Acta Cybernetica **18** (2008) 451–462.

Robust Clustering - Based Realtime Vowel Recognition

Dénes Paczolay[*], András Bánhalmi[*], and András Kocsor[*†]

Abstract

In the therapy of the hearing impaired one of the key problems is how to deal with the lack of proper auditive feedback which impedes the development of intelligible speech. The effectiveness of the therapy relies heavily on accurate phoneme recognition. Because of the environmental difficulties, simple recognition algorithms may have a weak classification performance, so various techniques such as normalization and classifier combination are applied to raising the overall recognition accuracy. In earlier work we came to realise that the classification accuracy is higher on a database that is manually clustered according to the gender and age of the speakers. This paper examines what happens when we cluster the database into a few groups automatically and then we train separate classifiers for each cluster. The results shows that this two-step method can increase the recognition performance by several percent.

Keywords: speech recognition, speech therapy, two-step classification method

1 Introduction

In the therapy of the hearing impaired one of the central problems is how to deal with the lack of proper auditive feedback that hinders the development of intelligible speech. Our Phonological Awareness Teaching System, the "SpeechMaster" package, seeks to apply speech recognition technology to speech therapy. It provides a visual phonetic feedback to supplement the insufficient auditive feedback of the hearing impaired. Our computer-aided training software package uses an effective phoneme recognizer and provides a realtime visual feedback in the form of flickering letters positioned over calling pictures.

[*]Research Group on Artificial Intelligence of the Hungarian Academy of Sciences and University of Szeged H-6720 Szeged, Aradi vértanúk tere 1., Hungary, E-mail: {pdenes, banhalmi, kocsor}@inf.u-szeged.hu

[†]Machine Intelligence Laboratory NPC., Petőfi S. Sgt. 43., H-6725 Szeged, Hungary, Applied Intelligence Laboratory Ltd., Petőfi S. Sgt. 43., H-6725 Szeged, Hungary, E-mail: kocsor@ail.hu

Since the system should work reliably for children of different ages and teachers as well, the recognizer has to be trained with the voices of users of both genders and of practically any age. The task is also special because the system has to recognize isolated phones, so it cannot rely on language models. Consequently, there is a heavy burden on the acoustic classifier, and we need to apply any helpful trick that might improve the overall performance.

During our previous work we found that the classification accuracy is generally higher on a homogeneous database (one whose gender and age are homogeneous) than a mixed database. This is probably because the variance of a homogeneous database is better than a mixed one. To train the latest version of SpeechMaster we applied training databases that were manually clustered according to speaker gender and age because we wished to achieve a higher recognition performance. This paper describes what happens when we cluster the database into several groups automatically and then train a separate classifier for each of these groups.

This paper is organized as follows. In the following section we will present our speech therapy system, then in Section 3 we will describe our previous study and experiences gained from it. In the next two sections we provide a brief description of the clustering and classification algorithms used in our tests. Section 6 then compares the performance of the various recognition methods. Lastly, we give some brief conclusions and ideas for the future.

2 Our therapy system: the SpeechMaster

The SpeechMaster package was developed for speech impediment therapy and teaching reading. The system is based on automatic speech recognition (machine learning [1, 3, 10]) and advanced signal processing methods. The developers co-operated with speech therapeutists and elementary school teachers, and tested the system with children in real environments. In the therapy of the hearing impaired one of the key problems is how to deal with the lack of proper auditive feedback - which, of course, impedes the development of intelligible speech. The idea is really to make the vocal sounds 'visible' for the hearing impaired. This way they are able to check their pronunciation by sight, that is, their hearing is supplemented

(a) (b) (c)

Figure 1: Screenshots from "SpeechMaster"

by visual input. The speech therapy of the hearing impaired traditionally requires enormous patience and the continuous presence of a teacher since, during the fixation of the correct sound-formation, a large amount of repetition and correction by the teacher are both needed. This automation process is significantly speeded up and simplified with our software, and also allows the students to practice on their own or with the teacher. In speech impediment therapy, at the beginning of the development of oral competence, it is recommended that young children concentrate mainly on their own voicing. This is supported by the creation of the many playful sound formation exercises. For each drill, skill and acceptance levels can be adjusted with a potentiometer. The software package also contains many useful features: customisable profiles, easily extendable word and image lists with sample utterances, half-speed sound replay, a web-camera serving as a "phonetic mirror" and so on. The program is available at our website and may be downloaded free of charge.

