# Using Neural Networks for a Discriminant Speech Recognition System

Daniela ŞCHIOPU, Mihaela OPREA
*Petroleum-Gas University of Ploieşti*
*Blvd. Bucureşti 39, RO-100680 Ploieşti*
*dschiopu@upg-ploiesti.ro, mihaela@upg-ploiesti.ro*

***Abstract* — The paper presents a case study of using neural networks for a discriminant automatic speech recognition system for the Romanian language. The uttered words by several speakers which the system must recognize can be commands for a robot. The study aims to examine the performance of the system by minimizing the error.**

***Index Terms — feature extraction, multilayer perceptrons, recurrent neural networks, speech processing, speech recognition***

## I. INTRODUCTION

Speech recognition is a major problem that needs to be solved during human-computer interaction in several applications (as e.g. automatic call processing in telephone networks, air traffic controllers training, translation and dictation, software control, industrial telematics, e-learning). The progress made in the development of automatic speech recognition (ASR) systems for many spoken languages (English, Japanese, German, Spanish, French, Swedish, Russian, Slovak, etc.), as well as different speech recognition methods and approaches were discussed and reported in the literature (see e.g. [1], [2], [3], [4]). In the last two decades, important achievements were brought to the development of some ASR systems for the Romanian language (examples of recent research results were reported in [5], [6], [7], [8], [9], [10], [11]). The main strategies used for the existing Romanian language ASR systems ([22], [23]) are (1) statistical strategies (based on hidden Markov models - HMMs), (2) connectionist strategies (based on artificial neural networks - ANNs), or (3) hybrid strategies (ANNs-HMMs, fuzzy ANN, fuzzy HMM, etc.).

Although the lack of a large database for the Romanian language has made research quite difficult in this area, recent progress has been made by different research groups. The results reported in the literature vary depending on the vocabulary, strategy used, task of the recognition, etc. For example, the tasks of the ASRS_RL system presented in [6] are digit recognition for telephone dial application and vowel recognition using HMMs, ANNs or hybrid structures. The results for ANNS in form of multilayer perceptron, Kohonen maps, or even hybrid ANN-HMM for these tasks were from 50% to 100%. The performance of the system presented in [22] was improved by adding discriminative training based on the minimum classification error and the result was the increasing the recognition rate from 92.2% to 94.7%.

Our research work is focused on the use of neural networks for a discriminant ASR system in the case of the Romanian language. The aim of the paper is evaluating the performance of the system by minimizing the error, and the results will be evaluated for using them in a control system. Previous work for this purpose is reported in [10] and [11].

The paper is organized as follows. Section II presents some generalities about automatic speech recognition. Section III describes the main neural networks techniques used in ASR, as well as training methods for improving the performance of these networks. Section IV presents a case study about a speech recognition system proposed and the results obtained. The last section is dedicated to the conclusions of the paper.

## II. AUTOMATIC SPEECH RECOGNITION

An automatic speech recognition system takes as input the human utterance (in acoustic form – the speech signal) and provides as output a sequence of meaningful words in a given language by using a computer program.

The main steps followed by a typical ASR system are:

(1) signal processing,

(2) feature extraction and analysis, and

(3) pattern recognition by using a specific recognition approach and some knowledge sources (such as a lexicon and a language model) that facilitates the recognition process.

The general structure of an ASR system is shown in figure 1. Signal processing performs an acoustic signal processing as for example, a white noise filtering (i.e. Fast Fourier Transform, Mels Scale Bank pass Filter and Cepstral Analysis). Feature extraction and analysis makes the extraction of a mel-frequency cepstral coefficients set along with the first and the second order derivatives of these features, and applies a specific feature analysis (e.g. a spectral analysis). During pattern recognition it is realized a pattern matching by using a recognition method, a lexicon, and the language model (the n-gram model, for example).

An ASR system can be discrete or continuous, speaker dependent, independent or adaptive. The performance of an ASR system is influenced by different factors: the speaking style (read speech, planned speech, spontaneous speech), the speaker specificity (pronunciation, variability, dialect), the size of the known vocabulary (from small to very large), the environmental conditions (environmental noise, acoustical distortions), the communication channel quality etc.

