Speaker Identification based on Hybrid Clustering and Radial Basis Function

Yap Teck Ann[1], Mohd Shafry Mohd Rahim[1], Ayman Altameem[2], Amjad Rehman[4], Ismail Mat Amin[1], Tanzila Saba[3]

[1.] Faculty of Computer Science and Information Systems University Teknologi Malaysia, 81310 Skudai Malaysia
[2.] College of Applied Studies and Community Services King Saud University Riyadh KSA
[3]College of Computer Science and Engineering Salman Abdul Aziz University Alkharj KSA
[4]MIS Department College of Business Administration Salman bin Abdul Aziz University Alkharj KSA
tanzilasaba@yahoo.com

Abstract: Speaker identification is the computing task to identify an unknown identity based on the voice. A good speaker identification system must have a high accuracy rate to avoid invalid identity. Despite of last few decades' efforts, accuracy rate in speaker identification is still low. In this paper, we propose a hybrid approach of unsupervised and supervised learning i.e. subtractive clustering and radial basis function(Sub-RBF).The proposed fused technique yields promising results because subtractive clustering is able to solve the initial guesses of cluster center and difficulty level to determine the number of cluster. Besides that, RBF has simple network structure and faster learning algorithm. In RBF input to output map uses the local approximations which will combine the linear approximations and causes the linear combinations with less weight. RBF neural network model uses subtractive clustering algorithm to select the hidden node centers for high training speed. In the meantime, the RBF network is trained with a regularization term so as to minimize the variances of the nodes in the hidden layer and to perform accurate prediction. Promising results are achieved to identify speaker using proposed fused approach.
[Ann Y. T, Rahim M.S.M., Altameem A, Rehman A, Amin, I, M. Saba T. **Speaker Identification based on Hybrid Clustering and Radial Basis Function.** *J Am Sci* 2012;8(10):71-75]. (ISSN: 1545-1003). http://www.jofamericanscience.org. 12

Keywords: Speaker identification, features extraction, clustering, radial basis function, K-means and fuzzy c-means.

1. Introduction

Speaker identification is the process of determining an identity from the registered speakers by providing an utterance. The utterance used to claim on the identity of the speaker by comparing the utterance with Ns trained speaker in its user database. So, the role of automatic speaker identification (ASI) is complex and the error of the system will increased if the Ns increase. The speaker verification is the process to decide the identity claim of a speaker either to accept or reject the claim. For the speaker verification, a speaker's recording is matched to previous recording, which is made during the voice registration period to produce the result. The result accepts or refuses the match between the speaker's recording and the previous recording [1-3].

Both of the speaker identification and speaker verification could be classified into text-dependent and text-independent. During the enrollment phase, the speaker's voice is recorded and the features of the voice are extracted and stored in the database. If this is text-dependent, the utterances use in the operational phase must be same with the utterances during the enrollment phase. On the other hand, for the text-independent, the utterances use in the operational phase is different with the utterances during the enrollment phase [4]. In this paper, we proposed a hybrid method to solve the text-independent speaker identification system.

Literature is replete with speaker recognition approaches. The most common and widely used techniques are based on Hidden Markov models, Gaussian mixture models, pattern models, pattern matching algorithm, neural network and so on [5]. Basically, there are two main approaches; unsupervised and supervised learning along with their own pros/cons. Unsupervised learning is suitable to learn large and complex models than supervised learning as if the models contain a large and complex dataset then supervised learning increase number of connections in the network. Additionally, supervised learning is time consuming. Besides that, K-means clustering is a famous unsupervised learning which has two critical problems. First, K-means clustering uses initial guesses to choose the first cluster centre. Second, it is hard to determine the number of clusters. Improper selection of cluster centre and incorrect estimate of cluster's number may degrade the learning process. However, speaker identification issue is still fresh due to its low accuracy rate. Accordingly, improved statistical modeling may produce better performance for the speaker identification. A better performance of the speaker identification for the speed and accuracy aspects will attract to use speaker

identification as the input method and replace the old ones.

