Tree Distribution Classifier for Automatic Spoken Arabic Digit Recognition

Nacereddine Hammami, Mokhtar Sellam
Department of Computer Science, LRI, Algeria
nacereddine.hammami@gmail.com, sellami@lri-annaba.net

Abstract

In this work we propose a novel method for automatic discrete speech recognition composed from two steps. In a first step, discrete speech features are extracted by means of Mel Frequency Cepstral Coefficients (MFCCs) followed by vector quantization (VQ). Then in a second step, the obtained features are fed to a Tree distribution classifier which provides the class-label associated with each feature by approximating the true class probability by means of an optimal spanning tree model. The experimental results obtained on a spoken Arabic digit dataset confirmed the promising capabilities of the proposed approach.

1. Introduction

Automatic Speech Recognition (ASR) is gaining a growing role for a variety of applications, such as hand-free operation and control (as in cars and airplanes), automatic query answering, telephone communication with information systems, automatic dictation (speech-to-text transcription), government information systems, etc. In fact, speech communication with computers, PCs, and household appliances is envisioned to be the dominant human machine interface in the near future.

In the literature of speech recognition many methods have been proposed. Generally, these are based on Hidden Markov Models (HMM) [1], neural networks [2], statistical analysis and vector quantization [3], and Gaussian mixture models (GMM) [4]. By contrast to other languages the Arabic language had limited number of research efforts [6] [7], although it is one of the oldest languages in the world and the fifth widely used language nowadays [5].

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. The choice of the discrete model is computationally efficient and represents a powerful base form by selecting appropriate and well trained observation symbols and incorporating parameter smoothing. Moreover, it performs very well for several tasks such as large vocabulary isolated word recognition. In this model type, 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. After this step, the derived features are fed to a Tree distribution classifier, 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 [8] [9] and object detection [11]. The experimental results obtained on a spoken Arabic digit dataset proved the promising capabilities of the proposed approach.

The remainder of this paper is organized as follows: in section two, we describe the feature extraction method. Section three details the Tree distribution classifier model and section four is devoted to experiments and results. Finally, conclusions and perspectives are drawn in section five.

2. Feature Extraction

In the signal analysis phase the input speech signal is transformed into feature vectors containing spectral and/or temporal information using Mel Frequency Cepstral coefficients (MFCCs) [6]. Table 1 shows some of the system parameters adopted for such task. The result of the feature extraction is a series of vectors, characteristic of the time varying spectral properties of the speech signal. These can then be mapped into discrete vectors by quantizing them using vector quantification (VQ). The latter is a potentially efficient representation of spectral information in the speech signal. It is based on the generation of a code of size M from a training set of vectors of size L. To this end, we purpose to adopt the well known k-means clustering algorithm, which is summarized in the following steps:

Initialization: we choose an arbitrarily *M* vectors to represent the initial set of code words in the codebook. Once the codebook of vectors has been obtained, the mapping between the observation vectors and codebook indices becomes a simple nearest neighbor computation, i.e. the observation vector is assigned the index.

TABLE1. System Parameters

Parameter	Value
Sampling rate	11025 Hz, 16 bits
Preemphased	0.97
Window type	Hamming

Centroid update: we update the codeword for each index using the centroid of the training vector assigned to the index. The distance used is the Euclidian distance, whose minimum value is used to update the centroid. If we consider $c(k)$ as the current centroid of the k^{th}cluster and $v(k)$ a vector in the cluster then:

$$D_{min} = \min_{1 \le n \le N} \left[\!\left[\sum_{k=1}^{K} (C_n(k) - v(k))^2 \right]\!\right] \qquad (1)$$

3. Tree Distribution Model

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 $j = 0, \dots, 9$. 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 $P(C_d = j|x_d)$ $q$uantifying the belief for the appurtenance of the spoken digit vector x_d to the class $C_d = j, j = 0, \dots, 9$. In practice for $x = (x_1, \dots, x_n)$ 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 $\{(x^{(1)}, C^{(1)}), \dots, (x^{(N)}, C^{(N)})\}$ where for each $1 \le i \le N$, $x^{(i)}$ 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. $x = (x_1, \dots, x_n)$ 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, we parameterize a tree 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 [12]:

$$q_T(x) = \prod_{(u \sim v) \in T} \frac{q_{T_{uv}}(x_u, x_v)}{q_{T_u}(x_u)\, q_{T_v}(x_v)} \prod_{u \in V} q_{T_u}(x_u) \qquad (2)$$

B. Learning of tree distribution

The learning problem is formulated as follows: given a set of observations $X = (x^{(1)}, \dots, x^{(N)})$, we want to find for each digit class j, $j = 0, \dots, 9$ one tree T_j in which the distribution probability is efficient.

We learn the model by maximizing the log-likelihood for the training data for each class. Chow and Liu [10] showed that the maximum weight spanning tree (MWST) using mutual information I_{uv} as the weight for the edge(u, v), maximizes the likelihood over tree distributions q_j for each class j. The algorithm is summarized on Table 2.

