COMBINING SPEAKER IDENTIFICATION AND BIC FOR SPEAKER DIARIZATION *

Xuan Zhu, Claude Barras, Sylvain Meignier† and Jean-Luc Gauvain

Spoken Language Processing Group (http://www.limsi.fr/tlp)
LIMSI-CNRS, BP 133, 91403 Orsay cedex, France

{xuan,barras,meignier,gauvain}@limsi.fr

ABSTRACT

This paper describes recent advances in speaker diarization by incorporating a speaker identification step. This system builds upon the LIMSI baseline data partitioner used in the broadcast news transcription system. This partitioner provides a high cluster purity but has a tendency to split the data from a speaker into several clusters, when there is a large quantity of data for the speaker. Several improvements to the baseline system have been made. Firstly, a standard Bayesian information criterion (BIC) agglomerative clustering has been integrated replacing the iterative Gaussian mixture model (GMM) clustering. Then a second clustering stage has been added, using a speaker identification method with MAP adapted GMM. A final post-processing stage refines the segment boundaries using the output of the transcription system. On the RT-04f and ESTER evaluation data, the improved multi-stage system provides between 40% and 50% reduction of the speaker error, relative to a standard BIC clustering system.

1. INTRODUCTION

Speaker diarization is the process of partitioning an input audio stream into homogeneous segments according to the speaker identity and the background and channel conditions. Unlike speaker identification or tracking tasks where a priori knowledge of the speaker's voice is provided and an absolute identification is required, the speaker diarization task is relative to a given show, and thus only a relative, show-internal speaker identification is output by the system.

Speaker diarization is a useful preprocessing step for an automatic speech transcription system. By separating out speech and non-speech segments, the recognizer only needs to process audio segments containing speech, thus reducing the computation time. By clustering segments of the same acoustic nature, condition specific models can be used to improve the recognition performance. By clustering segments from the same speaker, the amount of data available for unsupervised speaker adaptation is increased, which can significantly improve the transcription performance. Speaker diarization can also improve readability of an automatic transcription by structuring the audio stream into speaker turns and in some cases by providing the identity of the speakers. Such information can also be of interest for the indexation of multimedia documents.

For most speaker diarization tasks, the number of speakers and the speaker characteristics are unknown a priori, and need to be automatically determined. There are two predominant approaches to the speaker diarization problem. The first approach relies on a two step procedure [1, 2, 3]. First is the segmentation step, which locates segment boundaries based on acoustic changes in the signal. Second is the clustering step, which regroups segments coming from the same speaker into a cluster. A limitation of this method is that errors made in the segmentation step are not only difficult to correct later, but can also degrade the performance of the subsequent clustering step. An alternative is to optimize jointly the segmentation and the clustering, via, for example, an iterative segmentation and clustering procedure as described in [4] which uses a set of Gaussian Mixture Models (GMMs). An iterative method based on an ergodic hidden Markov model (HMM) is also proposed in [5, 6].

The remainder of this paper is organized as follows: Section 2 describes the baseline partitioning system, and Section 3 describes the Bayesian information criterion (BIC) clustering and speaker identification (SID) clustering used to improve the partitioning system. The experimental results are presented in Section 4 followed by some conclusions.

2. BASELINE PARTITIONING SYSTEM

The baseline data partitioning system is the first stage of the system developed for the LIMSI English broadcast news transcription system [4]. This baseline partitioner processes 38 dimensional feature vectors (12 Mel frequency cepstral coefficients, Δ and Δ-Δ coefficients plus the Δ and Δ-Δ energy), and is structured as follows (cf. Figure 1):

- Speech Activity Detection (SAD): Speech is extracted from the signal with a Viterbi decoding using Gaussian Mixture Models (GMM) for speech, speech over music, music, silence and noise. The GMMs, each with 64 Gaussians, were trained on about 1 hour of data.
- Chopping into small segments: Segmentation of the signal is performed by taking the maxima of a local Gaussian divergence measure between two adjacent sliding windows of 0.5 seconds, similar to [1]. A single diagonal Gaussian is used for each window.
- Iterative GMM segmentation/clustering procedure: Each initial segment is used to seed one cluster, and a 8 components GMM with diagonal covariance matrix is trained on the segment data. The algorithm alternates Viterbi resegmentation and GMMs reestimation steps until maximization of the objective function: $\log f(S|M) - \alpha N - \beta K$ where $f(S|M)$ is the likelihood of the N segments given the K models M, and αN and βK are segment and cluster penalties.