A variety of speech recognition methods were developed so far, which can be classified in four main classes: acoustic phonetic, pattern recognition, artificial intelligence based, and hybrid. The pattern recognition approaches include:



978-1-4799-5094-2/14/$31.00 ©2014 IEEE

template methods and stochastic methods (e.g. hidden Markov model), statistical pattern recognition, dynamic time warping (DTW), vector quantization, and support vector machine (SVM).

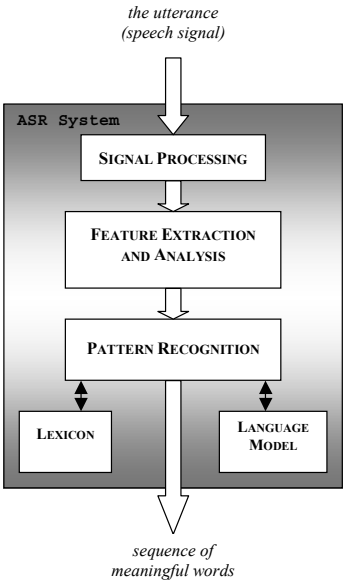


Fig. 1. The general structure of an ASR system

The artificial intelligence based approaches include artificial neural networks (such as Multi-Layer Perceptron – MLP, Kohonen networks, recurrent neural networks), knowledge based systems etc. Other speech recognition methods are provided by hybrid methods that combines different existing methods (e.g. fuzzy MLP, fuzzy HMM, HMM-MLP, etc).

## III. DISCRIMINATIVE TRAINING FOR NEURAL NETWORKS

### A. Neural Networks used in ASR

The process of recognition is to classify the patterns which produce the minimum distance to the input utterance.

Because of their ability to solve classification complex problems, artificial neural networks (ANN) have a wide applicability, including speech recognition.

Modeling the sequential nature of speech signal is one of the major problems of implementing speech recognition systems using neural networks. Thus, the mismatch between fixed size of input pattern and variable length of spectral vectors should be avoided.

A first solution is to set the task (recognition of words or parts of words) and to force the number of the input vectors to a fixed number. In this case there are used multilayer perceptron (MLP) and time delay neural networks (TDNN).

A second approach refers to the processing of each segment of speech into spectral vectors and implementing them sequentially in time, one by one, at the input of the network. For this solution there are used recurrent neural networks (e.g. the Elman network).

For a neural network to be suitable for speech recognition, it must have the following characteristics [21]:

- must contain enough nodes and weights in order to learn the diversity of the input vectors;
- able to retain the temporal relationship between events (spectral coefficients which model the speech signal);
- invariance to vectors' translation in time to achieve better recognition rate;
- the number of weights to be small comparing to the training set;
- training procedure should not be affected by time alignment.

### B. Discriminative Training

There are many techniques for improving performance of the ASR systems. Discriminative training has been used for speech recognition for three decades. The frameworks for discriminative training for ASR were maximum likelihood, mean squared error (MSE), misclassification error, minimum classification error (MCE), the recent research indicates the maximum mutual information (MMI) [12], [13], minimum phone error (MPE) [13], etc.

MSE is the average squared difference between outputs and targets (measures the average of the squares of the errors). Zero value means no error.

Misclassification error indicates the fractions of samples which are misclassified. A value of zero means no misclassification.

MCE is an approach of pattern classifier design known as generalized probabilistic descend.

In the maximum mutual information (MMI) methods, the training is made with a gradient learning approach, in which the parameters are modified so the most of them increase the mutual information [12].

The speech signal is often modeled like a random process due to many factors (the frequency response of audio channel, additive noise), so there are used statistical tools to analyze it. Thus, many models are used to compute the likelihood of the sequence of feature vectors, the likelihood of the corresponding data (e.g. HMM training with expectation maximization algorithm).

Discriminant training for sequence recognition systems is made such that the parameters for the classifier to be trained to distinguish between sample of different classes.

Although the model probabilities are usually estimated from likelihoods using the Bayes rule, some approach estimate the posterior probability directly and incorporate the maximization of these estimates directly in the training procedure. Such an approach is neural network (typically a MLP or a recurrent network).

It has been proved [20] that the outputs of gradient-training classification systems might be interpreted as posterior probabilities of output classes conditioned on the input, under the following conditions:

- The system must contain enough parameters to be trained to a good approximation of the outputs.
- The error criterion for gradient training is MSE or the relative entropy between outputs and targets.
- The system must be trained in the classification mode, so for a number of classes, the target is one for the correct class and zero for all the others.

