UNKNOWN-MULTIPLE SPEAKER CLUSTERING USING HMM

J. Ajmera *H.Bourlard* *I. Lapidot* *I. McCowan*

IDIAP, Martigny, Switzerland
{jitendra, bourlard, lapidot, mccowan}@idiap.ch

ABSTRACT

An HMM-based speaker clustering framework is presented, where the number of speakers and segmentation boundaries are unknown *a priori*. Ideally, the system aims to create one pure cluster for each speaker. The HMM is ergodic in nature with a minimum duration topology. The final number of clusters is determined automatically by merging closest clusters and retraining this new cluster, until a decrease in likelihood is observed. In the same framework, we also examine the effect of using only the features from highly voiced frames as a means of improving the robustness and computational complexity of the algorithm. The proposed system is assessed on the 1996 HUB-4 evaluation test set in terms of both cluster and speaker purity. It is shown that the number of clusters found often correspond to the actual number of speakers.

1. INTRODUCTION

For speech transcription problems with a large number of speakers, it is beneficial to adapt the *automatic speech recognition* (ASR) system for each speaker. It has been shown that such speaker adaptation leads to significant improvements in speech recognition performance [1, 2]. This speaker adaptation is, of course, dependent on an accurate speaker clustering system.

Several clustering schemes have been proposed in the literature, most of which first segment and then cluster the data. The segmentation is either assumed to be known [1, 3, 4] or performed automatically prior to clustering [5, 6]. These approaches have limitations: in the former case, the correct segmentation is rarely known *a priori* for practical applications, and in the latter case, the errors made in the segmentation step are not only difficult to correct later, but can degrade the performance of the subsequent clustering step. In the proposed technique, we perform clustering directly on the data, deriving the segmentation (according to the clusters) in the process. For applications other than speaker adaptation where finer segmentation is desired, such as speaker identification, this initial broad segmentation could then be further refined by applying acoustic change detection within clusters.

A *hidden Markov model* (HMM) based clustering scheme was proposed in [7, 8]; however the number of speakers was assumed to be known. In [9], a validity criterion was proposed to automatically determine the number of speakers for the purpose of speaker recognition; however the system was limited to a small number of speakers.

In this work, we investigate the use of an ergodic HMM with minimum duration constraints for speaker clustering. A similar approach has been used previously for speech/music discrimination in [10]. In the proposed system, we start by over-clustering (where the number of clusters is believed to be far greater than the number of speakers) and then refine this by merging mutually closest clusters according to a distance measure (log likelihood ratio in this work). The newly formed cluster is then retrained using the data belonging to the respective clusters. This process is repeated until a decrease in likelihood is observed. In this way, the system makes no assumptions regarding the type or number of speakers or their segmentation. To assess the technique, a new evaluation criterion taking into account both cluster and speaker purity is presented.

In the same framework, we also investigate clustering using only highly voiced frames. This has the effect of automatically rejecting most of the non-speech and silence frames, as well as basing the clustering on only the most reliable speech frames, as voiced frames have higher energy than unvoiced frames and are hence less susceptible to noise. Experiments indicate that using only voiced frames leads to similar performance, however the computational complexity is significantly reduced.

Experiments on the 1996 HUB-4 evaluation dataset demonstrate the effectiveness of the proposed technique compared to a baseline system assuming known number of speakers. Results for the proposed system are also presented when only the voiced frames are used.

2. SPEAKER CLUSTERING FRAMEWORK

In this section, we present the proposed HMM-based speaker clustering system. In addition, the use of only highly voiced frames for clustering is proposed and discussed.

2.1. System Overview

The proposed clustering system is based on an ergodic HMM with minimum duration constraints. Each state of the HMM represents a cluster and the *probability density function* (PDF) of each state (cluster) is represented by a *Gaussian mixture model* (GMM). The HMM is trained in an unsupervised manner using the *expectation maximization* (EM) algorithm. The initialization of the PDFs is done using the *k-means* algorithm.

We start by over-clustering the data. The term "over-clustering" means that at the initial clustering step, we deliberately cluster the data into a greater number of classes than the expected number of classes (speakers) in the data set. This reduces the probability that different speakers will be clustered into one class. This step is useful because different speakers may be very close in some features and tend to be under-segmented (grouped into same cluster). Also, when automatically determining the number of clusters to use, combining clusters that belong together is a much simpler task than splitting up those that do not.

Once the initial clusters are trained, the next step is to reduce the number of clusters by merging. The primary source of knowledge for this comes from the cluster distribution in the feature space. At the end of a segmentation process (using Viterbi algorithm), the mutually closest pair of clusters is identified using a likelihood ratio distance measure, and these are then merged to form a single new cluster. This new cluster is then represented by another GMM having a number of components equal to the sum of the components of the individual clusters. The parameters of this newly formed cluster are retrained using the EM algorithm using the features belonging to respective clusters.

