Efficient Feature Vector Clustering for Automatic Speech Recognition Systems

Lilia Lazli
Department of Electrical Engineering , ÉTS, University of Quebec, Montreal, Quebec Canada
lilia.lazli.1@ens.etsmtl.ca

Mounir Boukadoum
Department of Computer Science, UQAM, University of Quebec, Montreal, Quebec Canada
boukadoum.mounir@uqam.ca

Otmane Ait Mohamed
Department of Electrical Engineering and Computer Science, Concordia University, Montreal, Quebec Canada
ait@ece.concordia.ca

Mohamed-Tayeb Laskri
Department of Computer Science, UBMA, University of Badji Mokhtar, Annaba Algeria
laskri@univ-annaba.org

ABSTRACT

In this paper, we present an efficient algorithm for the clustering of speech data. The algorithm based on regulating a similarity measure to set the number of clusters and the cluster boundaries, thus overcoming the shortcomings of conventional clustering algorithms such as k-Means and Fuzzy C-Means, which require a priori knowledge of the number of clusters, the use of similarity measure that follows the data distribution, and are sensitive to the choice of initial configuration, The algorithm performance was tested in an HMM/MLP automatic speech recognition system, with the results were compared with those obtained when using a combination of Fuzzy C-Means and genetic algorithms to do the clustering, showing better performance.

KEYWORDS

Unsupervised speech clustering, genetic algorithm, fuzzy C-means algorithm, speech recognition system, HMM/MLP.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
BICT 2017, March 15-16, Hoboken, United States
ISBN 978-1-63190-148-5
DOI: 10.4108/eai.22-3-2017.152399
Copyright © 2017 EAI

1 INTRODUCTION

Cluster analysis is a main task of exploratory data mining, and a common technique for statistical data analysis. It is used in many fields, including machine learning, pattern recognition, image analysis, data compression, information retrieval, bioinformatics, and computer graphics, and it plays an important role in understanding various phenomena and exploring the nature of obtained data.

The widely used clustering algorithms (k-means [1], fuzzy c-means [2], and their variants) have many shortcomings, i.e., the need to set the number of clusters a priori, the sensitivity to initial conditions, and the definition of a suitable distance measure for the data. So, one major difficulty of these algorithms is how to set and initialize them, since the used parameters are crucial for successful clustering outcome [3].

In this paper, we describe a clustering approach for speech acoustic vectors that overcomes the previous limitations. Il is based on the unsupervised algorithm proposed by Wong et al. in [3], which will be referred to as UA. This algorithm is easier to implement than other unsupervised alternatives such as the self-organizing map (SOM) [4] or the adaptive resonance theory (ART) network [5].

The algorithm is integrated in an automatic speech recognition system using the J-RASTA-PLP (J-RelAtive SpecTrAl-Perceptual Linear Prediction) method [6] for acoustic parameters extraction, and using the hybrid HMM/MLP (Hidden Markov Models /Multi-Layer Perceptron) model [7] for learning the acoustic vectors quantified using the proposed clustering algorithm.

The balance of this paper is organized as follows: in Section 2, we describe the principle of the algorithm proposed for speech data clustering. Sections 3 presents the validation of the clustering algorithm, we will show that the unsupervised clustering of the speech data prior to decoding by a conventional HMM/MLP model leads to improved system performance in comparison to no clustering or to using clustering by a Fuzzy C-Means (FCM) algorithm, which the result used as initial population of genetic algorithm (GA). Finally, section 4 concludes this work.

2 UNSUPERVISED CLUSTERING ALGORITHM

Fig. 1, shows the block diagram of a typical automatic speech recognition system. We present the unsupervised algorithm for clustering the preprocessed input data next.

As mentioned, we used the unsupervised clustering algorithm of Wong et al. [3] to enhance the discriminative power of the speech feature vectors, and to reduce the input space size for faster processing. The algorithm requires no initial setting of the number of clusters or cluster centers, and no distance measure tuned to the data distribution (spherical, ellipsoidal, etc.).

