IDIAP RESEARCH REPORT

Improved Unknown-Multiple Speaker Clustering Using HMM

Jitendra Ajmera [1, 2]   Hervé Bourlard [1, 2]
Itshak Lapidot [1]

IDIAP–RR 02-23

September 2002

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

1 IDIAP—Dalle Molle Institute of Perceptual Artificial Intelligence, P. O. Box 592, CH-1920 Martigny, Switzerland, {jitendra, bourlard, lapidot}@idiap.ch
2 EPFL, Lausanne

IDIAP Research Report 02-23

Improved Unknown-Multiple Speaker Clustering Using HMM

Jitendra Ajmera Hervé Bourlard Itshak Lapidot

September 2002

Abstract. In this report, we build upon our previous work on automatic speaker clustering. In the previous work, we presented a HMM-based clustering framework where both the number of speakers and the segmentation boundaries are unknown *a priori*. Starting from over-clustering, we converge to a final clustering using an iterative merging and retraining process. The process consists of training a Gaussian Mixture Model (GMM) for each hypothesized speaker cluster, selecting the 'closest' pair of clusters for merging, and retraining the GMM of the merged cluster. Actually, the main contribution of this paper is to propose a new similarity measure between two probability density functions estimated by GMM. It is shown that this similarity measure can be used without the need for any threshold/penalty term as often used in information theoretic measures like Bayesian information criteria (BIC) and minimum description length (MDL). The merging and retraining are repeated until no possible pair of clusters for merging is left. The system is evaluated on 1996 Hub-4 evaluation set, and shows significant improvements over our previous results. In particular, it is shown that the system often converges to the correct number of clusters (that is, the correct number of speakers) and, consequently, a high average speaker purity is observed.

1 Introduction

In [1], we presented a novel approach for speaker clustering, where both the number of speakers (clusters) and the segmentation are unknown *a priori*. The approach ([1]) uses an ergodic hidden Markov model (HMM) with minimum duration constraints, and the number of clusters and the segmentation are found using an iterative procedure. The process is started by over-clustering the data, such that the initial number of clusters is much greater than the expected number of speakers. The probability density function (PDF) of each cluster is estimated using a Gaussian mixture model (GMM), whose parameters are found using expectation-maximization (EM) algorithm, and the segmentation is found using the Viterbi algorithm. Following this, given the current GMM parameters, the closest pair of clusters is merged using log likelihood ratio (LLR) as a measure of closeness of two clusters. The PDF of the new (merged) cluster is again estimated by a GMM, however, this GMM has the same number of parameters as the sum of the number of parameters of the GMMs of two individual clusters. Thus, the topology of the HMM changes as it has one less state (cluster), however, the number of parameters remain unchaned. A new segmentation is found using this new HMM topology and the Viterbi score (likelihood along the best path) of this segmentation is calculated. The iterative process is continued if this score is greater than the previous score, otherwise, terminated.

One disadvantage of this approach is that not all the possible pair of clusters are merged as the interative process is terminated when a decrease in the Viterbi score (likelihood along the best path) is observed. In simple words, at the time of termination, other (better) pair of clusters may exist which if merged, could have resulted an increase in this score. This was frequently observed in our experiments. Thus, a better criterion is required to identify all the possible pairs of clusters for merging.

In the present work, we address this problem by proposing a new smiliarity measure between two PDFs (estimated by GMM). It is observed that merging according to this measure always results in an increase in the Viterbi score of the data given the new HMM topology and parameters. All cluster pairs satisfying the criterion are merged and the process terminates when no more cluster pairs are available for merging. In this way, the system reaches a state of maximum likelihood for a given number of parameters. A further advantage of the proposed distance measure is that no penalty/threshold term is required and this issue is further explained in the paper.

The remainder of this report is organized as follows. Section 2 explains the new distance measure and its use for speaker clustering. In Section 3 the main steps of the proposed speaker clustering algorithm are presented. Finally, Section 4 presents an experimental evaluation of the proposed technique on the 1996 Hub-4 evaluation set.

2 Distance Measure

Let $\{D_1\}$ and $\{D_2\}$ be two data sets and θ_1 and θ_2 be the maximum likelihood estimates of the parameters of the PDF of $\{D_1\}$ and $\{D_2\}$ respectively. In the context of speaker clustering, a distance/similarity measure is required to decide if two clusters having data sets $\{D_1\}$ and $\{D_2\}$ should be merged or not. Symmetrical LLR and other information theoretic measures can be applied for this purpose and are explained below.

