IDIAP RESEARCH REPORT


# TUNING-ROBUST INITIALIZATION METHODS FOR SPEAKER DIARIZATION

David Imseng  Gerald Friedland

Idiap-RR-35-2010

OCTOBER 2010

**Centre du Parc**, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11  F +41 27 721 77 12  info@idiap.ch  www.idiap.ch

# Tuning-Robust Initialization Methods for Speaker Diarization

David Imseng, Gerald Friedland

**Abstract**

This paper investigates a typical speaker diarization system regarding its robustness against initialization parameter variation and presents a method to reduce manual tuning of these values significantly. The behavior of an agglomerative hierarchical clustering system is studied to determine which initialization parameters impact accuracy most. We show that the accuracy of typical systems is indeed very sensitive to the values chosen for the initialization parameters and factors such as the duration of speech in the recording. We then present a solution that reduces the sensitivity of the initialization values and therefore reduces the need for manual tuning significantly while at the same time increasing the accuracy of the system. For short meetings extracted from the previous (2006, 2007 and 2009) National Institute of Standards and Technology (NIST) Rich Transcription (RT) evaluation data, the decrease of the Diarization Error Rate is up to 50 % relative. The approach consists of a novel initialization parameter estimation method for speaker diarization that uses agglomerative clustering with Bayesian Information Criterion (BIC) and Gaussian Mixture Models (GMMs) of frame-based cepstral features (MFCCs). The estimation method balances the relationship between the optimal value of the seconds of speech data per Gaussian and the duration of the speech data and is combined with a novel non-uniform initialization method. This approach results in a system that performs better than the current ICSI baseline engine on datasets of the NIST RT evaluations of the years 2006, 2007 and 2009.

**Index Terms**

Gaussian mixture models (GMM), long-term acoustic features, machine learning, speaker diarization

## I. Introduction

THE goal of speaker diarization is to segment audio unsupervisedly into speaker-homogeneous regions trying to answer the question "who spoke when?". Knowing when each speaker is speaking in a meeting recording is useful as a pre-processing step for many tasks, such as speaker-attributed speech-to-text (vocal tract length normalization and/or speaker model adaptation) [1] or content indexing and retrieval. The task has been evaluated in the NIST RT evaluation for several years now and most state-of-the-art systems[1] use a combination of agglomerative hierarchical clustering (AHC) with Bayesian Information Criterion (BIC) [2] and Gaussian Mixture Models (GMMs) of frame-based cepstral features (MFCCs) [3]. While these approaches seem to currently dominate for their accuracy and speed, most, if not all of them, ultimately require a manual tuning of the initialization parameters such as initial the number of initial clusters. It is often claimed that these parameters (see for example [4]) are not very sensitive, i.e. small changes in the value of the parameters do not cause large changes in the system behavior. It is almost an "open secret" in the speaker diarization community, however, that this is generally not true: the behavior of many such algorithms can be unpredictable, even under small variations of the initialization parameter values [5]. By presenting a set of experiments on NIST meeting data, this article discusses the behavior of AHC under initialization parameter variation. We start with a discussion on which parameters are the most sensitive and what factors influence their optimal values. We realize that these parameters are inherently dependent on the amount of speech processed by the system – a fact easily overlooked when investigating length-standardized NIST benchmark data. Therefore, we also present experiments on randomly split NIST meeting data and show that there is a rather simple relation between the speech duration and the optimal values of the initial parameters. Based on these observations, we finally present a tuning-less speaker diarization approach in the form of an interpolation model on the initialization parameters in combination with a novel initialization method. A non-uniform initial segmentation, based on long-term acoustic features and a completely unsupervised parameter estimation significantly improve the system on all tested meeting lengths and generalizes to other meeting sets. The automatic estimation of the tunable initialization parameters makes the presented approach more robust and more accurate than current state-of-the-art speaker diarization engines. The approach was submitted and evaluated in the NIST RT Evaluation 2009. The article is organized as follows: Section II provides a quick overview of diarization research with a focus on existing tunable parameters and Section III presents the baseline system used for the experiments presented in this article. Section IV then presents an analysis of the tuneable parameters of the baseline system. Section V presents our solution supported by corresponding results in Section VI before Section VII concludes the article and presents future work.

This work was sponsored by the Swiss NSF through the National Center of Competence in Research (NCCR) on "Interactive Multimodal Information Management" (IM2) and the European Integrated Project on "Augmented Multiparty Interaction with Distance Access" (AMIDA).

D. Imseng is with Idiap Research Institute, P.O. Box 592, CH-1920 Martigny, Switzerland and Ecole Polytechnique Fédérale, Lausanne (EPFL), Switzerland e-mail: david.imseng@idiap.ch

G. Friedland is with International Computer Science Institute, 1947 Center Street, Suite 600, Berkeley, CA, 94704, USA e-mail: fractor@icsi.berkeley.edu

[1]Please also compare with the NIST RT evaluation of the past years, for example 2006, 2007 and 2009.

## II. Related Work

As previously explained, the goal of speaker diarization is answering the question "who spoke when?". While for the related task of speaker recognition, models are trained for a specific set of target speakers which are applied to an unknown test speaker for acceptance (target and test speaker match) or rejection (mismatch), in speaker diarization there is no prior information about the identity or number of the speakers in the recording. Conceptually, a speaker diarization system therefore performs three tasks: discriminate between speech and non-speech regions (speech activity detection), detect speaker changes to segment the audio data (segmentation) and group the segmented regions together into speaker-homogeneous clusters (clustering). The output consists of meta-data describing speech segments in terms of starting time, ending time, and speaker cluster name. For this study we disregard parameter tuning for speech/non-speech detection as this is usually seen as a separate task.

An analysis of current approaches to speaker diarization shows that the vast majority makes use of tunable parameters which may be sensitive to minor changes. That issue was also discussed in [5], where a study on audio files exhibiting hyper-sensitivity to tuning parameters is presented. It is reported, that factors such as the number of speakers and the number of turns affect the performance most. However, that study was done on bradcast news data and the result might not generalize to meeting data.

Most state-of-the-art speaker diarization systems, including the ICSI speaker diarization engine (see Section III) combine the segmentation and clustering steps into a single step. A very popular method of doing so is the combination of AHC with BIC and GMMs of frame-based cepstral features, as done for example in [6], [7], [8], or [9]. AHC starts with a certain number of initial clusters, each represented by a GMM that models the associated feature vectors with a certain number of Gaussians. Then, based on BIC, at each iteration step, two clusters are merged until a stopping criterion is met. In [7], where the BIC based clustering is combined with Speaker Identification (SID) techniques, the clustering makes use of a manually tuned penalty term and the SID technique depends on a threshold, optimized on a development set. The system described in [8] was manually tuned to tend to over-segment and under-cluster. That approach uses two sets of GMMs: one with a flexible number of Gaussians per cluster (fixed amount of data per Gaussian) for the Viterbi segmentation and one with a fixed number of Gaussians per cluster for determining the clusters to merge and the stopping criterion. The second set of GMMs with a fixed number of Gaussians is necessary to get rid of the penalty term that appears in the BIC comparison [1], [2]. Even when many different parameters are present, unfortunately, very few articles actually discuss parameter sensitivity and tuning. When discussed, the two important parameters seem to be the number of Gaussians used to model the data and the number of initial clusters. Most relevant for the work presented here is the discussion presented in [8] where a system is carefully designed around the notion that speech is best represented when $4.8$ seconds of speech data per Gaussian are used to train the system. In [8], the notion of "seconds per Gaussian", which is claimed to be constant, is introduced. Unfortunately, the authors do not provide empirical evidence for the claim. In [10], the notion of a "Cluster Complexity Ratio" is presented. While the idea is very similar to the one in [8], very little experimental evidence was provided and in further experiments the implementation of the method did not prove robust enough to withstand NISTs evaluation tasks.

