Journal of New Technology and Materials
JNTM
Vol. 05, N°01 (2015)26-30

OEB Univ. Publish. Co.

Voice Recognition Technology Using Neural Networks

Abdelouahab Zaatri[1], Norelhouda Azzizi[2] and Fouad Lazhar Rahmani[2]

[1]*Department of Mechanical Engineering, Faculty of Engineering Sciences,*
azaatri@yahoo.com
[2]*Department of Mathematics, Faculty of Exact Sciences,*
azzizinorelhouda@yahoo.fr
[2]*Department of Mathematics, Faculty of Exact Sciences,*
flrahmani@hotmail.com
University of Constantine1, Constantine, Algeria

Received date: April 14, 2015; revised date: May 25, 2015; accepted date: May 29, 2015

Abstract

This paper presents the use of a Multi-Layer Perceptron Neural Nets (MLP-NN) for voice recognition dedicated to generating robot commands. Our main goal concerns the estimation of the minimal number of elements required for the learning process in order to ensure an acceptable success of the neural nets recognition system. As the MLP requires references for the spoken words, we have provided these references by the means of a supervised classifier based on the mean square error.

An experimental approach has been followed for the design of experiments enabling to determine the minimal elements in the sample for each voice command. Satisfactory results have been obtained leading to a better understanding of variability of the system functioning. Finally, we have noticed that the success rate of the MLP and the minimal number of elements used for the learning process depend on the spoken word structure and of the variability of the situation (word length, noise, speaker, etc).

Keywords:design experiments, MLP, neural networks, speech recognition, supervised learning, VQ-LBG algorithm;

1. Introduction

Speech recognition is an important tool for control and interaction with modern robots. However, because of the complex nature of voice signal, the speech recognition still remains a hard issue. Most speech recognition systems use a learning process to identify the correct response of a spoken command. In this context, an interesting issue concerns the design experiments to reduce the data used for the learning phase. Compared to the design experiments in the case of discrete data, NN Model can be used for estimating the output of nonlinear systems in the case of noisy and sensitive process to various parameters such as speech recognition.

The field of automatic speech recognition (ASR)[1]is divided into four areas: recognition of isolated words recognition of chained words, continues recognition and speech understanding with a limited vocabulary and syntax. For our application; we are concerned with the recognition of isolated works that will be used as robot commands. [1].There exist different methods of speech recognition of isolated words using methods such as Hidden Markov Model [2,3], the Gaussian mixture models, VQ vector quantification [4,5],and NN(MLP)[6,7], etc. Concerning, the NN, we have remarked the use of self organizing Map[8], Waibel's Time Delay NN [9], Perceptron and Recurrent NN [10]. The multilayer Perceptron (MLP) is of a particular importance for acoustic modelling in ASR [7].

On the other hand, a survey of literature related to applications of NN applied to design experiments shows that they can be used to model complex non-linear and noisy processes[11,12].

In this paper, we intend to exploit MLP-NN for design experiments in order to determine the minimal number in a sample to reducing the data, time and cost used for the learning phase process[13]. The estimation of the reference words (robot commands) are obtained by a supervised classifier based on the minimization of the mean square error. These references words are stored into the dictionary and

used by the MLP to compare a pronounced word with a desired one.

We have tested this type of commands for a various kind robots including: mobile robot, serial robot manipulator and cable based robot.

2. The Initial Word Recognition System

The principle used for most Word recognition Systems can be illustrated in figure 1. It comprises two phases: the recognition phase and the learning phase. The learning phase consists of creating a list of words which are stored into a dictionary as reference words. The recognition phase consists of identifying a spoken unknown word to one of the reference words stored in the dictionary[14].