Given a set of feature vectors to cluster, UA starts with each feature vector taken as cluster center and replaces it by the weighted average of all similar vectors according to a starting distance threshold. Then the threshold is increased and the process is repeated until a single cluster is formed, at which point the results of the different iterations are compared in terms of a performance index to select the one offering the best cluster distribution. A Gaussian radial function measures vector similarity and its variance σ sets the range of data that may contribute to a cluster; as σ increases, the number of clusters decreases. Since each value of σ may lead to different results for the number of clusters and the cluster centers, a performance index is used to evaluate the obtained partitions.

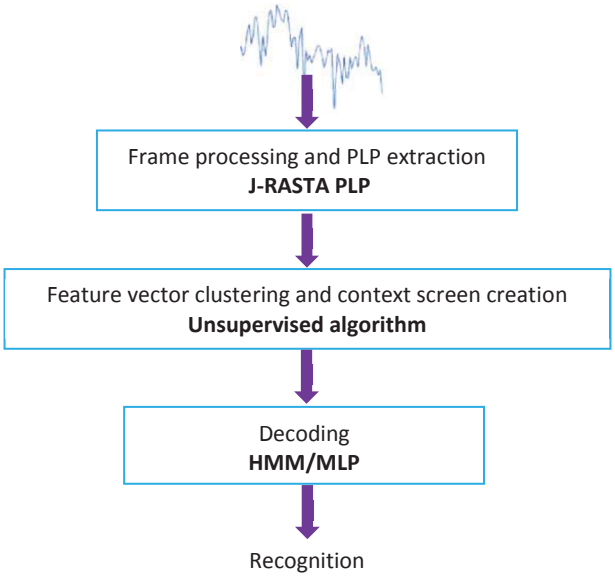

Figure 1: Block diagram of our ASR system showing the different stages and the main methods to implement them.

The performance index PI at the end of the k^{th} iteration is defined as:

$$PI^{k} = \frac{\sum_{d=1}^{h^{k}} n_d^k S_d^k}{N} \tag{1}$$

where d specifies a cluster, h is the number of created clusters, n_d is the number of feature vectors aggregated in the d^{th} cluster, N is the size of the dataset and S_d is a metric that evaluates cluster tightness, defined as:

$$S_d^k = \min_{j,j\neq d} \frac{\left\|\vec{m}_d^k - \vec{m}_j^k\right\|^2}{\sqrt{\frac{1}{n_d^k}\sum_{i=1}^{n_d^k}\left\|\vec{x}_i - \vec{m}_d^k\right\|}} \tag{2}$$

where $\vec{m}_d$ and $\vec{m}_j$ denote the centers of the d^{th} and j^{th} clusters and $\vec{x}_i$ is the i^{th} vector in cluster d. A close analysis of equation (2) reveals that a large value of S_d reflects a dense and well-isolated cluster. Thus, the larger PI^k is, the better the h clusters defined during the k^{th} iteration are. The final number of clusters and cluster centers, and the ensuing classification process are determined by the minimum width σ that has the best performance index. It should be emphasized again that the algorithm operates without the need to set the number of clusters a priori, or to set a distance measure that follows the data distribution.

After completion of the UA learning stage, the number of clusters that are determined sets the size of a binary vector where each bit position stands for a cluster. Then, each feature vector from a preprocessed acoustic frame will result in a '1' set at the bit position of the cluster it belongs to, and '0' otherwise.

3 VALIDATION EXPERIMENTS AND RESULTS

The clustering efficiency of the unsupervised algorithm was tested against à genetic algorithms alternative in a regular HMM/MLP model environment. The experiments were executed on a PC station equipped with an Intel core I7 CPU running at 4.0 GHz, 3.2 GB of RAM, and a SoundBlaster 64 AWE sound card, all running under Microsoft Windows 7, 64 bits edition, with service Pack 3. The MATLAB (R2014a) environment was used for coding the experiments.

We begin by describing the HMM/MLP model, followed by the speech data that was used and the parameters of the J-RASTA PLP preprocessing. Then, the various experiments will be covered.

