Retracted: A Novel Text-Independent Speaker Verification System Using Ant Colony Optimization Algorithm

Shahla Nemati[1], Reza Boostani[2], and Mohammad Davarpanah Jazi[1]

[1] Department of Electrical and Computer Engineering, Isfahan University of Technology(IUT), Isfahan 84156-83111, Iran
[2] Computer Sciences and Engineering, Campus No.2, Engineering School, Mollasadra St., Shiraz, Iran
s.nemati@ec.iut.ac.ir,
boostani@shirazu.ac.ir,
mdjazi@cc.iut.ac.ir

Abstract. Automatic speaker verification (ASV) has become increasingly desirable in recent years. This system in general, requires 20 to 40 features as input for satisfactory verification. In this paper, features size is reduced by Ant Colony Optimization (ACO) technique to increase the ASV performance. After feature reduction phase, feature vectors are applied to a Gaussian Mixture Model (GMM) which is a text-independent speaker verification Model. Experiments are conducted on a subset of TIMIT corpora. The results indicate that with the optimized feature set, the performance of the ASV system is improved. Moreover, the speed of verification is significantly increased because number of features is reduced over 73% which consequently decrease the complexity of our ASV system.

Keywords: Speaker Verification, Gaussian Mixture Model (GMM), Feature Selection, Ant Colony Optimization (ACO).

1 Introduction

Automatic Speaker Verification (ASV) refers to the task of verifying speaker's identity using speaker-specific information contained in speech signal. Speaker Verification methods are totally divided to text-dependent and text-independent applications. When the same text is used for both training and testing, the system is called to be text-dependent while for text-independent operation, the text used to train and test of the ASV system is completely unconstrained. Text independent speaker verification requires no restriction on the type of input speech. In contrast, Text independent speaker verification usually gives less performance than text dependent speaker verification, which requires test input to be the same sentence as training data [1].

Speech signals contain a huge amount of information and can be described as having a number of different levels of information. At the top level, we have lexical and syntactic features, below that are prosodic features, further below these

A. Elmoataz et al. (Eds.): ICISP 2008, LNCS 5099, pp. 421–429, 2008.

are phonetic features, and at the most basic level we have low-level acoustic features, which generally give information on the system that creates the sound, such as the speakers' vocal tract. Information solely about how the sound is produced (from low level acoustic features) should give enough information to identify accurately a speaker as this is naturally speaker dependent and independent of text [2].

Low level acoustic features also contain some redundant features, which can be eliminated using feature selection (FS) techniques. The objective of FS is to simplify a dataset by reducing its dimensionality and identifying relevant underlying features without sacrificing predictive accuracy. By doing that, it also reduces redundancy in the information provided by the selected features. Selected features should have high inter-class variance and low intra-class variability. Ideally they should also be as independent of each other as possible in order to minimize redundancy [3].

Feature selection has been rarely used in ASV systems. Day and Nandi [2] employed genetic programming (GP) for FS, also L plus-R minus feature selection algorithm is used by Pandit and Kittkr [4] for text-dependent speaker verification. Ant colony optimization (ACO) is a powerful method in many optimization methods [5] and has been employed here for feature selection in ASV systems. In this paper ACO algorithm has been applied to the problem of feature selection in ASV systems.

The rest of this paper is organized as follows. Section 2 presents a brief overview of ASV systems. Feature selection based ACO is described in Sections 3. Next section reports experimental results which include a brief discussion of the results obtained. Finally the conclusion and future research is offered in section 5.

2 An Overview of ASV Systems

The typical process in most proposed ASV systems involves: some form of preprocessing of the data (silence removal) and feature extraction, followed by some form of speaker modeling to estimate class dependent feature distributions (see Fig. 1). A comprehensive overview can be found in [6]. Adopting this strategy the ASV problem can be further divided into the two problem domains of:

1) Preprocessing, feature generation and selection.
2) Speaker modeling and matching.

