Improving Speaker Diarization

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

IMPROVING SPEAKER DIARIZATION

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

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

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

ABSTRACT

This paper describes the LIMSI speaker diarization system used in the RT-04F evaluation. The RT-04F system builds upon the LIMSI baseline data partitioner, which is 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. In the RT-03S evaluation the baseline partitioner had a 24.5% diarization error rate. Several improvements to the baseline diarization system have been made. A standard Bayesian information criterion (BIC) agglomerative clustering has been integrated replacing the iterative Gaussian mixture model (GMM) clustering; a local BIC criterion is used for comparing single Gaussians with full covariance matrices. A second clustering stage has been added, making use of a speaker identification method: maximum a posteriori adaptation of a reference GMM with 128 Gaussians. A final post-processing stage refines the segment boundaries using the output of the transcription system. Compared to the best configuration baseline system for this task, the improved system reduces the speaker error time by over 75% on the development data. On evaluation data, a 8.5% overall diarization error rate was obtained, a 60% reduction in error compared to the baseline.

1. INTRODUCTION

Acoustic diarization is the process of partitioning an input audio stream into acoustically homogeneous segments according to the speaker identity and the background and channel conditions. 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 performance of the transcription system. 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.

There are two predominant approaches to the speaker diarization problem. In most situations the number of speakers and the speaker characteristics are unknown a priori, and need to be automatically determined. The first approach relies on a two step procedure [6, 9, 13]. 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 clusters. 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 [7] which uses a set of Gaussian Mixture Models (GMMs). An iterative method based on an ergodic hidden Markov model (HMM) is also proposed in [2].

The remainder of paper is organized as follows: Section 2 briefly reviews the speaker diarization task and experimental conditions. Section 3 describes the baseline partitioning system, and Section 4 describes the BIC clustering and speaker ID clustering used to improve the partitioning system. The experimental results are presented in Section 5 followed by some conclusions.

2. EXPERIMENTAL SETUP

The experimental setup followed the RT-04F evaluation plan [1]. Here we briefly describe the task, the performance measures and the development and evaluation corpora.

Task

Diarization in RT-04F consisted of the "Who spoke when" speaker segmentation task, including gender classification (adult male, adult female, child). It was to be performed on English Broadcast News datasets only.

The "Who spoke when" task requires a system to identify all regions of time produced from the same speaker. Unlike the speaker identification or tracking tasks where a priori knowledge of the speaker voices 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.

Performance measures

The speaker diarization task performance is measured via an optimum mapping between the (absolute) reference speaker IDs and the (relative) 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. It is a time-based measure, computed on non-overlapping speech segments. Small pauses of less than 0.5 seconds are not considered to be segmentation breaks. Also, a time collar of 0.25 seconds is allowed at each speaker transition in the reference in order to take into account possible errors in the reference timing due to the automatic forced-alignment.

The overall speaker diarization error includes the missed and false alarm speaker times, thus taking speech/non-speech detection

*now with Laboratoire d'Informatique de l'Universite du Maine

errors into account. Speech activity detection (SAD) is of a different nature than speaker clustering, and it is of interest to provide separate error measures for SAD and speaker clustering, when analyzing a system's performance.

Other performance measures can provide better insight into the speaker clustering stage of the system than the single, global measure provided by the speaker error time. For some experiments, we report the average frame-level cluster purity and cluster coverage measures [7]. Similar to segment-level cluster purity proposed in [6], frame-level 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 is the dual measure, and accounts for the dispersion of a given speaker's data across clusters. Both measures are complementary, and the speaker error time can be interpreted as a combination of both.

Databases

Speaker diarization was assessed on the the Broadcast News transcription task for the English language. Development and training databases were provided by NIST along with a reference labeling determined by the LDC for system development. Evaluation references were made available after the evaluation. The data were taken from US radio or TV shows.