*This work was partially financed by the European Commission under the FP6 Integrated Project IP 506909 CHIL.

†now with the Laboratoire d'Informatique de l'Universite du Maine

- Viterbi resegmentation: The segment boundaries are refined using the last set of GMMs and an additional relative energy-based boundary, within a 1 second interval, so as to avoid cutting words.
- Bandwidth (studio or telephone) and gender (male or female) labeling is performed on the segments using 4 GMMs with 64 diagonal covariance matrices.

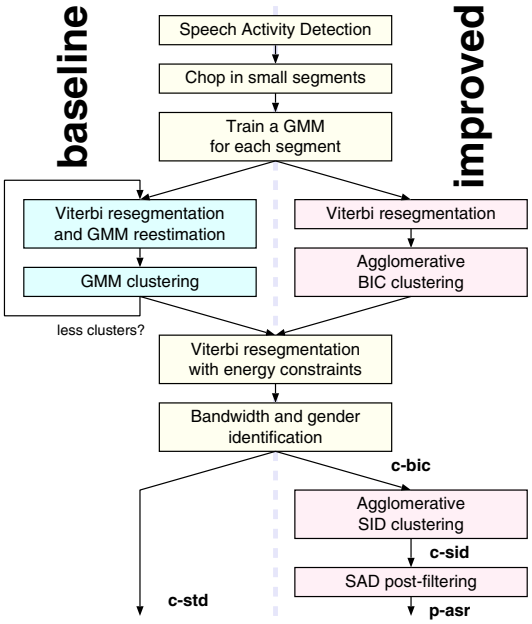

Figure 1: Standard baseline LIMSI partitioning system (c-std on the left side of the diagram) and improved speaker diarization system (p-asr to the right, along with c-bic and c-sid intermediate steps).

3. MULTI-STAGE PARTITIONING

In recent research on the speaker diarization task, BIC clustering methods have been widely used with a good performance [7, 8]. We therefore tested a modified system, replacing the iterative GMM clustering with BIC-based clustering (cf. Figure 1, (c-bic)). We also pipelined the output of the system into a second clustering stage which uses a speaker identification module (c-sid). Finally, a SAD post-filtering stage was added in order to take into account short pauses. The other parts of the system were kept unchanged.

BIC clustering

Agglomerative clustering is applied to the segments output by one iteration of GMM resegmentation. At the beginning, each segment seeds one cluster, modeled by a single Gaussian with a full covariance matrix trained on the 12 Mel cepstrum coefficients and the energy (but without the Δ coefficients). At each step, the two nearest clusters are merged until the stop criterion is reached. The BIC criterion [2] is used both for the inter-cluster distance measure and the stop criterion.

In order to decide whether to merge two clusters c_i and c_j, the ΔBIC value is computed as:

$$\Delta BIC = (n_i + n_j)\log|\Sigma| - n_i \log|\Sigma_i| - n_j \log|\Sigma_j| - \lambda P$$

where Σ is the covariance matrix of the merged cluster (c_i and c_j), Σ_i of cluster c_i, Σ_j of cluster c_j, and n_i and n_j are respectively the number of the acoustic frames in cluster c_i and c_j. The penalty P is: $P = \frac{1}{2}(d + \frac{1}{2}d(d+1))\log n$ where d is the dimension of the feature vector space. The merging criterion is that two clusters should be merged if $\Delta BIC < 0$. The clustering procedure terminates when $\Delta BIC > 0$.

In our BIC clustering procedure, the size of the two merged clusters, i.e. $n = n_i + n_j$, is used in the penalty P for the BIC criterion, as described in [9]. We refer to this as a local BIC penalty. But in general the size of the whole set of clusters, i.e. $n = \sum_{k=1}^{N} n_k$ has to be used in the penalty, which we refer to as a global BIC penalty. Since the BIC criterion is used as the distance measure for merging the clusters, using the total size makes the penalty constant, so the decision to merge two clusters is decided just by the increase in likelihood for the global BIC case. The local BIC thus seems to be a better choice for a merging criterion, even if it is not optimal as a stop criterion.

SID clustering

Speaker clustering methods performed by either the iterative GMM or the BIC agglomerative clustering procedures have to deal in the beginning of the process with short duration segments, and thus use a limited set of parameters per cluster. After several iterations, the amount of data per cluster increases, so a more complex model can be used. In addition, purely acoustic clustering tends to split a speaker's data into several clusters as a function of the various background conditions (clean speech, speech with noise, speech with music etc.), so an acoustic background normalization is necessary to regroup the data for a given speaker.