2. Related Work

2.1 Speaker Identification Based on Subtractive Clustering Algorithm with Estimating Number of Clusters

Subtractive clustering algorithm is an improved form of the mountain clustering that resolves the problem of high dependence on grid resolution and the data dimension. However, it is inefficient when applies in the high dimension dataset in the mountain clustering. Besides that, subtractive clustering has advantages of locating cluster centers and prior information about the number of clusters. The number of clusters is obtained by investigating the mutual relationship between clusters.

The proposed approach resolves the problems of K-means and fuzzy c-means (FCM). The major problems are the improper initial estimation of cluster centers that may degrade the performance and the number of clusters that cannot be defined in advance.

The proposed algorithm is based on subtractive clustering algorithm [6,7] and mutual relationship of clusters [8]. First, cluster centers are obtained incrementally by adding one cluster centre at a time through the subtractive clustering algorithm. Second, investigate the mutual relationship of a cluster with respect to all the other clusters, to obtain the estimation of the number of cluster. Two statistically dependent clusters are decided by the mutual relationship.

2.2 Speaker Identification Using Reduce RBF Networks Array

The purpose of speaker identification is to determine a speaker's identity from the speech utterances. Essentially, speaker identification is a pattern recognition problem of a speech signal [9]. Both of the training and recognition processes are important for the speaker identification. Those processes include the identification of discriminating features representing the specific characteristics of the speakers and the choice of the classifier. Radial basis function (RBF) networks are chosen due to its successful use of Gaussian Mixture Models in speaker identification.

The RBF network is a three-layer NN, which has the same underlying structure as the Gaussian Mixture Models (GMM) when Gaussian function is selected as the type of basis function in the RBF network[10]. Another reason to choose RBF is RBF network easy to use. The key point in design of radial basis function networks is to specify the number and the locations of the centers. Input training data (IC) and both the input and output data (IOC) are the ways to obtain the vectors from the clustering algorithm. In this paper, IOC is chosen because the structure of centers designed in the IOC is more effective. To select the suitable number or network centers, this method uses the recursive orthogonal least-squares (ROLS) algorithm after the training process. Mel's Frequency Cepstral Coefficients (MFCC) are extracted as the features for speaker characteristics. The system is composed of some binary classifiers, while the binary partitioned approach has been shown as an efficient solution for reducing training time [18].

3. Hybrid Subtractive Clustering With RBF In Speaker Identification (Sub-RBF)

3.1 Subtractive Clustering

Subtractive clustering [11] is an unsupervised clustering which is an upgraded version of the mountain clustering. Mountain clustering is proposed by Yage [12]. However, the mountain clustering face a critical problem, its computation grows exponentially with the dimension of the data. This problem occurs because the mountain clustering has to evaluate at each of the grid point. Subtractive clustering solves this problem by assuming all the data points as the candidates for cluster center, instead of grid points as in mountain clustering. This means that, subtractive clustering grows exponentially with the size of data instead of the data dimension. Besides that, the subtractive clustering also solves the problem of the initial guesses of the cluster center. A clustering method which does the initial guesses may face the problem of the improper initial guesses of the center point. The improper initial guesses of the cluster point will degrade the performance of the clustering method. The examples of the clustering algorithm which does the initial guesses are K-means and Fuzzy C-means (FCM).

3.2 Radial Basis Function

Radial basis function(RBF) neuron network is proposed by J.Moody and C.Darken in 1980's which has been successfully applied to solve many problems, such as pattern recognition, function approximation, system modeling, signal processing, time series prediction, and control.

RBF neural network is a multi-layer-feed-forward network which has a feed-forward structure with a modified hidden layer and training algorithm. RBF neural network uses radial basis function as activation function which allow Gaussian function as the type of basis function in the RBF network [13]. In pattern classification applications, the Gaussian function is preferred [13]. Using the Gaussian function as the activation function, RBF network is

capable to form an arbitrarily close approximation to any continuous function [14,15].

