IDIAP RESEARCH REPORT

SELF-ORGANIZING-MAPS WITH BIC FOR SPEAKER CLUSTERING

Itshak Lapidot

IDIAP-RR 02-60

Dalle Molle Institute
for Perceptual Artificial
Intelligence • P.O.Box 592 •
Martigny • Valais • Switzerland

phone +41 – 27 – 721 77 11
fax +41 – 27 – 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

IDIAP Research Report 02-60

SELF-ORGANIZING-MAPS WITH BIC FOR SPEAKER CLUSTERING

Itshak Lapidot

DECEMBER 2002

Abstract. A new approach is presented for clustering the speakers from unlabeled and unsegmented conversation, when the number of speakers is unknown. In this approach, each speaker is modeled by a Self-Organizing-Map (SOM). For estimation of the number of clusters the Bayesian Information Criterion (BIC) is applied. This approach was tested on the NIST 1996 HUB-4 evaluation test in terms of speaker and cluster purities. Results indicate that the combined SOM-BIC approach can lead to better clustering results than the baseline system.

1. Introduction

Most speaker recognition problems have been solved by using supervised methods. A less common problem is unsupervised speaker clustering, segmentation and indexing, where no labeled training data is available. The goal in this case is to assign the data to different clusters where each cluster represents a different speaker. Unlike most clustering approaches where each vector is associated with a specific cluster (static clustering), here a sequence of vectors has to be associated with the same cluster, which is known as a temporal data clustering.

Temporal data clustering has many applications in temporal data applications including speaker recognition [1]-[7], machine monitoring [8], switching chaos [9], [10], prediction of systems output [9], clustering of EEG signals [10], music clustering [11], and protein modeling [12].

Temporal data clustering must be used when there is a successive dependence between data vectors in a group. An additional problem of temporal data clustering is to determine the change points (segmentation and change detection problem). Sometimes the transitions between the models are not sharp (e.g., one model appears before the end of the previous one) which is known as drifting dynamics [10]. In this case, it is necessary to find the transients and to give membership weights to each cluster at every time point.

Many approaches have been applied for temporal data clustering, e.g., the Dendrogram [3]; the VQ algorithms [4], the expectation maximization (EM) algorithm [5], [7]; HMM [2], [6], [8], [12]; and neural networks (NN) [1], [9]-[11].

In the presented approach, a Code-Book (CB) is used to model each speaker. All the CBs are first trained such that each CB represents a different speaker. Then an iterative competitive algorithm is applied to all the CBs. The CBs are created using a SOM [13]. The convergence of the algorithm, in terms of distance minimization, is proved in [1]. Input data was an unsegmented and unlabeled conversation, with unknown number of speakers, R. BIC has previously been applied to validate the speaker clustering for a Gaussian cluster model [7]. In this report, we present a version of BIC for the distance-measure case and applied to validate the clustering.

The next sections are organized as follows: Section 2 describes the proposed systems. In section 3 we present the experiments and the results, and in section 4 the systems and results are discussed.

2. Systems Description

In general, given a conversation the goal is to estimate the number of clusters and to cluster the data into q clusters. The description of the VQ-based clustering system is summarized in subsection 2.1. In 2.2, the BIC validity criterion is presented. Sub-section 2.3 shows how BIC can be applied for a VQ-based clustering system.

2.1 VQ-Based clustering System

Assuming that the number of speakers and the segment boundaries are known, and in each conversation the data includes, in addition to speech data, non-speech events, the goal of the algorithm is to cluster the input data into $R+1$ clusters. The initial conditions for the system were determined as follows: segments classified by the crude speech/non-speech classifier as non-speech were used to train the non-speech network. Segments classified as speech segments were randomly and equally divided and used to train the R speaker models.

For the following temporal-data clustering algorithm it is necessary to know the start and end points of each segment. In reality this information is not usually available. For this reason we cut the data into segments of fixed length (this length was set to one second, 100 frames). It was found in [1] that 100 frames and a CB created using a SOM of size 6×10 are sufficient for speaker clustering.

The precise algorithm description and the proof of its convergence can be found in [1]. One iteration of the algorithm consists of following three steps:

1. Retrain the models with the new partition achieved by the previous iteration.
2. Regroup the data according to the defined distance measure.
3. Test for termination: if the termination criterion is met, exit; if not return to 1.