This work focuses on parameter tuning for AHC approaches, but other approaches not using AHC make also use of manually tuned parameters: in [11] for instance, where Direction of Arrival (DOA) estimates are combined with acoustic feature information, the DOA estimation uses an adjusted window length that determines the maximal possible microphone pair separation. The approach described in [12] is based on evolutive Hidden Markov Models (HMM) and uses a tuned threshold during the speaker turn detection, heuristic rules and a fixed number of Gaussians for a world model. In [13], it is shown that the main optimization criterion of the approach using the Information Bottleneck (IB) principle also requires manual parameter tuning and the model selection criterion of the IB approach makes use of a manually optimized threshold.

## III. ICSI speaker diarization Engine

We believe that the experiments in this article could be repeated with most GMM-based AHC approaches for speaker diarization and would yield similar results. For practicality, we performed our experiments using the ICSI speaker diarization engine, as illustrated in Fig. 1 and described as follows. At a high-level, the engine extracts MFCC features from a given audio track, discriminates between speech and nonspeech regions (speech activity detection), and uses an agglomerative clustering approach to perform both segmentation of the audio track into speaker-homogeneous time segments and the grouping of these segments into speaker-homogeneous clusters in one step. 19th-order MFCC features are extracted from the audio with a frame size of 30 ms and a step size of 10 ms. Speech activity regions are determined using a state-of-the-art speech/non-speech detector [9]. The detector performs iterative training and re-segmentation of the audio into three classes: speech, silence, and audible nonspeech. To bootstrap the process, an initial segmentation is created with an HMM trained on broadcast news data. The non-speech regions are then excluded from the agglomerative clustering, which is explained in the following paragraph.

The algorithm is initialized using $k$ clusters, where $k$ is larger than the number of speakers that are assumed to appear in the recording. Every cluster is modeled with a Gaussian Mixture Model containing $g$ Gaussians. Our rule of thumb prior to performing the experiments presented in this article was, that during NIST evaluations, we found empirically that for a 30-min broadcast news snippet $k = 64$ and for 10-min meetings $k = 16$ are good choices. For the number of Gaussians per initial cluster, $g = 5$ turned out to be a good choice.

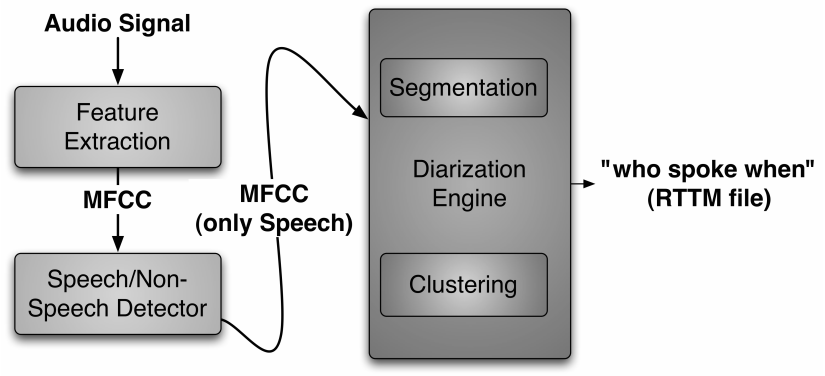


Fig. 1. The baseline speaker diarization Engine as described in Section III.

In order to train initial GMMs for the $k$ speaker clusters an initial segmentation is generated by uniformly partitioning the audio into $k$ segments of the same length. The algorithm then performs the following iterations:

- Re-Segmentation: run Viterbi alignment to find the optimal path of frames and models. As classifications based on 10 ms frames are very noisy, a minimum duration of 2.5 seconds is assumed for each speech segment.
- Re-Training: given the new segmentation of the audio track, compute new GMMs for each of the clusters.
- Cluster Merging: after several iterations (5 by default) of re-segmentation and re-training, given the new GMMs, try to find the two clusters that most likely represent the same speaker. This is done by computing the log-likelihood of each of the clusters (modeled with $g_n$ Gaussians) and the log-likelihood of a new GMM trained on the merged segments of two clusters (modeled with $g' = g_{n1} + g_{n2}$ Gaussians, $g_{n1}$, $g_{n2}$ being the number of Gaussians of the two individual clusters). If the log-likelihood of the merged GMM is larger than or equal to the sum of the individual log-likelihoods, the two models are merged and the algorithm continues at the re-segmentation step using the merged GMM. If no pair is found, the algorithm stops.

A more detailed description can be found in [1], [4]. As a result of different optimization approaches [14], our current implementation runs at about $0.6\times$realtime. This means that for 10 minutes of audio data, diarization finishes in roughly 6 minutes. Of course, other factors including CPU, memory, number of speakers in the meeting, speech/non-speech ratio and number of speaker turns affect the actual execution time.

The output of a speaker diarization engine is usually evaluated against manually annotated ground truth segments, refined by forced alignment techniques. A dynamic programming procedure is used to find the optimal one-to-one mapping between the hypothesis and the ground truth segments so that the total overlap between the reference speaker and the corresponding mapped hypothesized speaker cluster is maximized. The difference is expressed as Diarization Error Rate (DER), which is defined by NIST[2]. The DER can be decomposed into three components: misses (speaker in reference, but not in hypothesis), false alarms (speaker in hypothesis, but not in reference), and Speaker Errors (SE) (mapped reference is not the same as hypothesized speaker). The ICSI speaker diarization system has competed in the NIST evaluations of the past several years and established itself well among state-of-the-art systems[3].

## IV. Initialization Parameter Sensitivity Analysis

We started to investigate initialization parameter sensitivity by confirming anecdotal knowledge and reported experiences in the speaker diarization community. From that we know, that the initial amount of clusters $k$ and the initial number of Gaussians per cluster $g$ are very sensitive to small changes. Also, while it was shown in the past (for example in [15]) that speaker models can be successfully trained on about 50 seconds of speech per speaker for on-line diarization, we had anecdotally observed that agglomerative hierarchical clustering methods do not behave very well on short meetings. By analyzing the behavior of the engine under parameter and meeting length variation, we therefore expect to acquire knowledge about the sensitivity of the engine to different parameters. This section presents the most important parameters of the baseline engine and highlights the sensitivity of these parameters when the recording duration is varied.

