# Progress in the AMIDA speaker diarization system for meeting data

David A. van Leeuwen and Matej Konečný

TNO Human Factors, Postbus 23, 3769 ZG Soesterberg, The Netherlands
david.vanleeuwen@tno.nl, matej.konecny@gmail.com

**Abstract.** In this paper we describe the AMIDA speaker dizarization system as it was submitted to the NIST Rich Transcription evaluation 2007 for conference room data. This is done in the context of the history of this system and other speaker diarization systems. One of the goals of our system is to have as little tunable parameters as possible, while maintaining performance. The system consists of a BIC segmentation/clustering initialization, followed by a combined re-segmentation/cluster merging algorithm. The Diarization Error Rate (DER) result of our best system is 17.0 %, accounting for overlapping speech. However, we find that a slight altering of Speech Activity Detection models has a large impact on the speaker DER.

## 1 Introduction

The AMIDA speaker diarization system is an ongoing research effort to investigate different approaches to the challenging task of speaker segmentation and clustering of meeting recordings. This year's efforts have been concentrated on the use of multiple microphones recordings and exploring different modelling and initialization approaches.

The task of speaker diarization[1] is commonly summarized as determining *who* spoke *when*, where speakers can be given arbitrary labels, i.e, no absolute identification of speakers is required. This article describes the AMIDA system and its performance in the Spring NIST Rich Transcription 2007 (RT07s) speaker diarization task. As the successor of AMI, the EU-funded project AMIDA attempts to develop tools that allow more effective meetings in so-called smart meeting rooms, which are equipped with many microphones and cameras to record the meeting process. Automatic data processing tools extract and structure information, so that meeting participants, who are not available in place or time, can still benefit from and interact with the meeting process. The AMI consortium has donated evaluation data for the NIST RT series since 2005.

This paper is organized as follows. First, a recapitulation of earlier TNO/AMI work is made, and the 2007 system is described. Then, te evaluation results are discussed, and some experiments with speaker overlap detection are described.

[1] The meaning of the term *diarization* may not be very familiar outside this community. The word is related to *diary*, indicating an annotation of events with time marks.

## 2 System history and design goals

In 2005, TNO participated in the speaker diarization task for the first time [16]. The system consisted of a speech activity detector (SAD), followed by a Bayesian Information Criterion (BIC) based segmentation [8] and clustering [7] system. We had correctly identified the importance of a good SAD as a prerequisite for acceptable speaker Diarization Error Rates (DER), and obtained low SAD error rates. However, we had underestimated the sensitivity of the optimal setting of BIC parameters $\lambda$ to the test set. Merely tuning the two $\lambda$'s from the optimal setting for development test data (RT04s) to evaluation data (RT05s) reduced DER from the evaluation results 34.2 % to a post-evaluation result of 25.4 %.

In 2006, where we participated as AMI referring to the increased effort of co-operation, we attempted to remove the dependence of these parameters $\lambda$ by keeping the number of parameters in the speaker models before and after clustering the same [2]. The influence of $\lambda$, that penalizes such a difference in number of model parameters, is then effectively removed from the BIC. By moving from a full-covariance single-Gaussian speaker model to a diagonal covariance Gaussian Mixture speaker model the system became more in-line with other current approaches [5, 10, 20], allowing Viterbi re-segmentation to fine-tune speaker change boundaries during the clustering process. These GMM-based re-segmentation systems showed less sensitivity to the evaluation collection [18], and the drop in performance seen in going from development (RT05s) to evaluation (RT06s) data of 7–12 % could partially be attributed to the RT06s data being 'harder.'

In 2007 we re-designed the system and partially re-wrote the code base. This year, for the first time, we utilized more information than just the Single Distant Microphone (SDM) in the Multiple Distant Microphone (MDM) condition, an opportunity provided by ICSI by sharing their beamforming software with the research community [3]. This allowed modeling of the beamformer's delay-parameters as well as provided for better quality cepstral features, by improving the SNR.

The goals of the 2007 AMIDA system were to have almost no tunable parameters, no assumptions on the number of speakers in the meeting, reasonable speed and better utilization of the available microphones.