2.1 Learning the pronunciation of vowels

It is experimentally known [5] that the training of the utterance of vowels is more difficult than that of consonants because their phonation is not so easy to explain. The key feature of the therapy of the hearing impaired is the refined pronunciation of vowels in order to attain articulate/intelligible speech. It would be a real help in therapy if the computer were able to provide an objective rating of the quality of the uttered vowels. If it were reliable and matched the subjective opinion of the therapist, it would relieve teachers of the burden of the tedious work they have with traditional therapy. In SpeechMaster the role of effective real-time vowel recognition is essential. Real-time visual feedback helps improve the student's articulation because it aids the damaged or missing auditory feedback. The software package provides clear and simple forms of real-time visual feedback. It also has many feedback configurations: it can display the best individual vowel, all the vowels, diagrams and so on. The student can use a web-camera as a "phonetic mirror" to check his/her own articulation or compare his/her utterance with that of the teacher. The student's utterances are stored in separate directories in chronological order for analysis at same later time.

(a) (b) (c)

Figure 2: Screenshots of the learning the vowel pronunciation with SpeechMaster

2.2 Computer-aided therapy in practice

SpeechMaster has several target environments: (nursery) school, therapy and home. In most cases the children and the therapeutist use the recogniser during therapy in the following way: the therapeutist presents a vowel (word) and the child repeats it. The level of acceptance can be varied for each vowel separately and so the therapeutist can maintain the pupil's motivation. At home the child can play back the sample utterance and practice it. The child can work with his parents too, if he or she wishes. These activities require real-time "speaker adaptation, or normalization" techniques or a good user-independent recogniser.

3 The manually clustered vowel database

The variance of the data over the clusters of a database is smaller than that of the full database. Because of this, the classification error over the clusters is usually smaller. This gave us the idea of using a two-step recognition method where the first step identifies the cluster of the data item, and the second step classifies it using the cluster-specific classifier. This way if the error rate of clustering on unseen data is small enough, the final recognition performance can be increased. When we recorded the vowel database for SpeechMaster, we stored utterances separately according to the gender and age of the speakers. This manually clustered database was applied for training vowel recognition in the latest version of SpeechMaster. Table 1 shows the classification accuracies measured on the non-clustered and the clustered databases together with the results for each cluster. It is clear that the performance on the clusters "Men", "Women", "Children" as well as the performance of the two-step method were both better than that of the original one-step method (no pre-clustered). The results below were obtained from a database of 200 speakers [7] (CSCS 2004). We recorded the utterances of healthy hearing children, because wanted to like teach the hearing impaired to speak and not simply to allow them to recognise their vowels. Each speaker uttered 9 clearly formatted and pronounced, sustained, voiced Hungarian vowels. This task is easier than a general phoneme recognition task.

Speakers	Accuracy
Men	90.81 %
Women	91.32 %
Children	96.11 %

Method	Accuracy
1-step (No clusters)	88.32 %
2-step (3 clusters)	91.16 %

Table 1: Recognition accuracy for a manually clustered vowel database

As one can see, the 2-step method outperforms the 1-step classification when the clusters correspond to the manually clustered "Men", "Women", "Children" labels of the database. In the following we shall examine what happens when the clusters are created automatically. The clustering method here uses speaker-space vectors [6]. These vectors contains the mean feature vectors of 9 Hungarian vowels.

4 Clustering methods

Data clustering is a commonly applied technique in statistical data analysis. Clustering is a process where a data set is partitioned into subsets (clusters) so that the data in each subset (ideally) share some common trait - often approximately based on some pre-defined distance measure. Machine learning typically treats data clustering as a form of unsupervised learning. Actually there are two types of data clustering algorithms: hierarchical ones and partitioning ones. Hierarchical algorithms create successive clusters using previously established clusters, while partition algorithms find all clusters simultaneously. Hierarchical algorithms can be agglomerative (bottom-up) or divisive (top-down). Agglomerative algorithms start by considering each element as a separate cluster and successively merge them into larger clusters. Divisive algorithms start with the whole set and proceed to divide it into successively smaller clusters. In our experiments we investigated a partition clustering method and a bottom-up hierarchical clustering method.

4.1 K-Means

K-means clustering is an iterative partitional algorithm [1] the clusters the data points into K disjoint subsets $S_j (j = 1, ..., K)$ by minimizing the sum-of-squares criteria

$$\sum_{j=1}^{K} \sum_{i \in S_j} |x_i - \mu_j|^2,$$

where x_i is a vector representing the i^{th} data point and μ_j is the geometric centroid of the data points in S_j.

The method consists of the following steps:

1. Randomly generate K clusters and determine the cluster centres, or directly generate K seed points as cluster centres.
2. Assign each point to the nearest cluster centre.
3. Recompute the new cluster centres.
4. Repeat until some convergence criterion is met (e.g. the assignment does not change). It is guaranteed to stop, because number of ways the dataset can be partitioned is finite and the algorithm decreases the error criterion in every step.