3.1 Reference HMM/MLP model for testing

We used a ten-state HMM with a discrete observation symbol density. The number of states was determined empirically. A MLP with one hidden layer and 2880/1728 input neurons (numbers explained in the next sub-section), 170/131 hidden neurons (equ. 6 below) and ten output neurons ─one for each of the ten states of the HMM─ was trained by stochastic gradient descent, using the conditional entropy as error criterion. A sigmoid function was applied to the hidden layer units, and softmax (exponential of the unit's weighted sum normalized by the sum of exponentials for the entire layer) was used as the output nonlinearity. The number of hidden neurons n_h was chosen with the following heuristic:

$$n_h = (n_i * n_o)^{\frac{1}{2}} \tag{3}$$

where n_i and n_o stand for the numbers of input neurons and output neurons, respectively.

3.2 Test data

The test corpus, referred as (AD), contained continuous speech in Arabic and was composed of 4000 sentences pronounced by 50 speakers, using the following vocabulary:

- Db1: The ten digits 0, 1, 2, 3, 4, 5, 6, 7, 8 and 9.
- Db2: The last name, first name, city of birth and city of residence of each speaker.
- Db3: 13 control words (e.g.; View/new, save/save as/ save all).

The speech recordings were sampled by microphone at 11 kHz. The training and test data were defined as follows: two thirds of the sentences for training and the rest for testing. The training sentences were pronounced by 40 speakers and the test sentences by 10 speakers (5 men and 5 women).

3.3 Preprocessing and features extraction with J-RASTA PLP

Each processed frame represented 25 ms of speech, with 12.5 ms frame overlap. After pre-emphasis (factor 0.95) and application of a Hamming window, the twelve cepstral coefficients plus energy generated by J-RASTA PLP were normalized by the corresponding standard deviations measured on the training frames, and the frame's 26-dimensional feature vector was built with the obtained cepstral parameters, their first derivatives, and the first and second derivatives of the frame's energy.

J-RASTA was configured to incorporate high-pass filtering and slight spectral subtraction. A constant J of 1e-6 was used for training. Multiple-regression J mapping was used during testing.

Nine frames of contextual information were used as input to the MLPs after clustering.

3.4 UA clustering evaluation

We evaluated the merit of using UA to cluster the feature vectors by comparing it to a generic algorithm. The HMM/MLP model was used to decode both. The choice of GA as reference was motivated by the fact that, in a previous work to AD recognition, we found that clustering by GA yielded better results than other popular algorithms such as *k*-means and FCM [7,8].

3.4.1 Input clustering by GA. The result of FCM clustering was used as initial population. We varied the number of chromosomes from two to fifty and noticed that the correct recognition rate increased progressively before stabilizing for eight chromosomes. Therefore, a population of eight chromosomes was used. The acoustic feature vectors were quantized into four independent codebooks as done by [9] **Erreur ! Source du renvoi introuvable.** and others. These consisted of 128 clusters for the J-RASTA PLP vectors, 128 clusters for the first time derivative of the cepstral vectors, 32 clusters for the first time derivative of energy and 32 clusters for the second time derivative of energy (all values selected empirically). There was one set of codebooks for each of the 9 frames of quantized acoustic vectors used as input to the MLPs, leading to a 2880[1]-component, real valued input vector. The component values were the membership values of the acoustic vectors to the codebook classes as determined by the FCM algorithm, for example:

Chromosome 1

$Cluster_1$	$Cluster_2$	…	$Cluster_{2880}$
0.200	0.008	…	0.064

…

Chromosome 8

$Cluster_1$	$Cluster_2$	…	$Cluster_{2880}$
0.075	0.239	…	0.150

The following merit function was used to assess the fitness of a chromosome:

$$w = \sum_{l=1}^{M} \sum_{x_i \in C_l} p_i d^2 (x_i, g_l) \qquad (4)$$

where p_i is the weight of the i^{th} acoustic vector and g_l the center of gravity of cluster C_l, l referring to one of the M clusters. We have:

$$g_l = \frac{1}{|C_l|} \sum_{x_i \in C_l} x_i \qquad (5)$$

We followed [10] for parameter setting and chose a value ≥ 0.5 for the probability of crossover and a value inversely proportional to the size of the population for the probability of mutation. Since the fitness function reached a minimum between 90 and 100 iterations, we used the latter value (Fig. 2) as stop criterion and the chromosome with the lowest fitness value was then input to the decoding stage.