## 3 System description

The general design of the AMIDA 2007 speaker diarization system is depicted in Fig. 1.

### 3.1 Signal processing

For all signal processing steps, we used tools made available by third parties. For the MDM condition we processed the data using the `BeamformIt` tool [3]. We reduced the analysis window to 32 ms and step size to 16 ms, rather than the default 500 and 250 ms, in order have the delay feature stream synchronous to

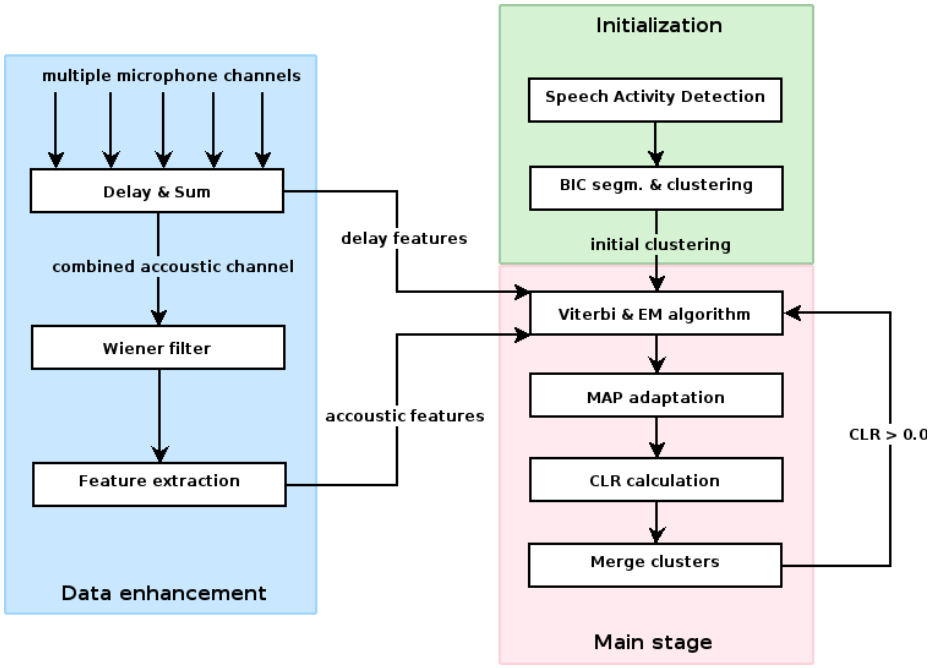


**Fig. 1.** Overview of the AMIDA speaker diarization system

the cepstral features. The single microphone signal output by the beamformer was further enhanced using a Wiener filter, by the Qualcomm-ICSI-OGI Aurora frontend tool [1].

This signal with improved SNR was used to extract 12 PLP coefficients and log energy, every 16 ms, as calculated over 32 ms windows by the ICSI implementation `rasta` [13, 12].

### 3.2 Speech Activity Detection

Only for this step, we augmented the features with 1st order derivatives estimated over 5 consecutive frames. A two-state HMM was used for SAD, with diagonal covariance GMMs (number of Gaussians $N_G = 16$) for a speech and a silence state. Because, since RT06s, the speech activity reference truth is determined from SRI's forced aligned decoding of the Individual Headset Microphones, we experimented with two sets of models for speech and silence. These are indicated in Table 1.

### 3.3 Initial segmentation and clustering

In order to speed up the clustering process we used our old BIC system [16] to perform an initial clustering. Respecting silence regions as segment boundaries,

**Table 1.** SAD models and their training data.

| Name | $N_G$ | Training data |
|---|---|---|
| AMI-dev | 16 | 10 AMI meetings distributed as development test for RT05s |
| RT-FA | 16 | RT05s and RT06s evaluation data, with force-aligned reference |

we set $\lambda = 1$ for both segmentation and clustering, which typically leads to over-segmentation and under-clustering. If the BIC system was to produce the final clustering, optimal values for segmentation would be $\lambda_s \approx 1.5$–$2$ and for clustering $\lambda_c \approx 6$–$14$. With the 'ideal' values of $\lambda_{c,s} = 1$ we ended up with typically 40 clusters after this step. The clusters found at this step were used to train initial diagonal covariance GMMs $\Lambda$ for each cluster. We used an occupancy driven approach [4, 18] for determining the number of Gaussians $N_G$,

$$N_G = \left\lfloor \frac{N_f}{R_{\mathrm{CC}}} + \frac{1}{2} \right\rfloor, \tag{1}$$

where $N_f$ is the number of frames in the cluster and $R_{\mathrm{CC}}$ is the 'cluster complexity ratio', which we set to 300. This corresponds to about 4.8 seconds/Gaussian.