- Development data: 6 shows of about 30 minutes each, recorded in February 2001 (sources: ABC, CNN, NBC, PRI, VOA), referred to as 'dev1', and 6 shows each of about 30 minutes, recorded in November and December 2003 (sources: ABC, CNBC, CNN, C-SPAN, PBS), referred to as 'dev2';
- training data: 23 shows lasting between 30 minutes and 2 hours, recorded between July 1997 and January 1998 (sources: ABC, CNN, C-SPAN, PRI);
- evaluation (test) data: 12 shows lasting about 30 minutes, recorded in Dec. 2003 (sources: ABC, CNBC, CNN, C-SPAN, PBS, WB17).

The training database is a subset of the LDC 1997 English Broadcast News Hub-4 corpus, completed with precise structural metadata annotations. We report some experimental results on this database; however we did not use it directly for model training. We used the entire 1996 Hub-4 (LDC97S44) and 1997 Hub-4 corpora for building several acoustic models, as detailed in the system description. The combined corpora have a total of about 150 hours of annotated audio data.

3. BASELINE PARTITIONING SYSTEM

Our baseline data partitioning system is the first stage of the system developed of the LIMSI English broadcast news transcription system [7]. It was shown to provide a high cluster purity (about 96%) and a cluster coverage slightly below 80% on 1996 and 1997 NIST evaluation data. The baseline partitioner is structured as follows (cf. Figure 1):

- Feature extraction: Mel frequency cepstral parameters are extracted from the speech signal every 10ms using a 30ms window. The 38 dimensional feature vector consists of 12 cepstrum coefficients, 12 delta and 12 delta-delta coefficients plus the delta and delta-delta energy. It is similar to the features used for speech transcription, except that the energy coefficient is discarded.
- Speech Activity Detection: 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 acoustic data extracted from the 1996/1997 Broadcast News 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 [13]. A single diagonal Gaussian is used for each window. The detection threshold was set on training data in order to provide small acoustically homogeneous segments lasting at least 0.25 seconds.
- Iterative GMM segmentation/clustering procedure: Each initial segment is used to seed one cluster, and a GMM with 8 Gaussians and a diagonal covariance matrix is trained by maximum likelihood estimation (MLE) on the segment data. Then, given a sequence of N non-overlapping segments $(s_1, \ldots, s_N)$ with their associated segment cluster labels $(c_1, \ldots, c_N)$, where $c_i \in [1, K]$ and $K \leq N$, the objective function used is a penalized log-likelihood of the form:

$$\sum_{i=1}^{N} \log f(s_i|M_{c_i}) - \alpha N - \beta K$$

  where $f(\cdot|M)$ is the likelihood given the model M, and $\alpha > 0$ and $\beta > 0$. The terms αN and βK can be seen as segment and cluster penalties. The algorithm alternates Viterbi resegmentation and GMMs reestimation steps, where $\sum_i \log f(s_i|M_{c_i}) - \alpha N$ is maximized, with GMM clustering steps, as long as the resulting log-likelihood loss per merge is less than β. The merging criterion between two GMMs is estimated as the log-likelihood loss for merging the 16 initial Gaussians of both GMMs into a final set of 8 Gaussians (cf. [7]).
- 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. This is done to locate the segment boundaries at silence portions, so as to avoid cutting words.
- Bandwidth and gender labeling: Band (studio or telephone) and gender (male or female) labeling is performed on the segments using 4 GMMs with 64 diagonal covariance matrices, trained on a subset of the 1996/1997 Broadcast News data.

4. IMPROVED SPEAKER PARTITIONING

The baseline partitioning system (c-std) results submitted in the RT-03S speaker diarization task, had an overall diarization error rate of 24.5%. Other approaches, e.g. BIC clustering methods, had a better performance on this task [14]. 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 to taking into account short pauses. The other parts of the system were kept unchanged.