After the BIC clustering stage, state-of-the-art speaker recognition methods [10, 11] were used to improve the quality of the speaker clustering. The feature vector consists of 15 Mel frequency cepstral coefficients plus delta coefficients and delta energy with a feature warping normalization [12]. For each gender (male, female) and each channel condition (studio, telephone) combination, a Universal Background Model (UBM) with 128 diagonal Gaussians is trained on the 1996/1997 English Broadcast News data. For each initial cluster, maximum a posteriori (MAP) adaptation of the means of the matching UBM is performed.

Agglomerative clustering is performed separately for each gender and bandwidth condition, using a cross log-likelihood ratio as in [13]. For each cluster c_i, its model M_i is MAP adapted from the gender and channel matched UBM R using the feature vectors x_i belonging to the cluster. Then, given two clusters c_i and c_j, the cross log-likelihood ratio is defined as:

$$clr(c_i, c_j) = log\frac{f'(x_i|M_j)}{f'(x_i|R)} + log\frac{f'(x_j|M_i)}{f'(x_j|R)}$$

where $f'(\cdot|M)$ is the likelihood of the acoustic frames given the model M, normalized by the length of the signal. This is a symmetric similarity measure. After each merge, a new model is trained for the cluster $c_{i\cup j}$. The clustering stops when the cross log-likelihood ratio between all clusters is below a given threshold δ estimated on development data.

SAD post-filtering

The output of the LIMSI Broadcast News Speech-To-Text system is used in a post-processing stage to filter out short-duration silence segments that are not detected by the initial speech detection step. Only inter-word silences lasting at least 1 second are filtered out, this value being determined on development data.

4. EXPERIMENTS AND RESULTS

Several configurations were tested for the systems. By default, the configuration used is the one that provided the best results on development data, i.e. $\alpha = \beta = 230$ for c-std, $\lambda = 5.5$ for c-bic and $\lambda = 3.5, \delta = 0.1$ for c-sid and p-asr. A local BIC merging and stop criterion was also used.

Databases

The experiments were conducted on the development database (dev1) and the evaluation database used in NIST RT-04f (Fall 2004 Rich Transcription Evaluation) [14] and on the databases of the French ESTER broadcast news evaluation [15].

The development database (dev1) used in English RT-04f consists of 6 audio files recorded in February 2001. The evaluation database consists of 12 audio files recorded in December 2003. All the audio files last 30 minutes and were extracted from different US radio and television broadcast news shows.

The development database for ESTER consists of 8 hours audio data from 4 radio broadcast news shows in French (France Inter, France Info, RFI, RTM); the evaluation database of ESTER consists of 10 hours data from the same 4 sources plus 2 new sources. The audio files last from 15 minutes to one hour.

Performance measures

The speaker diarization task performance is measured via an optimum mapping between the reference speaker IDs and the hypotheses. The primary metric for the task is the fraction of speaker time that is not attributed to the correct speaker, given the optimum speaker mapping. In addition to this speaker error, the overall speaker diarization error includes also the missed and false alarm speaker times, thus taking speech/non-speech detection errors into account [14].

In order to analyze better the performance of speaker clustering methods, average frame-level cluster purity and cluster coverage are used as defined by [4]. Cluster purity is defined as the ratio between the number of frames by the dominating speaker in a cluster and the total number of frames in the cluster. Cluster coverage accounts for the dispersion of a given speaker's data across clusters.

Results on the RT-04f development data

As expected, the standard partitioner c-std in its default configuration provides a high purity, but a relatively poor coverage, resulting in a high overall diarization error over 30% on dev1 data (cf. Table 1). Setting the penalty α and β to optimize these values reduces this error below 25%. The c-bic system also provides a high purity, with much better coverage (resp. 97% and 90%), reducing the overall error rate by almost 50%. The c-sid system achieves a large increase of the coverage without degradation of the purity, resulting in a global error rate about 7%, a reduction of almost 50% compared to c-bic system.

A global BIC merging and stop criterion was also tested, but always performed worse compared to the local BIC criterion in our experiments, as can be seen for c-bic system on RT-04f dev1 (cf. Table 2). A similar result was observed in [8]. This result remains to be further interpreted but may be due to an inadequacy between the BIC modelization and the real distribution of the data. Thus only the local criterion was used in the remaining experiments.