In order to systematically study the phenomenon on meeting data, we randomly split the recordings of the NIST RT-06 development set (2.05 hours of data, see Table V on page 9) into smaller pieces of different durations. The recordings were cut into 50, 100, 150, 200, 250, 300, 400, and 500 second segments and also processed uncut. The total durations of the meetings in this dataset are between 600 and 700 seconds. The diarization engine was then run on these recording segments with the number of initial clusters $k = 16$ and the number of initial Gaussians per cluster $g = 5$ (see Section III) and evaluated against

[2]http://www.itl.nist.gov/iad/mig/tests/rt/2009/index.html

[3]NIST rules prohibit publication of results other than our own. Please refer to the NIST website for further information: http://www.itl.nist.gov/iad/mig/tests/rt/2007/index.html

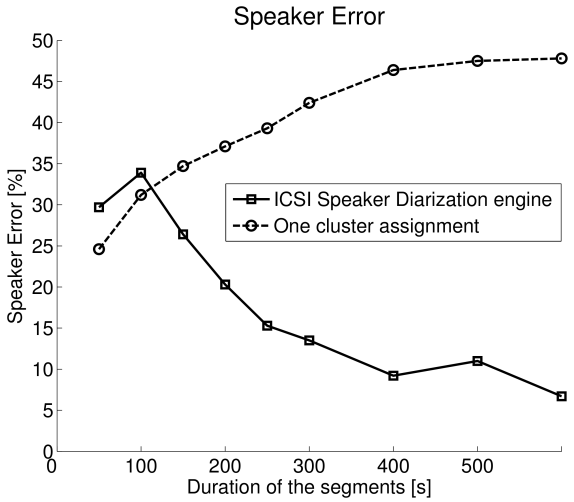


Fig. 2. The performance of the baseline engine on short recordings. For segments of 100 seconds and less, assigning a single speaker to all frames performs best. This underlines the very poor performance of agglomerative hierarchical clustering using fixed initialization parameters for short meetings.

the ground truth. This system is referred to as the baseline system ($k = 16$ and $g = 5$). The runtime is heavily dependent on the length of the audio file. Therefore splitting up the meetings and concentrating on the shorter duration recordings first, also allowed us to perform more experiments.

The accuracy of the algorithm is highly correlated with the length of the recorded audio file. Fig. 2 illustrates the issue. The speech activity detector works on-line, therefore the speech/non-speech error is almost constant even for shorter segments but the SE is clearly growing as the durations of the meeting segments become shorter. At first, it seems surprising that the SE gets smaller for segments of less than 100 seconds. This is due to the fact that in meetings shorter than 100 seconds, assigning all speech regions to one speaker starts to become a better heuristic than AHC with the wrong initialization parameters (as will be shown later in this article).

In order to find out if and which initialization parameters are responsible for the poor behavior of the engine on short meetings, we tested the behavior of four different parameters in the engine on the 100-second-segments. The four parameters were: the number of re-segmentation and re-training iterations (see Section III, default: 5 iterations), the minimum duration for a speech region (default: 2.5 seconds, as explained in Section III), the number of initial clusters $k$ (default: $k = 16$), and the number of Gaussians per initial cluster $g$ (default: $g = 5$). The results are plotted in Fig. 3. In each subfigure the same data is presented, each boxplot (for information about boxplots, see [16]) shows the SE when one parameter value is varied. We observe, that small changes in the number of re-segmentation and re-training iterations and the minimal duration do not have as much influence on the SE as the changes in the number of Gaussians per initial cluster ($g$) and the amount of initial clusters ($k$). High sensitivity can also be observed, when varying $k$ and $g$ while processing full-length meetings, as illustrated in Table I.

## V. Automatic Initial Parameter Estimation

### A. Linear Regression on the Number of Initial Clusters k

We have seen in Section IV, that the behavior of the speaker diarization engine seems to be sensitive to variations in $k$ and $g$. As already mentioned in Section II, in [8] the notion of seconds per Gaussian was introduced as the amount of speech available to train one single Gaussian in a GMM. It is measured by dividing the seconds of speech available by the total number of Gaussians in all of the GMM clusters in the meeting recording: $secpergauss = \frac{\text{speech duration in seconds}}{g \cdot k}$. In other words, seconds

| **Parameter Variation** | | | | | |
|---|---|---|---|---|---|
| Initial clusters $k$ ($g = 5$) | 14 | 15 | 16 | 17 | 18 |
| Rel. performance change | +41% | +10% | base | -13% | -9% |
| Gaussians $g$ ($k = 16$) | 3 | 4 | 5 | 6 | 7 |
| Rel. performance change | +29% | +16% | base | +42% | +29% |

TABLE I
SENSITIVITY OF THE PARAMETERS FOR THE UNCUT NIST RT-06 DEVELOPMENT SET: DURING ONE SET OF EXPERIMENTS $k$ IS VARIED AND $g = 5$ (SEE SECTION III), DURING ANOTHER SET OF EXPERIMENTS $g$ IS VARIED AND $k = 16$ (SEE SECTION III). EVEN SMALL VARIATIONS AFFECT THE PERFORMACE SIGNIFICANTLY.

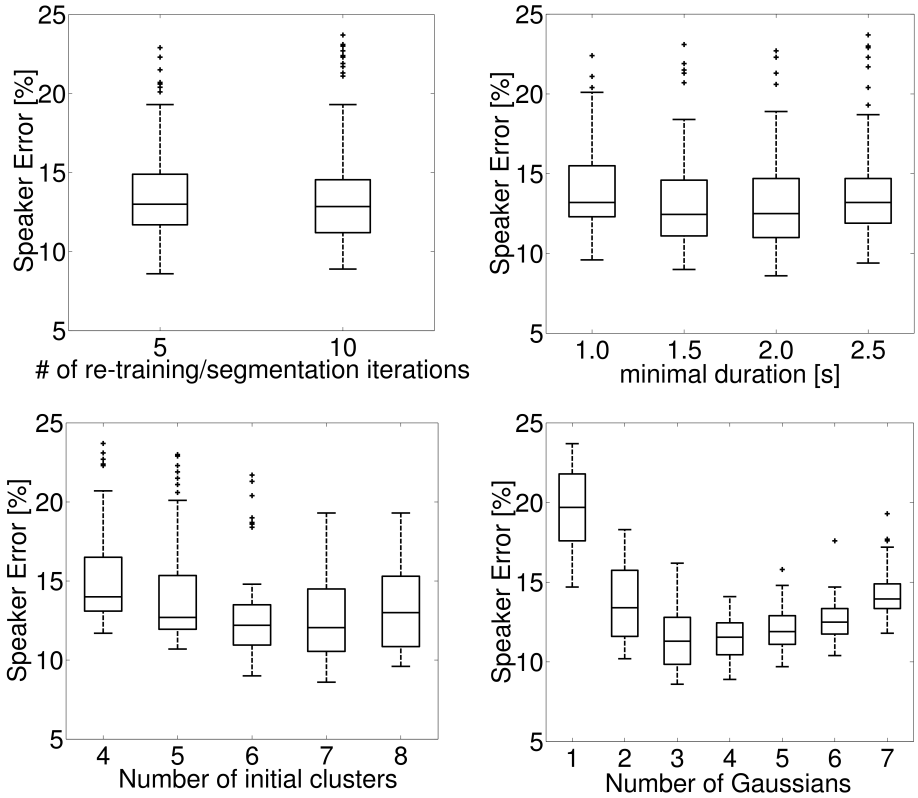


Fig. 3. Boxplots of the performance of the ICSI speaker diarization engine for 100-second-segments. One observes that the variations of the number of initial clusters and the number of Gaussians per initial cluster have most influence on the Speaker Error.