BIC clustering

A hierarchical clustering is applied to the segments output by the iterative GMM segmentation. At the beginning, each segment seeds one cluster, modeled by a single Gaussian with a full covariance matrix. At each step, the two nearest clusters are merged until the

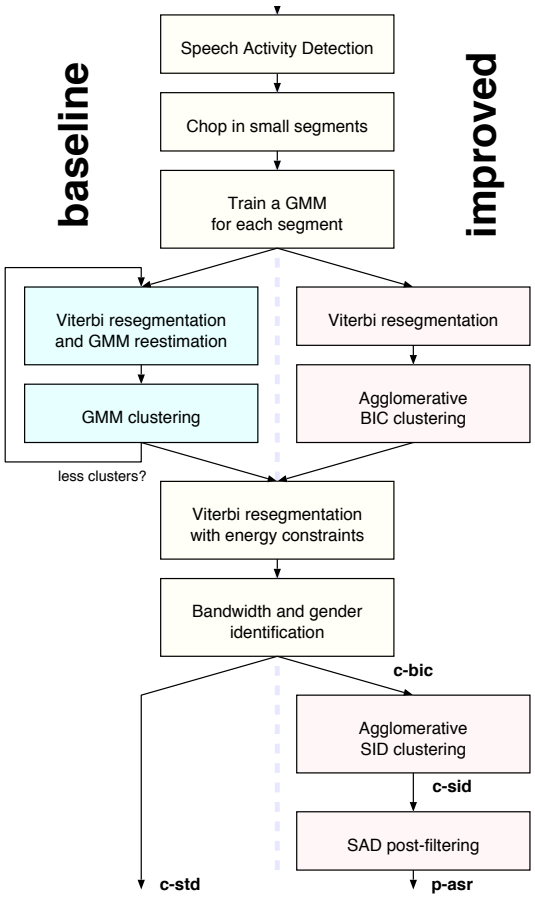

Figure 1: Standard LIMSI partitioning system (c-std on the left side of the diagram) and speaker partitioning system improved for RT-04F (p-asr to the right, along with c-bic and c-sid intermediate steps).

stop criterion is reached. The BIC criterion [6] is used both for the inter-cluster distance measure and the stop criterion. The BIC penalty weight was optimized on the dev1 and dev2 data.

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 $\Delta BIC < 0$. At each step, the two nearest clusters (i.e., those which have the most negative ΔBIC values) are merged into one cluster, and the ΔBIC value between the new cluster and all the other clusters is computed. 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 [5]. We refer to this as a local BIC penalty. But in general the size of the whole set of cluster, 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 we use the BIC criterion as the distance measure for merging the clusters, using the total size will make the penalty constant, so the choice of the two merged 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: a GMM with 8 diagonal components for the former, and a single Gaussian with full covariance matrix for the latter. After several iterations, the amount of data per cluster increases, and a more complex model can be used. Also, 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...). Acoustic background normalization is necessary to regroup the data for a given speaker.

A state-of-the-art speaker recognition methods [11, 3] were thus employed to improve the quality of the speaker clustering. The first clustering stage is tuned to provide the highest possible cluster purity, not the lowest speaker error, since wrong merges can not be canceled by a further agglomerative process. The SID clustering process is as follows:

- Front-end: The feature vectors consist of 15 Mel frequency cepstral coefficients plus delta coefficients and delta energy. Feature warping [10] is performed on each segment using a sliding window of about 3 seconds in order to reduce the effect of the acoustic environment.
- Models: 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 Broadcast News data. For each initial cluster, maximum a posteriori (MAP) adaptation [8] of the means of the matching UBM is performed.
- Clustering: Agglomerative clustering is performed separately for each gender and band condition, using a cross log-likelihood ratio as in [12]. For each cluster c_i, its model M_i is MAP adapted from the gender and channel matched UBM R using the set of feature vectors x_i belonging to the cluster. Then, given two clusters c_i and c_j, their 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 the development data sets.

system	*cluster purity (%)*	*coverage (%)*	*overall error (%)*
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

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 dev1 dataset.