We have implemented a word recognition system based on the following procedure: Any spoken word which is a continuous acoustic signal is translated by the microphone into an electric continuous signal.
This continuous electrical signal is then digitalized (sampled) by the sound card. Some digital operations are applied such as pre-emphasis, short-time Fourier analysis (FFT), power spectrum, filter bank integration (Mel's Filter), logarithmic compression, Discret Fourier transform. Some of these operations are applied to the spoken word START as shown in Figure 1. In this Figure 1-a represents the spoken word converted into an electrical signal by the microphone.
Figure 1-b represents the positive envelope of the electrical signal. Figure 1-c represents the detection of amplitude variation of this signal as on-off levels. Figure 1-d represents the detection of beginning and end of the spoken words.
The final output is a set of coefficients which are called Mel frequency Cepstral coefficients MFCC.The MFCCis technique to extract features from thespeech signal and compare the unknown words with some reference words stored in a database.
TheMFCC are based on the known variation of the human ear'scritical bandwidth frequencies with filters spaced linearly atlow frequencies and logarithmically at high frequencies usedto capture the important characteristics of speech. Studieshave shown that human perception of the frequency contentsof sounds for speech signals does not follow a linear scale.Thus for each tone with an actual frequency, *f*, measured inHz, a subjective pitch is measured on a scale called the Melscale. The Mel-frequencyscale is linear frequency spacingbelow 1000 Hz and a logarithmic spacing above 1000 Hz. Asa reference point, the pitch of a 1 kHz tone, 40 dB above theperceptual hearing threshold, is defined as 1000 Mels[15].
Vector quantization (VQ) is a lossy data compression method based on the principle of block coding. It is a fixed-to-fixed length algorithm. In the earlier days, the design of a vector quantizer (VQ) is considered to be a challenging problem due to the need for multi-dimensional integration. In 1980, Linde, Buzo, and Gray (LBG) proposed a VQ design algorithm based on a training sequence. The use of a training sequence bypasses the need for multi-dimensional integration. A VQ that is designed using this algorithm are referred to in the literature as an LBG-VQ [16]. The algorithm requires an initial codebook$C^{(0)}$. This initial codebook is obtained by the fractionation method (splitting) .In this method, an initial code vector is set as the average of the entire training sequence. This code vector is then split into two. The iterative algorithm is run with these two vectors as the initial codebook. The last two code vectors are divided in four and the process is repeated until the desired number of code vectors is obtained [14].

We used the VQ-LBG to reduce MFCC data from (12*128) to (12*32) coefficients. Figure 2 shows the electrical form of the spoken word START as well as its representation as MFC Coefficients and their compression into centroids[14].

The estimation of the reference words (robot commands) are obtained by a supervised classifier based on the minimization of the mean square error. These references words are stored into the dictionary and used by the MLP to compare with a pronounced word.

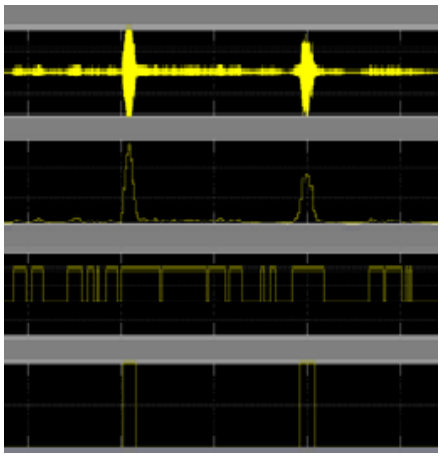

Figure 1: steps and procedure of treatment and detection of each spoken word

Figure1 represent an example of application implemented under matlab software.

3. MLP for Word Recognition

The technique of NN is used in several areas such as classification, pattern recognition (image, voice, ect.) and process control. In our work, we replaced the classifier by an MLP for voice recognition [11, 17,18].

The role of the MLP classifier is to select the most similar reference word with respect to an unknown word. The choice is based on the calculation of the distance between the unknown word and all the reference words (nearest neighbor) [17,7]. The scheme of a voice recognition system is given in Figure 2.