3.3 Sub-RBF

This process is termed as back-end processing. The main purpose of this process is to generate the speaker model for each of the sample speech in the training phase and stores into the database. First of all, the algorithm involved in this phase is subtractive clustering and the radial basis function. In this project, the subtractive clustering will be used as the pre-classifier for the radial basis function which use the vectors obtained from the front-end processing as the input and the produced result in subtractive clustering will be used as the input for the radial basis function.

4. Proposed Methodology

There are two phases in the speaker identification: enrolment phase and identification phase.

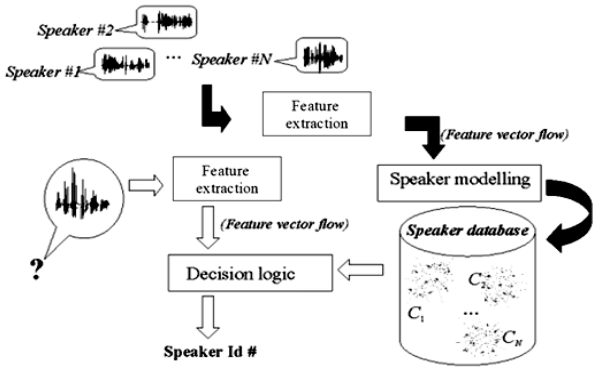

Fig. 1 Schematic diagram of the speaker identification system

In the enrolment phase, speech samples are collected from the speaker and convert the speech signal into set of feature vectors, which characterize the properties of speech that can separate different speakers. This objective is achieved by the proposed method to train the speech samples. The collections of the enrolment models are stored in the speaker database.

In the identification phase, the test sample of unknown speaker is analyzed and compared with all models from the speaker database. Both enrolment and identification phase has the same first step, feature extraction which converts the sampled speech signal into set of feature vectors, which characterize the properties of speech that can separate different speakers. The main purpose of this step is to reduce the amount of test data while retaining speaker discriminative information.

5. Implementation

Following are the main steps involved in Sub-RBF process to identify the speaker.

5.1. Select the cluster center

The main concept of the subtractive clustering is subtractive clustering assumed all of the data points as a potential center point, and select the cluster center according to the density of the surrounding of the data point. To select the cluster center, the density of the data points is defined as density measure (equ. i)

$$D_i = \sum_{j=1}^{N} \exp\left(- \frac{\| X_i - X_j \|^2}{(r_a / 2)^2} \right) \quad \text{(i)}$$

where r_a is a positive constant representing a neighborhood radius. The density measure for each of the data point are calculated, and the data points with the highest value of density measure, D_i will be selected as the first cluster center.

5.2. Remove all data points in the vicinity of the first cluster center

From the previous step, the first cluster center is obtained. Next, correct the density measure to avoid the next cluster center in the incinity of the cluster center. Calculate the density measure of each data point, Xi and the data point with the highest value of density measure, Dc1 will be selected as the next center point. The density measure of each data point is revised according to equation (ii).

$$D_i = D_i - D_{c1} \exp\left(- \frac{\| X_i - X_{c1} \|^2}{(r_b / 2)^2} \right) \quad \text{(ii)}$$

Where, r_b is a positive constant which defines a neighborhood that has measurable reductions in density measure. Therefore, the data points near the first cluster center X_{c1} will have significantly reduced density measure. [15,16].

5.3. Repeat iteration

After revising the density measure formula, the data points with the highest value of density measure is selected to ask the next cluster center. This process repeated until the sufficient number of clusters is attainted.

