International Journal of Scientific & Engineering Research, Volume 4, Issue 10, October-2013
ISSN 2229-5518

Text-Independent Egyptian Colloquial Speaker Recognition based on Hidden Markov Model and Sparse Coding

Sara Mohamed Mustafa, Mohamed Waleed Fakhr, Mohamed S. El-Mahallawy

Abstract—Hidden Markov Model (HMM) is one of the most popular techniques for speech and speaker recognition, while the Sparse Coding (SC) is widely used in face recognition and has not been used widely in speaker recognition. In this paper a comparison between the performances of Sparse Coding and Hidden Markov Model techniques is done on a text-independent speaker recognition task. Speaker recognition is applied on a closed set of 54 speakers speaking short sentences of the Egyptian Colloquial Arabic (ECA). An ergodic HMM is used with Mel Frequency Cepstral Coefficients (MFCCs) features. Sparse Coding is done on the same data; the used Sparse Coding classifiers are the non-negative least square (NNLS) and the linear regression classifier (LRC). The result of the comparison is that the Sparse Coding outperforms the Hidden Markov Model, in particular, when the LRC classifier is used.

Index Terms—Speaker Recognition, Sparse Coding (SC),Hidden Markov Model (HMM), Mel Frequency Cepstral Coefficients (MFFCs), Non-Negative Least Square (NNLS), Linear Regression Classifier (LRC), Egyptian Colloquial Arabic (ECA).

——————————◆——————————

1 INTRODUCTION

SPEAKER recognition refers to recognizing persons from their voice and it is a very important technique in a lot of the applications that we deal with every day like instance authorized check in, banking via telephone, in the telephone credit cards, used by police in tracking persons or criminals over voice networks and many other applications. No one has the same voice as any other one because their vocal tract shapes, larynx sizes, and other parts of their voice production system organs are different. In addition to these physical differences, each speaker has his or her characteristic manner of speaking, including the use of a particular accent, rhythm, intonation style, pronunciation pattern, speaking rate, choice of vocabulary and so on. The most recent speaker recognition systems use a number of these features in parallel, attempting to cover these different aspects and employing them in a complementary way to achieve more accurate recognition [1]. The accuracy of the speaker recognition system is very important achievement because the accuracy of any system evaluates its performance.

Hidden Markov Model encodes the temporal change of the features and efficiently model statistical variation of the features, to provide a statistical representation of how a speaker produces sounds. During enlistment process, HMM parameters are estimated from the speech using established automatic algorithms and during the verification process; the likelihood of the test feature sequence is computed against the speaker's HMMs [2].

Sparse coding has recently gained attention in speaker recognition [3]. SC is representation that account for most or all information of a signal with a linear combination of only a small number of elementary signals, called atoms. The collection of atoms that is used is called a dictionary [4]. In SC the signal is represented by a sparse linear combination of dictionary atoms followed by the dictionary learning, which is the method of constructing the dictionary from the training data. The performance of sparse coding relies on the quality of the dictionary.

To evaluate the performance of HMM and SC techniques on speaker recognition and to present this comparison a different number of cepstrums and different types of features are used. Also for HMM, different transition matrix structures are used.

The rest of the paper is organized as follows. Section 2 introduces a HMM overview on the speaker recognition. In subsections 3.1 and 3.2, the sparse coding concept and the dictionary learning methods are discussed respectively. In section 4, the HMM-SC comparison is explained. Experimental results are detailed in section 5 and conclusions are drawn in section 6.

- *Sara Mohamed Mustafa is currently pursuing masters degree program in electronics and communication engineering in Arab Academy for Science, Technology and Maritime Transport, Egypt ,E-mail: engs_mustafa@yahoo.com*
- *Mohamed Waleed Fakhr is currently a professor of Computer Science in Arab Academy for Science, Technology and Maritime Transport, Egypt , E-mail: waleedfakhr@yahoo.com*
- *Mohamed S. El-Mahallawy is currently an associative professor of Electronics and Communication Engineering in Arab Academy for Science, Technology and Maritime Transport, Egypt , E-mail: mahallawy@aast.edu*

2 Hidden Markov Model

Hidden Markov Model is the most popular technique used with the speech and speaker recognition applications. Speaker recognition system, like any other pattern recognition system, its task involves three stages, feature extraction stage, training stage and testing stage. Training is the process of familiarizing the system with the voice characteristics of a speaker, whereas testing is the actual recognition task [5].