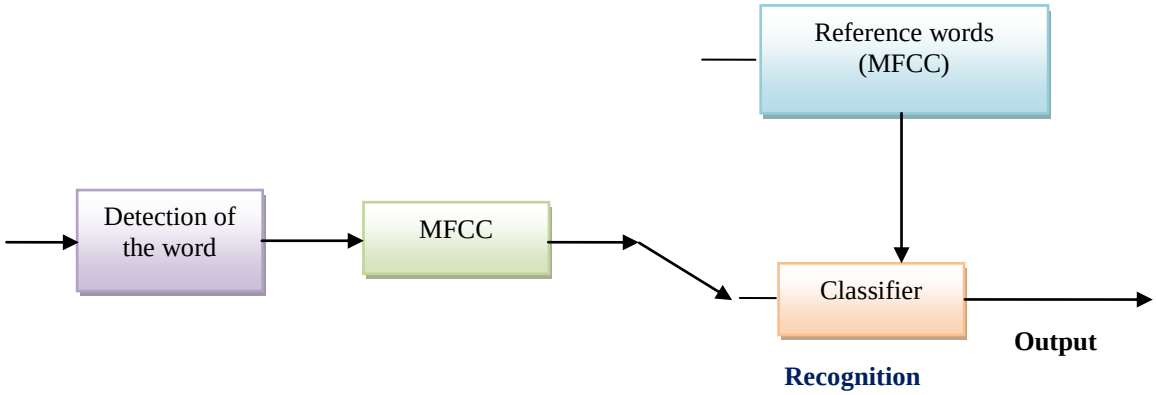

Figure 2: Voice recognition system

For our speech recognition system, we have chosen an FFT resolution of 1024 points. The result is an MFCC coefficients matrix of dimension $12 \times j$, where the value of j depends on the length of the spoken word, on the sampling frequency of the sound card and on the resolution of the FFT. The system is tested on a dictionary of four commands (START, STOP, UP, DOWN). The MFCC matrix is compressed into a The MFCC matrix is compressed into a matrix of (12 × 32) centroid coefficients. For the given commands;
We have the following dimensions (Table1):

Command	MFCC	MLP
START	12 × 128	12 × 32
STOP	12 × 128	12 × 32
UP	12 × 128	12 × 32
DOWN	12 × 128	12 × 32

Table 1: Dimensions Matrix of MFCC and MLP inputs

The implementation of the MLP was carried out by using the NN toolbox of Matlab software. Our MLP is a NN format; it is composed of an input layer and an output layer with one hidden layer in between (Figure 3). The input data of the MLP are the MFCC which are recorded into a file in a form of a matrix named "sepstr.mat". The MLP uses 12*32 neurons for the input layer.

The reference word was determined from the previous process. A supervised training was adopted comparing actual spoken words with those stored on the dictionary. After the achievement of the learning process, the obtain hidden layer derived from Matlab tool is constituted of 32 neurons.

The output layer is constituted of 4 neurons which corresponds to the reference words stored on the dictionary (START, STOP, UP, DOWN).

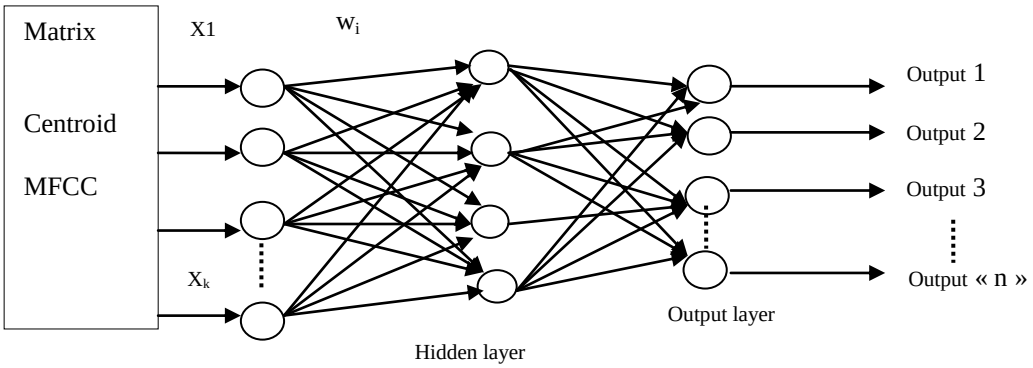

Figure 3: MLP voice recognition system

4. Experimental Results

It is important to analyze the evolution and the convergence of the learning process with respect to the number of experiments. To get an estimation of the required minimal number of elements in a learning tests for a given reference word during the learning phase; we have adopted an experimental approach. This leads to reduce the computation time.
Under Matlab software, the learning phase for the MLP was tested as follows. Each learning test corresponds to a certain number of trials N of learning experiments using the same word. For each reference word, the mean square error of MFCC is computed with respect to the number of trials N as

$$mse(N) = \frac{\Sigma(a_{ij}^{*} - a_{ij})^2}{N}$$ using the same word.