This algorithm has two significant advantages that allows it to be useable on large datasets, namely its simplicity and speed. Its main disadvantage is that it does not yield the same results with each run, since the resulting clusters depend on the initial random assignments. It maximises inter-cluster (or minimises intra-cluster) variance, but does not guarantee that each result will have a global minimum variance. An improved version of the algorithm is described in [2] which refines

the initial points. We did not try this improved version, because our database was quite small (only 300 speakers). We performed the partition several times with the base algorithm, and then selected the best partitions.

4.2 Unweighted Pair-Group Method with Arithmetic Mean

This is a simple hierarchical agglomerative (bottom-up) algorithm used in bioinformatics to create phylogenetic trees [4]. At each step this iterative algorithm merges the two nearest clusters and recalculates their distances from the remaining ones. The new distances can be calculated using the formula:

$$D_{(ij),k} = (\frac{N_i}{N_i + N_j})D_{i,k} + (\frac{N_j}{N_i + N_j})D_{j,k}$$

where the distance between the i^{th} and j^{th} cluster is $D_{i,j}$ and the i^{th} cluster contains N_i data points.

This method often leads to a degenerate tree (cluster), so we decided to introduce an extra criterion: we fuse the i^{th} and j^{th} clusters if and only if, for a given i and j, N_i or N_j, is less than some given threshold.

5 Classifiers

The classification problem is a supervised learning task. The learner is required to learn (to approximate the behaviour of) a function which chooses, for a sample represented by a feature vector, the right class by looking at several input-output examples of the function.

5.1 Artificial Neural Networks

Artificial Neural Networks (ANNs) is a well-known machine learning method. The basic idea behind ANNs is that many simple functional units (neurons) when combined in parallel produce effective models for learning [1]. A unit receives its input from several other units, or perhaps from an external source. Each input has an associated weight w, which can be modified so as to model synaptic learning. The unit next computes some function f of the weighted sum of its inputs:

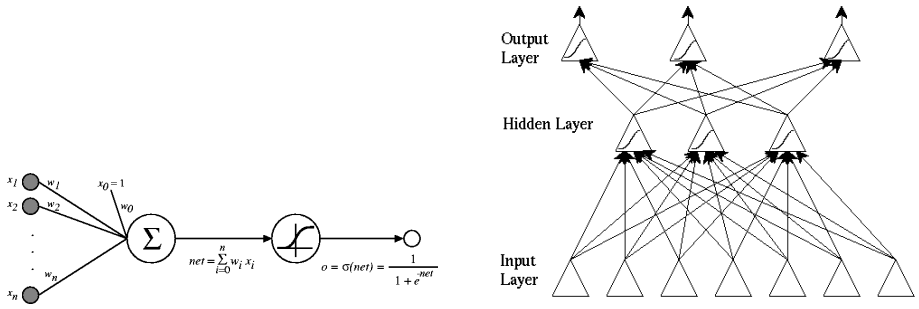
$$net_j = \sum_i w_{ij} y_i$$

$$y_j = f(net_j)$$

The function f is the activation function of the unit. A commonly used activation function is the Sigmoid function:

$$\frac{1}{1 + e^{-net}}$$

The input, output and hidden layer(s) contain many individual units and can model any function. The neurons on each layer are usually fully interconnected with other neurons on an adjacent layer (see Fig. 3). The ANN then learns by modifying the weights in the sigmoid unit. The back-propagation learning rule finds a local, but not necessarily global error minimum [1]. During the classification task the input will be the feature vector. The index belonging to the maximum value of the output vector will be the index returned as the class of the input sample.