In the last few decades, many methods have been proposed to extract the features of speech. The widely used features are Mel Frequency Cepstral Coefficient (MFCC) based on acoustical characters of human being, Linear Prediction Coefficient (LPC) based on auto-regression model, and RelAtive SpecTrA Perceptual Linear Prediction (RASTA-PLP) based on auditory perception and relative perception [5]. MFCCs are the most familiar features used to describe speech signal. MFCCs are based on the known evidence that the information carried by low frequency components of the speech signal is phonetically more important for humans than the information carried out by high frequency components [6]. MFCC makes use of the mechanism of hearing system effectively and has excellent performance when there is no noise interference; however, the performance of this feature is degraded in the presence of noise. The literatures had shown that the feature from LPC is sensitive to noise and performs worse than MFCC feature. In addition, RASTA-PLP can reduce the influence of channel distortion; however, this feature performs badly under noisy environments [7].

HMM is a doubly embedded stochastic process where the underlying stochastic process is not directly observable (hidden). HMMs have the capability of effectively modeling statistical variations in spectral features. In a various ways, HMMs can be used as probabilistic speaker models for both text-dependent and text independent speaker recognition. HMM not only models the underlying speech patterns but also the temporal sequencing among the sounds and this is a big advantage for text-dependent speaker recognition system. Left-to-Right HMM can model temporal sequence of patterns only, where as to capture the patterns of different type ergodic HMM is used [5]. Ergodic HMM means one must be able to travel from any state to any other state in finite time and that over time states are not visited in a periodic manner. For most ergodic HMM implementations, this constraint is relaxed to just allowing that any state may transition to any other state [8].

3 Sparse Coding

In this section we propose the concept of the sparse coding technique (subsection 3.1). Dictionary learning definition, types of dictionary learning and its recent approaches are introduced in subsection 3.2.

3.1 Sparse Coding concept

Sparse coding has mainly been used with face recognition applications but after a period of time SC has been spread on other applications along with the face recognition like image analysis, image denoising, audio classification, biological data classification and many other applications but it is still limited in speech and speaker recognition applications.

Face recognition (FR) is among the most visible and challenging research topics in computer vision and pattern recognition, and many methods, such as Eigenfaces, Fisherfaces and SVM, have been proposed in the past two decades. Recently, Wright et al. applied sparse coding to FR and proposed the sparse representation based classification (SRC) scheme, which achieves impressive FR performance. By coding a query image **y** as a sparse linear combination of the training samples via the l_1-norm minimization in (1), SRC classifies the query image **y** by evaluating which class of training samples could result in the minimal reconstruction error of it with the associated coding coefficients.

$$\overset{min}{\alpha} \|\alpha\|_1 \; s.t. \; \|y - D\alpha\|_2^2 \;\leq\; \varepsilon \qquad (1)$$

where **y** is a given signal, D is the dictionary of coding atoms, α is the coding vector of **y** overD, and ε> 0 is a constant [9].

In this paper the sparse coding concept is applied on voice via the text-independent speaker recognition application. The dictionary will be discussed in more details in the following subsection.

3.2 Dictionary Learning

Sparse coding relies on an over complete dictionary, i.e., a matrix D which has N columns and M rows, where ($N >> M$). M is the feature vector dimension and N is the number of atoms, or basis in the dictionary. SC assumes that any test frame y can be approximated by a linear combination of some atoms from the dictionary (y = D. α), and α is a length N sparse vector. The dictionaries used in our experiments were a matrices of dimensions ($14 \times 2538\,, 16 \times 2538, \text{and } 18 \times 2538$) where 14, 16 and 18 are the feature vector dimensions (M) and 2538 is the number of the dictionary atoms (N).

Dictionary learning is to construct a dictionary through learning over training data. Julien Mairal et al. [10] introduce a new online dictionary learning technique. The technique focuses on learning the dictionary to adapt it to a specific data and it is based on stochastic approximations which, scales up skillfully to large datasets with millions of training samples.

