

A COMPARATIVE STUDY ON CLOSED SET SPEAKER IDENTIFICATION USING RBF NETWORK AND MODULAR NETWORKS

[†]Özgür Devrim Orman, *Levent Arslan

*Boğaziçi Üniversitesi

Elektrik-Elektronik Mühendisliği Bölümü, Bebek, 80815 İstanbul

arslanle@boun.edu.tr

[†]TÜBİTAK

Ulusal Elektronik ve Kriptoloji Araştırma Enstitüsü, Gebze, 41470 Kocaeli

oorman@mam.gov.tr

ABSTRACT

This paper compares the speaker identification performances of Radial Basis Function Network (RBFN) and proposed Modular Network (MN) models. We use TIMIT speech corpus in our experiments. The difficulty of identification problem is increased by choosing the speakers from the same dialect region and in the same gender. According to our experimental results, generally both MN types are more robust than RBFN. On the other hand, if we consider the correctly classified vectors percentage as a performance measure, RBFN has the highest ratio in observations.

KEYWORDS

Neural Network Models and Learning Algorithms, Speech Processing.

1. INTRODUCTION

Speaker identification is an easy task for human auditory system. On the man machine interface perspective it is still a difficult problem to solve, because we can not generate such specific feature sets to ease and to make more robust the identification of speakers by computers. Also, popularity of the topic has increased parallel to the increasing demand of interactive services over the telephone and the Internet, such as telephone and Internet banking which require high levels of security.

Speaker identification process can be subdivided into three phases: i) Transformation of training set speaker records to feature vectors database, ii) Training of the system using these data iii) Identification performance test using free-text utterances of the speaker. In the first phase, we can use various methods to generate feature sets, such as LPC cepstrum [2] or mel cepstrum [3] representations. The process in the second phase depends on the choice of identification method. In this phase we can use Vector Quantisation (VQ) [4], Gaussian Mixture Models (GMM) [5], Hidden Markov Models (HMM) [6] or various types of Neural Network (NN) architectures such as RBFN [7]. In the last phase we basically test the speaker identification performance of the system using free-text test feature vectors database.

Neural Networks can be used to solve the classification problem in speaker identification systems. As a generalisation, we can define the identification problem as a classification problem in pattern recognition theory. Previous work [5] shows that RBF Network (RBFN) gives better performance than VQ and approximates the GMM method identification performance results. Another neural network type that we use in this paper is modular network. Modular

network is composed of three modules: i) a statistical pre-processing unit, ii) a neural network classifier and iii) a winner-takes-all output unit.

In this work we compare the speaker identification performances of RBF Network and two different Modular Networks on a closed set from TIMIT [1] speech corpus. Modular type approach is proposed to make the design of speaker identification systems more robust as compared with RBFN. According to the results, besides its robustness, proposed approach gives the same speaker identification performance with RBFN.

This paper includes five sections. In section two, the theoretical background of closed set speaker identification is given, section three includes the RBFN and Modular Network theory, in section four we give the experimental details and results. These results are discussed in section five.

2. CLOSED SET SPEAKER IDENTIFICATION THEORY

Closed set speaker identification system can be implemented using the steps that we mention in section one. In this system, presented speaker has to be known by the system. An open set system has one more parameter than a closed set system, this parameter represents whether the presented speaker is known or not. The solution of closed set speaker identification problem can be generalised to use in open set system with only an addition of a threshold parameter. However, for both systems, it is necessary to use an appropriate transformation to represent speaker characteristics well.

Speaker characteristics can be represented using cepstrum coefficients. It is possible to generate these cepstrum coefficients using various transforms. In this work we use mel cepstrum coefficients, these are calculated as follows. First step is segmentation of voice active regions of utterances to the overlapping frames. TIMIT database already includes the records of voice active regions in utterances, so in this work we do not need to use a voice activity detection mechanism. Second step is the multiplication of frames by a window function that is generally either a Hamming or a Hanning type. In the third step we calculate the magnitude spectrum of each frame via discrete Fourier transform. At the fourth step the logarithm of magnitude spectrum is calculated and passed through a mel scale filter bank. The relation between the Mel scale and the standard frequency scale is formulated as follows:

$$mel(f) = 2595 \cdot \log_{10} \left(1 + \frac{f}{7000} \right)$$

At the last step the cosine transform of coefficients is calculated, which gives us mel cepstrum coefficients.

The identification method can be chosen from various alternatives by considering the requirements and resources. In VQ based systems, training set is used to generate a codebook of speaker feature vectors using Linde Buzo Gray (LBG) algorithm [10], k-means clustering, learning vector quantisation algorithms [9] or their variations. On the other hand, training phase of a GMM based system includes determination of means, variances and observation probabilities of Gaussian nodes; we can do it using expectation maximisation (EM) algorithm. If we use RBFN, we adapt the weights and node variables (mean, variance) appropriately according to an error minimisation criterion.