- The system must be trained to a global error minimum (not achieved in practice).

There are several techniques for improving the performance of MLP [12]. One of these techniques is cross-validation, which determines if sufficient training has occurred during training and not to overtrain the network.

## IV. Case Study

We have implemented in Matlab® [14] a feed forward artificial neural network-based speech recognition system. The system inputs are six words in Romanian language (*start*, *stop*, *sus*, *jos*, *prinde*, *lasă*) uttered by 8 female and 6 male speakers. These Romanian words might be used for a specific domain of application (e.g. automatic control). The system outputs are the six classes; this problem is a classification matter. The database contains 90 samples. The system must recognize the spoken word with the minimum error.

### A. Speech Analysis and Feature Extraction

Essentially, the process of speech recognition for isolated words ASR systems with reduced vocabulary, consists of the following steps (see figure 2):

- collecting the database (speech corpus);
- speech processing;
- acoustic feature extraction;
- decoding (pattern or utterance classification).

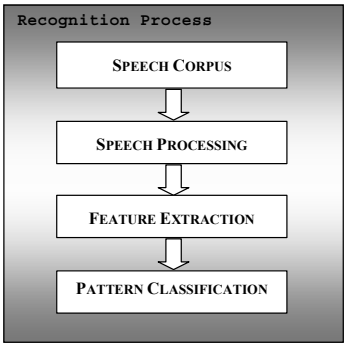


Fig. 2. The modules of an isolated words ASR system

The speech processing module refers to simple preprocessing that can be used to partially offset problems like position of the microphone or the channel characteristics (e.g. adaptively filters).

In our study, the speech data were recorded in a usual room and digitized at sample frequency 16 kHz.

The signal was pre-emphasis with a coefficient 0.97 for increasing high frequencies; then the signal was divided into fixed-duration frames. On these frames, the speech signal is considered quasi-stationary. Feature extraction is performed for each frame separately. From each frame we computed log-energy, each frame being windowed with a Hamming window (25 ms).

Acoustic feature extraction consists of computing representations of the speech signal that are robust to acoustic variations, but sensitive to linguistic content [12]. The role of this module is obtaining from speech a set of parameters (named features) that don't vary much when the same words are spoken many times. These parameters are usually computed at fixed-time intervals.

There are successfully used a vast range of feature extraction techniques for the speech recognition task, some of these techniques being: mel-frequency cepstral coefficients (MFCC) [15], linear predictive coding coefficients (LPC) [16], perceptual linear prediction coefficients (PLP) [17], as well as normalizations via cepstral mean substraction (CMS) [18], relative spectral filtering (RASTA), vocal tract length normalization (VTLN) [19]. We used for our study mel-frequency cepstral extraction. Thus every speech data was converted into a set of 13 feature vectors: 12 MFCC parameters along with their log-energy.

### B. Building the Network

For neural modeling, we build first a two-layer feed forward network, Multi-Layer Perceptron (MLP) total connected. The network has an input layer with 13 neurons (the number of parameters obtained in the feature extraction step), one hidden layer (80 neurons) and an output layer (6 neurons, each one for every class). The number of neurons of the hidden layer was determined experimentally. Also, we have tested a MLP with 60, 70 or 100 neurons on the hidden layer. The results are presented in Table I. The MLP has to classify the uttered word.

The structure of the MLP is shown in figure 3.

We used 60% utterances for training the MLP, 20% utterances for validation and 20% for testing the network.

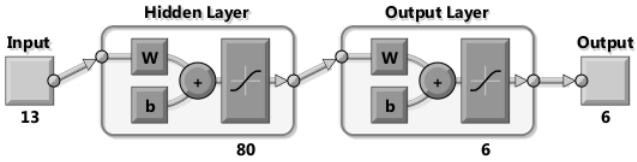


Fig. 3. The architecture of the feedforward neural network

The MLP parameters are:

- 13 acoustic vectors;
- training epochs;
- learning rate 0.05;
- scaled conjugate gradient backpropagation (for training);
- mean squared error MSE (for evaluating the performance).

Also, we built a simple recurrent neural network (Elman network). Elman neural networks are semi-recursive neural networks using the backpropagation learning algorithm for finding patterns in a sequence of value. The architecture for this network is defined in figure 4.