There is a variety of methods for dictionary learning for example Non-negative Matrix Factorization (NMF), compressed sensing, Kernel dictionary learning, and many other methods. In our SC experiments, kernel dictionary learning method is used.

Kernel-based methods for multimedia retrieval have shown their robustness for many applications, in shape recognition, image retrieval, or event detection for instance. Most methods first build a kernel function, and then train a classifier [11].

Kernel function has different types may be linear, sigmoid, Radial Basis Function (RBF), and polynomial. In our experiments, the built kernel function type used with NNLS classifier was linear function while the built kernel function type used with LRC classifier was RBF function.

4 The HMM-SC comparison

The comparison between HMM and SC was done on a closed set of 54 speakers; each speaker produces around 47 short sentence of the ECA. The environment surrounded the speakers may contain other sounds like opening a door, ringing phone, cough and other sounds. The data files were with bit rate 128 kbps and were not too long, the minimum data file length was less than 1 second and the maximum data file length was 14 seconds. The total number of the data files was 2538 file; we used 2027 data file for training and 511 data file for testing.

For HMM experiments, a model is made for each speaker. The used models had a continuous ergodic topology because it allows the transition between the different states and it is the most suitable topology for the speaker recognition applications. The number of the used states per model was 4 states and the number of the mixtures per state was one of the following values 2 or 5 or 10 mixtures. As well as different features types for example MFCC, MFCC with delta coefficients and, MFCC with delta and the acceleration coefficients are used. The number of cepstrums used was ranging from 14 up to 18 cepstrums. Also, different transition matrix structures are used. Hidden Markov Model Tool Kit (HTK) [12] is used for implementing the HMM experiments. A typical block diagram of speaker recognition task using HMM is shown in Fig2.

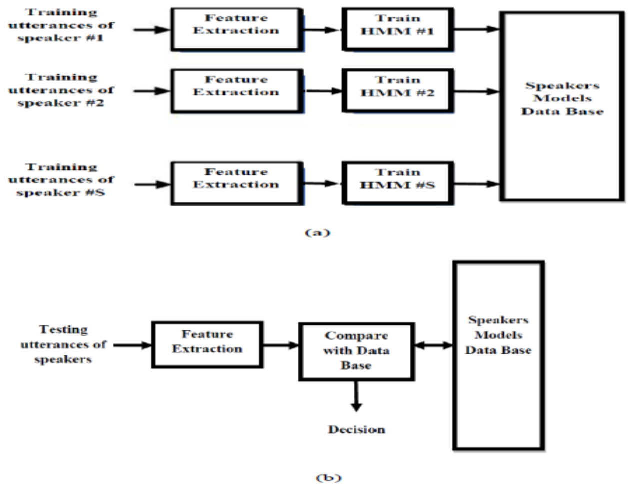

Training utterances of speaker #1
Feature Extraction
Train HMM #1
Training utterances of speaker #2
Feature Extraction
Train HMM #2
Training utterances of speaker #S
Feature Extraction
Train HMM #S
Speakers Models Data Base
(a)
Testing utterances of speakers
Feature Extraction
Compare with Data Base
Speakers Models Data Base
Decision
(b)

Fig. 2. A typical block diagram of speaker recognition task using Hidden Markov Model, (a) Training stage (b) Testing stage

For SC experiments, two types of SC classifiers; NNLS and LRC are used. The aim of the classifier is to find a suitable representation of the test sample, and classify it by checking which class can give better representation than other classes. MFCC features are used; the dimension of the feature vector used is ranging from 14 up to 18 cepstrums and the used dictionary learning is kernel dictionary learning method (as denoted in subsection 3.2). The MATLAB toolbox [13] is used for implementing the SC experiments. A typical block diagram of speaker recognition task using SC is shown in Fig3.