In the present work the following termination criterion was applied:

$$\frac{\text{Number of segments that change their assignment}}{\text{Total number of segments}} \le 0.01 \qquad (1)$$

2.2 The BIC Validity Criterion

The Bayesian Information Criterion for model selection was introduced by Schwarz in 1978 [14]. According to Schwarz, to select the best model for given data it is necessary to maximize the joint likelihood (log-likelihood $L\left(\mathbf{V},\hat{\Theta}\right)$) of the data $\mathbf{V}$ and the estimated parameters $\hat{\Theta}$. According to the Bayes rule, $L\left(\mathbf{V},\hat{\Theta}\right)=L\left(\mathbf{V}\middle|\hat{\Theta}\right)+L\left(\hat{\Theta}\right)$. Schwarz showed that under the assumption of continuous parameters in some range, $L\left(\hat{\Theta}\right)$ depends only on the number of estimated parameters $\left|\hat{\Theta}\right|$ and the number of data points that were used for parameter estimation, N. So the joint log-likelihood is:

$$L\left(\mathbf{V},\hat{\Theta}\right)=L\left(\mathbf{V}\middle|\hat{\Theta}\right)-\frac{1}{2}\left|\hat{\Theta}\right|\log(N) \qquad (2)$$

In practice the second term, often called a penalty term, is multiplied by a scaling factor λ to adjust the equation for a specific application. Then, a best model out of $R_{\max}$ estimated models can be obtained by maximizing the joint log-likelihood using this scale factor.

$$R^{*}=\underset{q=1,\ldots,R_{\max}}{\arg\max}\left\{L\left(\mathbf{V}\middle|\hat{\Theta}_q\right)-\frac{\lambda}{2}\left|\hat{\Theta}_q\right|\log(N)\right\} \qquad (3)$$

2.3 BIC Criterion and VQ

As our CB is a Euclidian distance-based model and for BIC a log-likelihood must be calculated, the following approximation is applied. For input vector $v_n \in CB_r$ we assume that each code-word in the codebook is the mean of a Gaussian probability-density-function (*pdf*) with a unit covariance matrix. Then the estimated log-likelihood of one input vector was calculated as:

$$L\left(v_n\middle|\hat{\Theta}\right)=-\frac{d}{2}\log(2\pi)-\frac{1}{2}\left(c_r^{l^*}-v_n^m\right)^T\left(c_r^{l^*}-v_n^m\right)\quad;\quad c_r^{l^*}=\min_{l=1,\ldots,L_r}\left\{\left(c_r^l-v_n^m\right)^T\left(c_r^l-v_n^m\right)\right\} \qquad (4)$$

and the joint log-likelihood, for a model that consists of q codebooks, is estimated as:

$$L\left(\mathbf{V},\hat{\Theta}\right)=-\frac{dN}{2}\log(2\pi)-\sum_{r=1}^{q}\sum_{v_n\in CB_r}\frac{1}{2}\left(c_r^{l^*}-v_n\right)^T\left(c_r^{l^*}-v_n\right)-\frac{\lambda}{2}\left|\hat{\Theta}_q\right|\log(N) \qquad (5)$$

In this work we applied input vectors of dimension 12, and 60 code-words per one cluster model, i.e., each cluster has $\left|\Theta_{Cluster}\right|=\left|\Theta\right|=60\times 12=720$ parameters. The estimation of the number of clusters was therefore according to the minimization of the following expression:

$$R^* = \underset{q=1,\ldots,30}{\arg\min} \left\{ \sum_{r=1}^{q} \sum_{v_n \in CB_r} \left(c_r^{l^*} - v_n \right)^T \left(c_r^{l^*} - v_n \right) + |\Theta| q \cdot \lambda \cdot \log(N) \right\} \qquad (6)$$

Where $\left|\hat{\Theta}_q\right| = q|\Theta| = 720q$.

3. Systems Evaluation

The system was tested on the NIST 1996 HUB-4 evaluation dataset. It is a broadcast news speech corpus, and the evaluation set consists of four datasets, each of approximately 30 minutes in duration. The four datasets are named File1, File2, File3, and File4 and the number of speakers is 7, 13, 15, and 20 respectively.

3.1 Feature Extraction

The features used were 12^{th} order LPCC. The features calculated from 30ms frames with a 10ms frame rate. In addition, for VQ-based system initialization, the mean absolute values for 50ms of accumulated frames were calculated for speech/non-speech evaluation. Preliminary segmentation of speech and non-speech data was performed by thresholding the absolute value feature. The threshold level was set at three percent of the maximum.

