A COMPARISON OF MODIFIED K-MEANS(MKM) AND NN BASED REAL TIME* ADAPTIVE CLUSTERING ALGORITHMS FOR ARTICULATORY SPACE CODEBOOK FORMATION

K.S.Ananthakrishnan
School of Electronic Engineering
University of South Australia, Levels, SA 5095

ABSTRACT

This paper proposes the use of a neural network based real time adaptive clustering* algorithm for the formation of a codebook of limited set of acoustical representation of finite set of vocal tract shapes from an articulatory space. Modified k-means algorithm (MKM) used for clustering nearly 10000 vocal tract shapes into 1000 cluster centers to form a codebook of articulatory shapes is computationally intensive for our application . An investigative study on the use of NN based algorithm over MKM algorithm at the peripheral level , for our application on Computer Aided Pronunciation-education, suggests the former for less intensive computation, with the possibility of improving the performance of the system by implementing the algorithm using a dedicated neural computer. In this paper, preliminary results of this study are reported.

1. INTRODUCTION

Vocal tract shapes are to be inferred from the speech signal for an application intended to provide meaningful feedback to the learner regarding the correct place of articulation in a Computer Aided Pronunciation-education System(CAPS) under development. As a first attempt , the proposed system envisages to derive the vocal tract shapes(VTS) for quantal vowels a , I , u [1] uttered by wide variety of speakers(male,female, children adoloescent, and people from different ethnic backgrounnds) for comparison with the vocal tract shape of a native Australian instructor for the same set of vowels stored in the data-base, to generate the necessary error vector in the articulatory domain for providing the *useful feedback* to the learner by indicating the *correct places* of articulation . This requirement necessitates the use of a clustering algorithm at the front end for the clustering of acoustical representation of VTS. In their paper, Larar et al[2], employed 'modified k-means (MKM) clustering algorithm[3] for clustering 10090 vocal tract shapes generated by a synthesizer based on Mermelstein's articulatory model into 10 percent of the number of original vectors and the intensive computational requirement was a major drawback for their application. Conventional clustering method such as k-means, based on partitional clustering , in addition to intensive computation also require prior knowledge about the number of clusters. In our application, it is envisaged that the training vocabulary of the instructor would expand in future and any clustering algorithm at the front end must meet the adaptability demands too , in addition to less intensive computational complexity. Recently neural network has been used in the estimation of motion of articulatory organs directly from speech waves [5] and has been proved to be a stable high speed technique for applications where accurate articulatory positions are not of paramount importance. Neural network based clustering techniques have also been in use in other fields such as medicine for cytometry data analysis and recently Real Time Adaptive Clustering(RTAC) technique has been successfully used for the analysis of flow Cytometric data in an application to leukemic diagonosis[6]. In this paper, we attempt to apply a modified version of this neural network based real time adaptive clustering (RTAC) to speech data for the geneartion of a codebook of limited vocal tract shapes and study the feasibility of using the same tehnique for clustering physiologically psssible shapes in the articulatory space , by taking more samples in the articulatory domain. RTAC is simple to implement, has reduced computational intensity (besides training the network) and reliable.

2 CODE BOOK GENERATION USING MKM ALGORITHM

The philosophy of modified k-means(MKM) is basically that used in [7] and this clustering algorithm as applied to isolated word recognition is explained in [3]. Larar et al [4] , in their work , used this algorithm for the formation of codebook consisting of voiced , fricative and closed-tract sub-codebooks. Nearly 1029 clusters from 9990 shapes were generated for voiced sub-code book ,300 clusters from 3002 shapes were generated for fricative sub-codebook and 100 clusters for closed-tract shapes were added directly to the codebook. The clustering operations were performed at two levels, by performing initial and final clustering and the estimated time of nearly 1000 hours,was reduced to 23 hours by using the above algorithm on a super minicomputer. Due to its nature of parallel processing capabilities, a neural network based algorithms could potentially be less intensive in computation for such an application (possibly, in real time).

***** the term ' Real Time' is used by the author in the sense of potentially realizing the algorithm using a neural computer. As the algorithm implemented for our application is the same as the one used by Limin Fu[4], the term is retained without any modification

In the following section, a neural network based algorithm is described.

3. REAL TIME ADAPTIVE CLUSTERING ALGORITHM (RTAC)*

3.1 RTAC Architecture

RTAC architecture used in our application, employs two layers, one input layer of 10 nodes corresponding to LPC-10 vectors (acoustical representation of vocal tract shape) and one output layer in which each node is designated by a cluster. The network is feedforward (i.e no recurrent connections) and fully connected(i.e every input node connecting to every output node).