2.1 Symmetrical Log Likelihood Ratio (LLR)

Symmetrical LLR can be used for the above mentioned problem in the following form:

$$D_{llr} = \sum_{x \in \{D_1\}} \log \frac{p(x|\theta_1)}{p(x|\theta_2)} + \sum_{x \in \{D_2\}} \log \frac{p(x|\theta_2)}{p(x|\theta_1)} \tag{1}$$

where $p(x|\theta)$ is the likelihood of data-point x given the PDF parameterized by θ.

Since θ_1 and θ_2 are the maximum likehood estimates on $\{D_1\}$ and $\{D_2\}$ respectively, D_{llr} calculated using equation (1) is always non-negative. Thus, a threshold is required to decide if $\{D_1\}$

and $\{D_2\}$ should be merged or not. This threshold can be found empirically, however, it may require re-tuning for different databases.

2.2 Information Theoretic Measures

Approaches based on information/coding theory such as Rissanen's minimum description length (MDL) [2] which formally coincides with Bayesian information criterion (BIC) [3], the minimum message length (MML) criterion [4], Akaike information criterion (AIC) [5] try to handle the above mentioned problem in a different way. A comprehensive study and comparison about these criteria is presented in [6].

In general the rationale behind these criteria is: if you can build a short code for your data, that means you have a good data generation model. More specifically, consider some data-set X, known to have been generated according to $p(Y|\theta)$, which is to be encoded and transmitted. Following Shannon theory [7], the shortest code lenght is $\lceil -\log p(Y|\theta) \rceil$. If $p(Y|\theta)$ is fully known to both the transmitter and reciever, they can both build the same code and communication can proceed. However, if θ is a priori unknown, the transmitter has to start by estimating and transmitting θ. This leads to a two part message, whose total length is given by:

$$Length(\theta, Y) = Length(\theta) + Length(Y|\theta) \tag{2}$$

All minimum encoding lenght criterion state that the parameter estimate is the one minimizing $Length(\theta, Y)$. minimising the second term in Eq. 2 is equivalent to maximising likelihood $p(Y|\theta)$. Thus, the *theta* is estimated so as to maximise $p(Y|\theta)$, i.e. maximum likelihood (ML) approach. However, maximizing likelihood is not enough when we are forced to choose among nested classes of parametric models, i.e. when we have variable $Length(\theta)$. In this case, the resulting selection rule takes the form of a penalized log-likelihood, where the penalty is the extra cost incurred for sending the parameters in θ (where, θ is the ML estimate).

Accordingly, BIC (and similarly MDL) can be used in the present context by comparing the penalized log-likelihoods for two different hypotheses. In the first hypothesis, the data sets $\{D_1\}$ and $\{D_2\}$ come from two different sources, and hence are modeled by individual models (with parameters θ_1 and θ_2 respectively). In the second hypothesis, they come from the same source and hence should be modeled together. Let θ be the maximum likelihood estimate of the parameters of the PDF of the complete data $\{D\}$ $(= \{D_1\} \cup \{D_2\})$. The decision is made according to the score calculated as follows:

$$D_{bic} = \sum_{x \in \{D_1\}} \log p(x|\theta_1) + \sum_{x \in \{D_2\}} \log p(x|\theta_2) - \sum_{x \in \{D\}} \log p(x|\theta) - \frac{1}{2}\lambda K \log N \tag{3}$$

where λ is ideally equal to one, N is the number of data points in $\{D\}$ and K is the difference of the number of parameters used to model the data in two hypothesis. It should be noted that this is essentially a penalized log-likelihood ratio [8] with the penalty term being $\frac{1}{2}\lambda K \log N$. The factor of λ in the penalty is ideally one, however, in practical applications it is always needed to tune this parameter [9]. To our knowledge, there is no formalized way of finding this parameter.

2.3 Proposed Similarity Measure

In this paper, we propose another similarity measure. The main motivation behind this measure is to eliminate the need for any thresholding or penalization.

In the case when the PDF is modeled by GMMs, let θ_1 and θ_2 be the parameters of GMMs having M_1 and M_2 Gaussian components respectively. Let θ be the maximum likelihood estimate of the parameters of the PDF of the complete data $\{D\}$ $(= \{D_1\} \cup \{D_2\})$. In our approach, θ are the paramters of a GMM having $M_1 + M_2$ component.