Speaker identification performance of the system is tested using free-text utterances. Using these data, corresponding mel cepstrum feature vector set is calculated. In VQ based systems, identification of a speaker can be done using majority voting or Mahalanobis distance criteria. To decide on a speaker in GMM systems, we compare the likelihood of test utterances. These likelihood values are calculated for each speaker in the set and we choose the one with maximum likelihood value. On the other hand the NN speaker identification systems used in this work have a winner-takes-all unit at the output layer. The winner-takes-all unit is used to decide on a speaker who has the maximum output value in the previous layer. Decision results on presented test vectors at the output of winner-takes-all unit are accumulated. When the presentation of the whole utterance is completed, accumulated decisions are processed using majority voting method.

3. RBFN AND MODULAR NETWORKS

RBFN has more simple network structure when compared with other multilayer neural network architectures such as multilayer perceptrons. There are two layers in an RBFN. First layer is a hidden layer that includes a nonlinear transform to make the patterns linearly separable at the next layer [8]. In the second layer, outputs of the previous layer are weighted and summed. A schematic representation of RBFN architecture is given in Figure 1. As can be seen from Figure 1, input vector is n-dimensional and there are m nodes in the hidden layer.

Output layer can be composed of more than one neuron. The input-output relation of the RBFN that is represented in Figure 1 is given as:

$$y = f(\underline{x}) = \sum_{i=1}^m w_i \cdot G(\|\underline{x} - \underline{c}_i\|) \quad (1)$$

$$\underline{x} = [x_1, x_2, \dots, x_n]^T \text{ and } \underline{c}_i = [c_{i1}, c_{i2}, \dots, c_{in}]^T$$

If we look at (1), $G(\cdot)$ represents the radial basis function of the network. We use Euclidean norm in our experiments but the norm can be defined in any metric, w_i represents the i^{th} connection weight and \underline{c}_i is the i^{th} centre vector.

Since the layers of RBFN are not required to be trained together, we can use different methods to optimise the parameters in these two layers. Hidden layer optimisation requires more learning time as compared with the output layer. If we know the centres, the optimisation problem reduces to a least mean square type. These centres can be chosen using various methods, such as random selection from training vectors, Orthogonal Least Square, self-organization or gradient descent.

In this work we use an RBFN architecture that has multiple outputs in the second layer. The number of neurons in second layer is the same as the number of speakers in the speaker set. The outputs of this layer are fed to a winner-takes-all unit which makes speaker decisions. Figure 2 represents the RBFN structure that we use in this work. Input-output relation of this architecture is given in below formulas.

$$y_i = \sum_{j=1}^m w_{ij} G(\underline{x} - \underline{c}_j), \quad 1 \leq i \leq k$$

$$D = \text{Arg max}\{y_1, y_2, \dots, y_k\}$$

Modular network type that we use in this work is based on the idea of using the NN classifiers on likelihood data. General schematic representation of this architecture is given in Figure 3. Statistical pre-processing unit is composed of sub modules; each includes a GMM system. The number of sub modules in the first unit is equal to the number of speakers in the set. These GMM based sub-modules calculate the likelihood values of presented pattern for each speaker. Each sub-module represents a different speaker and its parameters are determined by this speaker's training data. The parameters of sub-modules are optimised using the EM algorithm. Outputs of this layer are fed through a NN classifier that can be either a linear network or a RBFN. Number of outputs of the second layer is equal to the number of speakers in the set and these outputs are fed to the last layer which is a winner-takes-all unit. It works the same as in the RBFN based speaker identification system that we mention before. The detailed representations of proposed network architectures are given in Figure 4 and Figure 5. Formulation of the input-output relation of the proposed modular network architecture that has linear classifier unit is given in the following formulas. We call this network as Type-1.

$$y_i = \sum_{j=1}^m w_{ij} \cdot \log_{10}(p(\underline{x}/\lambda_j)), 1 \leq i \leq k$$

$$\lambda_j = \{p_j, \underline{\mu}_j, \Sigma_j\}, 1 \leq j \leq m$$

$$D = \text{Arg max}\{y_1, y_2, \dots, y_k\}$$

On the other hand, the same relations can be defined for the proposed modular architecture that has a RBFN unit as given below; which we call as Type-2.

$$\underline{p}(\underline{x}) = \{p(\underline{x}/\lambda_1)p(\underline{x}/\lambda_2)\dots p(\underline{x}/\lambda_k)\}$$

$$y_i = \sum_{j=1}^m w_{ij} \cdot \log_{10}(G(\|\underline{p}(\underline{x}) - \underline{c}_j\|)), 1 \leq i \leq k$$

$$D = \text{Arg max}\{y_1, y_2, \dots, y_k\}$$

Logarithms in the preceding equations are used to guarantee the numerical stability of weight calculation on the computer.