5.4. Compute the spreads

In this step, spread, for the RBF function will be computed based on the distance of each of the cluster centers. The spread is computed as follows:

$$\sigma = \frac{\text{Maximum distance between any 2 centers}}{\sqrt{\text{number of centers}}} = \frac{d_{max}}{\sqrt{m_1}} \quad \text{(iii)}$$

Then the activation function of hidden neuron is

$$\phi_i\left(\|x-t_i\|^2\right)=\exp\left(-\frac{m_1}{d_{max}^2}\|x-t_i\|^2\right) \text{ - (iv)}$$

Where, x is the input and t_i is the i^{th} center.

5.5. LMS algorithm finding weights

Once obtains the cluster centers and the values of the spread, the weight of the neuron could be computed by using the supervised learning. In this study, the supervised learning is based on Least Mean Squares (LMS) algorithm. Assume there are N training samples:

{x(1),d(1)}, {x(2),d(2)}, {x(3),d(3)},..., {x(N),d(N)} (v)

Where d(k) is the target value of x(k).

Since the center of each neuron in the hidden layer is obtain from the previous step, thus the weights in the LMS algorithm are computed as below:

$$f[x(k)] = \sum_{j=1}^{m} w_j \phi_j(k) \text{ ---------- (vi)}$$

5.6 Decision phase

From the previous phase, the outputs of the RBF network are obtained. The next step of the proposed method is to determine the tested data belongs to training data. To make the decision, the distortion distance between the tested data and the training data is measured using equation (vii)

$$\sqrt{(p_1-q_1)^2+(p_2-q_2)^2+\cdots+(p_n-q_n)^2}=\sqrt{\sum_{i=1}^{n}(p_i-q_i)^2} \text{ (vii)}$$

The lower of the distortion distance of the second data exhibits that second data is more similar. The speaker with the lowest distortion distance is identified as the speaker of the tested data.

5.7 Speech database

The data set involves the selection from the TIMIT Acoustic-Phonetic Continuous Speech Corpus. There are 16 speakers selected from 8 dialect regions of the United States (1 male and 1 female from each dialect region) which contains a 130 sentences, 10 sentences are spoken by each speaker. The TIMIT corpus includes time-aligned orthographic, phonetic and word transcriptions as well as a 16-bit, 16 kHz speech waveform file for each utterance.

6. Experimental Results

Each of the speakers provide one of the speech sample as the training set and the speech samples are converted into set of feature vectors. The MFCC is used to get the feature vectors from all the speech samples and the sub-RBF algorithm is used to produce the speaker model from the feature vectors. For the identification phase, ten utterances for each of the speakers are analysed and compared with all the speaker models. The vector set from each of the utterance is measure the distortion distance with all the speaker models and the lowest distortion distance is chosen to be identified as the unknown person.

Based on the result, fusion of subtractive clustering and radial basis function has a better performance from the aspect of accuracy rate. Since the subtractive clustering is an unsupervised learning which is highly suitable to learn larger and complex models in scenario of pattern matching. So application of only subtractive clustering approach will not perform well and results into incorrect speaker identification.

Table 1. Speaker identification results based on Subtractive clustering

Speaker	Correct
dr1-fvmh0	0/10
dr1-mcpm0	1/10
dr2-faem0	1/10
dr2-marc0	0/10
dr3-falk0	0/10
dr3-madc0	0/10
dr4-falr0	7/10
dr4-maeb0	3/10

Speaker	Correct
dr5-ftlg0	0/10
dr5-mbgt0	0/10
dr6-fapb0	0/10
dr6-mbma1	0/10
dr7-fblv0	0/10
dr7-madd0	0/10
dr8-fbcg1	0/10
dr8-mbcg0	6/10

Table 2. Speaker identification results based on hybrid approach

Speaker	Correct
dr1-fvmh0	10/10
dr1-mcpm0	3/10
dr2-faem0	2/10
dr2-marc0	10/10
dr3-falk0	1/10
dr3-madc0	8/10
dr4-falr0	510
dr4-maeb0	7/10