BIC criterion	λ	*overall error (%)*
local	5.0	13.32%
	6.0	12.77%
	7.0	13.78%
global	5.0	16.39%
	6.0	15.46%
	7.0	18.22%

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

SAD post-filtering

The output of the LIMSI Broadcast News Speech-To-Text system is used in a post-processing stage for filtering out short-duration silence segments not detected by the initial speech detection. Only inter-word silences lasting at least 1 second are filtered out. This duration was chosen to be the sum of the minimal 0.5 sec inter-segment gap plus two collars, and its relevance was verified on development data.

5. EXPERIMENTAL 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.

Results on the development data

As expected, the standard partitioner c-std in its default configuration provides a high purity, and 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 obtains 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.

A global BIC merging and stop criterion was also tested, but always performed worse than the local BIC criterion in our experiments, as can be seen for c-bic on dev1 (cf. Table 2), thus only the local criterion was used in the remaining experiments.

Looking in more detail at the performance of c-sid system, we can see that the speech detection error rate is 1.7% for dev1 and 3.6% for dev2 (cf. Table 3). The speaker clustering error is 5.4% for dev1 and 4.1% for dev2; but this average value hides a large variation across shows, ranging from the lowest error of 0.1% for the C-SPAN show to over 12% for the ABC and NBC shows.

data set	*missed speech (%)*	*false alarm speech (%)*	*speaker error (%)*	*overall error (%)*
dev1	**0.4**	**1.3**	**5.4**	**7.1**
ABC	1.6	1.3	12.4	15.2
VOA	0.3	1.2	2.2	3.7
PRI	0.1	0.9	2.8	3.8
NBC	0.1	1.1	12.0	13.2
CNN	0.5	1.4	5.6	7.6
MNB	0.2	1.8	0.8	2.8
dev2	**0.5**	**3.1**	**4.1**	**7.6**
CSPAN	0.3	2.9	0.1	3.3
CNN	0.6	4.2	5.0	9.8
PBS	0.1	2.8	7.4	10.3
ABC	2.1	6.7	12.5	21.2
CNNHL	0.0	1.4	0.5	1.9
CNBC	0.2	1.0	0.9	2.1

Table 3: Performance of c-sid system on the dev1 and dev2 data sets.

δ	*speaker error (%)*	*overall error (%)*
0.1	12.9%	16.1
0.2	12.5%	15.7
0.3	9.0%	12.2
0.4	8.9%	12.1
0.5	7.8%	11.0
0.6	8.1%	11.3
0.7	8.6%	11.8

Table 4: Speaker diarization error on the training corpus for c-sid system, as a function of the SID clustering threshold δ.

Results on the training data

On the training corpus, the c-sid system has a much higher overall speaker diarization error: 16.1% compared to 7.1% and 7.6% on dev1 and dev2 respectively. The setting of the SID clustering threshold δ (0.1) on the development data is not optimal for the training corpus (cf. Table 4): an optimal value $\delta = 0.5$ provides a 40% relative reduction of the speaker error rate, from 12.9% to 7.8%. A possible reason is the variability observed in the duration of the shows of the training set, between 30 minutes and 2 hours. By restricting the training set to the subset of shows with a matching duration (30 minutes), the speaker error rate and overall error in the standard configuration are 9.7% and 12.3% respectively, which shows that the SID clustering threshold δ is dependent on the show duration.

Results on the evaluation data

On the evaluation test set, the trend observed on the development data was confirmed, with a slight increase to 17% overall diarization error for c-bic and 9.1% for c-sid. The final SAD post-processing stage gives an improvement of 0.6%, mainly by reducing false alarms in speech detection.