| Dur. | $k$ | $g$ | $k \cdot g$ | $secpergauss$ | # configs |
|---|---|---|---|---|---|
| 100 | [4;16] | [2;7] | [8;112] | [0.7;9.4] | 78 |
| 150 | [4;16] | [2;7] | [8;112] | [1.0;14.1] | 78 |
| 200 | [4;25] | [2;7] | [8;112] | [1.4;19.1]* | 88 |
| 250 | [4;25] | [2;7] | [8;112] | [1.7;24.2]* | 92 |
| 300 | [4;25] | [2;7] | [8;112] | [2.1;28.8]* | 99 |
| 400 | [6;25] | [2;17] | [32;102] | [3.0;9.6] | 110 |
| 500 | [6;25] | [2;21] | [32;128] | [3.0;9.8] | 139 |
| Total number of configurations | | | | | 684 |
| Total hours of experiments | | | | | 1402 |

TABLE II

TEST INTERVALS OF THE PARAMETERS $k$ AND $g$ FOR THE EXHAUSTIVE SEARCH EXPERIMENTS. THE INTERVALS REPRESENT MINIMAL AND MAXIMAL VALUES OF THE CORRESPONDING PARAMETERS (THE INTERVALS WITH STARS HAVE A LOWER BOUND OF 2.9 FOR $k > 16$, BECAUSE OF LIMITED RESSOURCES). IN TOTAL MORE THAN 1400 HOURS OF MEETING DATA WAS DIARIZED.

per Gaussian is a combination of the two parameters $k$ and $g$. The authors claim that the seconds per Gaussian is a constant of value 4.8. To automate the process of parameter value selection, we ran the following experiments.

We performed an exhaustive parameter search on recording lengths of $\{100, 150, 200, 250, 300, 400, 500\}$ seconds. In total, we diarized more than 1400 hours of meeting data. The analyzed parameter ranges are summarized in Table II, where the effective count of tested configurations is also shown. We reduced the search space of $g$ for $k > 16$, because of limited ressources and raised the lower limit of the intervals marked by a star (in Table II) to 2.9 seconds per Gaussian. By limiting the search interval of $k \cdot g$, we constrain our approach to find the best configuration in that interval only.

| Duration<br>Speech duration | 100<br>74.86 | 150<br>112.89 | 200<br>152.49 | 250<br>190.84 | 300<br>230.52 | 400<br>308.12 | 500<br>384.14 | Correlation with<br>Speech duration |
|---|---|---|---|---|---|---|---|---|
| $k \cdot g$ | 26 | 32 | 36 | 45 | 45 | 45 | 48 | 0.88 |
| $secpergauss$ | 2.88 | 3.52 | 4.23 | 4.24 | 5.12 | 6.84 | 8.00 | 0.99 |

TABLE III

THE OPTIMAL VALUES FOR $k \cdot g$ AND $secpergauss$ ON THE WHOLE DATASET INCLUDING THE CORRELATION WITH THE SPEECH DURATION.

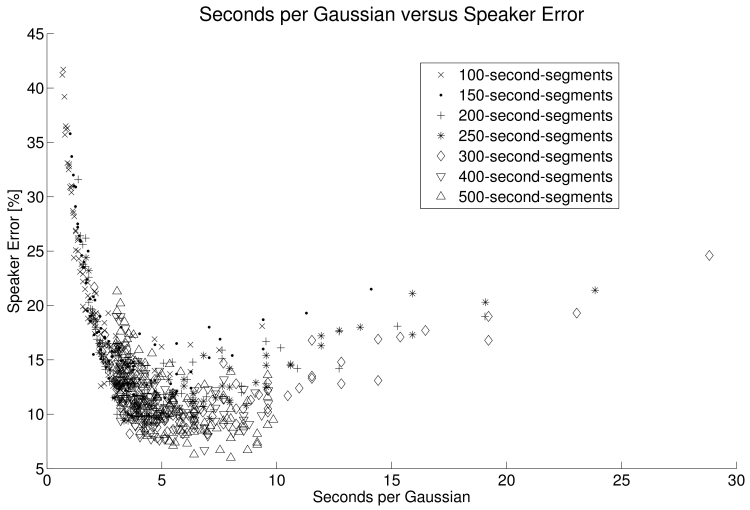


Fig. 4. Speaker Error versus seconds per Gaussian. Each data point corresponds to the average SE of 12 meetings (2.05 hours of data) for one particular configuration. Configurations for all tested segment durations are shown in the same plot. One can recognize a combination of curves, the minimum seems to be similar for different recording durations.

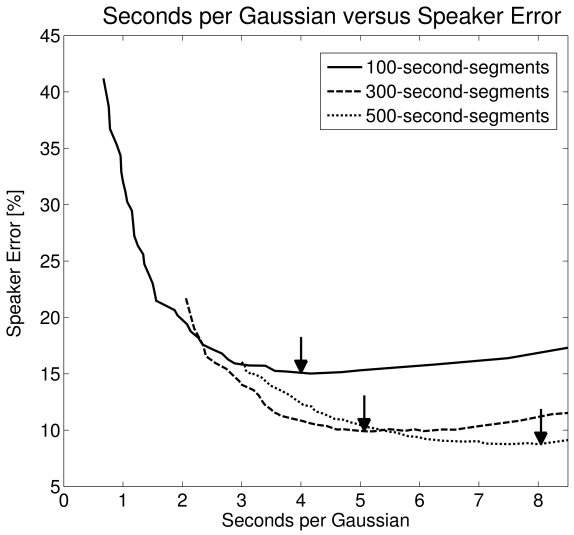


Fig. 5. Speaker Error versus seconds per Gaussian. For three different segment durations, the data points from Fig. 4 are drawn as a smoothed curve. The arrows mark the minimal Speaker Errors for different segment durations. For longer segments, the arrow moves down and to the right.

Fig. 4 presents the results plotted as the number of seconds per Gaussian vs the Speaker Error. For the purpose of visualization, Fig. 5 shows a smoothed curve (because of noisy data, the smoothed minima might not exactly match the data reported in Table III) of the data points of three different segment durations. Two major observation can be made:

1) By tuning the seconds per Gaussian parameter, it is possible to obtain a low Speaker Error even on short meetings.
2) It can be observed that the optimal amount of speech per Gaussian used for the training procedure seems to roughly follow a curve that has a global minimum.

In Table III, the best parameter configurations per segment duration (the same configuration for the whole dataset) are displayed and the correlation between the best parameter choices and the speech durations is shown. The parameter $secpergauss$ has a correlation of 0.99 what confirms the visual observations. The correlation is based on only seven data points, thus this value might be too high, but we believe that there is a significant correlation.