Speaker	Correct
dr5-ftlg0	7/10
dr5-mbgt0	6/10
dr6-fapb0	2/10

dr6-mbma1	4/10
dr7-fblv0	10/10
dr7-madd0	10/10
dr8-fbcg1	10/10
dr8-mbcg0	10/10

7. Conclusion

This paper has presented a hybrid approach of subtractive clustering and radial basis function to identify speakers. Based on the results of the subtractive clustering, fusion of subtractive clustering and RBF, it is observed that hybrid approach yields better accuracy than individual approaches. Additionally, hybrid of subtractive clustering and RBF has promising performance of speaker identification system in comparison with the subtractive clustering. However, the proposed method has the advantages of the unsupervised and supervised learning that results into better accuracy. Finally, the proposed approach uses subtractive clustering which can proceed faster to get the cluster center. Hence, in the decision phase, proposed method is based on the supervised learning to identify the speaker using least processing time and higher accuracy rate.

Corresponding Author:
Dr. Tanzila Saba
College of Engineering and Computer Sciences
Salman bin Abdul Aziz University Alkharj KSA
E-mail:tanzilasaba@yahoo.com

References

1. Rahim, MSM, Razzali, N. Sunar, M.S. Altameem, A and Rehman, A. Curve interpolation model for visualising disjointed neural elements. Neural Regeneration Research 2012;7(21):1637-1644.
2. Rehman, A. and Saba, T. Features extraction for soccer video semantic analysis: current achievements and remaining issues Artificial Intelligence Review, 2012; doi: 10.1007/s10462-012-9319-1.
3. Rehman, A. and Saba, T. Off-line Cursive Script Recognition: Current Advances, Comparisons and Remaining Problems. Artificial Intelligence Review Springer, 2012; 37(4): 261-268.
4. Saba, T. Sulong, G. and Rehman, A. Document Image Analysis: Issues, Comparison of Methods and Remaining Problems. Artificial Intelligence Review, 2011; 35(2): 101-118.
5. Rehman, A. and Saba, T. Performance Analysis of Segmentation Approach for Cursive Handwritten Word Recognition on Benchmark Database. Digital Signal Processing, 2011;21: 486-490.
6. Chiu, S.L. Fuzzy model identification based on cluster estimation, J. of Intelligent and Fuzzy sys. 1994.
7. Kothari, R., Pittas, D. On finding the number of clusters, Pattern Recognition Letters. 1999.
8. Yang, Z.R., Zwaliuski, M.: Mutual information theory for adaptive mixture model, IEEE Trans. Pattern and Machine Intelligence. 2001.
9. Xicai, Y., Datian, Y., Chongxun, Z.: Neural networks for improved text independent speaker identification. IEEE Engineering in Medicine and Biology Magazine, 2002; 21.
10. Reynodls D. Rose R. Robust Text-independent Speaker Identification using Gaussian Mixture Speaker Models. IEEE Trans on Speech and Audio processing, 1995; 3.
11. Chiu, S. Fuzzy Model Identification Based on Cluster Estimation. Journal of Intelligent & Fuzzy Systems, 1994; 2(3).
12. Yager, R. and Filev, D. Generation of Fuzzy Rules by Mountain Clustering, Journal of Intelligent & Fuzzy Systems, 1994; 2(3): 209-219.
13. Bors A.G. Pitas I. Median Radial basis Functions Neural Network. IEEE Trans. On Neural Networks, 1996; 7(6): 1351-1364.
14. Lian, H. Wang, Z. Wang, J. and Zhang, L. Speaker Identification Using Reduced RBF Networks Array. Dept. E.E, Fudan University, Shanghai 200433, China ,2004.
15. Mohamad, I. Rahim, MSM, Bade, A. Rehman, A. and Saba, T. Enhancement of the Refinement Process for Surface of 3D Object. Journal of American Science. 2012; 8(4):358-365.
16. Norouzi,A. Saba,T. Rahim, MSM, Rehman, A. Visualization and Segmentation of 3D Bone from CT images, International Journal of Academic Research, 2012, 4(2): 202-208.

8/8/2012