For different learning test, the mse(N) is recorded.
On the other hand, we have applied various approximations functions in order to obtain an appropriate form of the evolution of the learning process. As a result, we have noticed that the most appropriate approximation of the learning process is a bi-exponential function in the form of: f(x) = a*exp(b*x) + c*exp(d*x).

We present graphically two examples illustrating the learning process. The first example shows the evolution of the learning process with respect to the number of trials for the word STOP. Figure 4-a shows the electrical signal of the word STOP. Table 2 shows the experimental results of 9 learning tests
with respect to the number of trials for the word STOP.
As shows in Figure 4-b, the analysis of these experiments shows that the mse(N) decreases with the number of trials; improving therefore the learning process.
For the example at hand, the bi-exponential approximation function is given by the Curve Fitting toolbox of Matlab Software as:f(x) = a*exp(b*x) + c*exp(d*x) with the following coefficients:
a = 2.019e+004 (-2.852e+014, 2.852e+014)
b = -0.0634 (-5413, 5413)
c = -2.019e+004 (-2.852e+014, 2.852e+014)
d = -0.0634 (-5413, 5413)
(with 95% confidence bounds), Goodness of fit: SSE(0.01536), R-square(0.984),Adjusted R-square(0.9787), RMSE(0.04131).

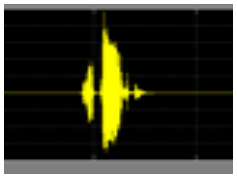

Figure4-a: example of the spoken word STOP

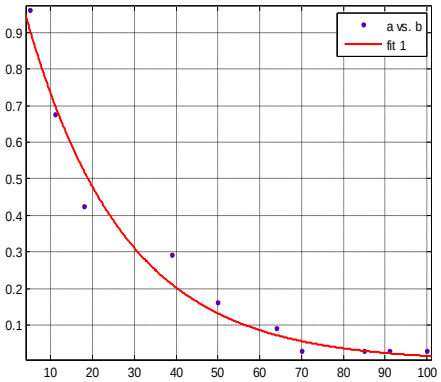

Figure 4-b: word STOP Error distribution with the growth in the number of test

The second example concerns the training process for identifying the word UP with respect to the number of trials. We have used 10 learning tests. As shows in Figure 5, the analysis of these experiments shows that the mse(N) decreases with the number of trials,

Learning tests	mse(N)	N (trials)
1	0,9094	10
2	0,3313	23
3	0,1686	36
4	0,1195	41
5	0,0674	54
6	0,0490	56
7	0,0244	61
8	0,0209	59
9	0,0191	60

Table2: Experimental learning tests mse(N)

the bi-exponential approximation function is given by the Curve Fitting toolbox of Matlab Software as the following coefficients:
General model:Exp2: f(x) = a*exp(b*x) + c*exp(d*x)
a =2.073e-008 (-6.453e-006, 6.495e-006)
b =0.1348 (-3.013, 3.282)
c =1.129 (0.9298, 1.329)
d = -0.04309 (-0.05608, -0.03011)
(With 95% confidence bounds), Goodness of fit: SSE: 0.02093, R-square: 0.9778, Adjusted R-square: 0.9667, RMSE: 0.05906

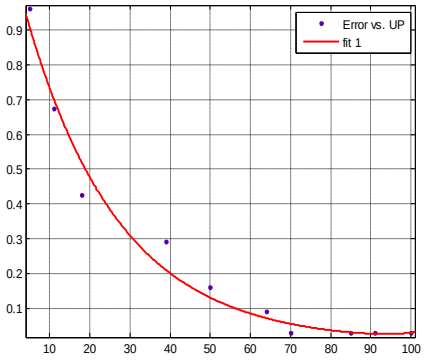

Figure 5: Word UP Error distribution with the growth in the number of test

After testing experimentally the MLP with some spoken robot commands (START, STOP, UP, DOWN), We have remarked that this minimal number depends on the structure of the spoken word itself, on the speaker, on the used equipment and on the environment noise.