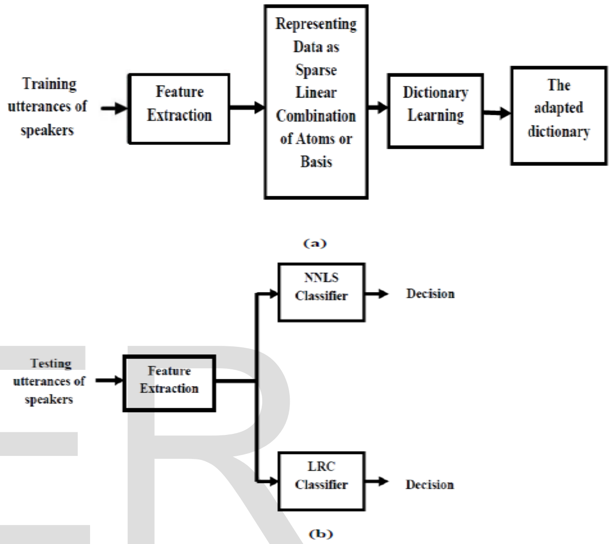


Fig. 3. A typical block diagram of speaker recognition task using Sparse Coding, (a) Training stage (b) Testing stage

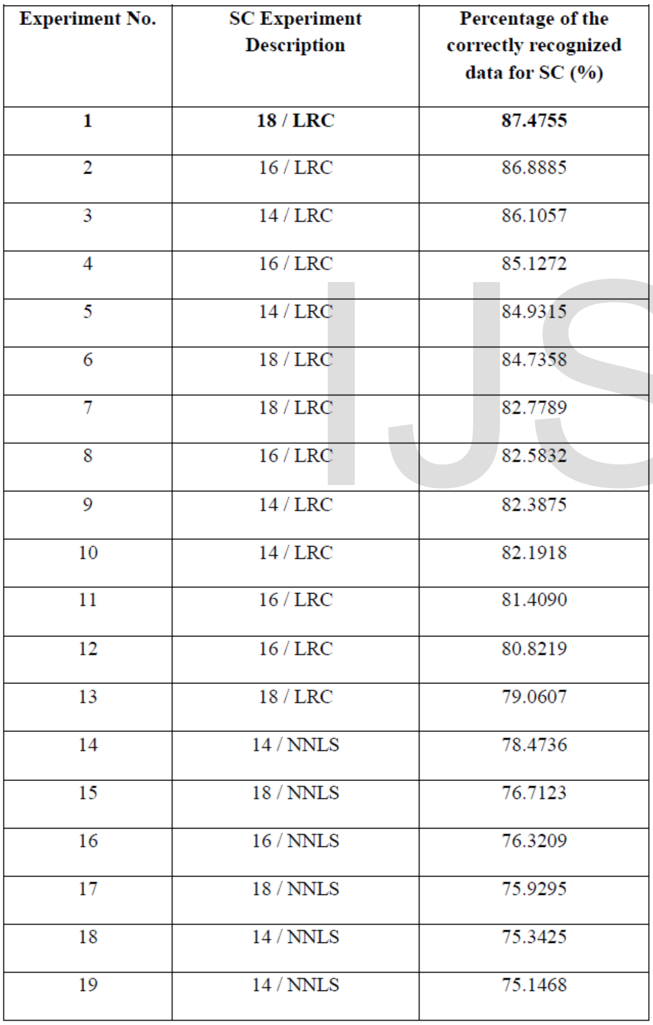
5 Experimental Results

The results of the comparison are shown in tables 1 and 2, which are arranged in descending order w.r.t the higher percentage of correctly recognized data. Table 1 shows the results of the SC technique where, the middle column represents the experiment description (no. of cepstrums/classifier type) and the right one shows the percentage of the correctly recognized data. Table 2 shows the results of the HMM technique where, the middle column represents the experiment description (features type/no. of cepstrums/no. of mixtures per state) and the right one shows the percentage of the correctly recognized data.

As shown on tables 1 and 2, SC technique achieves better results than HMM technique in text-independent speaker recognition system, where the maximum percentage of the correctly recognized data was (87.4755%) for SC and (85.5186%) for HMM and the minimum percentage of the correctly recognized data was (75.1468%) for SC and (73.7769%) for HMM. The best result for the SC (87.4755%) is achieved from the LRC sparse coding classifier with feature vector dimension equals to 18 while, the best result for HMM (85.5186 %) is achieved from an ergodic HMM with number of states equals 4, the number of mixtures per state equals 10 mixtures, the number of cepstrums is 14 and, the features are MFCC. The worst result for SC (75.1468%) is resulted from the NNLS sparse coding classifier with feature vector dimension equals to 14 while, the worst result of HMM (73.7769%) is resulted from an ergodic HMM with number of states equals 4, the number of mixtures per state equals 2 mixtures, the number of cepstrums is 16 and, the features are MFCC.