In the next iteration of the procedure, the segmentation is again found using the updated HMM topology with one less cluster and the likelihood of the data based on this segmentation is observed. This likelihood increases if the two merged clusters are valid candidates for merging (the data in two clusters is from the same source/speaker). If, however, clusters having data from two different sources are merged, this likelihood decreases. We stop the merging process when a decrease in likelihood is observed. One possible limitation of this criterion is that the likelihood may decrease after merging a particular pair of clusters (found using LLR), while some other possible pairs may have resulted in increase in likelihood. In [11], we propose a a different merging and termination criterion, which overcomes this limitation.

2.2. Clustering using only voiced frames

In the same framework, we also investigate using only "highly voiced" frames. A frame is identified as being either voiced or unvoiced by observing the auto-correlation sequence. If an explicit pitch frequency exists for a given frame, we regard it as being voiced. The number of such frames is less than half the total number of frames. The selected voiced frames are then used in the framework described above. The smaller clusters (having only voiced features) are then projected onto the whole audio streams.

We list a number of motivations for this approach :

- The voiced frames are high energy frames and are thus less susceptible to noise.
- In this work, we have used *linear predictive cepstral coefficients* (LPCC) features. The all-pole model of the vocal tract (given by LPC analysis) fits the voiced events better compared to unvoiced events, and hence the features for voiced frames should be more reliable for speaker discrimination.
- using our voiced/unvoiced criterion, there are sufficient (more than 50%) voiced frames during speech segments and very few during non-speech (depending on the kind of non-speech signal). Thus we automatically remove a lot of non-speech and silence regions.
- The system becomes at least 4 times faster as the number of frames and minimum duration is reduced to half (approximately).

3. EVALUATION EXPERIMENTS

3.1. Evaluation Criterion

We use the *purity* concept explained in [4] and extend it to calculate both the *average cluster purity* (acp) and *average speaker purity* (asp), as explained below.

First we define:

n_{ij}: Total number of frames in cluster i spoken by speaker j.

N_s: Total number of speakers.

N_c: Total number of clusters.

N: Total number of frames.

$n_{.j}$: Total number of frames spoken by speaker j.

$n_{i.}$: Total number of frames in cluster i.

The purity of a cluster $p_{i.}$ can then be defined as:

$$p_{i.} = \sum_{j=1}^{N_s} n_{ij}^2 / n_{i.}^2 \tag{1}$$

and the average cluster purity acp is:

$$acp = \frac{1}{N} \sum_{i=1}^{N_c} p_{i.} . n_{i.} \tag{2}$$

Similarly, we calculate the speaker purity $p_{.j}$ and asp as:

$$p_{.j} = \sum_{i=1}^{N_c} n_{ij}^2 / n_{.j}^2 \tag{3}$$

$Test$	N_s	asp	acp	K
$File1$	7	0.91	0.68	0.79
$File2$	13	0.59	0.74	0.66
$File3$	15	0.75	0.73	0.74
$File4$	20	0.50	0.64	0.57

Table 1. Results for baseline system, using $N_c = N_s$

$$asp = \frac{1}{N} \sum_{j=1}^{N_s} p_{.j}.n_{.j} \quad (4)$$

The asp gives a measure of how well a speaker is limited to only one cluster, and the acp gives a measure of how well a cluster is limited to only one speaker. We found it neccessary to calculate asp because, it is easy to achieve high value of acp with more number of clusters than really required. However, note that non-speech frames are not taken into account while calculating asp.

In order to facilitate comparison between systems, we propose multiplying these two numbers to obtain an overall evaluation criterion :

$$K = \sqrt{acp * asp} \quad (5)$$

3.2. Results

The system was tested on 1996 HUB-4 evaluation data. HUB-4 is a broadcast news speech corpus, and the evaluation set consists of four datasets, each of approximately 30 minutes duration. Table 1 shows the results for a baseline system, in which the number of clusters (speakers) is assumed known *a priori* and all feature frames are used.

The results for the proposed system are presented in Table 2 for the case when all data frames are used, and the case when only voiced frames are used. In each case, the iterative algorithm was initialized with 30 clusters and the number of GMM components was set to 5. The table indicates the final number of clusters determined by the algorithm(N_c), and presents results in terms of the average speaker purity (asp), average cluster purity (acp) and the overall evaluation criterion (K).

In the following, we discuss the results for each test set.

File1: There are 7 speakers with a few large non-speech segments. In this case, we finish having many extra clusters. However, a high value of asp indicates that most of the speaker frames were clustered correctly and so the extra clusters are mostly occupied by non-speech frames. To verify this, an experiment was run only on the speech segments, and the system converged to 9 clusters. The performance of voiced and all frames is similar in this case.

File2: There are 13 speakers in this data set, with practically no non-speech segments. For both cases (voiced and all) we finish having 13 clusters with a high speaker and cluster purity. The results for using all the frames are slightly better in this case compared to using only the voiced frames.