The clusters are also used to train single Gaussian GMMs for the delay parameters.

### 3.4 Agglomerative clustering

This is the main step in the speaker diarization system. It consists of several smaller steps.

1. First, all silence frames are removed from the feature stream, making it appear continuous. The silence regions are recovered at the end of the clustering process. Note that the SAD for determining silence frames does not need to be the same as in the BIC segmentation/clustering step.
2. A GMM $\Theta$ with $N_G = 64$ was trained using all acoustic speech frames of the meeting, for later usage in the clustering process. In speaker recognition such a model is known as a 'Universal Background Model' (UBM), which is perhaps too grand a name for a model containing only meeting speakers.
3. Using the $N_C$ initial GMMs $\Lambda$, do a Viterbi decode using GMMs as single states in an $N_C$ parallel state topology. The new segmentation is used to train new GMMs for the clusters, changing $N_G$ according to (1) if necessary. This step is repeated several times, and both the cepstral and delay GMMs are used. We mix the per-frame log likelihoods $\log L$ for the acoustic model $\Lambda_A$ and delay model $\Lambda_D$ using linear interpolation,

$$\log L(a_t, d_t|\Lambda_A, \Lambda_D) = \alpha \log L(a_t|\Lambda_A) + (1 - \alpha) \log L(d_t|\Lambda_D), \tag{2}$$

where $a_t$ and $d_t$ are acoustic and delay parameters at frame $t$. We used a fixed value $\alpha = 0.9$.

4. Using Maximum A Posteriori adaptation [11] of the GMM in step 2 to the data of each cluster, build adapted 64-Gaussian GMMs. For each pair $(i, j)$ of these $N_C$ models $\Theta_i$, calculate the cross likelihood ratio [15]

$$R_{\mathrm{CL}}(i,j) = \frac{1}{N_i} \log \frac{L(x_i|\Theta_j)}{L(x_i|\Theta} ) + \frac{1}{N_j} \log \frac{L(x_j|\Theta_i)}{L(x_j|\Theta)}, \tag{3}$$

where $N_{i,j}$ are the number of frames contributing to the clusters $i$ and $j$, respectively. Then, determine the pair $(I, J)$ that maximizes $R_{\mathrm{CL}}$

$$(I, J) = \arg\max_{i,j} R_{\mathrm{CL}}(i,j). \tag{4}$$

If $R_{\mathrm{CL}}(I, J) > 0$, merge the data from $I$ and $J$ to a single cluster, train new GMMs $\Lambda$ and continue with step 3. Otherwise, go to the final step.
5. On finalization, insert the silence frames that were removed in step 1.

### 3.5 Differences from other approaches

All steps described here have been used elsewhere [5, 20], but with slightly different implementation details. The order of the delay-and-sum beamforming and Wiener filtering is traditionally reversed [19]. We applied the filtering after the beamforming, because we were uneasy about phase difference the Wiener filtering might introduce. We did experiment with the order reversed, but did not see a performance difference. Applying Wiener filtering to only one signal is less computationally expensive.

The current ICSI system [19] uses a linear initial clustering with a fixed number of initial clusters, and many re-segmentation iterations for initial clustering. Our BIC initial clustering does not assume a maximum number of speakers. This approach is similar to the LIMSI system [20].

The use of the UBM/GMM cross likelihood ratio is similar to the LIMSI diarization system for both Meeting [20] and Broadcast News [6] data, but we do not use a tunable threshold as stopping criterion for $R_{\mathrm{CL}}$, but rather 0.