The proposed measure $D_{proposed}$ is then calculated as:

$$D_{proposed} = \sum_{x \in D_1} \log p(x|\theta_1) + \sum_{x \in D_2} \log p(x|\theta_2) - \sum_{x \in D} \log p(x|\theta) \tag{4a}$$

$$= \sum_{x \in D_1} \log \frac{p(x|\theta_1)}{p(x|\theta)} + \sum_{x \in D_2} \log \frac{p(x|\theta_2)}{p(x|\theta)} \tag{4b}$$

In the form of equation (4b), the proposed measure has a similar form as the LLR (equation (1)), i.e. it is also essentially a log-likelihood ratio. however, the major difference is that the values of $D_{proposed}$ range from negative to positive values for very close to very distant distributions respectively. Thus, it gives much clearer indication of the closeness of the two distributions and in many cases, zero can serve as a natural threshold.

Furthermore, in the form of (4a), $D_{proposed}$ can be compared to D_{bic} (Eq. (3)) with $K = 0$ (i.e. using the same number of parameters in two hypothesis on which BIC works. In other words, since $K = 0$, the "threshold" is set to zero.

The proposed similarity measure is successfully used for the purpose of speaker clustering in the present paper. We have also verified the use of this measure for the purpose of speaker change detection.

3 Speaker Clustering Algorithm

The system is based on an ergodic HMM with minimum duration constraints (MDC). Each state of this HMM represents a cluster (speaker) and is composed of several sub-states (tied to the same PDF) to impose the MDC. The PDF of each state is represented by a GMM.

The process starts by over-clustering the data, that is, we deliberately divide the data into a larger number of classes than the expected number of classes (speakers). The motivation for this approach was explained in [1]. The parameters of the HMM are then trained in an unsupervised manner using the iterative EM (Viterbi) algorithm. First, a new segmentation of the data based on the current set of parameters is found using the Viterbi algorithm. Then, the parameters are updated based on this segmentation. Several iterations of the procedure are performed to train the initial number of clusters.

In the next step, all possible cluster pairs are tested as potential candidates for merging. We use the decision criterion explained in section 2.3 to decide if the two clusters are to be merged or not. More specifically, two clusters are considered for merging if $D_{proposed}$ calculated using Eq.(4b) or (4a) is negative. Finally, the pair of clusters resulting in minimum and negative ΔL is selected for merging. The merging is done in such a way that the number of parameters used in the PDF of the new cluster is same as the sum of the number of parameters needed to define the individual PDF of the two clusters. In this way, while the HMM topology is changed since it has one state (cluster) less, the number of parameters in the HMM remains unchanged. Thus, the likelihoods in subsequent iterations are comparable without the need for any penalization.

After the merging, a new segmentation is found using the new HMM topology. It is consistently observed that the merging done using the proposed criterion always results in an increase in likelihood of the data given the new HMM topology and parameters.

The process is repeated until there are no remaining candidates for merging, that is, no pair of clusters exists for which $D_{proposed} < 0$.

The algorithm can thus be summarized as:

1. Start by over-clustering, i.e. clustering the data into larger number of clusters than the hypothesized number of speakers. The PDF of each cluster is estimated by a GMM and the parameters of this GMM are trained in an unsupervised way using the iterative EM algorithm.

2. Obtain the segmentation (using the Viterbi algorithm) using the current HMM topology and parameters.

3. Retrain the parameters of all clusters based on this segmentation.

4. Search for all possible candidate pairs satisfying $D_{proposed} < 0$, and select the best pair.
5. If a candidate pair exists, repeat steps 2 through 4, otherwise terminate.

4 Evaluation Experiments

The system was evaluated on the 1996 Hub4 evaluation data. This dataset comprises of 4 different speech datasets, each of approximately 30 minutes duration. We use the same evaluation criterion as in [1], that is, calculating both speaker (*asp*) and cluster (*acp*) purities. The *asp* indicates the degree with which a speaker is limited to one cluster and the *acp* indicates how well a cluster is limited to one speaker. Finally, $K = \sqrt{asp * acp}$ indicates the percentage of total frames falling in their correct clusters.

