Improved Tree Model for Arabic Speech Recognition

Nacereddine Hammami
Faculty of Computer and Information Sciences
University of Al Jouf
Sakaka, Kingdom of Saudi Arabia
n.hammami@ju.edu.sa

Mouldi Bedda
Faculty of engineering
University of Al Jouf
Sakaka, Kingdom of Saudi Arabia
Mouldi_bedda@yahoo.fr

Abstract—**This paper introduces a fast learning method for a graphical probabilistic model for discrete speech recognition based on spoken Arabic digit recognition by means of a new proposed spanning tree structure that takes advantage of the temporal nature of speech signal. The experimental results obtained on a spoken Arabic digit dataset confirmed that for the same rate of recognition the proposed method, in terms of time computation is much faster than the state of art algorithm that use the maximum weight spanning tree (MWST).**

Keywords- Arabic Speech recognition; optimal–spanning-tree; dependency tree; discrete probability distributions; graphical model

I. INTRODUCTION

Speech recognition technology plays important role in many applications such as speech-to-text, language translation and speech input interface. It realizes human machine interfaces such as TV remote control, navigation system, telephone, and so on in practical world. In literature many methods have been proposed. Hidden Markov Model (HMMs) [1] is the one of the most popular techniques in recognition task. HMM models are very rich in mathematical structure and hence can form the theoretical basis for use in wide range of application. Word speech recognition can be adopted to a word HMM [2] [3] or phoneme HMM [4] [5]. In addition to HMMs, many authors have proposed recognition algorithms based on artificial neural networks (ANNs) [6]. The results of many recent studies on HMM and ANN hybrids for automatic speech recognition (ASR) indicate that a hybrid system is superior to either approach [7] [8]. Adapted Gaussian mixture model (GMM) [9] [10] approach to this problem is used. Support vector machines (SVMs) have proven to be a new effective method for speaker recognition [11] [12]. By contrast to other languages the Arabic language had limited number of research efforts, although it is one of the oldest languages in the world and the fifth widely used language nowadays [13]. In precedent work we proposed a new graphical model for discrete speech recognition [14], discrete speech features are extracted by means of Mel Frequency Cepstral (MFCCs) coefficients followed by vector quantization (VQ). Then in a second step, the obtained features are fed to a Tree Distribution Classifier (TDC) which provides the class-label associated with each feature by approximating the true class probability by means of an optimal spanning tree model, the latter are increasingly and successfully used to deal with the probability estimation in different pattern recognition problems such as skin detection [15] [16] and object detection [17]. In contrast the problem of learning the structure of a graph from data is significantly harder. In practice, most structure learning methods are heuristic methods that perform local search by starting with a given graph and improving it by adding or deleting one edge at a time. There is an important special case in which both parameter learning and structure learning are tractable, namely the case of graphical models in the form of a tree distribution. As shown by Chow and Liu [18], the tree distribution that maximizes the likelihood of a set of observations on M nodes as well as the parameters of the tree

978-1-4244-5540-9/10/$26.00 ©2010 IEEE

can be found in time quadratic in the number of variables in the domain. To deal with this issue, we propose in this paper an automatic method for discrete recognition of spoken Arabic digits using a tree distribution classifier with a proposed tree structure which provides a fast learning. In this model type as in [14], the extracted feature vector is mapped into one of a set of prototype vectors through a vector quantization process. Prototype vectors are constituted during the training phase using a clustering algorithm. The experimental results obtained on a spoken Arabic digit dataset gives similar results to the maximum weight spanning tree structure with very large gains in complexity on the learning step.

The remainder of this paper is organized as follows: in section two, we describe the feature extraction method. In section three, we outline the vector quantization (QV). Section four details the tree distribution classifier model for the MWST and the proposed structure for the tree graph and section five is devoted to experiments and results.

II. Feature Extraction

The purpose of this module is to convert the speech waveform to some type of parametric representation (at a considerably lower information rate).

A wide range of possibilities exist for parametrically representing the speech signal for the speaker recognition task, such as Linear Prediction Coding (LPC), MFCCs, and others.

MFCCs are the most commonly used acoustic features in automatic speech recognition, and this feature has been used in this paper. MFCC's are based on the known variation of the human ear's critical bandwidths with frequency.