Also note, that we use two sets of GMMs: one set (with lower $N_G$, depending on the amount of data available for the cluster) for Viterbi re-segmentation, and one set (with $N_G = 64$) for determining the cluster to merge and the stopping criterion. This is computationally expensive, but we found this to give us best results for development test data.

### 3.6 Initialization of GMMs

We experienced problems with random $k$-means initialization of the GMMs. It had a strong effect in the segmentation/clustering and led to badly reproducing results. We therefore reverted to a more deterministic estimation of the GMM parameters, starting with a single GMM and doubling Gaussians until the highest power of two below the desired $N_G$, followed by sequentially adding a single Gaussian until reaching $N_G$. In the re-segmentation process, we used existing

GMMs. When $N_G$ had to grow (according to (1)), this was carried out though adding Gaussians by splitting the Gaussian with highest variance of all feature dimensions. In reducing $N_G$, the GMM was retrained from scratch.

## 4 Results

We tabulated the results for development and evaluation data in Table 2, at RT submission time. Results are reported in the primary evaluation measure defined by NIST [9], the Diarization Error Rate (DER), evaluated including overlapping speech.

**Table 2.** Speaker Diarization Error rate (DER) for different data sets (MDM), including overlapped speech. First line is our original BIC-based system from 2005. SAD1 and SAD2 refer to speech activity detection used for initial BIC segmentation and final agglomerative clustering, respectively, see Table 1. The last column shows the SAD error for the evaluation data.

| system | SAD1 | SAD2 | RT05s | RT06s | RT07 | SAD (RT07) |
|---|---|---|---|---|---|---|
| TNO'05 | AMI-dev | AMI-dev | 21.7 % | 32.4 % | 26.2 % | 6.4 % |
| AMIDA'07 | AMI-dev | AMI-dev | 16.3 % | 18.1 % | **22.0 %** | 6.7 % |
| AMIDA'07 | AMI-dev | RT-FA | - | 20.1 % | 17.0 % | 2.9 % |
| AMIDA'07 | RT-FA | RT-FA | - | - | 18.6 % | 2.9 % |

The result of the primary system (in bold), 22 %, is not particularly good. Once again we observe that development test results (16.3 and 18.1 %, respectively) do not generalize very well to the evaluation data. We blame this partly to the unexpectedly bad SAD performance, 6.7 %, which is probably related to the fact that the SAD models used (AMI-dev) are trained on Single Distant Microphone data, without Wiener filtering, and using manually-annotated speech/non-speech labeling. For the RT07s reference, forced-aligned speech/non-speech labeling is used, and models trained on beamformed, Wiener filtered data from RT05 and RT06. This would suggest that the RT-FA SAD models should work better. Indeed, we find lower SAD error for the evaluation data, and a much more improved DER or 17.0 %.

We used RT06s data[2] primarily for development testing, but at a later stage looked at RT05s as well to check for dataset dependence. In retrospect, we should have combined all available proper meetings for development testing, since per-meeting DER tends to vary a lot. Hence, our 'wrong choice' to use AMI-dev SAD models, based on RT06s performance where AMI-dev models scored better with 18,1 % than the RT-FA models, with 20.1 %.

[2] Excluding the meeting recorded at TNO, because the wrong MDM microphones had been included in the test set.

In Table 3 we show the influence of the Delay parameter modeling on DER for RT05s and RT06s. We can clearly observe an improvement including these parameters.

**Table 3.** Influence of delay parameters on development test DER.

| Delay parameters | RT05s | RT06s |
|---|---|---|
| no | 20.5 % | 24.3 % |
| yes | 16.3 % | 18.1 % |

## 5 Overlapping speech

Most implementations so far [5, 10, 20, 18] have always interpreted the speaker diarization task as a speaker segmentation and clustering task. Although theoretically possible [18], the segmentation/clustering implementations do not consider the possibility of overlapping speech, i.e., speakers speaking simultaneously. Since the NIST Rich Transcription evaluation in Spring 2006, the primary evaluation measure accounts overlapping speech.