Then, among all tested parameter configurations, the best performing ones per segment duration were picked (on a per meeting basis, not on the total dataset, thus $7 \cdot 12 = 84$ configurations). We calculated the correlation between the duration of every processed segment versus the corresponding seconds per Gaussian. The correlation value for the speech duration of the

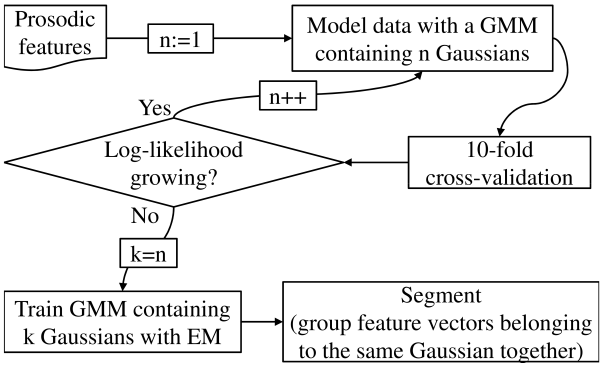


Fig. 6. A schematic view of the algorithm to estimate $k$ and perform a non-uniform initialization.

segments versus the Gaussians per second is $0.68$. Given the definition of the parameter $secpergauss = \frac{\text{speech duration in seconds}}{g \cdot k}$ and knowing the speech duration after the speech activity detection, we are able to use linear regression as an automatic parameter selection mechanism that depends on the speech duration of a recording. For that purpose we calculate the least-square linear regression over the best performing configurations and use the resulting Equation (1) afterwards to estimate the optimal amount of speech per Gaussian. One problem that remains, however, is that we are actually in need of estimating two parameters. As a start, we decided to fix one parameter, namely $g$ and then adjust $k$ because we observed that the optimal $k$ varies more and correlates better with the speech duration than the optimal $g$ (see [17, Ch. 5.2 and Ch. 5.3]). This system is summarized in Equation 1 to 3.

$$secpergauss = 0.01 \cdot \text{speech in seconds} + 2.6 \quad (1)$$

$$g = 4 \quad (2)$$

$$k = \frac{\text{speech in seconds}}{secpergauss \cdot g} \quad (3)$$

We found that fixing $g = 4$ and then using the linear regression to estimate $k$ using (3), results in relative improvements of up to 50 % for very short meeting segments (100 seconds) while maintaining the performance of the system for long recordings (600-700 seconds). We were also able to apply the same linear regression formula successfully to other data sets with relative improvements in the same range [18].

### *B. New initialization for the Number of Initial Clusters $k$*

We have seen that the number of seconds of data available per Gaussian for training, $secpergauss$ can be estimated based on a linear regression depending on the duration of speech in a meeting. However, $secpergauss$ is a combination of the two initialization parameters $k$ and $g$. Anecdotal evidence suggests that $k$ is more related to the number of different speaker in the meeting, whereas $g$ is more related to the total amount of available speech. Therefore, instead of fixing $g$ and using the linear regression to estimate $k$, another method to estimate the number of initial clusters, based on features with good speaker discriminability, is presented in this section (see Fig. 6). Having an estimate for $k$, the linear regression can then be used to determine $g$. The presented method estimates the number of initial clusters and also provides a non-uniform initialization for the AHC procedure based on the long-term feature study and ranking presented in [19], where 70 different suprasegmental features have been studied according to their speaker discriminability. Derived from the ranking in [19], the 12 top-ranked features (listed in Table IV) are extracted on all the speech regions in the recording. Some features such as mean or standard deviation have statistical character. In our configuration, Praat calculates 100 pitch values and 80 formant values per second (see [20] for more information about Praat). The features can be extracted on the segments found by the speech/non-speech detector, what may result in very few feature vectors for the clustering because the segments are relatively large compared to typical window size choices. In [19] for example, a 500-ms Hamming window with overlap is used to extract the features. Because of the statistical nature of some feature, the calculations are more accurate if there are more samples (the window is longer), but for the estimation of the number of initial clusters $k$ and the clustering itself, a certain amount of feature vectors is needed to result in a good estimation of $k$ and a reasonable non-uniform initialization. In the experiments for this article, the Hamming windowing function is used and a minimum window size of 1000 ms is chosen. A minimal window size of 1000 ms is defined as follows: Every segment (output of the speech/non-speech detector) of less than 2000 ms is untouched and the larger ones are split into segments of at least 1000 ms (effective window length $w \in [1000, 2000[$ for a minimal window size of 1000 ms). Fortunately, the minimal window size is not a very sensitive initialization parameter because even if the initial segmentation and $k$ varies, we can still interpolate $g$ accordingly. Section VI (Fig. 9 on page 11) provides further

| Category | Feature ID | Short description |
|---|---|---|
| pitch | f0_median | median of the pitch |
| pitch | f0_min | minimum of the pitch |
| pitch | f0_mean_curve | mean of the pitch tier (a time-stamped pitch contour) |
| formants | f4_stddev | standard deviation of the 4th formant |
| formants | f4_min | minimum of the 4th formant |
| formants | f4_mean | mean of the 4th formant |
| formants | f5_stddev | standard deviation of the 5th formant |
| formants | f5_min | minimum of the 5th formant |
| formants | f5_mean | mean of the 5th formant |
| harmonics | harm_mean | mean of the harmonics-to-noise ratio |
| formant | form_disp_mean | mean of the formant dispersion |
| pitch | pp_period_mean | mean of the pointprocess of the periodicity contour |

TABLE IV
THESE 12 LONG-TERM ACOUSTIC FEATURES HAVE A GOOD SPEAKER DISCRIMINATE ABILITY ACCORDING TO THE RANKING METHOD PROPOSED IN [19]. THE FEATURES ARE EXTRACTED WITH THE HELP OF PRAATLIB, A LIBRARY THAT IS USING PRAAT [20], ON ALL THE SPEECH REGIONS OF THE RECORDINGS AND AFTERWARDS USED TO ESTIMATE THE NUMBER OF INITIAL CLUSTERS TO PERFORM THE AGGLOMERATIVE CLUSTERING. FOR MORE INFORMATION ABOUT THE FEATURES REFER TO THE DOCUMENTATION OF PRAAT.

