS'/70.5/4	'DIREITA'	'DIREITA'/32/9	PARA FRENTE'	'PARA FRENTE'/48/14
'SETE'	'SETE'/ 89.5/4	'ESQUERDA'	'MAXIMO'/56/12	'PARAR'	'SETE'/33.5/4
'OITO'	'OITO'/79/5	'FECHAR'	'FECHAR'/83/10	'REPOUSAR'	'REPOUSAR'/62.5/15
'NOVE'	'NOVE'/78/5	'FINALIZAR'	'FINALIZAR'/68.5/10	'SALVAR'	'SALVAR'/85/15

Table 6. Preclassification of test C_{16}^{fm} patterns.

the RBF are fixed. Therefore, after preclassification, the next level is the final classification of the selected expert network. The expert makes the patterns classification mapped into high dimensional space by the RBFs on preclassification stage. The obtained classification result by the expert is compared to the classification threshold of the respective class, determined in individual test step. At this point, the decision rule for final result of the system is carried out as shown in **Figure 5**. The preclassification results of the test patterns generated by DCT matrices of order 4 are presented in **Table 6**, where *RBF* indicates the Gaussian RBF preclassification and *%MaxPr* means maximum probability value in percent.

From preclassification results shown in **Table 6**, it is observed that this step selects (in great majority) the correct experts in second level of classification. Hence, the hit average rates in the

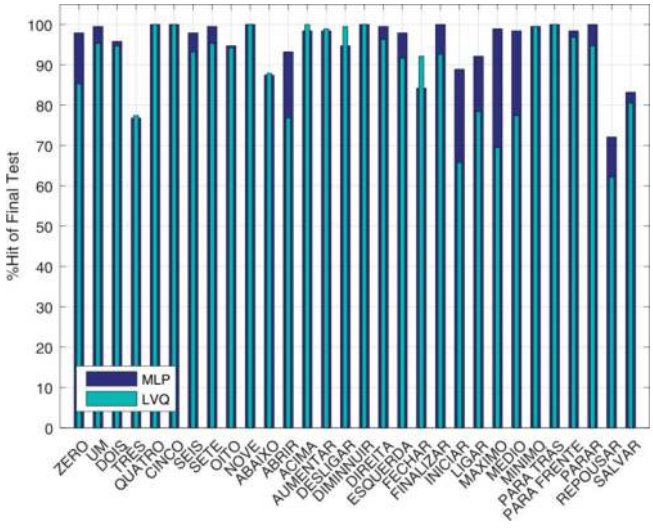

Figure 6. Comparison between MLP and LVQ using C_{16}^{fm} in final test of multilevel speech recognition system.

preclassification step by radial basis functions are for test patterns in the low-dimensionality space C_4^{jm}, C_9^{jm} e C_{16}^{jm} of 83.33, 86.33 and 86.33%, respectively. The test algorithm solved problem for the classes that presented error in preclassification through decision rule. In **Figure 6**, the performance analysis is observed between the MLP and LVQ configurations for composition of expert set using C_{16}^{jm}. Similar results were obtained for the other patterns used.

5. Conclusion

In this chapter, it was proposed to evaluate the performance between the MLP and LVQ neural networks configurations to determine the set of expert classifiers to compose a multilevel recognition system. The developed methodology associates the efficient coding of the speech signal through DCT two-dimensional time matrix of low order with integration between MLP and LVQ expert neural networks and Gaussian radial basis functions to develop a speech recognition system of high performance. In view of the presented results, it was concluded that the parameterization of the speech signal through the generation of the DCT two-dimensional time matrix proposed in the methodology proved to be efficient in the formation of the set of input patterns. They were modified by a Gaussian radial basis functions set parameterized with centroid and variances of the classes and they are the inputs presented to the neural networks during the training and validation step. It was verified that despite the small number of parameters that constitute a speech signal pattern, the two-dimensional time matrix can represent the long-term variations of the locutions spectral envelope to be recognized and these characteristics are reproduced in proposed multidimensional space. The versatility of the Gaussian radial basis function set in proposed recognition system structure demonstrates the potential of these functions. It is emphasized that the parameters of the RBF models were adequately determined, since hit rate in preclassification step was higher than 80%. It was verified that the increase in neurons number of the MLP and LVQ neural networks did not show significant improvements in the global validation hit, which was the criterion used to select the best topologies for the application of the tests. Based on the tests carried out, it was verified that the LVQ network can be used satisfactorily in pattern recognition problems, specifically for multilevel speech recognition system proposed in this chapter. This is evidenced by the very close performance of the MLP Network, which is widely used in pattern classification. Finally, the performance in the multiclass task of speech signal patterns given by the integration between Gaussian radial basis functions and set of expert neural networks is highlighted.

Author details

Washington Luis Santos Silva[1]*, Priscila Lima Rocha[2] and Allan Kardec Duailibe Barros Filho[2]

*Address all correspondence to: washington.silva@ifma.edu.br

1 Federal Institute of Maranhão, São Luís, Brazil

2 Federal University of Maranhão, São Luís, Brazil

References

[1] Dougherty G. Pattern Recognition and Classification: An Introduction. Illustrated ed. New York, USA: Springer Science & Business Media; 2012. 196 p. DOI: 10.1007/978-1-4614-5223-9

[2] Duda R, Hart P, Stork D. Pattern Classification. 2nd ed. New York, USA: John Wiley & Sons; 2012.680 p. ISBN: 111858600X, 9781118586006

[3] Husnjak S, Perakovic D, Jovovic I. Possibilities of using speech recognition systems of smart terminal devices in traffic environment. Procedia Engineering. 2014;**69**:778-787. DOI: 10.1016/j.proeng.2014.03.054