2.1 Feature Extraction

Most previous works relied on the use of low-level acoustic features. Mel-frequency cepstral coefficients (MFCCs) have been particularly popular for ASV systems in recent years. This transform gives a highly compact representation of the spectral envelope of a sound. Many proposed systems have relied solely on these features and good results have been reported [1,7]. MFCCs have usually been associated

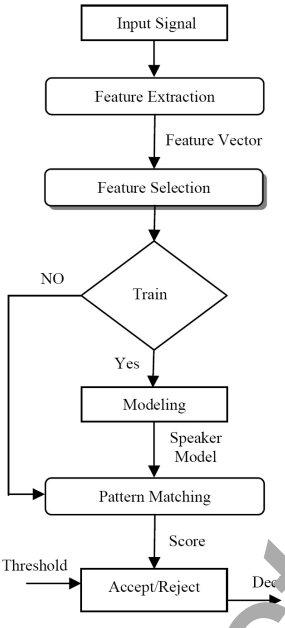

Fig. 1. Overview of the speaker verification process [8]

with speech analysis and processing, and can be used as a compact representation of the spectral envelope.

The process of MFCC analysis is summarized as follows [8]:

- ✓ Segmentation of a speech signal into frames with a fixed window size and a fixed shift period.
- ✓ For each frame:
  - Compute its fast Fourier transform (FFT).
  - Wrap the frequencies according to the Mel scale.
  - Compute the log of the magnitude.
  - Compute the discrete cosine transform (DCT).

2.2 Speaker Modeling

The speaker modeling stage of the process varies more in the literature. The purpose of speaker modeling is characterizing an individual which is enrolled into an ASV system with the aim of defining a model (usually feature distribution values). The two most popular methods in previous works are Gaussian mixture models (GMM) [7] and vector quantization (VQ) [9]. Other techniques such as mono-Gaussian models [10], decision trees [11], SVM [12], and ANN [13] have also been applied. In this paper GMM is used for speaker modeling.

The GMM method is a parametric method that consists of M Gaussian distributions parameterized by their priori probabilities, mean vectors and covariance

matrices. The parameters are typically estimated by maximum likelihood (ML) estimation [7].

Let $X = \{x_1, x_2, ..., x_T\}$ be a set of T vectors, each of which is a d-dimensional feature vector extracted by digital speech signal processing. Since the distribution of these vectors is unknown, it is approximately modeled by a mixture of Gaussian densities, which is a weighted sum of M component densities, given by the equation:

$$p(x_t|\lambda) = \sum_{i=1}^{M} w_i N(x_t, \mu_i, \Sigma_i) \tag{1}$$

where λ denotes a prototype consisting of a set of model parameters $\lambda = \{w_i, \mu_i, \Sigma_i\}$, w_i are the mixture weights and $N(x_t, \mu_i, \Sigma_i)$ are the d-variate Gaussian component densities with mean vectors μ_i and covariance matrices Σ_i

$$N(x_t, \mu_i, \Sigma_i) = \frac{\exp\{-\frac{1}{2}(x_t - \mu_i)' \Sigma_i^{-1} (x_t - \mu_i)\}}{(2\pi)^{d/2} |\Sigma_i|^{1/2}} \tag{2}$$

In training the GMM, these parameters are estimated such that they are fitted with the distribution of the training vectors. The most widely used training method is the ML estimation. For a sequence of training vectors X, the likelihood of the GMM is:

$$p(X|\lambda) = \prod_{t=1}^{T} p(x_t|\lambda) \tag{3}$$

The classification for speaker verification may be performed based on the log likelihood ratio (LLR). Given the observation X, the LLR is defined as:

$$LLR(X) = \log p(X|\lambda) \tag{4}$$

Let Θ be a given threshold, associated with the claimed speaker model λ. A discriminant function $LLR(X)$ for the unknown X and model λ used to reject or accept the claimed speaker is as follows:

$$LLR(X) \begin{cases} \geq \Theta & Accept \\ < \Theta & Reject \end{cases} \tag{5}$$

Where threshold Θ is taken as an equal error rate (EER) threshold [8].

3 Application of ACO for Feature Selection

In the early 1990s, ant colony optimization (ACO) was introduced by M. Dorigo and colleagues as a novel nature-inspired meta-heuristic for the solution of hard combinatorial optimization (CO) problems [5]. ACO belongs to the class of meta-heuristics, which includes approximate algorithms used to obtain good enough solutions to hard CO problems in a reasonable amount of computation time.

The paradigm is based on the observation made by ethologists about the medium used by ants to communicate information regarding shortest paths to

food by means of pheromone trails. A moving ant lays some pheromone on the ground, thus, a path is made by a trail of this substance. While an isolated ant moves practically at random (exploration), an ant encountering a previously laid trail can detect it and decide with high probability to follow it and consequently reinforce the trail with its own pheromone (exploitation)[5].