(a) A sigmoid unit (b) A three layered network

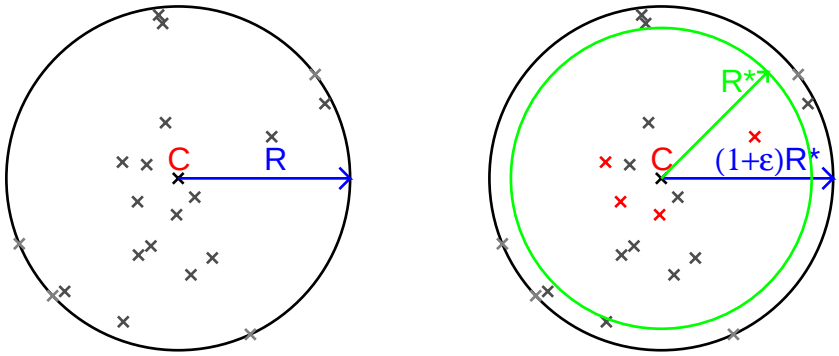
Figure 3: A sigmoid unit and a full ANN

5.2 Core Vector Machine

The Core Vector Machine (CVM) method [9] is a variant of the Support Vector Machine (SVM) [10] approach. The Support Vector Machine performs the following task: it maps the input vectors into a high dimensional feature space through a non-linear mapping. In this space a linear decision surface has high generalization ability. The standard Support Vector Machine training algorithm is of $O(n^3)$ in time complexity and $O(n^2)$ in space complexity, where n is the size of the training database. CVM only approximates the optimal solution via an iterative algorithm, but it has $O(n)$ time complexity and its space complexity is independent of n. The basic aim here is to find, using the notion of core sets, an efficient approximation for the solution of the minimum enclosing ball (MEB) problem (see Fig. 4). This iterative algorithm works by selecting the furthest point outside the current estimated ball until all the points are covered. The CVM technique essentially combines the method of core sets and nonlinear kernels.

6 Experiments and evaluation

Firstly we will describe the corpus and the feature extraction technique, followed by the clustering and classifier algorithms used in the tests. After that we will specify the task of the recognition test, and list the results in tables.



(a) The MEB problem (b) ϵ approximation

Figure 4: Solving the MEB problem by an efficient approximation.

6.1 Conditions

- **Corpus:** For training and testing purposes we recorded samples from 300 speakers, namely 75 women, 75 men, 75 girls and 75 boys. (The ages of the children were between 6 and 9.) The speech signals were recorded and stored at a sampling rate of 22050 Hz in 16-bit quality. Each speaker uttered all the Hungarian vowels, one after the other, separated by a short pause. Since we decided not to discriminate their long and short versions, we only worked with 9 vowels altogether.

- **Feature set:** The signals were processed in 10 ms frames, the log-energies of 24 critical-bands being extracted by using FFT and triangular weighting [8]. The energy of each frame was normalized separately, so only the spectral shape was used for classification.

- **Speaker-space:** The speaker-space database contained the 24 critical-bands of the 9 vowels for each speaker. Hence the dimension of the speaker-space was 9x24.

- **The K-Means clustering method:** We tested it with values of k between 3 to 6. In the evaluation $k = 4$ was chosen because this is the maximum value of k for which the size of the clusters did not become unusably small. The applied distance metric was the Euclidean one.

- **The UPGMA clustering method:** We used the modified UPGMA method with a threshold value of 10. It produced 3 clusters. During experiments we applied the Euclidean distance as a distance metric.

- **The ANN classifier:** We employed the well-known three-layer feed-forward MLP networks trained with the back-propagation learning rule. The number of hidden neurons was 16, which performing some preliminary tests.

- **The CVM classifier:** For the CVM we used the Radial Basis Function

$$k(x,y) = \exp\left(-\frac{\|x-y\|^2}{\gamma}\right)$$

with

$$\gamma = \frac{1}{m^2}\sum_{i,j=1}^{m}\|x_i - x_j\|^2$$

where m is the size of the train set.

6.2 The recognition tests

The experiments were conducted as follows. First we divided the database into train and test sets. The ratio of the data in the train and test sets was 80% to 20%, keeping the ratio of boys, girls, men and women the same in each set.

We clustered the training part of the speaker-space into k blocks using K-Means and UPGMA. Since the speaker-space was not available for the test set (because the features of speaker-space contains all 9 vowel), an ANN (CVM) learner (denoted by M_0) was trained to separate the clusters of the speakers. The training of the the M_0 was then performed on the vowel training database. The speaker clusters defined a clustering of this database at the same time. K separate ANNs (CVMs) learners (denoted $M_{1...k}$) were afterwards trained to classify the vowels within each cluster.

The testing was performed only on a previously unseen vowel database. In the first step the M_0 machine learner chose the proper cluster for each test data (features of a single vowel). The test items were then classified by the corresponding $M_{1<=j<=k}$ learner. In this part classifier combination methods were be used as well.

7 The test results

7.1 Recognition accuracy on the clusters $M_{1...k}$

Table 2 shows that the vowel classification accuracy over the automatically formed clusters "$K_{1..4}$", "$U_{1..3}$" of the data sets turned out to be better than that over the manually clustered "Men", "Women", "Girls", "Boys", "Children" data sets. But the reader should note that this classification accuracy was calculated on the train set, and over-learning may influence the results.