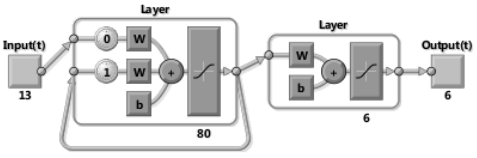


Fig. 4. The architecture of the Elman network

### C. Results

The results obtained in the experiments are summarized in Table I, figure 5 and figure 6. The training criterion was mean squared error. MSE was minimum in the case of training with 80 neurons on the hidden level. In this case the performance was obtained after 223 iterations. Figure 4 shows how MSE decreases with the epochs.

TABLE I. PERFORMANCE OF THE SYSTEM

| Number of neurons on the intermediary level | Classification errors | |
|---|---|---|
| | *Misclassification error* [%] | *Mean squared error (MSE)* |
| 60 | 41.86 | 0.0966 |
| 70 | 44.48 | 0.1010 |
| 80 | 37.72 | 0.0889 |
| 100 | 42.76 | 0.0978 |

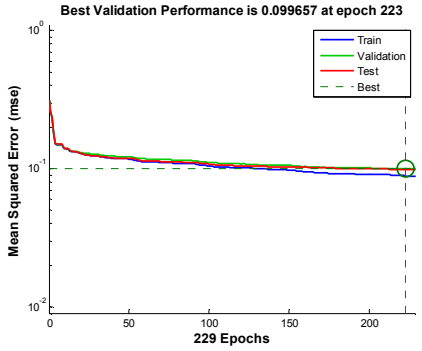


Fig. 5. Neural network performance

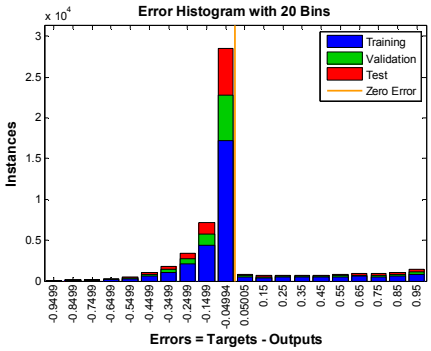


Fig. 6. Error histogram

We compared the results obtained with this type of network (MLP) with those obtained with the Elman network. In both cases, the used training method was scaled conjugated gradient (SCG) method.

We used both MSE as performance function, and mean squared error with regularization (MSEREG). MSEREG measures the network performance as the weight sum of two factors: MSE and the mean squared weights and biases.

The comparisons between the results obtained are presented in Table II. We observed that the performance of the network was better for the Elman network in the case of MSEREG and for MLP in the case of MSE.

TABLE II. PERFORMANCE OF DIFFERENT ANNS

| **Network type** | Feed forward backpropagation | | Elman backpropagation | |
|---|---|---|---|---|
| **Training** | SCG | | SCG | |
| **Performance function** | MSE | MSEREG | MSE | MSEREG |
| **Error (80 neurons)** | 0.0889 | 0.1080 | 0.110 | 0.0929 |
| **Error (90 neurons)** | 0.1050 | 0.1050 | 0.116 | 0.0995 |

## V. CONCLUSIONS AND FUTURE WORK

The paper evaluates how discriminative training methods can be used in an automatic speech recognition system based on neural networks.

We emphasized how neural networks can be suitable for speech recognition and the characteristics that they must contain for the speech recognition case. Also, we have presented different discriminative training methods.

For the system presented, we used MLP and Elman network and the error criterion for training was both MSE and MSE with regularization. The minimum error was obtained for MLP with MSE 0.0889.

As a future development, we are going to enhance the performance of the system by using neural networks together with other techniques, e.g. hidden Markov models. Also, the results obtained are preliminary ones and as other future development, we want to extend this research with a greater database. In this way, the results will be more conclusive for applying them in controlling a robot by voice commands.

## REFERENCES


[1] M. A. Anusuya, S.K. Katti, "Speech Recognition by Machine: A Review," International Journal of Computer Science and Information Security, vol. 6, no. 3, pp. 181-205, 2009.

[2] N. Desai, K. Dhameliya, V. Desai, "Feature Extraction and Classification Techniques for Speech Recognition: A Review," International Journal of Emerging Technology and Advanced Engineering, vol. 13, no. 12, pp. 367-371, 2013.