We report in Table 1 the values of the posterior probability $P(\mathbf{O}|\lambda)$, where $\mathbf{O} = O_1 O_2 \ldots O_T$ is the observation sequence and λ the HMM model, obtained for different values of crossover probability P_c and a fixed value for mutation probability P_m. The GA parameters were as follows:chromosome size = 2880 (the number of clusters), stopping criterion = 100 iterations, P_m = 0.01 and P_c between 0.5 and 0.9. From Table 1, we see that the maximum value of the posterior probability was obtained for P_c = 0.9.

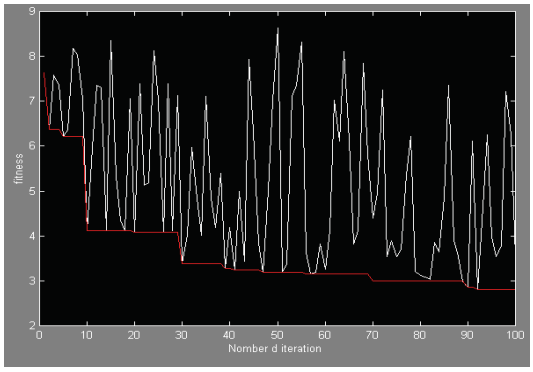

Figure 2: Convergence of the classification process.

3.4.2 Input clustering by UA. In this case, the best parameter values, arrived at empirically, were 64 clusters for the J-RASTA PLP vectors, 64 clusters for the first time derivative of cepstral vectors, 32 clusters for the first time derivative of energy and 32 clusters for the second time derivative of energy.

[1] 2880 = (128+128+32+32)*9

With 9 frames of quantized acoustic vector as before, clustered input to the decoding stage had 1728 components[2].

To illustrate UA's operation, Fig. 3, plots the performance index PI^k with respect to $\sigma = k\,d\sigma$ for three values of the increment, when clustering three digit sounds from AD[3].

Table 1: GA parameters for HMM training.

Number of clusters	Number of Iterations	P_m	P_c	**P(O/ λ)**
2880	100	0.01	0.5	0.3630
			0.6	0.5838
			0.7	0.1423
			0.8	0.9134
			0.9	0.9422

As the curve shows, using the second and third choices for dσ leads to fewer clusters for the same performance index. The curves indicate that an optimal partition of the data was obtained with three clusters.

Fig. 4, illustrates the progression of the cluster centers (white triangles) at each iteration when dσ =0.1241. Fig. 5, shows the necessity of gradually increasing the value of σ for the algorithm to work. In Fig. 5a, the acoustic vectors are classified correctly when using an incremental value of σ for the clustering process. On the other hand, in Fig. 5b, several data points are misclassified when using the approach with a fixed width.

Fig. 6, shows the result of clustering the same dataset by GA, with an initial partition of three clusters by FCM and the use of a Euclidean distance. As one can see, many patterns were clustered improperly.

The previous figures clearly show the effectiveness of UA to cluster complex speech patterns such as those of AD. The next subsection shows that UA clustering also yields better ASR accuracy in comparison to using GA clustering, regardless of the database considered.

3.4.3 UA versus GA clustering results. Table 2 reports the average recognition accuracy obtained by the basic HMM/MLP model when the feature vectors are clustered by UA or by GA, and when there are not clustered (WC). The clustering by UA consistently offered better recognition performance, leading to its adoption in subsequent experiments.

Table 2: Average recognition accuracy (%) by the basic HMM/MLP model with no clustering (WC), and with GA and UA clustering of the acoustic feature vectors.

	WC	GA	UA
Train	78.5	85.1	**86.0**
Test	77.3	83.1	**83.4**

[2] 1728 = (64+64+32+32)*9. Notice that 64 clusters led to the best results in our experiments, in comparison to the 128 cluster and 2880 components when using the GA clustering approach.

[3] The data was from the digits sub-corpus (Db1) and consisted of 579 acoustic vectors belonging to 9 sounds: 3 occurrences of digit "One", 3 occurrences of digit "Two", 3 occurrences of digit "Three" (the vectors were selected randomly).