Looking in more detail at the performance of the c-sid system, a large variation of the speaker error across shows is observed, ranging from the lowest error of 0.8% for MNB show to over 12% for ABC and NBC shows (cf. Table 3).

system	*cluster purity (%)*	*coverage (%)*	*overall error*
RT-04f dev1 dataset			
c-std ($\alpha = \beta = 160$)	95.0	71.6	32.3
c-std ($\alpha = \beta = 230$)	90.6	82.1	24.8
c-bic ($\lambda = 5.5$)	97.1	90.2	13.2
c-sid ($\lambda = 3.5, \delta = 0.1$)	97.9	95.8	7.1
ESTER development dataset			
c-bic ($\lambda = 5.5$)	92.8	89.4	15.8
c-sid ($\lambda = 3.5, \delta = 1.5$)	95.3	94.8	8.0

Table 1: The cluster purity, cluster coverage and the overall diarization error from the systems c-std (both in initial configuration and best configuration), c-bic and c-sid on the RT-04f dev1 dataset and the ESTER development dataset.

BIC criterion	λ	*overall error*	*BIC criterion*	λ	*overall error*
local	5.0	13.3%	global	5.0	16.4
	6.0	12.8%		6.0	15.5
	7.0	13.8%		7.0	18.2

Table 2: The overall diarization error for c-bic system on the RT-04f dev1 database, as a function of the penalty weight λ for the local and global BIC criterion.

Results on the ESTER development data

For ESTER, SAD was performed using the same acoustic models for RT-04f plus an additional speech over music model trained on French broadcast news data. The optimal threshold for the SID clustering on development data was $\delta = 1.5$. Similar to the results obtained on the RT-04f dev1 dataset, the c-sid system provides also high cluster purity and coverage (resp. 95.3% and 94.8%) on the ESTER development dataset (cf. Table 1). A 50% reduction of overall error rate is gained by adding the c-sid system to the c-bic system.

Results on the evaluation data

On the RT-04f evaluation dataset, the trend observed on the development data was confirmed, with a slight increase in overall diarization error to 17% for the c-bic system and to 9.1% for the c-sid system. The final SAD post-processing stage gives an improvement of 0.6%, mainly by reducing false alarms in speech detection (cf. Table 4). As mentioned in [16], the p-asr system had the best performance in all the participants of the RT-04f evaluation. Postevaluation experiments show that the speaker error for the c-sid system could be reduced from 6.9% to 6.0% with the SID clustering threshold $\delta = 0.4$.

On the ESTER evaluation dataset, with the setting optimized

show	*REF*	*SYS*	*MS*	*FA*	*SPK*	*DIA*
average	-	-	**0.4**	**1.3**	**5.4**	**7.1**
ABC	27	37	1.6	1.3	12.4	15.2
VOA	20	22	0.3	1.2	2.2	3.7
PRI	27	30	0.1	0.9	2.8	3.8
NBC	21	35	0.1	1.1	12.0	13.2
CNN	16	21	0.5	1.4	5.6	7.6
MNB	10	16	0.2	1.8	0.8	2.8

Table 3: Performance of c-sid system on the RT-04f dev1 dataset, scores are given for miss (MS), false alarm (FA), speaker error (SPK) and overall diarization error (DIA), REF and SYS are respectively the reference and system speaker number.

system	*missed speech*	*false alarm speech*	*speaker error*	*overall error*
RT-04f test dataset				
c-bic	0.4%	1.8%	14.8%	17.0%
c-sid($\delta = 0.1$)	0.4%	1.8%	6.9%	9.1%
p-asr*	0.6%	1.1%	6.8%	8.5%
ESTER test dataset				
c-bic	0.7%	1.0%	12.1%	13.8%
c-sid($\delta = 1.5$)*	0.7%	1.0%	9.8%	11.5%
c-sid($\delta = 2.0$)	0.7%	1.0%	7.4%	9.1%

Table 4: Performances of c-bic, c-sid and p-asr systems on the evaluation data of RT-04f and ESTER (*these systems are the primary systems submitted to the evaluations.).

on the development database, overall diarization error was reduced from 13.8% for the c-bic system to 11.5% for the c-sid system (cf. Table 4). Postevaluation experiments on the evaluation data illustrate that the c-sid system has a still better result (9.1% overall diarization error) with $\delta = 2.0$.