The feature selection task is reformulated based on the ACO which requires a problem to be represented as a graph. Here nodes represent features, with the edges between them denoting the choice of the next feature. The search for the optimal feature subset is then an ant traversal through the graph where a minimum number of nodes are visited that satisfies the traversal stopping criterion. The heuristic desirability of traversal and edge pheromone levels are combined to form the so-called probabilistic transition rule, denoting the probability of ant k at feature i choosing to travel to feature j at time t [14]:

$$P_{ij}^{k}(t) = \begin{cases} \frac{[\tau_{ij}(t)]^{\alpha}.[\eta_{ij}]^{\beta}}{\sum_{l \in J_i^k}[\tau_{il}(t)]^{\alpha}.[\eta_{il}]^{\beta}} & j \in J_i^k \\ 0 & otherwise \end{cases} \tag{6}$$

Where, η_{ij} is the heuristic desirability of choosing feature j when at feature i (η_{ij} is optional but often needed for achieving a high algorithm performance), J_i^k is the set of neighbor nodes of node i which have not yet been visited by the ant k. $\alpha > 0, \beta > 0$ are two parameters that determine the relative importance of the pheromone value and heuristic information (the choice of α, β is determined experimentally) and $\tau_{ij}(t)$ is the amount of virtual pheromone on edge (i,j). The pheromone on each edge is updated according to the following formula:

$$\tau_{ij}(t+1) = (1-\rho).\tau_{ij}(t) + \sum_{k=1}^{m} \Delta_{ij}^{k}(t) \tag{7}$$

where:

$$\Delta_{ij}^{k}(t) = \begin{cases} \gamma'(S^k)/|S^k| & if\ (i,j) \in S^k \\ 0 & otherwise \end{cases} \tag{8}$$

The value $0 \leq \rho \leq 1$ is decay constant used to simulate the evaporation of the pheromone, S^k is the feature subset found by ant k. The pheromone is updated according to both the measure of the "goodness" of the ant's feature subset ($\acute{\gamma}$) and the size of the subset itself. By this definition, all ants can update the pheromone. The overall process of ACO feature selection for ASV is shown in Fig. 2.

4 Experimental Results

4.1 TIMIT Dataset

The TIMIT corpora [15] is used in this paper. This corpus contains 630 speakers (438 male and 192 female) representing 8 major dialect regions of the United States, each speaking ten sentences. There are two sentences that are spoken by

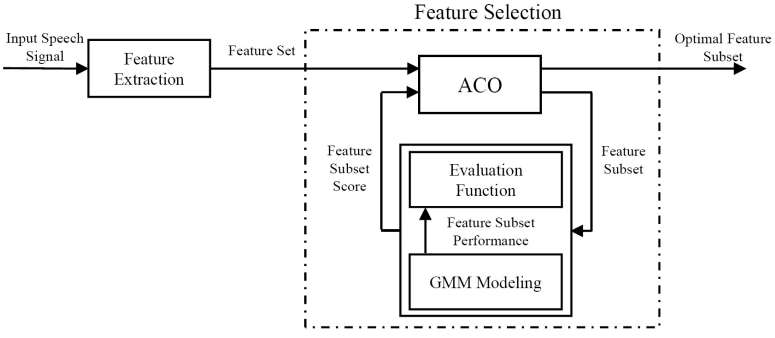

Fig. 2. Overall process of ACO feature selection for ASV

all speakers and the remaining eight are selected randomly from a large database. The speech signal is recorded through a high quality microphone with a sampling frequency of 16 kHz in a quiet environment, with no session interval between recordings.

Eight sentences (SX, SI) were used to develop each speaker model, and the remaining 2 SA sentences were used to test each speaker. The 35 speakers included in both the test and train directories were used during the TIMIT(35) trials.

4.2 Evaluation Measure

The evaluation of the speaker verification system is based on detection error tradeoff (DET) curves, which show the tradeoff between false alarm (FA) and false rejection (FR) errors. Typically equal error rate (EER), which is the point on the curve where $FA = FR$, is chosen as evaluation measure. We also used detection cost function (DCF) defined as [7]:

$$DCF = C_{miss}.E_{miss}.P_{target} + C_{fa}.E_{fa}.(1 - P_{target}) \tag{9}$$

where P_{target} is the priori probability of target tests with $P_{target} = 0.01$, E_{miss} and E_{fa} are false rejection rate and false acceptance rate respectively at a operating point and the specific cost factors $C_{miss} = 10$ and $C_{fa} = 1$. Hence, the point of interest is shifted towards low FA rates.