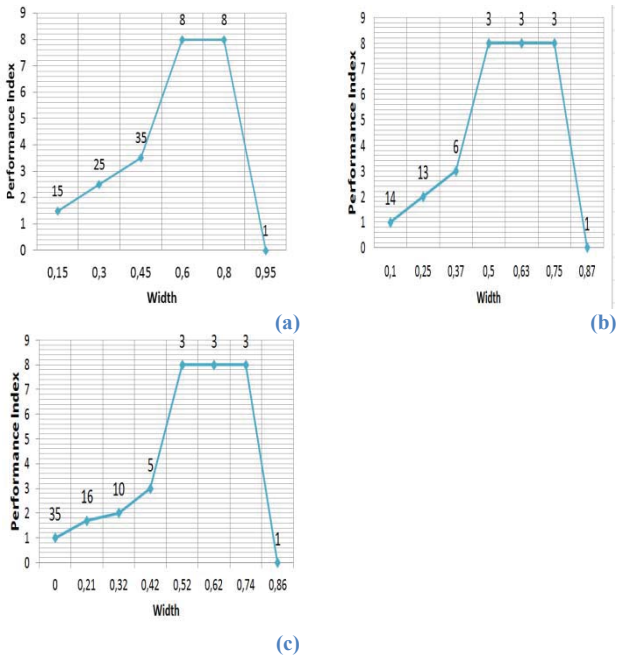

Figure 3: Plot of the performance index PI^k at the k^{th} Iteration with respect to width σ =kdσ for: a) dσ =0.1544, b) dσ =0.1241, and c) dσ =0.1066 (The number next to each data point indicates the number of clusters created).

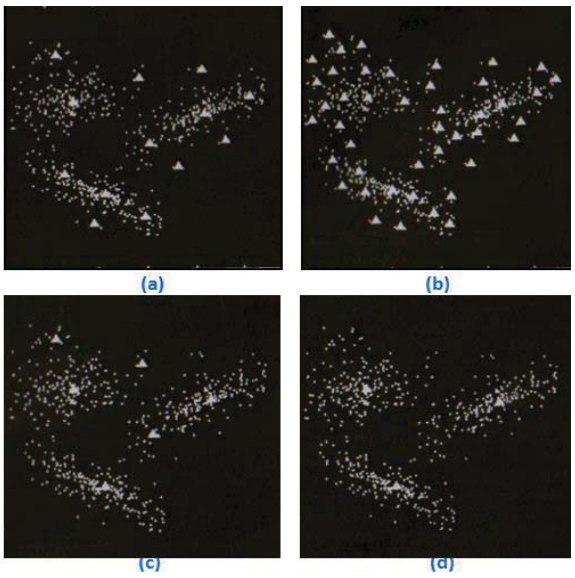

Figure 4: Estimated cluster centers at the end of each iteration for dσ =0.1241: a) k=1, b) k=2, c) k=3, and d) convergence k = 12.

3.5 Discussion

Using a more basic perspective, this work addressed two questions related to developing more efficient HMM/MLP systems:

- How to improve reducing the number of inputs to the MLP component?
- How to improve the performance of the HMM/MLP model to make it closer to state-of-the-art HMM models– especially for large speech corpora?

Our results show that clustering the acoustic data by UA is a better choice than using FCM alternative clustering technique. Unfortunately, even with an optimization of the FCM results with GA, its performance is always lower than that of the UA. However, in the validation section, it was shown that the improvement in performance brought by the UA over the combined FCM/GA clustering approach is not as large, only 0.3% for test corpora and 0.9% for train data. Better yet, the results shown in Table 2 point to a new direction to improve the recognition accuracy, where UA may provide better results.

In this work, we only applied the GA with FCM clustering to a HMM/MLP model, but we intend to apply the optimization process of GA to the UA clustering in the near future. A finding that a genetic optimization of UA clustering results yields better performance than a regular UA. This allows training the GA with a population of empirically generated UA chromosomes and not randomly initialized.