4. EXPERIMENTAL RESULTS

Our experiments can be grouped into two categories, speaker identification using RBFN and speaker identification using the modular network types given above. We use TIMIT database in our experiments. TIMIT database has eight different dialect regions and there are various numbers of speakers in each region. Furthermore, in each speaker's directory, there are both free-text and text-dependent records. In order to increase the difficulty of identification process, we compose the speaker set of the speakers from the same dialect region. In addition, we choose only female speakers in the set, which makes the task more complex.

The speaker set used in these experiments includes twelve female speakers from dialect region one. We generate the training set using the files with "sa" and "si" prefixes, and the files with "sx" prefix are used in the test set. In section 4.1. we give the details and performance results of implementation of the former system. Also, implementation and performance of second system is explained in section 4.2.

4.1. Speaker Identification Using RBFN

To implement a RBFN based speaker identification system, it is necessary to determine the centres using training data set. In this work we determine these centres using LBG method and we generate 32 centres for each speaker. We optimise the network weights using a least mean square type method. After the training phase is complete, we present the test set to the system and observe the performance in two different categories. These categories are percentage of the correctly classified vectors and percentage of the correctly identified speakers. Vector identification performance of RBFN method is equal to %47.1 but speaker identification performance is equal to %100.

4.2. Speaker Identification Using Modular Networks

In the training phase of both Type-1 and Type-2 systems, first step is to determine each GMM sub-modules parameters using the training data. In each sub-module, there are 32 Gaussian nodes and three parameters of each node (observation probability, mean vector and variance matrix that is assumed diagonal) are optimised using the EM algorithm. Centres in Type-2 systems are calculated using the LBG algorithm. Calculation of weights is the same in both types of systems, which is a least mean square method. In Type-1 system, correct vector classification percentage is equal to %31.8 and correct speaker identification percentage is %100. Also, the observed percentages for Type-2 system are % 34.3 in correct vector classification and % 100 in speaker identification.

5. CONCLUSION

If we compare vector classification performances of all types of systems that we discuss in the preceding sections, we see that RBFN has the highest performance according to this criterion. This is the result of the major difference between RBFN and proposed modular networks. RBFN network weights are optimised using all training data, which results in embedding of a discriminant function in the network. It means that each y_i value not only represents the correct classification of the presented speaker's test data but also it includes the rejection of an incorrect speaker decision. On the other hand, in both proposed modular network types, GMM sub modules are optimised in the maximum likelihood (ML) sense. This optimisation increases the likelihood of the whole test utterance. Therefore it is possible to increase the likelihood of an incorrect speaker's vectors, because there is no rejection information in the training phase. However, it does not result with incorrect identification of speakers, because decisions are based on looking at the whole test utterance.

Proposed architectures have the same speaker identification performance as compared with RBFN. Moreover, proposed types are more robust than RBFN, because the addition of a new speaker to the system requires less computation time in these architectures.

REFERENCES

1. "Getting started with darpa TIMIT CD-ROM: an acoustic phonetic continuous speech database", National Institute of Standards and Technology (NIST), Gaithersburg, MD (prototype as of Dec. 1988).
2. Atal, B.S., "Automatic recognition of speakers from their voices", Proc. IEEE, Vol. 64, pp. 460-474, 1976

3. S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-28, pp. 357-366, 1980.
4. A.E. Rosenberg, and F. K. Soong, "Evaluation of a vector quantization talker recognition system in text independent and text dependent modes", *Computer Speech and Language*, Vol. 22, pp. 143-157, 1987.
5. D. A. Reynolds, R.C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models", *IEEE Trans. Speech and Audio Processing*, Vol. 3, pp. 72-83, 1995.
6. N. Z. Tishby, "On the application of mixture AR hidden Markov models to text independent speaker recognition", *IEEE Trans. Signal Processing*, Vol. 39, pp. 563-570, 1991.
7. J. Oglesby and J. Mason, "Radial basis function networks for speaker recognition", in *Proc. ICASSP*, May 1991, pp. 393-396.
8. S. Haykin, *Neural Networks –A Comprehensive Foundation–*, Pr. Hall, 1994.
9. T. Kohonen, "The self-organizing map", *Proc IEEE*, Vol. 78., pp. 84–95, 1980
10. Y. Linde, A. Buzo ve R.M. Gray, "An algorithm for vector quantizer design", *IEEE Trans. Comm.*, Vol. 20, pp. 84-95, 1980.