[3] J. Pylkkönen, "Towards Efficient and Robust Automatic Speech Recognition: Decoding Techniques and Discriminative Training," Aalto University, Doctoral Dissertations 44/2013, Helsinki, Finland, 2013.

[4] U. Shrawankar, V. Thakare, "A Hybrid Method for Automatic Speech Recognition Performance Improvement in Real World Noisy Environment," Journal of Computer Science, vol. 9, no. 1, pp. 94-104, 2013, doi:10.3844/jcssp.2013.94.104.

[5] C. Burileanu, V. Popescu, A. Buzo, C.S. Petrea, D. Ghelmez-Haneş, "Spontaneous Speech Recognition for Romanian in Spoken Dialogue Systems," Proceedings of the Romanian Academy, Series A, vol. 11, no. 1, pp. 83-91, 2010.

[6] C.-O. Dumitru, I. Gavăţ, "Progress in Speech Recognition for Romanian Language, Advances in Robotics, Automation and Control," J. Aramburo, A.R. Trevino (Eds.), InTech, 2008.

[7] C. Chivu, "Systems of Continuous Speech Recognition for Romanian Language," Control Engineering and Applied Informatics, vol. 7, no. 4, pp. 63-68, 2006.

[8] H.-N. Teodorescu, "AI Tools for Speech Analysis Applied to the Romanian Language, Proceedings of the 4th European Computing Conference," pp. 272-279, 2010.

[9] S.M. Feraru, "Emotional Speech Classification for Romanian Language – Preliminary Results, Proc. of the 11th International Conference on Development and Application Systems (DAS)," Suceava, pp. 158-161, 2012.

[10] M. Oprea, D. Şchiopu, "An Artificial Neural Network-Based Isolated Word Speech Recognition System for the Romanian Language," Proceedings of 16th International Conference on System Theory,

Control and Computing - ICSTCC2012, Sinaia, Romania, October 2012.
[11] D. Şchiopu, "Using Statistical Methods in a Speech Recognition System for Romanian Language," Proceedings of 12th IFAC/IEEE International Conference on Programmable Devices and Embedded Systems (PDeS), Velke Karlovice, Czech Republic, pp. 252-256, 2013.
[12] B. Gold, N. Morgan, D. Ellis, "Speech and Audio Signal Processing. Processing and Perception of Speech and Music," John Wiley and Sons, 2011.
[13] E. McDermott et al., "Discriminative Training for Large-Vocabulary Speech Recognition Using Minimum Classification Error," IEEE Trans. On Audio, Speech and Language Processing, 2006.
[14] Matlab ®, Available: http://www.mathworks.com/ .
[15] S. Davis, P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," IEEE Trans. Acoust. Speech Signal Processing, vol. 28, no. 4, pp. 357-366, 1980.
[16] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech," J. Acoust. Soc. Amer., vol. 87, no. 4, pp. 1738-1752,1990.
[17] A.E. Rosenberg, C.H. Lee, F.K. Soong, "Cepstral Channel Normalization Techniques for HMM-Based Speaker Verification," Proceedings of IEEE ICASSP, pp. 1835-1838, 1994.
[18] H. Hermansky, N. Morgan, "RASTA Processing of Speech," IEEE Trans. Speech Audio Processing, vol. 2, no. 4, pp. 578-589, 1994.
[19] E. Eide, H. Gish, "A Parametric Approach to Vocal Tract Length Normalization," Proceedings of IEEE ICASSP, pp. 346-349,1996.
[20] M. Richard, R. Lippmann, "Neural Network Classifiers Estimate Bayesian a Posteriori Probabilities," Neural Comput., vol. 3, pp. 461-483, 1991.
[21] G. Toderean, M. Costeiu, M. Giurgiu, Reţele neuronale artificiale, Albastră Publish., Cluj-Napoca, 1995.
[22] Z. Valsan, et al., "Statistical and Hybrid Methids for Speech Recognition in Romanian," International Journal of Speech Technology 5, pp. 259-268, 2002.
[23] H. Cucu, et al. "Recent Improvements of the SpeeD Romanian LVCSR System," Proc. Int. Conf. on Communications (COMM), Bucharest, Romania, 2014.