Now, with the new definition of speech/non-speech in the reference, using SRI's forced aligned segmentation, the amount or overlapped speech has reduced from 22 % to 8 % for RT06s, in going from manual to forced alignment. Also the average duration of overlap reduces from 1.57 s to 0.53 s. This appears to make the challenging task of determining the identity of overlapping speakers of less priority. However, in the light of ICSI's very good performance [19], overlapped speech might gain renewed attention. For RT07s, there was a difference in DER with and without accounting overlapping speech of about 3.5 %-point.

Last year we tried to generate, as a post-processing step, 'two-speaker models' by adding the probability density functions of pairs of clusters, and include these $\binom{N_C}{2}$ models in the decoding process, including restrictions for transitions to/from the two-speaker models. Then, we could not obtain better DER values, so this year we tried other approaches.

As a cheating experiment, we used the reference transcription to detect overlapping speech. Then, as a post-processing of normal diarization output, we included the 'most talkative speaker' as a second speaker in the output. This led to a reduction of the DER of 2 %-point. Although not a dramatic improvement, this simple 'guess' helps, if we know the regions of overlapping speech. In order to detect overlapping speech, we tried the following.

- Use the output of the `beamformit` tool [3]. The beamformer can give an indication if there is overlapping speech, presumably because of confusion of the location of the sound source. We determined the detection capability, in terms of False Alarm (FA) and missed time, and obtained 6.65 % and 85.6 % respectively. Having small FA time is good for the DER [16],

but the detection capability is really too low to use as input for the 'most talkative speaker guess' algorithm. Assuming equal variance of overlap and non-overlap score distributions, this corresponds to a detector with an EER of 41 %, or $d' = 0.44$ [17].

– Train an overlap/non-overlap speech detector, by training GMMs on RT06s development data, and use a Viterbi decoding to detect overlapping speech. This did not seems to give any reasonable overlapping speech detection capability either.

## 6 Discussion

Our systems that included the newly trained SAD models showed an improvement over our 'baseline' BIC speaker diarization system. Overall, the performance of our system improved a lot since last year by inclusion of the delay parameters and signal to noise ration improvement due to the beamforming and Wiener filtering. However, it is quite unsatisfactory that small changes in the application of SAD give dramatic differences (from 17.0 to 22.0 %) in DER. Another unsatisfactory fact is that our development data set (RT06s) did not indicate correctly what the right SAD models were. Possibly, the development data set was too small (and hence the DER too noisy), and also the 'algorithmic tuning' which had been carried out on the development data with the old AMI-dev SAD models will have had an effect. Even though we have been striving towards as few tunable parameters as possible, there still are the SAD models and $R_{\mathrm{CC}}$, as well algorithmic parameter choices and design decisions (number of re-alignment iterations, number of Gaussians in UBM, clustering criterion, etc.) that may overtrain on the development data.

Although the importance of proper SAD is clear from the definition of DER (SAD error is a lower bound to the DER), we have some indication that the SAD used in the clustering process should not necessarily be the same as the SAD used in producing the speaker diarization result. Indeed, this is what has been realized by the ICSI team [14, 19]. where very strict speech acceptance thresholds are applied for selecting speech frames that take part inthe speaker clustering process. Later, for the final postprocessing of speaker segment timing, the speech/silence boundaries are smoothed.

With the beamforming tools made available by ICSI we now have a chance to work seriously at the problem of overlap detection. We feel that this is an intersting task, and useful by itself: the moments of overlap can be indicative of 'hot spots,' disagreement or social cohesion, in meetings. In the context of diarization, good ovelap detection can already be helpful by simply guessing the most talkative speaker as the second speakers. Actually identifying the overlapping speakers from the acoustics and/or delay parameters will remain an even more challenging task.