In the experiment, the coefficients MFCCs are computed with the following conditions;

- Sampling rate :11025 Hz, 16 bits
- Window applied: hamming
- Filter preemphasized : $1 - 0.97\mathrm{Z}^{-1}$

III. Vector quantization (VQ)

The discrete Tree distribution classifier, system requires a scalar sequence *X*. VQ [19] divides *D*-dimensional Euclidian space R^D into unempty subsets $V_i, i = 1, \dots, U$,

$$\bigcup_{i=1}^{U} V_i = R^D, V_i \cap V_j = \emptyset \ (i \neq j), \qquad (1)$$

Where $\emptyset$ is empty set. The operation which maps an input vector $x \in V_i$ into a predetermined vector Y_i in subset V_i is called VQ. Denoting the operation of VQ as Q,

$$Q(\boldsymbol{x}) = y_i \ if \ x \in V_i, \qquad (2)$$

Where y_i is called code, $\boldsymbol{C} = \{y_1, \dots, y_{|C|}\}$ is called codebook, and $|\boldsymbol{C}|$ $(= U)$ is the codebook size. Codebook is predesigned and stored in system memory. To this end, we purpose to adopt the well known *k-means* clustering algorithm, the subset V_i in "(1)" is obtained by the following nearest neighbor condition,

$$V_i = \left\{x \in R^D, d(x, y_i) \leq d(x, y_j) ; j = 1, \dots, |C|\right\} \quad (3)$$

Where $d(.)$ denotes the Euclidian distance.

We propose the Cross-Validation (CV) method to estimate the optimal number of clusters (initial set of code words in the codebook).

The procedure of *m*-fold cross-validation is as follows:

1) Split the training data into *m* roughly equal-sized parts.

2) For the i^{th} part, train the classifier model to the other $(m - 1)$ parts of the data using k clusters, and test classifier for the i^{th} part of the data.

3) Do the above for $i = 1, \cdots, m$, and take the average of m results.

IV Tree Distribution Model

Let us consider D the set of spoken digits and x_d the n-dimensional vector representing the spoken Arabic digits. The class of the spoken digit is C_d with $C_d = j$ if x_d belongs to class j with $j = 0, \ldots, 9$.

Let us assume that we know the joint probability distribution $P(x_d, C_d)$ of the vector (x_d, C_d). Then the Bayesian analysis tells us that, whatever the cost function the user might think of, all that is needed is the a-posterior distribution $P(x_d | C_d)$. The useful information is contained in the one spoken digit vector marginal of the a-posterior probability. That is for each spoken digit vector, the quantity $P(C_d = j | x_d)$ quantifying the belief for the appurtenance of the spoken digit vector x_d to the class $C_d = j, j = 0, \ldots, 9$. In practice for $x = (x_1, \ldots, x_n)$ the model $P(x, C_d)$ is unknown. Instead, we have spoken Arabic digits database. It is a collection of pronounced Arabic digit (zero to nine) from dependent speaker. The collection samples noted $\{(x^{(1)}, C^{(1)}), \ldots, (x^{(N)}, C^{(N)})\}$ where for each $1 \leq i \leq N$, $x^{(i)}$ is a vector representation of the spoken digit and $C^{(N)}$ is the associated class. We assume that the samples are independent each other with the distribution $P(x, C_d)$.

The collection of samples is referred later as the training data. Our objective is to find for each class a non oriented acyclic graph (tree) modeling $P(x, C_d = j)$ noted $P_j(x)$ and construct a probabilistic classifier.

A. Tree model

In this section we introduce the Tree model. Let V denotes a set of n discrete random variables of interest. For each random variable $v \in V$, let $\delta(v)$ represent its range, $x_u \in \delta(v)$ a particular value. $x = (x_1, \ldots, x_n)$ denotes an assignment to the variables in V. Let's consider a complete non oriented graph $G(V, E)$ corresponding to the n variables, where E is a set of edges. Two neighbor vertices u and v are noted $u \sim v$.

Proposition

If the graph G was a tree (a connected graph without loops) which we note T parameterized in the following way:

For $u, v \in V$ and $(u, v) \in E$, let $q_{T_{uv}}$ denote a joint probability distribution on u and v. We require these distributions to be consistent with respect to marginalization, denoting by $q_{T_u}(x_u)$ the marginal of $q_{T_{uv}}(x_u, x_v)$, or $q_{T_{vu}}(x_v, x_u)$, with respect to x_u for any $v \neq u$. We now assign a distribution q_T to the graph $G(V, E)$ as follows [20]:

$$q_T(x) = \prod_{(u \sim v) \in T} \frac{q_{T_{uv}}(x_u, x_v)}{q_{T_u}(x_u)\, q_{T_v}(x_v)} \prod_{u \in V} q_{T_u}(x_u) \quad (4)$$

B. Learning of tree distribution based on MWST

The learning problem is formulated as follows:

Given a set of observations $X = (x^{(1)}, \ldots, x^{(N)})$, we want to find for each class digit j, $j = 0, \ldots, 9$ one tree T_j in which the distribution probability is efficient. To do this, we have adopted the well known spanning tree algorithm proposed by Chow-Liu [14] [18].

For instance, Fig.1 shows a maximum weight spanning tree using Chow-Liu algorithm, utilizing the computed mutual information as edge weights.