6. CONCLUSIONS

The baseline partitioning system provides a high cluster accuracy, but may split data from a single speaker into several clusters according to the background acoustic conditions. This behavior is desirable as a preprocessing stage of a speech transcription system, where unsupervised adaptation of the acoustic models is performed

system	*missed speech (%)*	*false alarm speech (%)*	*speaker error (%)*	*overall error (%)*
c-bic	0.4	1.8	14.8	17.0
c-sid	0.4	1.8	6.9	9.1
p-asr	0.6	1.1	6.8	8.5

Table 5: Performances of c-bic, c-sid and p-asr systems on the RT-04F evaluation data.

using the clustering output. The lower cluster coverage is not an issue, and has only a small impact on the quality of the transcription. On the other hand, the speaker diarization task gives equal value to cluster purity and coverage, which led us to improve upon the baseline partitioner.

We have thus explored several modifications to the baseline system. First, the iterative GMM clustering has been replaced by an agglomerative BIC clustering, using mono Gaussians with full covariance matrices. A local BIC merging and stop criterion was shown to outperform the global criterion which would be more in agreement with the theory. 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. A similar result was observed in [14]. A second clustering module has been applied to the output of the system, relying on techniques used for speaker identification and verification: acoustic channel normalization, and MAP adaptation of a reference GMM with a large number of Gaussians.

On the development data, the overall speaker diarization error was reduced from 24.8% for the best setting of the baseline system to 13.2% using BIC clustering, and to 7.1% with the additional SID clustering step. These figures include the speech/non-speech detection errors, which remain constant at about 1.7%. This corresponds to a relative speaker error time reduction of over 75%. Consistent, but somewhat higher overall error rates are observed on the evaluation data: 17% for c-bic and 9.1% for c-sid. 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 systems and models, each one focusing on a different acoustic aspect. The final post-processing filtering using the ASR output provides a further reduction of the overall error rate from 9.1% to 8.5% on the evaluation data, mainly due to a reduction of false alarm speaker time on long pauses.

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. However the speaker error does not provide a stable and continuous measure of a clustering system: splitting a speaker in two classes, which is a single decision, results in doubling of the error rate for this speaker. Only with a large amount of files can statistically consistent results be obtained. No specific optimizations were made for speed in the systems described, and for a 30 minutes show, c-bic system speed is between 0.2 and 0.3xRT, and c-sid speed is between 0.5 and 1.5xRT on a single 2.4GHz CPU. Finally, most speaker diarization systems rely on a purely acoustic segmentation and clustering. An essential part of the information in speech is of a linguistic nature, and obviously in TV and radio shows most speakers are presented and identified. Combining the acoustic with the linguistic layer as explored in [4] would clearly improve the robustness of a speaker diarization system and make it more exploitable by a human reader.

REFERENCES

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

[2] J. Ajmera, C. Wooters, "A robust speaker clustering algorithm" *IEEE ASRU. Automatic Speech Recognition and Undertanding Workshop*, Virgin Islands, Nov. 2003.

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

[4] L. Canseco-Rodriguez, L. Lamel, J.-L. Gauvain, "Speaker diarization from speech transcripts," *Proc. of the International Conference on Speech and Language Processing*, Jeju, Oct. 2004.

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

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

[7] J.L. Gauvain, L. Lamel and G. Adda, "Partitioning and Transcription of Broadcast News Data," *Proc. of the International Conference on Speech and Language Processing*, vol. 4, pp. 1335-1338, Sydney, Dec 1998.

[8] J.-L. Gauvain and C.H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, Apr. 1994.

[9] 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. of DARPA Broadcast News Transcription and Understanding Workshop*, Landsdowne, VA, Feb. 1998.

[10] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. ISCA Workshop on Speaker Recognition - 2001: A Speaker Oddyssey*, June 2001.

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

[12] 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. of the International Conference on Speech and Language Processing*, Sydney, Dec 1998.

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

[14] S. Tranter and D. Reynolds, "Speaker diarisation for broadcast news," *Proc. ISCA Odyssey 2004 Workshop on speaker and language recognition*, Toledo, June 2004.