4.3 Experimental Setup

Experiments were conducted on a subset of TIMIT corpora consists of 22 male and 13 female speakers of different accent that were selected randomly. Data were processed in 30 ms frames (480 samples) with 50% overlaps. Frames were segmented by Hamming window and pre-emphasized with $\alpha = 0.97$ to compensate the effect of microphone's low pass filter. At first, for each frame MFCCs were extracted from silence removed data. Moreover, delta coefficients were calculated based on the MFCCs and appended to them. Furthermore, two energy terms were also extracted to form input vectors. ACO-based feature selection

Table 1. Selected Features of ACO-GMM

Method	Number of Original Features	Number of Selected Features	Selected Features
ACO-GMM (32)	26	7	1,2,3,4,6,9,12
ACO-GMM (32)	42	7	1,4,5,7,9,18,24
ACO-GMM (64)	26	7	1,3,4,6,7,8,10
ACO-GMM (64)	42	6	1,4,5,6,8,13

Table 2. Equal Error Rate (EER) Results (%) and DCF for GMM and ACO-GMM

Number of MFCCs	GMM-32		ACO-GMM-32		GMM-64		ACO-GMM-64	
	EER	DCF	EER	DCF	EER	DCF	EER	DCF
26	6.79	0.0689	**3.76**	0.0388	6.99	0.0731	**7**	0.0691
42	9.65	0.0937	**8.16**	0.0615	8.64	0.0814	**7.89**	0.0943

was applied to input vectors which was described earlier. Then, verification process was performed using the GMM approach. The performance criterion is due to EER and DCF according to an adopted decision threshold strategy.

Table 1 shows the number of selected features and selected features which were chosen by ACO. As we can see in table 1, ACO can degrade dimensionality of features over 73%. EER and DCF for GMM and ACO-GMM with different number of Gaussian (32, 64) were shown in Table 2.

DET curves for GMM and ACO-GMM with 32 Gaussian are shown in Figure 3, and those for 64 Gaussian are shown in Figure 4. From the results, it can be seen that ACO-GMM yields a significant improvement in speed than the baseline GMM approach. The improvement is due to the selection of optimal feature set by ACO algorithm.

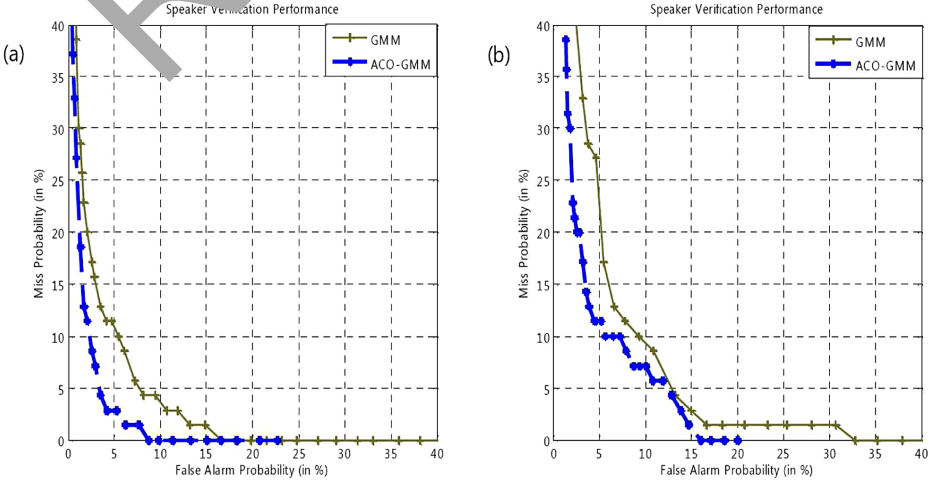

Fig. 3. DET curves for GMM and ACO-GMM with 32 Gaussians (a)original features 26 (b) original features 42