TABLE 2. The Chow and Liu Algorithm for Maximum Likelihood Estimation of Tree Structure and Parameters

```
**Algorithm Chow_ Lui** $(P_X)$
**Input:** Distribution $P_X$ over domain V
Procedure MWST (Weights) that outputs a maximum weight spanning tree over V
1 Compute marginal distributions $P_{X_v}, P_{X_{uv}}$
for $u, v \in V$
2 Compute mutual information values $I_{uv}$
for $u, v \in V$
3 $T_j = MWST(\{I_{uv}\})$
4 Set $q_{j_{uv}} \equiv P_{X_{uv}}$ for $u, v \in V$
```

C. Inference

We would denote the class of a vector $x = (x_1, \dots, x_n)$, 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:

$$P(C_d = j|x) = \frac{P(x|C_d = j)P(C_d = j)}{P(x)} \qquad (3)$$

Moreover,

$$P(x) = \sum_{j=0}^{j=9} P(x|C_d = j)P(C_d = j) \qquad (4)$$

In which

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

$$= q_j(x) \qquad (5)$$

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

Therefore

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

All the elements of "(4)" and "(5)" are previously computed in learning setup.

4. Experimental Results

A. Dataset Description

The experiments were performed using the Arabic digit corpus database from the national laboratory of automatic and signals at the University of Badji-Mokhtar in Annaba, Algeria. This data base was created from all ten Arabic digits. A number of 40 individual (20 males and 20 females) Arabic native speakers were asked to utter all digits ten times. Hence, the database consists of 10 repetitions of every digit produced by each speaker. Depending on this, the database consists of 4000 tokens (10 digits x 10 repetitions x 40 speakers). In this research, speaker-independent mode is considered.

B. First experiment

In this experiment, before applying the tree model to the above dataset, we perform feature extraction by means of VQ as described previously. The result of the clustering is a codebook of 16 (optimum size obtained using 2-fold cross validation on the training set). Figure 1 shows accuracy results for different values of k.

The discretized vectors with 16-means are the final features used throughout, and it's the input-set that is used in classification step. Table 3 shows the classification results obtained by the proposed Tree model. As can be seen, the overall accuracy of the system is 90.35%.

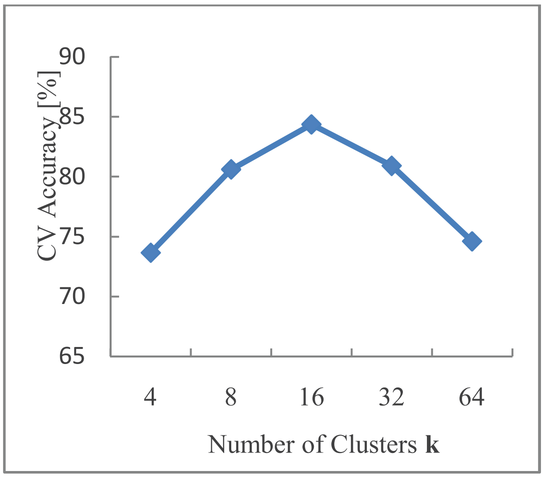

Figure 1. Cross Validation (CV) Accuracy Versus the Number of Clusters

C. Second Experiment

In this experiment, we assess the sensitivity of the Tree classifier with respect to the training set size. For such purpose, we reduce the training set and repeat the experimental scenario of experiment one. The obtained results shown in Table 4 confirm the robustness of the tree classier with respect to the training set size.

TABLE 3. Recognition by Dependence Tree Model

Arabic Digits Classes	Dependence Tree Model Success Rate %
0	91.00
1	99.00
2	91.50
3	88.00
4	81.50
5	94.50
6	84.50
7	89.50
8	92.50
9	91.00
OA	90.35

TABLE 4. Sensitivity for the Size of Training Data Set

(Test set is fixed at 2000 samples)

Training data size	Dependence Tree Model Success Rate %
400	72.25
800	81.45
1200	87.40
1600	87.75
2000	90.35

This result compared to result obtained by alternative methods [6] [7] showed the benefit of using tree distribution model.

5. Conclusion

In this paper, we have presented a Tree distribution model for discrete speech recognition. The experimental results obtained on spoken Arabic digits confirm the promising capabilities of the proposed approach. Future developments adopted for more than one tree in classifier design, will hopefully lead to more robust classification results.

6. References

[1] X.Huang, A. Acero, and H. Hon, “Spoken Language Processing”, *Prentice Hall PTR*, 2001.

[2] J.P Hosom, R.Cole and M.Fanty, "Speech Recognition Using Neural Networks at the Center for Spoken Language Understanding", *NSFGraduate Research Traineeships project*, Center for Spoken Language Understanding (CSLU), Oregon Graduate Institute of Science and Technology, USA, Jul. 1999.

[3] F. Bimbot and L. Mathan. "Second-order statistical measures for text independent speaker identification", In Proc. *ESCA Workshop on Automatic Speaker Recognition Identification and Verification*, pp. 51-54, 1994.

[4] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B.Dunn, "Speaker verification using adapted gaussian mixture models", *Digital Signal Processing*, vol. 10, no. 1-3, 2000.

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

[6] Alotaibi,Y.A, “Spoken Arabic digits recognizer using recurrent neural networks”, *Signal Processing and Information Technology*, 2004. Proceedings of the Fourth IEEE International.

[7] Khalid Saeed, Mohammad Kheir Nammous, “Speech-and-Speaker Identification System: Feature Extraction, Description, and Classification of Speech-Signal Image”, *IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS*, VOL. 54, NO. 2, APRIL 2007.

[8] 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

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

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

[11] 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.

[12] J. Pearl, “Probabilistic reasoning in intelligent systems: networks of plausible inference”. *Morgan Kaufmann*, 1988.