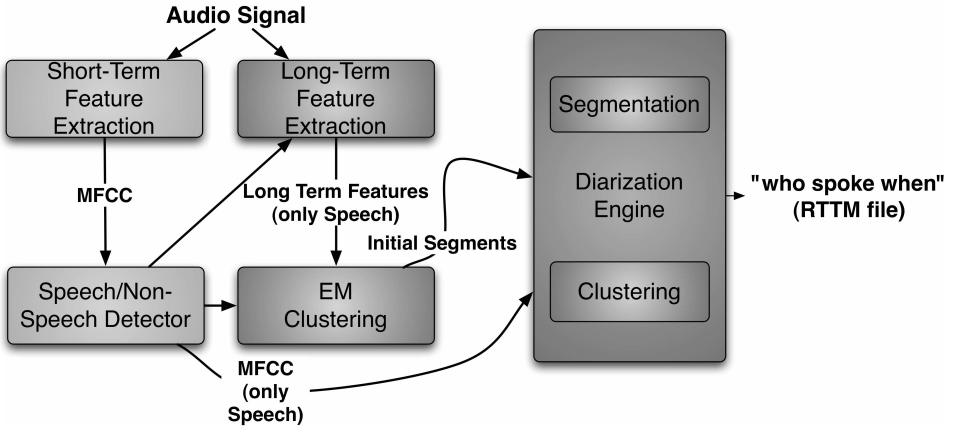


Fig. 7. The modified speaker diarization Engine as described in Section V-C.

details. The 12-dimensional feature vectors are then clustered with the help of one GMM with diagonal covariance. As this clustering serves only as initialization for an agglomerative clustering algorithm, it is desired that the model selection tends to over-estimate the number of initial clusters. The agglomerative clustering algorithm will merge redundant clusters whereas it is not able to split clusters. To determine the number of Gaussians, a 10-fold cross-validation (see [21, p. 150]) is used to calculate the log-likelihood of GMMs with different number of Gaussians. Then, expectation maximization (see [22, p. 65]) is used to train the GMM (consisting of the previously determined number of Gaussians) on all the feature vectors. Finally, every feature vector is assigned to one of the Gaussians in the GMM. We can group all the feature vectors belonging to the same Gaussian into the same initial segment. The clustering thus results in a non-uniform initialization where the number of initial clusters is automatically determined.

### *C. Estimation of the parameters $k$ and $g$*

Combining this estimation of the number of initial clusters with the linear regression to estimate the number of Gaussians per initial cluster, results in an AHC approach where the two most sensitive parameters are unsupervisedly estimated as summarized in Equations 4 to 6.

$$secpergauss = 0.01 \cdot \text{speech in seconds} + 2.6 \quad (4)$$

$$k = \text{estimated with long-term features} \quad (5)$$

$$g = \frac{\text{speech in seconds}}{secpergauss \cdot k} \quad (6)$$

The modified version of the agglomerative clustering approach is schematically shown in Fig. 7. Short-term (MFCC) and

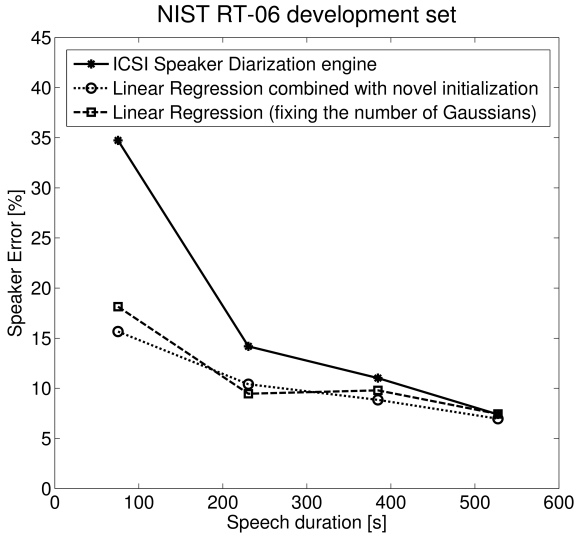


Fig. 8. Performance of the linear regression model in combination with the proposed initialization method vs the baseline and the linear regression with a fixed number of Gaussians on the NIST RT-06 development set.

| data set | file count | speech dur mean | # spkrs mean | overlap mean | # turns mean | baseline (MDM) SE/Spnsp Error | baseline (SDM) SE/Spnsp Error |
|---|---|---|---|---|---|---|---|
| NIST RT-06 dev | 12 | 615 s | 4.42 | 9.65% | 193.67 | 7.10% / 4.60% | - |
| NIST RT-06 eval* | 8 | 909 s | 5.25 | 11.65% | 347.63 | 15.10% / 8.20% | 17.10% / 8.70% |
| NIST RT-07 eval | 8 | 1'160 s | 4.38 | 10.28% | 399.25 | 11.00% / 6.00% | 15.80% / 6.80% |
| NIST RT-09 eval | 7 | 1'485 s | 5.43 | 17.14% | 622.43 | 18.20% / 10.00% | 24.80% / 10.50% |
| Correlation with SE (MDM) | | 0.80 | 0.92 | 0.88 | 0.88 | | |
| Correlation with SE (SDM) | | 0.83 | 0.73 | 1.00 | 0.95 | | |

TABLE V

SUMMARY OF ALL THE DATA SETS THAT ARE USED DURING THIS WORK. SOME CHARACTERISTICS AND THEIR CORRELATION WITH THE SPEAKER ERROR OF THE BASELINE SYSTEM ARE SHOWN AS WELL. FOR THE SAKE OF COMPLETENESS, THIS TABLE ALSO SHOWS THE SPEECH/NON-SPEECH ERROR. THE NIST RT-06 EVAL SET ORIGINALLY CONTAINED 9 MEETINGS. ONE OF THE MEETINGS (TNO_20041103_1130) WAS REMOVED FROM THE EVALUATION DUE TO CHANNEL CONFUSIONS CAUSED BY RENAMING PROCEDURES.

long-term (suprasegmental) acoustic features are extracted from the audio signal. The speech/non-speech detector output and the long-term acoustic features are used to determine the initial segmentation for the diarization engine as described previously in this section. Then, the diarization engine is run as described in Section III. The performance of the modified engine is shown in Fig. 8, where it is compared to the engine using the linear regression and fixing $g = 4$ [18] and to the baseline engine. The engine that combines the linear regression with the proposed initialization method has the best overall performance.

## VI. APPLICABILITY OF THE AUTOMATIC PARAMETER ESTIMATION

One of the main concerns when developing a parameter estimation method as presented in the previous sections is that it might not generalize over all data sets and is too specialized on a particular training set. This section presents our experiments that provide evidence that the automatic parameter estimation, and especially the linear regression, is indeed generalizeable. We tested the presented algorithms on all meetings ever provided by NIST for the RT Evaluations since 2005 in different recording conditions and chunk sizes. NIST distinguishes between recordings with multiple distant microphones (MDM) and recordings with one single distant microphone (SDM). In the case of MDM, beamforming is typically performed to produce a single channel out of all available ones and often the delay between different channels is used as a feature and combined with MFCCs as in [9]. In this article we present results for both, SDM and MDM recordings. In the case of MDM we are using the enhanced channel but we do not use the delays between channels as an additional feature stream.

The modified diarization engine as presented in this article and shown in Fig. 7 is compared to the baseline system, see Fig. 1. For the comparison, several data sets from different NIST RT evaluations are used. Table V gives an overview over all the data sets including some characteristics. To show the robustness of the modified engine against varying recording durations, the different data sets are also split into segments of 100, 300 and 500 seconds in the same manner as described in Section IV.