FIGURES

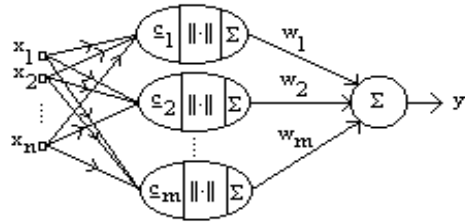


Figure 1: RBFN with a single output.

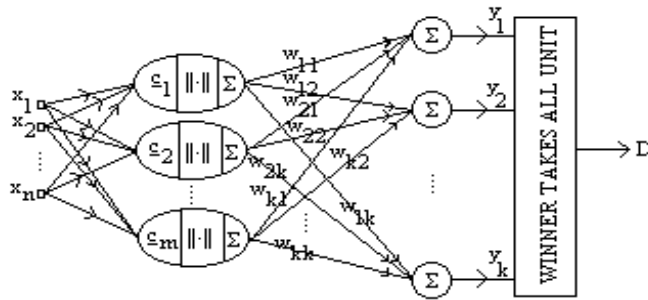


Figure 2: RBFN architecture that is used in this work.

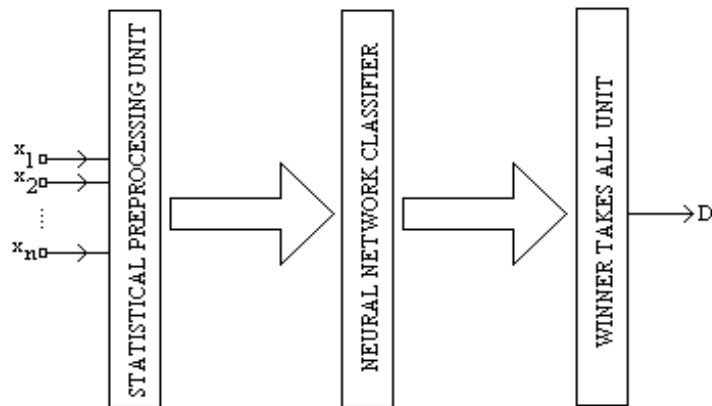


Figure 3: General representation of modular networks.

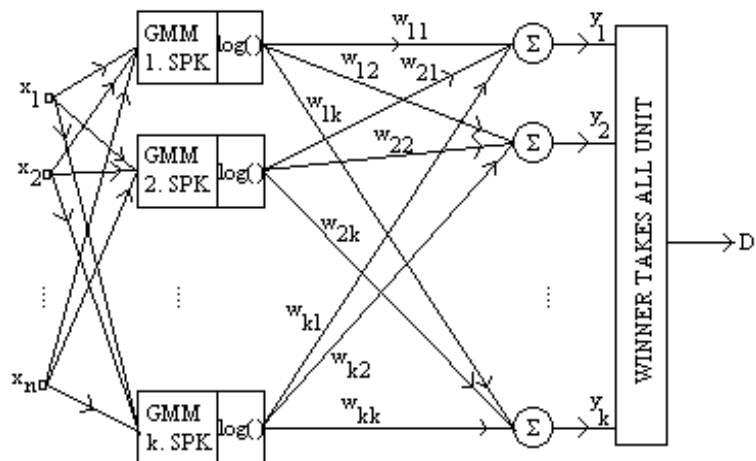


Figure 4: Modular Network that has a linear classifier.

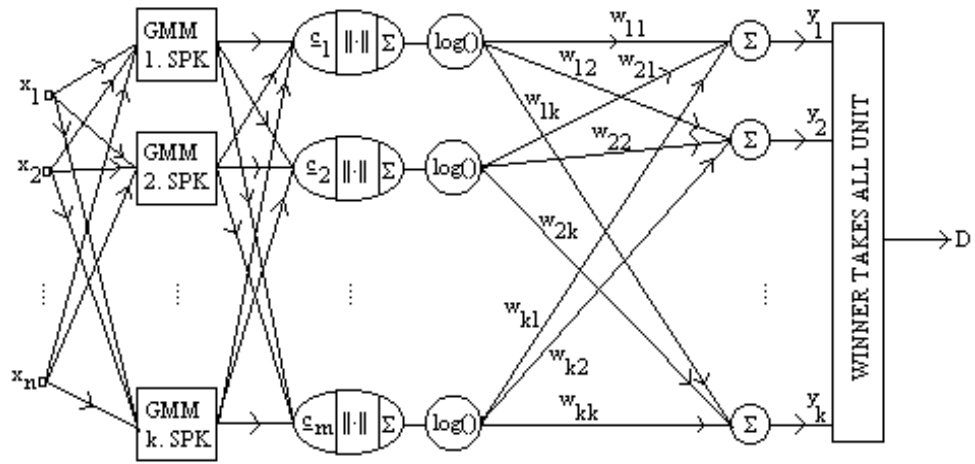


Figure 5: Modular Network that has an RBFN unit.