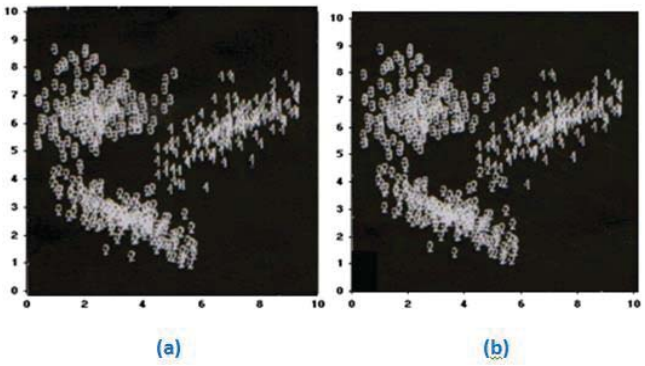

Figure 5: Classification Results by: (a) incremental σ, (b) $\sigma = 0.5$.

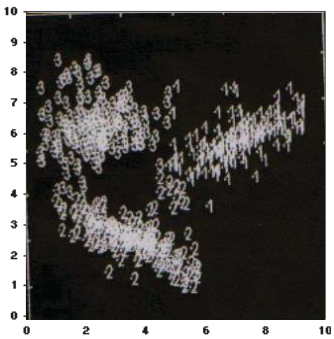

Figure 6: Classification results by AG/FCM algorithm.

4 CONCLUSION

This paper proposes an unsupervised clustering algorithm for a hybrid HMM/MLP speech recognition system. The features of the proposed algorithm are: (1) It relies an unsupervised algorithm to cluster a data set based on the underlying data structure; (2) It can efficiently process the acoustic vectors with clusters of various sizes, shapes and orientations; (3) It does not need to determine a suitable similarity measure according to the shapes of the data; (4) It does not require a predetermination of the number of clusters; and (5) It does not need to determine the appropriate cluster centers in the initial step like conventional clustering algorithms. The validation experiment with an HMM/MLP model for ASR, and a comparison with an alternative clustering approach that uses a hybrid FCM/GA showed improved recognition accuracy with the proposed clustering technique for the speech data.

ACKNOWLEDGMENTS

This research was supported by grants from UNESCO for women in Science and ReSMiQ of Quebec.

REFERENCES

[1] H. Ralambondrainy, "A conceptual version of the k-means algorithm. *Pattern Recognition Letters*," (1995) 16:1147–1157, DOI : 10.1016/0167-8655(95)00075-R.

[2] J. Bezdek. Pattern Recognition With Fuzzy Objective Function Algorithms *(Second Edition). New-York: Plénum,* (1987), DOI : 10.1007/978-1-4757-0450-1.

[3] C-C. Wong, C-C. Chen, M-C. Su, "A novel algorithm for data clustering," Pattern recognition 34(2) (2001) 425- 442, DOI: 10.1016/S0031-3203(00)00002-9.

[4] A. Mingoti-Sueli , Lima Joab O, "Comparing SOM neural network with Fuzzy c -means, K -means and traditional hierarchical clustering algorithms". Stochastics and Statistics, in science direct, European Journal of Operational Research 174: (2006) 1742–1759.

[5] G. Carpenter, Grossberg S,"The ART of adaptive pattern recognition by a self-organizing neural network". IEEE Computer 21(3): (1988) 77-88.

[6] H. Hermansky, N. Morgan, "RASTA processing of speech," IEEE Trans. on Speech and Audio Processing 2(4) (1994) 578-589, DOI: 10.1109/89.326616.

[7] L. Lazli, M. Boukadoum, A. Chebira, K Madani, M-T. Laskri, "Connectionist probability estimators in HMM using genetic clustering: Application for speech recognition and medical diagnosis," Int. J. digital information and wireless communications proc. IJDIWC 1(1) (2011) 14-31.

[8] L. Lazli, A. Chebira, M-T. Laskri, K. Madani, "Hybrid HMM-ANN system using a fuzzy clustering for speech and medical pattern recognition," Int. Conf. Digital Information and Communication Technology and Its Applications proc. DICTAP 167, chapter Springer LNCS (2011) 557- 570, DOI: 10.1007/978-3-642-22027-2_46.

[9] J-M. Boite, H. Bourlard, B. D'hoore, S. Accaino, J. Vantieghe, "Task independent and dependent training: performance comparison of HMM and hybrid HMM-MLP approaches," IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP I(1) (1994) 617-I620, DOI: 10.1109/ICASSP.1994.389218.

[10] D-E. Goldberg, Algorithmes génétiques : Exploration, optimisation et apprentissage automatique, (1994) Addison Wesley.