3.2 RTAC Theory

The theory of Real -Time Clustering Algorithm is given below:

The basic equation of RTAC is given by

$$\frac{dW_{j_i}}{dt} = rO_j(-W_{ji} + O_i) \qquad (1)$$

where

W_{j_i} = Weight of the connection from input unit I to output j

r = learning rate

O_i = activation of unit I

O_j = activation of unit j

The learning algorithm used is winner-take-all strategy and the weight W_{j_i} learns by minimizing the squared error between itself and O_i in the direction of steepest descent. As the activation O_j is either 1 or 0 , when the activation O_j becomes 0 , from equation (1) , it is obvious that dW_{j_i}/dt = 0 i.e, no weight change. Thus, only the input weights of the winner node can be modified. The learning rate 'r' is inversely proportional to the size of the cluster 'j' incremented by 1. The details of the algorithm is given in the following section.

3.3 RTAC Algorithm

The RTAC algorithm used is an incremental algorithm and in terms of clustering, is related to two other neural-network based algorithms : Kohonen network[8] and ART network[9]. The RTAC algorithm is more closer to the ART network and implementation of RTAC is simpler than ART. The algorithm works as follows:

- **Weight Initialization**

  According to the first instance of the input vector , the connection weights are initialized.

- **Calculation of Activation level of Input and Output**

  1. The activation level of an input is determined by the instance presenting to the network.

  2. The activation level of output unit O_{jk} when an instance k is presented is determined by

$$O_{j_k} = \{1 \quad if \quad s(W_j, O_k) > s(W_l, O_k)$$

for all $l \neq j$

$$and\ s(W_j, O_k) > \theta$$

O_{jk} = 0 otherwise

where

W_j = weight vector associated with output unit 'j'

O_k = input vector (feature vector) prsented to the network at the instant 'k'

θ = threshold for activation. This parameter is the threshold for exciting a cluster node. (similar to *vigilance* parameter of ART).

's ' is a function which measures the similarity between two vectors.Other similarity functions can be used. This function is defined as follows:

$$s(x,y) = 1 - \frac{||x-y||}{||x||+||y||} \qquad (2)$$

where

$$||x-y|| = \sqrt{\sum (x_i - y_i)^2}$$

and

$$||x|| = \sqrt{\sum x_i^2}$$

- The range of 's' is between '0' and '1' as the difference vector in the numerator of (2) is always smaller than the denominator of (2). The similarity function defined on the basis of perceptual acoustical distance has been used by Larar et al[10].

- **Weight training**

1. For the 'excited ' cluster node, the weights are adjusted as follows:

$$W_{ji}(t+1) = W_{ji}(t) + \Delta W_{ji} \qquad (3)$$

where

$W_{ji}(t)$ = weight from unit 'I' to unit 'j' at time 't'

and

ΔW_{ji} is the weight adjustment factor.

The weight adustment factor is given by

$$\Delta W_{ji} = \frac{1}{n_{jk}} O_{jk}(O_{ik} - W_{ji}) \qquad (4)$$

where

n_{jk} = size of cluster 'j' incremented by 1.

O_{ik} = activation of unit 'I' at the instant 'k'

O_{jk} = activation of unit 'j' at the instant 'k'

2. Create a new cluster node , if no cluster node is excited and initialize its weights.

3. Repeat until no more instances are available.

4. RESULTS AND DISCUSSION

The derivation of acoustical representation of vocal tract shapes was done by extracting 10th order LPC vectors using autocorrelation analysis for 32 msec speech segments sampled at 8000 Hz. VTS vectors corresponding to the segments were computed using inverse filtering [11]. The VTS-SS(spectral shape) pairs were generated by a program implemented in 'Matlab' and a set of dat data files were created for quantal vowels . The algorithm was implemented in 'C' and ran on Sun Sparc station under SunOs version 5.4. The weights file created during 'training ' was used to cluster the' test set'.

In the MKM algorithm , cluster centres were formed and in the nn based algorithm , the weights file data need to be interpreted for obtaining the codebook statistics such as average minimmum distance between any cluster center and the next closest center to it , average intracluster distance and stability of the system.

In the next part of our investigation, LPC vectors derived from a articulatory synthesizer (guarenteeing physiologically possible vocal tract shapes) and the corresponding spectral shape vector will be used as the ' training set' to the algorithm so as to evaluate the codebook for its adequacy. The preliminary results obtained will be presented at the conference.