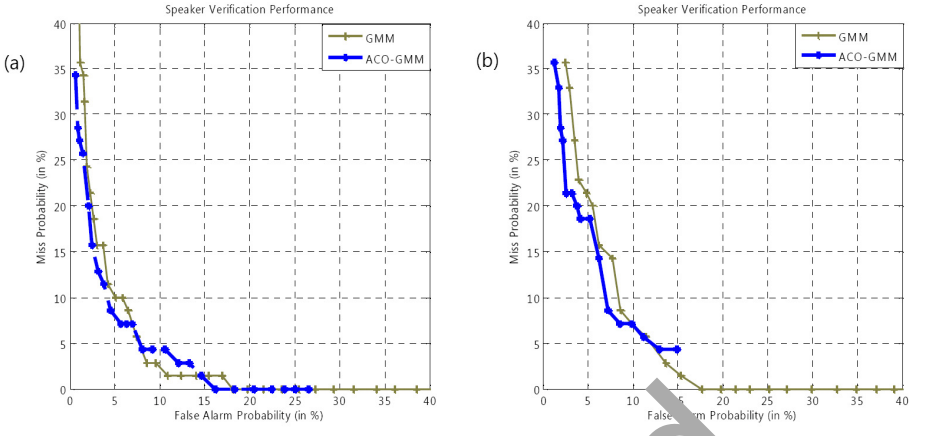

Fig. 4. DET curves for GMM and ACO-GMM with 64 Gaussians (a)original features 26 (b) original features 42

5 Conclusion and Future Research

In this paper, we have addressed the problem of optimizing the acoustic feature set by ACO technique for text-independent speaker verification system based on GMM. ACO selected the relevant features among all Mel-Cepstrum coefficients in order to increase the performance of our ASV system. The experimental results on subset of TIMIT database showed that ACO is able to select the most informative features without loosing the performance. The feature vectors size reduced over 73% which led to a less complexity of our ASV system. Moreover, verification process in the test phase speeds up because less complexity is achieved by the proposed system in comparison with current ASV systems.

For future work, the authors plan to investigate the performance of proposed ASV system by taking advantage of using GMM-UBM and other models instead of GMM. Finally, another research direction will involve experiments with different datasets.

Acknowledgments. The authors wish to thank the Office of Graduate studies of the Isfahan University of Technology for their support.

References

1. Xiang, B., Berger, T.: Efficient Text-Independent Speaker Verification with Structural Gaussian Mixture Models and Neural Network. IEEE Transactions On Speech And Audio Processing 11(5) (2003)
2. Day, P., Nandi, A.K.: Robust Text-Independent Speaker Verification Using Genetic Programming. IEEE Transactions On Audio, Speech, And Language Processing 15(1), 285–295 (2007)

3. Jensen, R.: Combining rough and fuzzy sets for feature selection. Ph.D. Thesis, University of Edinburgh (2005)
4. Pandit, M., Kittkr, J.: Feature Selection for a DTW-Based Speaker Verification System, pp. 796–799. IEEE, Los Alamitos (1998)
5. Dorigo, M., Caro, G.D.: Ant Colony Optimization: A New Meta-heuristic. In: Proceeding of the Congress on Evolutionary Computing (1999)
6. Campbell, J.P.: Speaker recognition: a tutorial. Proc. IEEE 85(9), 1437–1462 (1997)
7. Reynolds, D.A., Rose, R.C.: Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Transactions On Speech And Audio Processing 3(1), 72–83 (1995)
8. Cheung-chi, L.: GMM-Based Speaker Recognition for Mobile Embeded Systems, Ph.D. Thesis, University of Hong Kong (2004)
9. Linde, Y., Buzo, A., Gray, R.M.: An Algorithm for Vector Quantizer Design. IEEE Trans. Comm. 28, 84–95 (1980)
10. Bimbot, F., Magrin-Chagnolleau, I., Mathan, L.: Second-order statistical measures for text-independent speaker identification. Speech Commun 17(12), 177–192 (1995)
11. Navratil, J., Jin, Q., Andrews, W., Campbell, J.: Phonetic speaker recognition using maximum-likelihood binary-decision tree models. In: Proc. ICASSP, Hong Kong (2003)
12. Wan, V.: Speaker Verification Using Support Vector Machines, Ph.D. disseration, Univ. Sheffield, U.K (2003)
13. Wouhaybi, R., Al-Alaou, M.A.: Comparison of neural networks for speaker recognition. In: Proc. 6th IEEE Int. Conf. Electronics, Circuits Systems (ICECS), vol. 1, pp. 125–128 (1999)
14. Kanan, H.R., Faez, K., Hosseinzadeh, M.: Face Recognition System Using Ant Colony Optimization-Based Selected Features. In: CISDA 2007. Proceeding of the First IEEE Symposium on Computational Intelligence in Security and Defense Applications, pp. 57–62. IEEE Press, Los Alamitos (2007)
15. Garofolo, J., et al.: DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM. National Institute of Standards and Technology (1990)