We can also notice promising results while testing this type of commands for various kind of robots such as those experimental systems developed in our laboratory including: mobile robots, serial robot manipulators and cable based robots.

5. Conclusion

We have presented an experimental technique for design experiments to estimate the minimal number that should compose a learning test to ensure an acceptable performance for a learning process of supervised neural networks dedicated to speech recognition used for robot commands.

We have initially developed a system of word recognition based where any spoken word is processed and translated into a set of coefficients which are the Cepstral coefficients (MFCC). Then, these MFCC coefficients are compressed to centroids by the VQ-LBG algorithm based on the mean square error.

Neural networks are a technique to analyze and make an estimate of the output of a nonlinear system in the case of a random process. However; the MLP requires reference words. For each spoken word, its reference has been obtained by calculating the mean value of its MFCC Centroids Coefficients. These reference words have been used as words models to train a supervised learning NN of type MLP.

We have experimentally tested the MLP with some spoken robot commands and we have obtained an estimation of the required minimal number of elements in a learning test to ensure an acceptable learning process. We have remarked that this minimal number depends on the structure of spoken word itself, on the used equipment and on the environment noise. We have remarked that we can approximate the learning process by a bi-exponential function.

References

[1] D. Paul, R. Parekh, Automated speech recognition of isolated words using neural networks, International Journal of Engineering Science and Technology, (IJEST), 3(6), 2011, 4993-5000.

[2] N. Shokhirev, Hidden Markov Models, 2010.

[3] L.R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in Speech Recognition", Proceedings of the IEEE Journal, Feb 1989, vol. 77, Issue: 2.

[4] R. M. Gray,Vector Quantization, IEEE ASSP Magazine, pp. 4--29, (April 1984).

[5] Y. Linde, A. Buzo, and R. M. Gray,An Algorithm for Vector Quantification Design, IEEE Transactions on Communications, January 1980, pp. 702--710

[6] B. GOSSELIN, Application de réseaux de neurones artificiels a la reconnaissance automatique de caractères manuscrits, Doctoral, Thesis, 1996.

[7] J. Praveen Pinto Multilayer Perceptron Based Hierarchical Acoustic Modelling for Automatic Speech Recognition, These N°4649,Lausanne, EPFL (2010).

[8] T. Kohonen, Self-Organizing Maps, Springer-Verlag, 1995.

[9] K. J. Lang and A. H. Waibel, A Time-Delay Neural network Architecture for Isolated Word Recognition, Neural networks, vol. 3, 1990.

[10] K-I Funahashi and Y. Nakamura, "Approximation of Dynamical Systems by Continuous Time Recurrent Neural Networks," Neural Networks, vol. 6, 1993.

[11] S. Haykin, Neural Networksa Comprehensive Foundation, Second edition, Canada.

[12] P. Zegers ,"speech recognition using neural networks" , Master's Thesis,The university of Arizona, 1998.

[13] J. Freidman, T. Hastie & R.Tibshirani, the elements of Statistical Learning, (September 30, 2008).

[14]http://voce.sourceforge.net/files/VoceWhitePaper.pdf

[15] K. Patel, R.K. Prasad, Speech Recognition and Verification Using MFCC & VQ , International Journal of Emerging Science and Engineering (IJESE), ISSN: 2319–6378, Volume-1, Issue-7, May 2013.

[16] http://www.data-compression.com/vq.html

[17] C.M. Bishop, Neural Networks for Pattern Recognition, Aston University, Birmingham, UK, (1995).

[18]R. Low, R. Togneri, Speech recognition using the probabilistic neural network, Proc. 5thInt. Conf. on Spoken Language Processing, Australia, 1998.