In [5], the behavior of different speaker diarization engines on broadcast news was studied to determine features that influence the DER most. The authors show, that factors such as the number of speakers and the number of turns affect DER most, whereas the show duration has less effect. In order to demonstrate that this work is not generalizable to meeting data, we present some

| | | NIST RT-06/RT-07 evaluation sets | | | | NIST RT-09 evaluation set | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | MDM | | SDM | | MDM | | SDM | |
| Duration | Configuration | SE | Rel. +/- | SE | Rel. +/- | SE | Rel. +/- | SE | Rel. +/- |
| Entire Meeting | baseline | 12.80% | - | 16.40% | - | 18.20% | - | 24.80% | - |
| | new initialization | 10.50% | -17.97% | 12.80% | -21.95% | 16.10% | -11.54% | 19.00% | -23.39% |
| 500 | baseline | 16.40% | - | 20.40% | - | 18.30% | - | 23.80% | - |
| | new initialization | 11.80% | -28.05% | 11.80% | -42.16% | 15.50% | -15.30% | 19.40% | -18.49% |
| 300 | baseline | 23.80% | - | 27.40% | - | 23.60% | - | 27.30% | - |
| | new initialization | 14.00% | -41.18% | 14.80% | -45.99% | 17.20% | -27.12% | 18.90% | -30.77% |
| 100 | baseline | 44.00% | - | 50.10% | - | 41.10% | - | 41.40% | - |
| | new initialization | 18.00% | -59.09% | 16.60% | -66.87% | 19.70% | -52.07% | 19.80% | -52.17% |

TABLE VI
COMPARISON OF THE NEW INITIALIZATION TO THE BASELINE ON THE NIST RT-06, RT-07 AND RT-09 EVALUATION SETS (MDM AND SDM). ENTIRE MEETINGS AND SHORTER SEGMENTS ARE COMPARED. ONLY THE SPEAKER ERROR IS SHOWN.

| Configuration | RT-06/07/09 MDM | RT-06/07/09 SDM | English broadcast news |
|---|---|---|---|
| baseline ($k = 16$, $g = 5$) | 14.80% | 19.50% | 18.10% |
| new initialization | 12.60% | 15.10% | 15.50% |
| manually tuned on broadcast news ($k = 40$, $g = 5$) | 21.40% | 25.60% | 14.20% |

TABLE VII
CROSS-DOMAIN: BEHAVIOR OF THE PROPOSED INITIALIZATION ON ENGLISH BROADCAST NEWS DATA (NIST RT-04). THE NEW INITIALIZATION IS COMPARED TO THE BASELINE AND TO A SYSTEM, MANUALLY TUNED ON ENGLISH BROADCAST NEWS. RESULTS ON MEETINGS (NIST RT-06, RT-07 AND RT-09) ARE SHOWN AS WELL.

characteristics and correlation values of all data sets used for this work in Table V. We observe that the correlation values vary considerably for different recording conditions (MDM and SDM). This may be caused by the higher noise level in the SDM recordings and the missing SDM data for the NIST RT-06 development set. Further, it can be seen that overlapping speech is a significant challenge in meeting data. One of the listed features, that has a lower correlation value than other features, is the speech duration (based on our speech/non-speech detector [9]). But the correlation is considerably high and this feature is the only one, that we are aware off after the speech activity detection and therefore we believe that it can be used to estimate initialization parameters. The statistics show that the Speaker Error of the baseline system is growing for longer meetings.

Table VI compares the performace of the new initialization to the baseline on the NIST RT-06, RT-07 and RT-09 evaluation sets. These sets were not used for any training or tuning. The results from NIST RT-09 are shown separately. This set was considered to be different from previous ones because it contains more overlapped speech (up to 37% per meeting) and more speakers (up to 11). Nevertheless, the proposed approach behaves robustly on it as well and lowers the DER. It can be seen that the novel method improves the performace for shorter meetings by up to 67 % relative. Even for entire meetings (610 - 1525 seconds for these sets), the novel initialization method performs better than the baseline system (up to 23 % relative improvement). This can be explained by the fact, that not only the parameter estimation but also the non-uniform initialization affects the behavior of the diarization engine positively (see [17, p. 63]). The average relative improvement by the presented approach measured on all meetings (7.5 hours) is 15% for the MDM recordings and 23% for the SDM recordings compared to the baseline approach.

The presented approach has only one remaining "pseudo initialization parameter", the minimal window length (see Section V) used during the suprasegmental feature extraction. In another experiment, this parameter is varied to show that the diarization accuracy is not sensitive to it. For this purpose, all the meetings contained in the NIST RT-06, RT-07 and RT-09 evaluation sets are processed using different minimal window lengths. The results are shown in Fig. 9. In [19], the window length for the suprasegmental features is set to 500 ms, therefore we are not considering smaller minimal window sizes. The model constraint "minimum speech duration of a segment", presented in Section III, is set to 2500 ms, therefore larger minimal window lengths are not considered. In the end though, whatever reasonable minimum window length is chosen, the presented approach performs better than the baseline. Overall, the results for the SDM recordings vary more, which may be explained by the fact that the beamforming in the MDM case positively affects feature extraction [23].

Finally, despite of the fact that our proposed approach behaves robust on about 7.5 hours of test data, we have performed some experiments with English broadcast news data from the NIST RT-04 evaluation, to underline the robustness of the proposed approach. In Table VII, the baseline system, the proposed approach and a system, manually tuned on English broadcast news, are compared. All the systems are run on the NIST RT-06, RT-07 and RT-09 MDM and SDM data (meetings) and on English broadcast news from the NIST RT-04. The proposed approach, running completely automatically without manual adaption, yields 2.6% absolute improvement compared to the baseline and performs 1.3% absolute worse than a manually tuned system on the cross-domain task. The performance of the system, manually tuned on broadcast news, however, drops off on meeting data and performs almost 10% absolute worse than our approach.

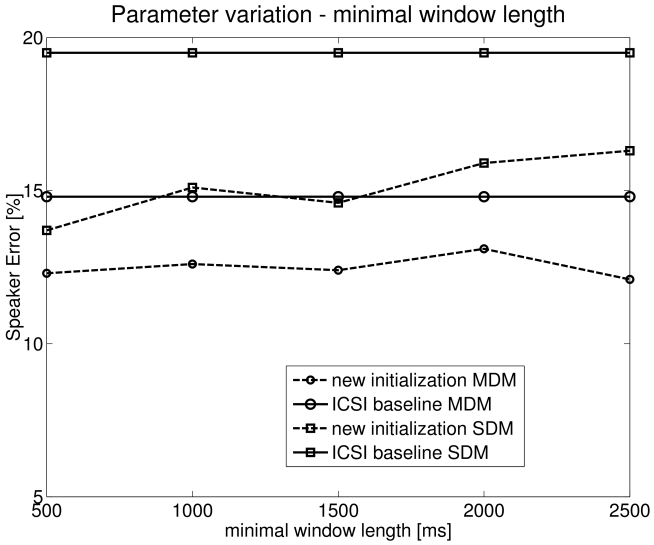


Fig. 9. Variation of the minimum window length used for the suprasegmental feature extraction. The performace is consistently better than the one of the baseline. The behavior on the MDM data is more stable.