5. CONCLUSIONS

In this paper, we have presented the LIMSI improved speaker diarization system. Several modifications to the baseline system have been explored by replacing the iterative GMM clustering with the combination of an agglomerative BIC clustering and a second clustering using the state-of-the-art speaker identification techniques.

The improved system performs much better for the diarization task. On the RT-04f development data, the improved system has a relative speaker error reduction of over 75% compared to the baseline system. An overall diarization error under 10% was obtained with the c-sid system on the RT-04f evaluation data, while the performance of BIC-based systems was at least 15%. In the ESTER evaluation, the c-sid system has 8.0% overall diarization error on the development dataset and 9.1% overall error could be obtained on the evaluation dataset with an optimal δ threshold. This dramatic improvement over the baseline system results from several changes: a model complexity which increases with the average amount of speech data per cluster, and the combination of two different clustering stages and models, each one focusing on a different acoustic aspect.

Several issues remain to be investigated in order to improve the robustness and the efficiency of the system. It was observed that the clustering threshold needs to be set according to the length of the audio document, and that the system still has a large variability across individual shows. Only with a large amount of files can statistically consistent results be obtained. Finally, as most speaker diarization systems rely on a purely acoustic segmentation and clustering, combining the acoustic with the linguistic layer as explored in [17] would improve the robustness of a speaker diarization system and make it more exploitable by a human reader.

REFERENCES

[1] M. Siegler, U. Jain, B. Raj and R. Stern, "Automatic Segmentation, Classification, and Clustering of Broadcast News Audio," *Proc. DARPA Speech Recognition Workshop*, pp. 97-99, Chantilly, Virginia, Feb. 1997.

[2] S.S. Chen and P.S. Gopalakrishnan, "Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion," *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Landsdowne,VA, Feb. 1998.

[3] T. Hain, S.E. Johnson, A. Tuerk, 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*, Landsdowne, VA, Feb. 1998.

[4] J.L. Gauvain, L. Lamel and G. Adda, "Partitioning and Transcription of Broadcast News Data," *Proc. ICSLP*, pp. 1335-1338, Sydney, Dec 1998.

[5] S. Meignier, J.-F. Bonastre, and S. Igounet, "E-HMM approach for learning and adapting sound models for speaker indexing," *Proc. ISCA Odyssey Workshop*, Chania, Crete, June 2001, pp. 175–180.

[6] J. Ajmera, C. Wooters, "A robust speaker clustering algorithm" *Proc. IEEE ASRU Workshop*, Virgin Islands, Nov. 2003.

[7] Y. Moh, P. Nguyen, and J.-C. Junqua, "Towards domain independent speaker clustering," *Proc. IEEE ICASSP*, Hong Kong, April 2003.

[8] S. Tranter and D. Reynolds, "Speaker diarisation for broadcast news," *Proc. ISCA Odyssey Workshop*, Toledo, June 2004.

[9] M. Cettolo, "Segmentation, Classification and Clustering of an Italian Broadcast News Corpus", *Proc RIAO*, Paris, Apr. 2000.

[10] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[11] C. Barras and J.-L. Gauvain, "Feature and score normalization for speaker verification of cellular data," *Proc. IEEE ICASSP*, May 2003.

[12] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," *Proc. ISCA Odyssey Workshop*, June 2001.

[13] D. Reynolds, E. Singer, B. Carlson, G. O'Leary, J. McLaughlin and M. Zissman, "Blind clustering of speech utterances based on speaker and language characteristics," *Proc. ICSLP*, Sydney, Dec 1998.

[14] "Fall 2004 Rich Transcription (RT-04f) Evaluation Plan," 2004, `http://nist.gov/speech/tests/rt/rt2004/fall/`.

[15] G. Gravier, J.-F. Bonastre, S. Galliano, E. Geoffrois, K. Mc Tait, and K. Choukri, "The ESTER evaluation campaign of rich transcription of french broadcast news," *Proc. LREC*, Lisbon, Portugal, May 2004.

[16] D. Reynolds and P. Torres-Carrasquillo, "Approaches and Applications of Audio Diarization," *Proc. IEEE ICASSP*, Philadelphia, March 2005.

[17] L. Canseco-Rodriguez, L. Lamel, J.-L. Gauvain, "Speaker diarization from speech transcripts," *Proc. ICSLP*, Jeju, Oct. 2004.