3.2 Evaluation Criterion

Clustering evaluation was based on the purity concept explained in [5]. In [5], only Average Cluster Purity (acp) was calculated. The criterion used in this paper calculates the Average Speaker Purity (asp) as well, as was used in [2]. The reason for the second coefficient is to penalize the splitting of one speaker into several clusters. Two coefficients calculation means that one cluster can include several speakers and one speaker may get split between several clusters. It is important to have a confidence measure taking both factors into account. In the case of speaker purity non-speech data was ignored, because it is not relevant to relate it to a particulate cluster. The notation used is:

R: Number of speakers
q: Number of clusters
n_{ij}: Total number of frames in cluster i spoken by speaker j
$n_{.j}$: Total number of frames spoken by speaker j, $j = 0$ means non-speech frames
$n_{i.}$: Total number of frames in cluster i

The acp, based on cluster purity $\{p_{i.}\}_{i=0}^{q}$, can be defined as:

$$acp = \frac{1}{N} \sum_{i=0}^{q} p_{i.} \cdot n_{i.} \quad ; \quad p_{i.} = \sum_{j=0}^{R} \frac{n_{ij}^2}{n_{i.}^2} \qquad (7)$$

Similarly, asp based on the speaker purity, $\{p_{j.}\}_{j=1}^{R}$, but without the non-speech data, is:

$$asp = \frac{1}{N - n_{.0}} \sum_{j=1}^{R} p_{j.} \cdot n_{.j} \quad ; \quad p_{.j} = \sum_{i=0}^{q} \frac{n_{ij}^2}{n_{.j}^2} \qquad (8)$$

In order to compare between the systems, we calculate the evaluation criterion as the geometric mean of the acp and asp:

$$K = \sqrt{acp \cdot asp} \tag{9}$$

It is important to note that the values of acp, asp and K are always between zero and one without dependence on the number of speakers. Higher acp means that the cluster consists mostly of one speaker. Higher asp means that the speaker data does not split between many clusters. The optimal case is to maximize K, ideally with both acp and asp equal to one.

3.3 Experiment and Results

As was shown at [1], a model with SOM size 6×10 (720 parameters per cluster) and 100 frames per segment are sufficient for speaker clustering. In all the experiments in this report we use these sizes.

The first row in Table 1 shows the results for the baseline system. In this experiment, the number of clusters (q) was set to $R+1$, where R is the number of speakers, plus an additional cluster for non-speech event, as in [1].

The same experiment was conducted again with different initial conditions (SOM's initial weights and initial segmentation) to verify the robustness of the system. The values of K in the second repetition were 0.80, 0.78, 0.80 and 0.71 for the four files. As in [1], it can be seen that the clustering performances are very robust to the systems' initial conditions. The system has a maximum difference of 0.03 between each pair of K values. It is clear that performance degraded where the number of speakers increased due to the higher complexity of the data and the smaller amount of data per speaker.

The second experiment includes clustering and estimation of the optimal number of clusters. The initial number was 30. At each stage the data was clustered and the validity was calculated, for $|\Theta| = 720$, according to $\sum_{r=1}^{q}\sum_{v_n\in CB_r}\left(c_r^{l^*}-v_n\right)^T\left(c_r^{l^*}-v_n\right)-|\Theta|\,q\cdot\lambda\cdot\log(N)$. Different values of scaling factor were applied: $\{\lambda_k\}_{k=0}^{6}=\{0.5k\}_{k=0}^{6}$. After validity calculation the cluster with the minimum amount of data was removed and the system was retrained with the reduced number of clusters. The process was continued until the number of clusters reduced to one. The penalty term influences the validation criterion, as bigger λ leads to a smaller number of clusters and vice versa. It was found that the best scaling factor was $\lambda = 1.5$. Table 1 shows the result of the clustering of the four files according to their scores:

- Second row: the score for the correct number of speakers, $R+1$ (one for non-speech events).
- Third row: the score for the best clustering result achieved according to the best K value as described in Section 3.2.
- Forth row: the score according to the estimated number of clusters, with penalty term $\lambda = 1.5$.

From the result analysis following conclusions can be made:

1. The results for *a-priori* known number of speakers are the same as for clustering with the clustering reducing approach for $R+1$ clusters. This means that starting with a high number of clusters does not influence the clustering performance of the reduced number of clusters.
2. Results using the VQ-BIC approach are close to the best results, usually better than with $R+1$ clusters. As non-speech data can come from different sources, several clusters can be attached to this data. Speakers with close characteristics in the feature space can be attributed to the same model while speakers with high variability in their voice can be split into more than one cluster. For these reasons the optimal number of clusters may differ from $R+1$.

Table 1: Clustering results for ($\lambda = 1.5$).