5. COMPARISON OF MKM AND RTAC FOR CODEBOOK FORMATION

In the following section we give a comparison of the above two algorithms at the peripheral level from the point of view of code book generation .

5.1.1 Modified K-Means Algorithm :

- Levels of clustering: In general two(Initial and Fine)
- Basis of Algorithm: conventional partitional clustering.
- Training: (equivalent to 'learning' in Neural Network based algorithm) is slow.
- Time Complexity: dependent on the fine clustering(usually hours).
- Similarity measure : Perceptual based *acoustical distance*
- Stability :
- Prior Knowledge of Number of Clusrers : To be given(may lead to unnatural clustering).

5.1.2 Real-Time Adaptive Clustering Algorithm

- Levels of clustering: Single Level
- Basis of Algorithm: Neural Network method
- Learning : (equivalent to 'Training' in Conventional Clustering algorithm) can be fast, achieving real-time performance.
- Time Complexity: is dependent on O(nm), where
  n - number of instances presented to the network
  m- the number of clusters formed.

- Similarity measure: Novel function 's' measuring both quantitative and qualitative differences between vectors.
- Stability: relies on the fact that clusters formed are well separated from each other (i.e) Intercluster distance need to be very high.
- Prior Knowledge of Number of Clusrers: Not to be given as the algorithm is adaptive(much the same was as ART)

SUMMARY

In this paper, a new simple potentially computational-efficient technique for quantization of limited set from articulatory space has been described. A comparison of MKM and RTAC clustering algorithms were made at the surface level. Samples of articulatory space corresponding to vowels were carefully derived from real speech samples and the algorithms were studied. Besides the training time of the neural network , seemingly the algorithm could be a candidate for codebook formation . Stabilty of the system needs further study and this involves more samples from articulatory space so as to cover small constrictions (1 cm2 or less) and complete tract closure(near zero constrictions). As the algorithm is based on neural network concepts, the performance can be greatly improved by the use of a well designed neural computer in a real system for fast responses.

ACKNOWLEDGMENT

The author would like to acknowledge the help given by John Asenstorfer and Lakmi Jain. The author also would like to thank Malcolm Raymond for suggesting this project.

REFERENCES

1. Stevens, K.N., " Quantal Nature of Speech", Human Communication: A unified view, McGraw -Hill, New York, 1972.

2. Larar,J.N .,Schroeter,J., and Sondhi,M.M., " Vector quantization of the articulatory space", *IEEE Trans. Acoustic., Speech, Signal Processing*, vol . ASSP-36, pp 1812-1818 , June 1988.

3. J.G.Wilpon and L.R.Rabiner, " A modified k-means clustering algorithms for use in isolated word recognition," *IEEE Trans. Acoustic., Speech, Signal Processing*, vol . ASSP-33, pp 587-594, June 1985.

4. Limin Fu, " Neural Networks in Computer Intelligence", McGraw-Hill Inc, 1994.

5. T. KObayashi.,M.Yagyu., and K.Shirai., " Application of neural networks to articulatory motion estimation", *IEEE Trans. Acoustic., Speech, Signal Processing*, vol . ASSP-33, pp 587- 594, June 1985.

6. Frankel, D.S.,Olson,R.J., Frankel,S.L., and Chisholm,S.W., " Use of a neural net computer system for analysis of flow cytometry data of phytoplankton populations", *Cytometry*, pp 540-550, 1989.

7. Y.Linde., A.Buzo., and R.M.Gray., " An algorithm for vector quantizer design, *IEEE Trans. Communication*, vol. COM-28, pp 84-95, Jan 1980.

8. Kohonen, T., "Self-Organization and Associative Memory, Springer-Verlag, New-York, 1988.

9. Carpenter,G.A., and Grossberg , S ., " The ART of adaptive pattern recognition by a self-organizing neural network", *Computer*, March, pp 77-78.

10. J.Schroeter,J.N.Larar and M.M. Sondhi, "Speech parameter estimation using a vocal tract/cord model ",*Proc.IEEE InternationalConference,Acoustics,Speech and Signal Processing*, pp 308-311,1987.

11. H.Wakita, "Direct estimation of the vocal tract shape by inverse filtering of the acoustic speech waveforms", *IEEE Trans. Audio Electroacoustic*., vol AV-21, pp 417-427,1973.