[4] Spale J, Schweize C. Speech control of measurement devices. IFAC-Papers OnLine. 2016;**49**:13-18. DOI: 10.1016/j.ifacol.2016.12.003

[5] Weng F et al. Conversational in-vehicle dialog systems: The past, present, and future. IEEE Signal Processing Magazine. 2016;**33**:49-60. DOI: 10.1109/MSP.2016.2599201

[6] Youcef C, Elemine M, Islam B, Farid B. Speech recognition system based on OLLO French corpus by using MFCCs. In: Chadli M, Bououden S, Zelinka I, editors. Recent Advances in Electrical Engineering and Control Applications. Cham: Springer; 2017;**411**:326-331. DOI: 10.1007/978-3-319-48929-2_25

[7] Bellegarda J, Monz C. State of the art in statistical methods for language and speech processing. Computer Speech & Language. 2016;**35**:163-184. DOI: 10.1016/j.csl.2015.07.001

[8] Picone J. Signal modeling techniques in speech recognition. Proceedings of the IEEE. 1993;**81**:1215-1247. DOI: 10.1109/5.237532

[9] Kautz T, Eskofier B, Pasluosta C. Generic performance measure for multiclass-classifiers. Pattern Recognition. 2017;**68**:111-125. DOI: 10.1016/j.patcog.2017.03.008

[10] Song Q, Jiang H, Liu J. Feature selection based on FDA and F-score for multi-class classification. Expert Systems With Applications. 2017;**81**:22-27. DOI: 10.1016/j.eswa.2017.02.049

[11] Kheradpisheh S, Sharifizadeh F, Nowzari-Dalini A, Ganjtabesh M, Ebrahimpour R. Mixture of feature specified experts. Information Fusion. 2014;**20**:242-251. DOI: 10.1016/j.inffus.2014.02.006

[12] Shih P, Chen C, Wu C. Speech Emotion recognition with ensemble learning methods. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); New Orleans, LA; 2017. pp. 2756-2760

[13] Xie F, Fan H, Li Y, Jiang Z, Meng R, Bovik A. Melanoma classification on dermoscopy images using a neural network ensemble model. IEEE Transactions on Medical Imaging. 2017;**36**:849-858. DOI: 10.1109/TMI.2016.2633551

[14] Haykin S. Neural Networks and Learning Machines. 3rd ed. New Jersey, USA: Pearson Education; 2011. 936 p. ISBN: 9780131471399

[15] Silva I, Spatti D, Flauzino R. Redes Neurais Artificiais para Engenharia e Ciências Aplicadas - Curso Prático. São Paulo, Brasil: Artliber; 2010. 399 p. ISBN: 9788588098534

[16] Rocha P, Silva W. Artificial neural networks used for pattern recognition of speech signal based on DCT parametric models of low order. In: IEEE 14th International Conference on Industrial Informatics (INDIN); 19-21 July 2016; Poitiers. New York. IEEE; 2017

[17] Rocha P, Silva W. Intelligent system of speech recognition using neural networks based on DCT parametric models of low order. In: International Joint Conference on Neural Networks (IJCNN); 24-29 July 2016; Vancouver

[18] Silva W. Intelligent genetic fuzzy inference system for speech recognition: An approach from low order feature based on discrete cosine transform. Journal of Control, Automation and Electrical Systems. 2014;**25**:689-698. DOI: 10.1007/s40313-014-0148-0

[19] Bresolin A. Reconhecimento de Voz através De Unidades Menores Do Que a Palavra, Utilizando Wavelet Packet e SVM, Em Uma Nova Estrutura hierárquica de decisão [Thesis]. Natal: Universidade Federal do Rio Grande do Norte; 2008

[20] Siniscalchi S, Svendsen T, Lee C. An artificial neural network approach to automatic speech processing. Neurocomputing. 2014;**140**:326-338. DOI: 10.1016/j.neucom.2014.03.005

[21] Buhmann M. Radial Basis Functions: Theory and Implementations. Vol. 12. Cambride, United Kingdom: Cambridge University Press; 2003. 259 p. ISBN: 0-521-63338-9

[22] Hu Y, Hwang J editors. Handbook of Neural Networks for Speech Processing. New York, USA: CRC Press; 2014. 408 p. ISBN: 9780849323591

[23] Priddy K, Keller P. Artificial Neural Networks: An Introduction. Illustrated ed. Washington, USA: SPIE Press; 2005. 165 p. ISBN: 0819459879

[24] Rokach L. Pattern Classification Using Ensemble Methods. Vol. 75. Singapore: World Scientific; 2010. 225 p. ISBN: 13 978-981-4271-06-6

[25] Zhou Z. Ensemble Methods: Foundations and Algorithms. Illustrated ed. New York, USA: CRC Press; 2012. 236 p. ISBN: 978-1-4398-3003-1

[26] Fissore L, Laface P, Ravera F. Using word temporal structure in HMM speech recognition. In: EEE International Conference on Acoustics, Speech, and Signal Processing; 21-24 April 1997; Munich. New York: IEEE; 2002. pp. 975-978

[27] Bhardwaj A, Tiwari A, Bhardwaj H, Bhardwaj A. A genetically optimized neural network model for multi-class classification. Expert Systems with Applications. 2016;**60**:211-221. DOI: 10.1016/j.eswa.2016.04.036

[28] Sousa C. An overview on weight initialization methods for feedforward neural networks. In: Proceedings of the international joint conference on neural networks (IJCNN 2016); 24-29 July 2016; Vancouver. New York: IEEE; 2016. pp. 52-59