## VII. Conclusion and Future Work

In this article, we make the following contributions. First we show that the current standard approach for speaker diarization depends heavily on its initialization parameters. We measure how slight variations of the values can have a large impact on the accuracy of the result. We then identify the two most important parameters to be the number of initial clusters and the number of Gaussians, thereby confirming an assumption that there is an optimal number for the duration of speech that is represented per Gaussian for speaker diarization. We then explore this relation experimentally and find it to be a function with a single global minimum in our search space. This function is then used to infer the number of Gaussians from the speech duration of the recording. We also present a method to estimate the number of initial clusters based on a method that leverages an earlier study on the speaker discriminability of 70 different suprasegmental features. The two estimation methods are combined to a novel initialization method which is then confirmed to work on a set of different corpora. Even though this was not an initial goal, the resulting system outperforms the current ICSI GMM/HMM-based approach using AHC significantly. The Diarization Error Rate on short meetings is improved by over 50 % relative. The automatic method even improves over manually tuned parameters on standard-length recordings, as was measured by comparing the novel methods on past evaluation sets. Finally, even on broadcast news data, the performance of the system is considerably better than the baseline and is competitive with a manually tuned system.

Although we believe that the actual parameters for the linear regression might be dependent on intrinsic parameters in the baseline AHC approach, we think the methods presented in this article could be easily adapted and used for any speaker diarization system based on the described techniques.

In the presented work, only MFCC features were used, but the presented approach was also implemented in the ICSI speaker diarization multistream engine that successfully uses MFCCs in combination with other feature streams to perform the AHC. This multistream engine also participated in the NIST RT-09 evaluation.

From a speed/performance point of view, the extraction of the long-term features and the clustering procedure to estimate the number of initial clusters and to perform a non-uniform initialization adds about realtime to the processing time. On the other hand, the AHC algorithm is sequentially merging clusters and statically choosing 16 initial clusters may be superfluous, especially in the case of shorter meeting durations. Thus, by choosing a more appropriate, smaller value for the number of initial clusters (as the presented parameter estimation is doing), fewer merging iterations need to be performed and the AHC procedure could be sped up.

We have shown that the approach is robust against meeting length variation in the range from 100 to 1500 seconds. We believe that the approach also applies to longer meeting durations and meetings with more speakers. The results of preliminary experiments are very promising and we will further explore this research direction. We could also implement the approach in a different engine in order to further demonstrate its generalizability. The idea of estimating the amount of speech in seconds per Gaussian depending on the quantity of available data that results in choosing an accurate model complexity might also be beneficial for Speaker Identification purposes. In commercial applications for example, the amount of training data can probably be reduced if the model complexity is optimally chosen.

## ACKNOWLEDGMENT

The authors wish to thank H. Bourlard, M. M. Doss, B. Favre, C. Oei and N. Morgan for helpful comments on this work.

## REFERENCES

[1] J. Ajmera, "A robust speaker clustering algorithm," in *In Proceedings of IEEE Workshop on Automatic Speech Recognition Understanding*, 2003, pp. 411–416.

[2] S. S. 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*, Lansdowne, Virginia, USA, February 1998. [Online]. Available: http://www.nist.gov/speech/publications/darpa98/pdf/bn20.pdf

[3] *Approaches and Applications of Audio Diarization*, vol. 5. IEEE, 2005. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1416463

[4] X. Anguera, C. Wooters, B. Peskin, and M. Aguilo, "Robust speaker segmentation for meetings: The ICSI-SRI spring 2005 diarization system," in *Proceeding of the NIST MLMI Meeting Recognition Workshop, Edinburgh*. Springer, 2005.

[5] N. Mirghafori and C. Wooters, "Nuts and Flakes: A Study of Data Characteristics in Speaker Diarization," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2006. Proceedings.(ICASSP'06)*, vol. 1, May 2006, pp. 1017–1020.

[6] J. Luque, X. Anguera, A. Temko, and J. Hernando, "Speaker diarization for conference room: The upc rt07s evaluation system," in *Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007, Baltimore, MD, USA, May 8-11, 2007, Revised Selected Papers*. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 543–553.

[7] X. Zhu, C. Barras, L. Lamel, and J.-L. Gauvain, "Multi-stage Speaker Diarization for Conference and Lecture Meetings," in *Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007, Baltimore, MD, USA, May 8-11, 2007, Revised Selected Papers*. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 533–542.

[8] D. A. Leeuwen and M. Konečný, "Progress in the AMIDA Speaker Diarization System for Meeting Data," in *Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007, Baltimore, MD, USA, May 8-11, 2007, Revised Selected Papers*. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 475–483.

[9] C. Wooters and M. Huijbregts, "The ICSI RT07s Speaker Diarization System," in *Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007, Baltimore, MD, USA, May 8-11, 2007, Revised Selected Papers*. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 509–519.

[10] A. Miró, F. J. H. Pericás, and C. Wooters, "PhD Thesis: Robust Speaker Diarization for Meetings," Ph.D. dissertation, Universitat Politecnica de Catalonia, 2006.

[11] Koh, Eugene Chin and Sun, Hanwu and Nwe, Tin Lay and Nguyen, Trung Hieu and Ma, Bin and Chng, Eng-Siong and Li, Haizhou and Rahardja, Susanto, "Speaker Diarization Using Direction of Arrival Estimate and Acoustic Feature Information: The I2R-NTU Submission for the NIST RT 2007 Evaluation," in *Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007, Baltimore, MD, USA, May 8-11, 2007, Revised Selected Papers*. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 484–496.

[12] C. Fredouille and N. Evans, "The LIA RT'07 Speaker Diarization System," in *Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007, Baltimore, MD, USA, May 8-11, 2007, Revised Selected Papers*. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 520–532.

[13] D. Vijayasenan, F. Valente, and H. Bourlard, "Agglomerative information bottleneck for speaker diarization of meetings data," in *IEEE Automatic Speech Recognition and Understanding Workshop*, 2007, iDIAP-RR 07-31.

[14] Y. Huang, O. Vinyals, G. Friedland, C. Müller, N. Mirghafori, and C. Wooters, "A fast-match approach for robust, faster than real-time speaker diarization," in *Proceedings of the IEEE Automatic Speech Recognition Understanding Workshop*, 2007.

[15] *Live speaker identification in conversations*. ACM Press, 2008. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1459359.1459558

[16] D. Massart, J. Smeyers-Verbeke, X. Capron, and K. Schlesier, "Visual Presentation of Data by Means of Box Plots," *LCGC Europe*, vol. 18, pp. 215–218, April 2005. [Online]. Available: http://www.lcgceurope.com/lcgceurope/Column%3A+Practical+Data+Handling/Visual-Presentation-of-Data-by-Means-of-Box-Plots/ArticleStandard/Article/detail/152912

[17] D. Imseng, "Novel initialization methods for speaker diarization," Idiap, Idiap-RR Idiap-RR-07-2009, May 2009, master's thesis.

[18] D. Imseng and G. Friedland, "Robust speaker diarization for short speech recordings," in *Proceedings of the IEEE workshop on Automatic Speech Recognition and Understanding*, December 2009, pp. 432–437.

[19] G. Friedland, O. Vinyals, Y. Huang, and C. Muller, "Prosodic and other long-term features for speaker diarization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 5, pp. 985–993, Jul 2009. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5067417

[20] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 5.0.32) [computer program]," retrieved August 12, 2008. [Online]. Available: http://www.praat.org/

[21] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*, 2nd ed. Morgan Kaufmann, San Francisco, 2005.

[22] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[23] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 7, pp. 2011–2022, Sep 2007. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4291588