$Test$	N_s	N_c		asp		acp		K	
		$Before$	**Now**	$Before$	**Now**	$Before$	**Now**	$Before$	**Now**
$File1$	7	13	**17**	0.88	**0.84**	0.79	**0.88**	0.84	**0.86**
$File2$	13	13	**14**	0.82	**0.80**	0.75	**0.79**	0.79	**0.80**
$File3$	15	21	**16**	0.77	**0.92**	0.77	**0.80**	0.78	**0.86**
$File4$	20	21	**16**	0.58	**0.68**	0.55	**0.64**	0.57	**0.66**

Table 1: Evaluation results *now* on 1996 Hub-4 evaluation set compared with the results obtained *before* [1]. The results are presented in terms of number of clusters found N_c, average speaker purity *asp*, average cluster purity *acp* and the overall evaluation criterion K. N_s is the number of speakers in the dataset

In Table 1, speaker clustering performance on the 1996 Hub4 set is reported for four test conditions:

- File1: This file has 7 speakers and large non-speech segments. The number of clusters found, N_c, is much higher than the number of speakers. However, a high value of *asp* shows that the speech/speaker data is limited to the correct clusters. To further verify this fact, the algorithm was run only on the speech segments of the dataset, and the system converged to 9 clusters (with K= 0.88). A higher value of N_c in this case reflects the fact that most of the non-speech clusters could not be merged using the proposed distance measure.
- File2: This dataset has 13 speakers with practically no non-speech segments. For this dataset, the performance in the two cases is comparable. This test also shows that, in the case of a limited number of speakers and no non-speech data, the algorithm converges to the correct number of clusters.
- File3: This dataset has 15 speakers. The performance of the system in this case improves significantly using the proposed distance measure. An improvement in both *asp* and *acp* shows that some of the speaker clusters which could not be merged in [1] have now been successfully merged.
- File4: This dataset has 20 speakers. As mentioned in [1], a large number of speakers degrades the performance in two ways. First, the amount of data available for each speaker decreases, making it difficult to model the speakers. Second, the possibility of overlap in the feature space increases as the number of speakers increases. Results indicate a clear improvement in the system performance using the proposed distance measure.

From the above results, it is clear that the clustering performance has been significantly improved by using the proposed distance measure and the new termination criterion.

The algorithm was also tested on several monologues, and in each case the system successfully converged to one cluster.

5 Conclusion

A new distance measure to assess the closeness of two PDFs has been presented and successfully applied in a speaker clustering framework. The approach is compared with the LLR, as well as information theoretic approaches, like the BIC and the MDL. An important advantage of the technique is that no thresholding is required. The measure is used in a speaker clustering application, using an ergodic HMM with minimum duration constraints. Each HMM state represents a cluster (speaker) and the PDF of each state is estimated by a GMM. The parameters of the HMM are trained in a completely unsupervised manner using the iterative EM algorithm. Initially, since no information about the number of speakers (clusters) is known, the data is clustered into large number of clusters. In subsequent iterations, the closest clusters (where the closeness of clusters is decided using the proposed distance measure) are merged. The merging is done in such a way that the total number of parameters in the HMM remain constant. The HMM topology, however, is changed to contain one less cluster (state). The merging is continued until there are no cluster pairs satisfying the merging criterion. The algorithm was evaluated on 1996 Hub-4 evaluation set and improvements in the clustering performance were observed. In these experiments, the number of clusters found often corresponds to the correct number of speakers.

References

[1] J. Ajmera, H. Bourlard, I. Lapidot, and I. McCowan, “Unknown-multiple speaker clustering using HMM,” *International Conference on Spoken Language Processing*, 2002, to be published.

[2] J. Rissanen, “Stochastic complexity in statistical inquiry,” *Singapore: World Scientific*, 1989.

[3] G. Schwarz, “Estimating the dimension of a model,” *Annals of Statistics*, vol. 6, pp. 461–464, 1978.

[4] J. Oliver, R. Baxter, and C. Wallace, “Unsupervised learning using MML,” *Proc. 13th Int'l Conf. Machine Learning*, pp. 364–372, 1996.

[5] M. Whindham and A. Cutler, “Information ratios for validating mixture analysis,” *Journal of American Statistical Association*, vol. 87, pp. 1188–1192, 1992.

[6] Mark H. Hansen and Bin Yu, “Model selection and the principle of minimum description length,” *Journal of the American Statistical Association*, vol. 96, no. 454, 2001.

[7] T. Cover and J. Thomas, *Elements of Information Theory*, New York: John Wiley & Sons, 1991.

[8] S. S. Chen and P. S. Gopalakrishnan, “Speaker, environment and channel change detection and clustering via the Bayesian information criterion,” Tech. Rep., IBM T.J. Watson Research Center, 1998.

[9] Alain Tritschler and Ramesh Gopinath, “Improved speaker segmentation and segments clustering using the Bayesian information criterion,” *Eurospeech*, pp. 679–682, 1999.