K-Means

	Accuracy	
Cluster	ANN	CVM
K_1	96.76%	100.00%
K_2	98.26%	100.00%
K_3	97.74%	99.83%
K_4	98.64%	100.00%

UPGMA

	Accuracy	
Cluster	ANN	CVM
U_1	97.99%	99.87%
U_2	97.12%	100.00%
U_3	97.35%	99.94%

Manually

	Accuracy	
Cluster	ANN	CVM
Women	90.12%	91.13%
Men	92.84%	92.51%
Girls	99.51%	98.01%
Boys	92.83%	92.54%
Children	96.05%	95.59%

Table 2: Vowel recognition accuracy on the clusters $M_{1..k}$ expressed in percentage terms.

7.2 Clustering accuracy on the train database (M_0)

Table 3 shows the results of the cluster identification test on the train database using the M_0 machine learner. The classification accuracy of the clusters was between 93% and 96%.

		Accuracy	
Method	**#clust.**	**ANN**	**CVM**
Baseline*	1	100.00%	100.00%
Manually	3	94.12%	94.48%
Manually	4	93.32%	94.24%
K-Means	4	94.25%	95.63%
UPGMA	3	93.53%	93.70%

*This corresponds to the original, one-step recognition method

Table 3: Cluster classification accuracy on the train database M_0 (in percent)

7.3 Recognition accuracy on the test database

Table 4 lists the final test results, that is the vowel classification accuracy on the test database. As can be seen, the performance of the two-phase recogniser was better than that of the original one-step method. On the other hand the performance of the ANN and the CVM methods were quite similar. Still, we found that this database was not large enough to show the full advantages of using CVM.

		Accuracy	
Method	**#clust.**	**ANN**	**CVM**
Baseline	1	89.57%	90.58%
Manually	3	92.03%	92.48%
Manually	4	90.97%	91.01%
K-Means	4	92.59%	93.26%
UPGMA	3	91.33%	90.97%

Table 4: Vowel classification accuracy on the test database (in percent)

8 Conclusions and future suggestions

This paper described a computer-aided speech therapy system where, of course, effective real-time vowel recognition is essential. We presented a simple idea for increasing the recognition performance based on our previous experiences that the training part is more efficient when the database is homogeneous in some way. During our study we found that the classification accuracy was higher when we used a database that was separated according to speaker gender and age than when it was not. This suggested the idea of using a two-step recognition method where the data is automatically clustered, and separate classifiers are trained over the clusters. We found experimentally that the classification error over these clusters is actually smaller than that over the full database. In the proposed two-step recognition process the algorithm first identifies the cluster of the data item, and then, in the second step, the item is classified by applying the cluster-specific classifier. We found that with this method the recognition performance improved, so the clustering step can indeed improve the recognition performance. However, the error from the clustering part (namely, that of the learner M_0) during testing seemed to cause a significant loss in performance. Hence, in the future, we plan to do more experiments to find a better method for choosing the right kind of cluster.

9 Acknowledgements

The project described in paper was financially support by the Hungarian Ministry of Education.

References

[1] Bishop, C. M. *Neural Networks for Pattern Recognition.* Oxford University Press, 1995.

[2] Bradley, P. S. and Fayyad, U. M. Refining initial points for K-Means clustering. In *Proc. 15th International Conf. on Machine Learning*, pages 91–99. Morgan Kaufmann, San Francisco, CA, 1998.

[3] Duda, R. O., Hart, P. E., and Stork, D. G. *Pattern Classification.* John Wiley and Son, New York, 2001.

[4] Fitch, W. M. and Margoliash, E. Construction of phylogenetic trees. *Science*, 155:279–284, 1967.

[5] Hégely, G. and Kocsor, A. A vizuális beszédértékelés alkalmazásának magyar vonatkozású történeti áttekintése a hallássérültek beszédoktatásában. *Alkalmazott Nyelvtudomány*, 2005.

[6] Hu, Z., Barnard, E., and Vermeulen, P. Speaker normalization using correlations among classes. In *Image, Speech, Signal Processing and Robotics*, volume II, pages 223–228, The Chinese University of Hong Kong, Hong Kong, 1998.

[7] Paczolay, D., Felföldi, L., and Kocsor, A. Classifier combination schemes in speech impediment therapy systems. Submitted to Periodica Polytechnica Electrical Engineering, 2005.

[8] Rabiner, L. R. and Juang, B. H. *Fundamentals of Speech Recognition.* Englewood Cliffs, NJ, Prentice Hall, 1993.

[9] Tsang, I. W., Kwok, J. T., and Cheung, P. Core vector machines: Fast svm training on very large data sets. *The Journal of Machine Learning Research*, 6:363–392, 2005.

[10] Vapnik, V. N. *Statistical Learning Theory.* John Wiley and Son, 1998.