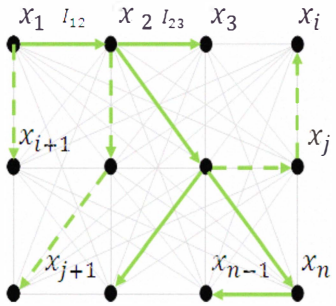

Figure 1. Maximum Weight Spanning Tree Using The Mutual Information As Weight.

C. Learning of tree distribution based on proposed structure

One of the main drawbacks of Chow-Liu algorithm is the enormous computational complexity needed to obtain the optimal spanning tree. As a natural way to alleviate the before mentioned drawback, we propose a new graphical tree structure inspired from the temporal nature of the speech signal, in which, only the linear dependencies between the features are considered, see Fig.2. In this case , the "(4)" can be simplified to the following;

$$q(x) = \prod_{i=1}^{n-1} \frac{q_{i\,i+1}(x_i, x_{i+1})}{q_i(x_i)\, q_{i+1}(x_{i+1})} \prod_{i=1}^{n} q_i(x_i) \qquad (5)$$

In contrast to Chow-Liu algorithm, the proposed method doesn't require any computation of mutual information to find the (MWST).

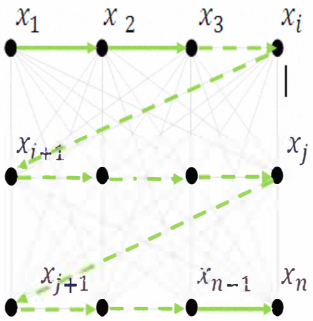

Figure 2. Proposed Tree Structure To Learning the Classifier Model.

D. Inference

We would denote the class of a vector $x = (x_1, \ldots, x_n)$, which represents a spoken Arabic digit. The expected classification error can be minimized by choosing $Argmax_j\big(P(C_d = j|x)\big)$.

According to Bayes's theorem:

$$P(C_d = j|x) = \frac{P(x|C_d = j)P(C_d = j)}{P(x)} \qquad (6)$$

Moreover,

$$P(x) = \sum_{j=0}^{j=9} P(x|C_d = j)P(C_d = j) \qquad (7)$$

In which

$$P(x|C_d = j) \approx \prod_{(u\sim v)\in T_j} \frac{q_{j_{uv}}(x_u, x_v)}{q_{j_u}(x_u)\, q_{j_v}(x_v)} \prod_{u\in V} q_{j_u}(x_u)$$

$$= q_j(x) \qquad (8)$$

$$\forall\, i, j = (0, \ldots, 9)\ \ P(C_d = j) \approx P(C_d = i)$$

Therefore

$$P(C_d = j|x) \approx \frac{q_j(x)\, P(C_d = j)}{P(x)} \qquad (9)$$

All the elements of "(7)" and "(8)" are previously computed in learning setup.

V. Experimental Results

A. Dataset Description

The experiments were performed using the Arabic digit corpus collected by the laboratory of automatic and signals , University of Badji-Mokhtar - Annaba, Algeria.. A number of 88 individual (44 males and 44 females) Arabic native speakers were asked to utter all digits ten times.. Depending on this, the database consists of 8800 tokens (10 digits x 10 repetitions x 88 speakers). In this experiment, the data set is divided into two parts: a training set with 75% of the samples and test set with 25% of the samples. In this research, speaker-independent mode is considered.

B. Experiment

In this experiment, before applying the tree model to the above dataset, we perform feature extraction for the both trees structure by means of VQ as described previously. The result of the clustering is a codebook of 32 for the tree model (optimum size obtained using 3-fold cross validation on the training set). Fig.3 shows the accuracy results for different values of k.

In classification step. Table.3 shows the classification results obtained by the MWST and the proposed Tree model.

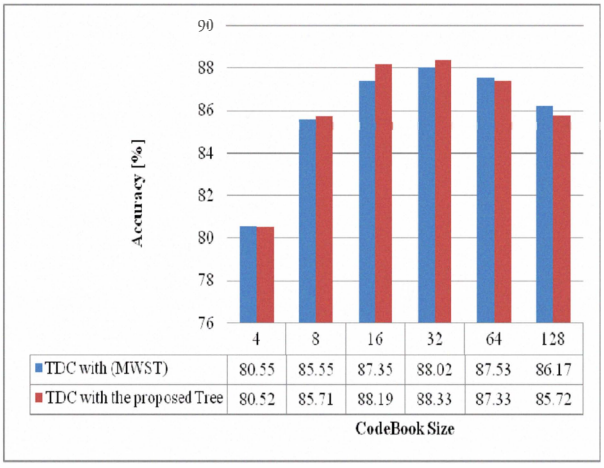

Figure 3. Cross Validation (CV) Accuracy versus the Number of Clusters

As can be seen, the overall accuracies of the two methods are similar.