File3: There are 15 speakers in this data set, with regions of non-speech and silences. When we use only voiced frames, we finish having several extra clusters. Again, a high value of asp indicates that these extra clusters mostly occupy non-speech regions. On the other hand, we obtain the correct number of clusters while using all the frames, but the cluster purity is very low. In this case, using voiced frame features gives better results compared to using all the frames.

File4: There are 20 speakers in this data set as well as regions of non-speech. Although, we finish with the correct number of clusters, the overall performance for both voiced and all cases is poor. This is because of the presence of too many speakers in limited (30 minutes) audio data, making the modelling of every speaker and hence clustering more difficult.

The system was also tried on some monologues (long speech segments from the same speaker) and all the cases, it converged to a single cluster.

3.2.1. Remarks

- In general, the performance of the proposed system is better than that of the baseline system. This means that, even if we know the number of speakers, training those many clusters is not an optimal solution.
- The presence of non-speech data produces many extra clusters, especially when the non-speech comes from different sources like, music, noise, clapping etc. It would be better to remove these segments before clustering by using a speech/non-speech discrimination system.
- We note that, the clustering performance also depends on the initial over-clustering, especially if we do not start with sufficient clusters (compared to the actual number of speakers in the data).
- The efficiency of the system decreases as the number of speakers increases. This was especially true for these four tests as the total data size was the same for all. Thus the amount of data available per speaker decreases as the number of speakers increases, making the modelling of each speaker, and hence clustering, poorer. Also, the possibility of overlap between different speakers in the feature space increases as the number of speakers increase.
- We observe that using only voiced frame features gives similar results to that of using all the features, resulting however in much reduced computational complexity. This also indicates that unvoiced regions do

$Test$	N_s	N_c		asp		acp		K	
		$Voiced$	All	$Voiced$	All	$Voiced$	All	$Voiced$	All
$File1$	7	13	17	0.88	0.83	0.79	0.85	0.84	0.84
$File2$	13	13	13	0.82	0.87	0.75	0.77	0.79	0.84
$File3$	15	21	15	0.77	0.82	0.77	0.36	0.78	0.55
$File4$	20	21	22	0.58	0.62	0.55	0.52	0.57	0.57

Table 2. Results for proposed system, in terms of number of clusters found N_c, average speaker purity asp, average cluster purity acp and the overall evaluation criterion K

not carry additional speaker specific informations (though we checked it only for the case of LPCC features).

- On average asp and acp are greater than 0.7. This means that more than 70% of the time, the speakers are in their right clusters and the clusters occupy data from the same source. This performance would make this system beneficial in the speaker adaptation process for applications like broadcast news transcription.

4. CONCLUSION

An HMM based framework for speaker clustering was presented for the case where both the number of speakers and segmentation is unknown. The system was tested on broadcast news data with different number of speakers and regions of non-speech. Experiments indicate that our system creates a correct number of clusters to fit the data. In the same framework, using only highly voiced frame features was compared against using all the features. Although, a high reduction in computation complexity was observed, the results were similar to that of using all the feature.

5. ACKNOWLEDGEMENTS

This work was supported by the Swiss National Science Foundation grant 2100-65067.01 (AudioSkim), and the European Commission ASSAVID project (IST-1999-13082).

6. REFERENCES

[1] 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.

[2] S. E. Johnson and P. C. Woodland, "Speaker clustering using direct maximisation of the MLLR adapted likelihood," *International Conference on Spoken Language Processing*, vol. 5, pp. 1775–1779, 1998.

[3] M. Sugiyama, J. Murakami, and H. Watanabe, "Speech segmentation and clustering based on speaker features," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 395–398, 1993.

[4] A. Solomonoff, A. Mielke, M. Schmidt, and H. Gish, "Clustering speakers by their voices," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 757–760, 1998.

[5] M . A. Siegler, U. Jain, B. Raj, and R. M. Stern, "Automatic segmentation, classification and clustering of broadcast news audio," *DARPA Speech Recognition Workshop, Chantilly*, pp. 97–99, Feb 1997.

[6] T Hain, S. E. Johnson, A. Turek, P. C. Woodland, and S. J. Young, "Segment generation and clustering in the HTK broadcast news transcription system," *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, pp. 133–137, 1998.

[7] J. O. Olsen, "Seperation of speakers in audio data," *EUROSPEECH*, pp. 355–358, 1995.

[8] J. Murakami, M. Sugiyama, and H. Watanabe, "HMM based unknown multiple signal source clustering problem," *Techical Report of ASJ Speech Committee*, Oct. 1992, (In Japanese).

[9] I. Voitovetsky, H. Guterman, and A. Cohen, "Validity criterion for unsupervised speaker recognition," in *Proc. of the first workshop on Text, Speech, Dialogue*, Brno, Czech Republic, September 1998.

[10] J. Ajmera, I. McCowan, and H. Bourlard, "Robust HMM based speech/music segmentation," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002, to be published.

[11] J. Ajmera, H. Bourlard, and I. Lapidot, "Improved unknown-multiple speaker clustering using HMM," *IDIAP Reseach Report RR-02-23*, 2002, http://www.idiap.ch/publications.