Model Type	File 1 – $R = 7$				File 2 – $R = 13$				File 3 – $R = 15$				File 4 – $R = 20$			
	N_c	acp	asp	K	N_c	acp	asp	K	N_c	acp	asp	K	N_c	acp	asp	K
Baseline	8	0.74	0.94	0.83	14	0.73	0.78	0.75	16	0.75	0.80	0.77	21	0.74	0.67	0.70
$R+1$ clusters	8	0.92	0.76	0.84	14	0.72	0.84	0.78	16	0.78	0.83	0.80	21	0.68	0.78	0.73
Best score	12	0.84	0.88	0.86	10	0.82	0.79	0.81	16	0.78	0.83	0.80	13	0.78	0.71	0.74
R^* clusters	12	0.84	0.88	0.86	11	0.79	0.81	0.80	12	0.82	0.74	0.78	13	0.78	0.71	0.74

4. Conclusions

The temporal data clustering approach based on VQ, which was presented at [1], was applied for long conversations with different numbers of speakers. It can be seen from the results that as the number of speakers increases, the performance of the system degrades. This is logical due to the fact that the number of estimated parameters that had to be estimated increases linearly with the number of speakers. Another reason is that as the number of speakers increases the overlapping between the clusters become bigger and the shapes that should be learned are more complex. The SOM-based system results are robust to initial parameters, as was already shown in [1].

Estimation of the number of the participants (validity problem) is very important. A BIC version for a distance measure based algorithm was presented. In the presented validity criterion, the scaling factor for BIC is important but the results for $\lambda = 1.0$ and $\lambda = 2.0$ gave comparable results. For instance $\lambda = 2.0$ gave the same result in File2, while $\lambda = 1.0$ was slightly better in File 3.

Acknowledgment

The author wants to thank the Swiss Federal Office for Education and Science (OFES) in the framework of both EC/OFES "MultiModal Meeting Manager (M4) project" and the Swiss National Science Foundation, through the National Center of Competence in Research (NCCR) on "Interactive Multimodal Information Management (IM2)" for supporting this work.

References

[1] I. Lapidot (Voitovetskz), H. Guterman, and A. Cohen, "Unsupervised Speaker Recognition Based on Competition Between Self-Organizing-Maps," *IEEE Trans. on Neural Networks*, vol. 13, no.4, pp. 877-887, July 2002.

[2] J. Ajmera, H. Bourlard, and I. Lapidot, "Improved unknown-multiple speaker clustering using HMM," IDIAP, Martigny, Switzerland, Tech. Rep. IDIAP-RR02-23, August 2002.

[3] M. H. Kuhn, "Speaker recognition accounting for different voice conditions by unsupervised classification (cluster analysis)," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-10, no. 1, pp. 54-57, January 1980.

[4] A. Cohen and V. Lapidus, "Unsupervised text independent speaker classification," *Proc. of the Eighteenth Convention of Electrical and Electronics Engineers in Israel*, 1995, pp. 3.2.2 1-5.

[5] A. Solomonoff, A. Mielke, M. Schmidt, and H. Gish, “Clustering speakers by their voices,” *ICASSP’98*, vol. 2, 1998, pp. 557-560.

[6] A. Cohen and V. Lapidus, “Unsupervised, text independent, speaker clussification,” *Proc. of the Int. Conf. on Signal Processing Application and Technology*, 1996, pp. 1745-1749.

[7] S. S. Chen and P. S. Gapalakrishnan, “Clustering via the Bayesian criterion with applications to speech recognition,” *ICASSP’98*, vol. 2, 1998, pp. 645-648.

[8] L. M. D. Owsley, L. E. Atlas, and G. D. Bernard, “Self-organizing feature maps and hidden Markov models for machine-toll monitoring,” *IEEE Trans. on Signal Processing*, vol. 45, no. 11, pp. 2787-2798, November 1997.

[9] K. Pawelzik, J. Kohlmorgen, and K.-R. Muller, “Annealed competition of expert for segmentation and classification of switching dynamics,” *Neural Computation*, vol. 8, no. 2, pp. 340-356, February 1996.

[10] J. Kohlmorgen, K.-R. Muller, and K. Pawelzik, “Segmentation and identification of drifting dynamical systems,” *Proc. Neural Networks for Signal Processing VII IEEE Workshop*, 1997, Amalia Island, USA, pp. 326-335.

[11] O .A. S. Carpinteiro, “A hierarchical self-organising map model for sequence recognition,” *Pattern Analysis and Applications*, vol. 3, no. 3, pp. 289-287, 2000.

[12] A. Krogh, M. Brown, I. Saira Mian, K. Sjolander, and D. Haussler, “Hidden Markov models in computational biology applications to protein modeling,” *J. Mol. Biol.*, vol. 235, no. 5, pp. 1501-1531, February 1994.

[13] T. Kohonen, “The self-organizing map,” *Proc. IEEE*, vol. 78, no. 9, pp. 1464-1480, September 1990.

[14] G. Schwarz, “Estimating the dimension of a model,” *The Annals of Statistics*, vol. 6, no. 2, pp. 461-464, 1978.