TABLE 3. Recognition by Dependence Tree Model

Arabic Digits Classes	Success Rate %	
	Dependence Tree Model for MWST	Dependence Tree Model for proposed structure
Syfr - 0	85.55	84.09
Wahid- 1	98.36	98.27
Ethnan- 2	92.91	92.55
Thalath'a- 3	94.09	94.09
Arb'a- 4	89.91	91.19
Khams'a- 5	94.00	94.45
Syt'a- 6	93.82	94.18
Sab'a-7	90.18	89.45
Thamany'a- 8	99.00	99.00
Tes'a- 9	93.36	93.72
Average	**93.12**	**93.10**
Learning Time (s)	**480.36**	**19.80**

VI. CONCLUSION

This paper has presented a way speed up graphical model learning for speech recognition. The advantage of the proposed technique over Chow-Liu algorithm which is considered as a state of art is the impressive gain in time computation. In fact, this method is capable of achieving speed ups to up 25 orders of magnitude in the experiment (19.80 seconds for the proposed technique against 480.32 seconds obtained by [14]). Moreover, the experiment has also shown that that the proposed method is capable of keeping the same accuracies as the ones obtained by using the Chow-Liu Algorithm.

VI. REFERENCES

[1] L. Rabiner and B. H. Juang, Fundamentals of Speech Recognition.Englewood Cliffs, NJ: Prentice-Hall, 1993.

[2] W. Han, K. Hon, and C. Chan, "An HMM-based speech recognition IC," Proc. IEEE ISCAS'03, vol. 2, 2003, pp. 744–747.

[3] F. Vargas, R. Fagundes, and D. Barros, "A FPGA-based Viterbi Algorithm implementation for speech recognition systems," Proc. IEEE ICASSP'01, vol. 2, May 2001, pp. 1217–1220.

[4] J. Pihl, T. Svendsen, and M. H. Johnsen, "A VLSI implementation of pdf computations in HMM based speech recognition," Proc. IEEE TENCON'96, 1996, pp. 241–246.

[5] S. J. Melnikoff, S. Quigley, and M. J. Russell, "Implementing a Simple continuous speech recognition system on an FPGA," Proc. IEEE Symp. (FCCM'02), 2002, pp.275–276.

[6] R. P. Lippmann, "Review of neural networks for speech recognition," Neural Comput., vol. 1, pp. 1–38, 1989.

[7] Q. Liang and J. G. Harris, "The feature of artificial neural networks and speech recognition," in Intelligent Systems: Technology and Applications: Vol. III Signal, Image, and Speech Processing, C. T. Leondes, Ed. Boca Raton, FL: CRC Press, 2003, pp. 215–236.

[8] H. A. Bourlard and N. Morgan, Connectionist Speech Recognition—A Hybrid Approach. Norwell, MA: Kluwer, 1994, ch. 7.

[9] D. A. Reynolds, T. F. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," Dig. Signal Process., vol. 10, no. 1-3, pp. 19–41, 2000.

[10] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," IEEE Trans. Speech Audio Process,, vol. 13, no. 3, pp. 345–354, May 2005.

[11] W. M. Campbell, "Generalized linear discriminant sequence kernels for speaker recognition," in Proc. Int. Conf. Acoustics, Speech, Signal Processing, 2002, pp. 161–164.

[12] V.Wan and S. Renals, "Speaker verification using sequence discriminant support vector machines," IEEE Trans. Speech Audio Processing, vol. 13, no. 2, pp. 203–210, Mar. 2005.

[13] M.Al-Zabibi, "An Acoustic–Phonetic Approach in Automatic Arabic Speech Recognition", the British Library in Association with UMI, 1990.

[14] N. Hammami, M. Sellami ,"Tree distribution classifier for automatic spoken Arabic digit recognition", Proc. IEEE ICITST'09 conference, 2009 , PP 1 – 4.

[15] S. El Fkihi, M. Daoudi, D. Aboutajdine, "Probability Approximation Using Best-Tree Distribution for Skin Detection", Advanced Concepts for Intelligent Vision Systems, University of Antwerp, Antwerp,Belgium, September 18–21, 2006, pp. 767–775

[16] Sanaa El Fkihi , Mohamed Daoudi, Driss Aboutajdine ,"The mixture of K-Optimal-Spanning-Trees based probability approximation: Application to skin detection", Elsevier 2008.

[17] S.Ioffe, D.Forsyth, "Mixtures of trees for object recognition", Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference.

[18] C. Chow, C. Liu, "Approximating discrete probability distributions with dependence trees", Fifteenth IEEE Transactions on Information Theory 14 (3) (1968) 462–467, May.

[19] R. M. Gray, ``Vector Quantization," IEEE ASSP Magazine, pp. 4--29, April 1984.

[20] J. Pearl, "Probabilistic reasoning in intelligent Systems: networks of plausible inference". Morgan Kaufmann,1988.