TABLE1

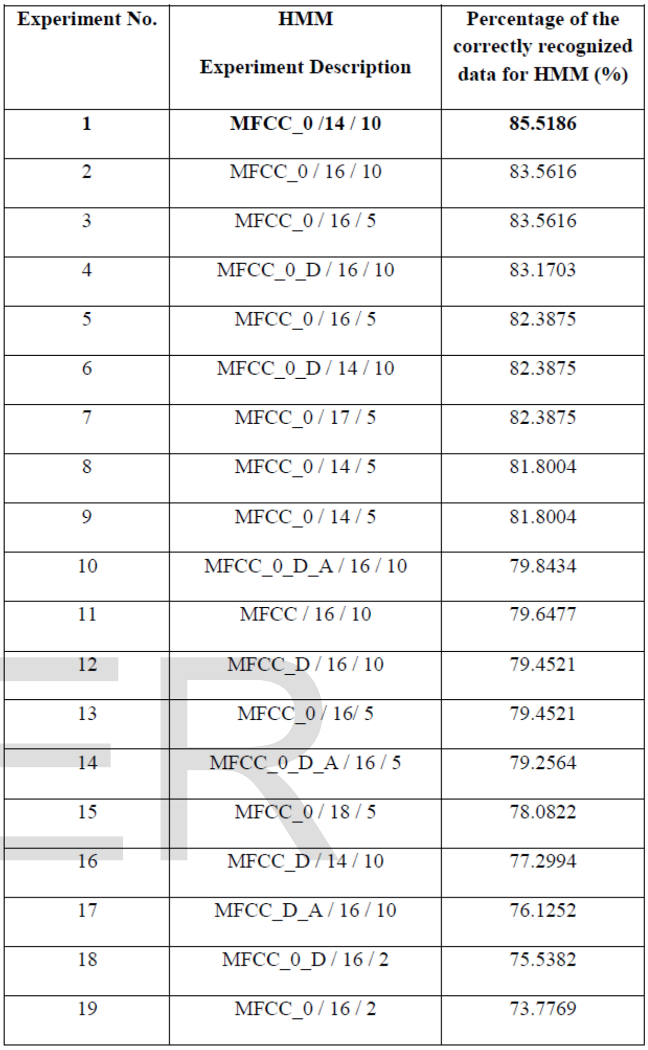
SPARSE CODING TECHNIQUE RESULTS



Experiment No.	SC Experiment Description	Percentage of the correctly recognized data for SC (%)
1	**18 / LRC**	**87.4755**
2	16 / LRC	86.8885
3	14 / LRC	86.1057
4	16 / LRC	85.1272
5	14 / LRC	84.9315
6	18 / LRC	84.7358
7	18 / LRC	82.7789
8	16 / LRC	82.5832
9	14 / LRC	82.3875
10	14 / LRC	82.1918
11	16 / LRC	81.4090
12	16 / LRC	80.8219
13	18 / LRC	79.0607
14	14 / NNLS	78.4736
15	18 / NNLS	76.7123
16	16 / NNLS	76.3209
17	18 / NNLS	75.9295
18	14 / NNLS	75.3425
19	14 / NNLS	75.1468