## References


1. Andre Adami, Lukáš Burget, and Hynek Hermansky. Qualcomm-ICSI-OGI noise-robust front end. http://www.icsi.berkeley.edu/Speech/papers/qio/, September 2002.
2. Jitendra Ajmera, Iain McCowan, and Hervé Bourlard. Robust speaker change detection. *IEEE Signal Processing Lettres*, 11(8):649–651, 2004.
3. Xavier Anguera. BeamformIt, the fast and robust acoustic beamformer. http://www.icsi.berkeley.edu/~xanguera/BeamformIt/, 2006.
4. Xavier Anguera, Chuck Wooters, Barbara Peskin, and Mateu Aguiló. Robust speaker segmentation for meetings: The ICSI-SRI spring 2005 diarization system. In *Proc. RT'05 Meeting Recognition Evaluation Workshop*, pages 26–38, Edinburgh, July 2005.
5. Xavier Anguera, Chuck Wooters, Barbara Peskin, and Mateu Aguiló. Robust speaker segmentation for meetings: The ICSI-SRI spring 2005 diarization system. *Lecture Notes in Computer Science*, pages 402–414, 2006.
6. Claude Barras, Xuan Zhu, and Sylvain Meignier. Multistage speaker diarization of broadcast news. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5):1505–1512, September 2006.
7. Scott Shaobing Chen and P. S. Gopalakrishnan. Clustering via the Baysian Information Criterion with applications in speech recognition. In *Proc. ICASSP*, 1998.
8. Scott Shaobing Chen and P. S. Gopalakrishnan. Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion. In *Proceedings of the Darpa Broadcast News Transcription and Understanding Workshop*, 1998.
9. Jonathan Fiscus, Nicolas Radde, John S. Garofolo, Audrey Le, Jerome Ajot, and Christophe Laprun. The rich transcription 2005 spring meeting recognition evaluation. In *Machine Learning for Multimodal Interaction*, Lecture Notes in Computer Science, pages 369–389. Springer, 2005.
10. Corinne Fredouille and Grégory Senay. Technical improvements of the e-hmm based speaker diarization system for meeting records. In *Machine Learning for Multimodal Interaction*, volume 4299 of *Lecture Notes in Computer Science*, pages 359–370. Springer Berlin/Heidelberg, 2006.
11. J.-L. Gauvain and C.-H. Lee. Maximum a posteriori esitimation for multivariate gaussian mixture observations of markov chains. *IEEE Trans. Speech Audio Processing*, 2:291–298, 1994.
12. H. Hermansky and N. Morgan. Rasta processing of speech. *IEEE Transactions on Speech and Audio Processing, special issue on Robust Speech Recognition*, 2(4):578–589, 1994.
13. Hynek Hermansky. Perceptual linear predictive (plp) analysis of speech. *JASA*, 87(4):1738–1752, 1990.
14. Marijn Huijbregts, Chuck Wooters, and Roeland Ordelman. Filtering the unknown: Speech activity detection in heterogeneous video collections. In *Proc. Eurospeech*, Antwerpen, 2007. Accepted for publication.
15. D. A. Reynolds, E. Singer, and B. A. Carlson. Blind clustering of speech utterances based on speaker and language characteristics. In *Proceedings of International Conference Spoken Language Processing (ICSLP 98)*, pages 3193–3196, November 1998.

16. David A. van Leeuwen. The TNO speaker diarization system for NIST rich transcription evaluation 2005 for meeting data. In *Machine Learning for Multimodal Interaction*, volume 3869 of *Lecture Notes in Computer Science*, pages 400–449. Springer Berlin/Heidelberg, 2006.
17. David A. van Leeuwen and Niko Brümmer. An introduction to application-independent evaluation of speaker recognition systems. In Christian Müller, editor, *Speaker Classification*, volume 4343 of *Lecture Notes in Computer Science / Artificial Intelligence*. Springer, Heidelberg - New York - Berlin, 2007.
18. David A. van Leeuwen and Marijn Huijbregts. The AMI speaker diarization system for NIST RT06s meeting data. In *Machine Learning for Multimodal Interaction*, volume 4299 of *Lecture Notes in Computer Science*, pages 371–384. Springer Berlin/Heidelberg, 2006.
19. Chuck Wooters and Marijn Huijbregts. The ICSI RT07s speaker diarization system. In *Machine Learning for Multimodal Interaction*, Lecture Notes in Computer Science. Springer Berlin/Heidelberg, 2007.
20. Xuan Zhu, Claude Barras, Lori Lamel, and Jean-Luc Gauvain. Speaker diarization: From broadcast news to lectures. In *Machine Learning for Multimodal Interaction*, volume 4299 of *Lecture Notes in Computer Science*, pages 396–406. Springer Berlin/Heidelberg, 2006.