TABLE2

HIDDEN MARKOV MODEL TECHNIQUE RESULTS



Experiment No.	HMM Experiment Description	Percentage of the correctly recognized data for HMM (%)
1	**MFCC_0 /14 / 10**	**85.5186**
2	MFCC_0 / 16 / 10	83.5616
3	MFCC_0 / 16 / 5	83.5616
4	MFCC_0_D / 16 / 10	83.1703
5	MFCC_0 / 16 / 5	82.3875
6	MFCC_0_D / 14 / 10	82.3875
7	MFCC_0 / 17 / 5	82.3875
8	MFCC_0 / 14 / 5	81.8004
9	MFCC_0 / 14 / 5	81.8004
10	MFCC_0_D_A / 16 / 10	79.8434
11	MFCC / 16 / 10	79.6477
12	MFCC_D / 16 / 10	79.4521
13	MFCC_0 / 16/ 5	79.4521
14	MFCC_0_D_A / 16 / 5	79.2564
15	MFCC_0 / 18 / 5	78.0822
16	MFCC_D / 14 / 10	77.2994
17	MFCC_D_A / 16 / 10	76.1252
18	MFCC_0_D / 16 / 2	75.5382
19	MFCC_0 / 16 / 2	73.7769

6 CONCLUSION

It can be concluded that Sparse Coding outperforms Hidden Markov Model, in particular, when the LRC classifier is used and it is obvious from the results where the maximum percentage of the correctly recognized data for SC is (87.4755%) and for HMM is (85.5186%).

From the experiments, at the same time taken to implement one HMM experiment a multiple number of SC experiments are implemented. The training time for HMM is longer than that of SC; however, the testing time SC takes longer time than HMM especially with NNLS classifier. Thus, it can also be concluded that SC performs its task faster than HMM.

REFERENCES

[1] T. Kinnunen and H. Li, "An Overview of Text-Independent Speaker Recognition: From Features to Supervectors," *Speech Communication,* vol. 52, no. 2010, pp. 12-40, Aug. 2009.

[2] D. A. Reynolds, "An Overview of Automatic Speaker Recognition Technology," IEEE International Conference in Acoustics, Speech, and Signal Processing (ICASSP '02), pp. 4072 - 4075, 2002.

[3] R. Saeidi, A. Hurmalainen, T. Virtanen and D. A. van Leeuwen,"Exemplar-based Sparse Representation and Sparse Discrimination for Noise Robust Speaker Identification," *Odyssey speaker and language recognition workshop,* 2012.

[4] J.F. Gemmeke, T. Virtanen and A. Hurmalainen, "Exemplar-Based Sparse Representations for Noise Robust Automatic Speech Recognition," IEEE *Trans. Audio, Speech, and Language Processing,* vol. 19, no. 7, pp. 2067-2080, 2011.

[5] R. R. Rao, A. Prasad and Ch. K. Rao, "Performance Evaluation of Statistical Approaches for Automatic Text-Independent Speaker Recognition Using Robust Features," *IJCSI International Journal of Computer Science Issues,*Vol. 9, no. 1, pp. 468-476, Jan. 2012.

[6] A. A. Moustafa, "Speaker Recognition Enhancement in Multimedia Environments," MSc thesis, Dept. of Electronics and Communications Eng., Arab Academy for Science and Technology and Maritime Transport, Cairo, 2004.

[7] Y. XIE, J. HUANG and X. WANG," A Robust Feature Based on Sparse Representation for Speaker Recognition," *Journal of Computational Information Systems,* Vol. 9, no. 9, pp. 3553–3561, May 1, 2013.

[8] L. Terry,"Ergodic Hidden Markov Models for Visual-Only Isolated Digit Recognition,"MSc thesis,Dept. of Electronical Eng., Northwestern Univ., Evanston, Illinois, 2007.

[9] M. Yang, L. Zhang, J. Yang, and D. Zhang,"Robust Sparse Coding for Face Recognition,"IEEE Conference in Computer Vision and Pattern Recognition (CVPR '11), pp. 625-632, 2011.

[10] J. Mairal, F. Bach, J. Ponce, G. Sapiro, "Online Dictionary Learning for Sparse Coding,"Proc. Twenty-Sixth Ann. International Conf. MACHINE LEARNING, pp. 689-696, 2009.

[11] P.H. Gosselin, F. Precioso, S. Philipp-Foliguet, "Incremental Kernel Learning for Active Image Retrieval without Global Dictionaries," *Pattern Recognition,* vol. 44, no. 10, pp. 2244-2254, 2011.

[12] "What is HTK?,"http://htk.eng.cam.ac.uk/.

[13] Y. Li, and A. Ngom,"The Sparse Representation (SR) Toolbox in MATLAB,"https://sites.google.com/site/sparsereptool/home.2013