Robust Speaker Diarization for Short Speech Recordings

David Imseng [#][†][1], Gerald Friedland [*][2]

[#] *Idiap Research Institute*
P.O. Box 592, CH-1920 Martigny, Switzerland
[†] *Ecole Polytechnique Fédérale, Lausanne (EPFL)*
CH-1015 Lausanne, Switzerland
[1] david.imseng@idiap.ch

[*] *International Computer Science Institute*
1947 Center Street, Suite 600, Berkeley, CA, 94704, USA
[2] fractor@icsi.berkeley.edu

***Abstract*—We investigate a state-of-the-art Speaker Diarization system regarding its behavior on meetings that are much shorter (from 500 seconds down to 100 seconds) than those typically analyzed in Speaker Diarization benchmarks. First, the problems inherent to this task are analyzed. Then, we propose an approach that consists of a novel initialization parameter estimation method for typical state-of-the-art diarization approaches. The estimation method balances the relationship between the optimal value of the duration of speech data per Gaussian and the duration of the speech data, which is verified experimentally for the first time in this article. As a result, the Diarization Error Rate for short meetings extracted from the 2006, 2007, and 2009 NIST RT evaluation data is decreased by up to 50 % relative.**

I. Introduction

The goal of Speaker Diarization is to segment audio into speaker-homogeneous regions with the goal of answering the question "who spoke when?". Most state-of-the-art systems use a combination of agglomerative clustering with Bayesian Information Criterion (BIC) [1] and Gaussian Mixture Models (GMMs) of frame-based cepstral features (MFCCs) [2]. While these approaches seem to currently dominate, most, if not all, of them ultimately require a certain level of manual tuning of the initialization parameters, such as the initial amount of clusters and the initial number of Gaussians per cluster ([3], [4], [5]). Even though it is often claimed that small changes in the value of the parameters do not cause large changes in the system behavior (see for example [6]), in practice, the robustness of these systems can depend heavily on the manual tuning of the above mentioned parameters. The reason for this is that the performance can drop off dramatically when there is not enough data to train the mixture models. Too few Gaussians, on the other hand, may be unable to model the different speakers appropriately.

By presenting a set of experiments on randomly partitioned NIST meeting data, this article contains a discussion of the behavior of agglomerative hierarchical clustering given different meeting lengths (from 500 seconds down to 100 seconds) that indeed shows that the performance of a state-of-the-art system with manually tuned static parameters is much worse for shorter meeting segments. Using a series of experiments varying the initialization parameters, the correlation between the amount of speech per Gaussian and the speech duration is investigated. We demonstrate that these two parameters are inherently dependent on the length of the meeting recording processed by the system – a fact easily overlooked when investigating length-standardized NIST benchmark data. Based on this analysis we then present an approach in the form of a linear interpolation model on the initialization parameters, which is build on the RT-06 development set. The resulting model is then verified to generalize to other test sets (RT-06, RT-07 and RT-09 evaluation sets) and compared to previous ideas on this topic.

The remainder of this paper is organized as follows: Section II provides a quick introduction to Diarization and related work before Section III presents the baseline system. Section IV then illustrates the behavior of the baseline on short meetings. Section V presents our new approach and Section VI concludes the article with thoughts on future work.

II. Speaker Diarization

As previously mentioned in Section I, the goal of Speaker Diarization is trying to answer 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 involves three tasks: separate speech from non-speech (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 and is usually evaluated against manually annotated ground truth segments. 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[1]. 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 (mapped reference is not the same as hypothesized speaker). For this study we focus on Speaker Errors and disregard parameter tuning for speech/non-speech detection as this is usually seen as a separate task.

Most state-of-the-art Speaker Diarization systems, including the ICSI Speaker Diarization engine [5] (see Section III, Figure 1) combine the segmentation and clustering steps into a single step. As mentioned earlier, a very popular method of doing so is the combination of agglomerative clustering with Bayesian Information Criterion and Gaussian Mixture Models of frame-based cepstral features, as done in [4], [5], [7], [8].

Even though Diarization on meetings shorter than NIST length-standardized benchmark data has not been extensively studied, people have already done some work on automatically adapting initialization parameters based on the recording duration. Most relevant for the work presented here is the discussion presented in [4] 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 [4], the notion of "seconds per Gaussian", which is claimed to be constant, is introduced. Therefore this approach is referred to as "constant seconds per Gaussian" (CSPG). Unfortunately, the authors do not provide empirical evidence for the claim. In [9], the notion of a "Cluster Complexity Ratio" is presented. While the idea is very similar to the one in [4], very little experimental evidence was provided. In contrast to related work, our system uses "adaptive seconds per Gaussian" (ASPG), which increases the overall robustness of the system. A comprehensive evaluation (RT-06, RT-07 and RT-09 evaluation sets) compares the proposed approach to related work (CCR and CSPG) and to the state-of-the-art baseline system. Our approach decreases the Diarization Error Rate on different meeting lengths, which may be generalized to other data sets.

III. Baseline System

For the experiments presented in this article, we used the ICSI Speaker Diarization engine (illustrated in Figure 1). This study investigates the behavior of the agglomerative clustering algorithm, which is described very briefly in this section. The main target of the description is to give an overview over the tunable parameters. For more details about the baseline Speaker Diarization engine, the reader is referred to [5] and [10].

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. 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. 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 15-min meetings with 4-6 speakers $k = 16$ are good choices. For the number of Gaussians per initial cluster, $g = 5$ turned out to be a good choice. After initialization, the algorithm performs the following iterations:

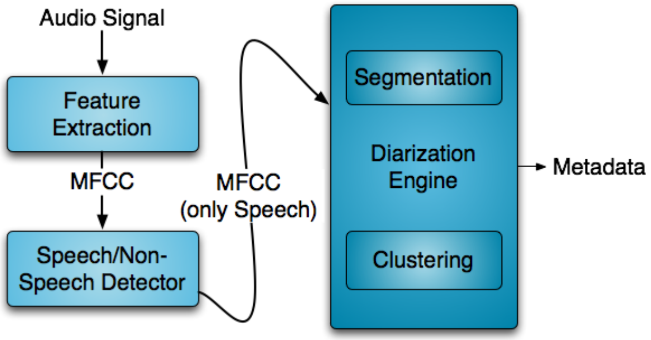

Fig. 1. The baseline ICSI Speaker Diarization Engine as described in Section III.

- Re-Segmentation: Run Viterbi alignment to find the optimal path of frames and models. The 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 Gaussian Mixture Models for each of the clusters.
- Cluster Merging: Given the new GMMs, try to find the two clusters that most likely represent the same speaker. This is done by computing a score based on the Bayesian Information Criterion (BIC) of each of the clusters and the BIC score of a new GMM trained on the merged segments for two clusters. If the BIC score of the merged GMM is larger than or equal to the sum of the individual BIC scores, 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.

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[2]. 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 [5]. In this article we present

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

[2] NIST rules prohibit publication of results other than our own. See: http://www.itl.nist.gov/iad/mig/tests/rt/2009/index.html

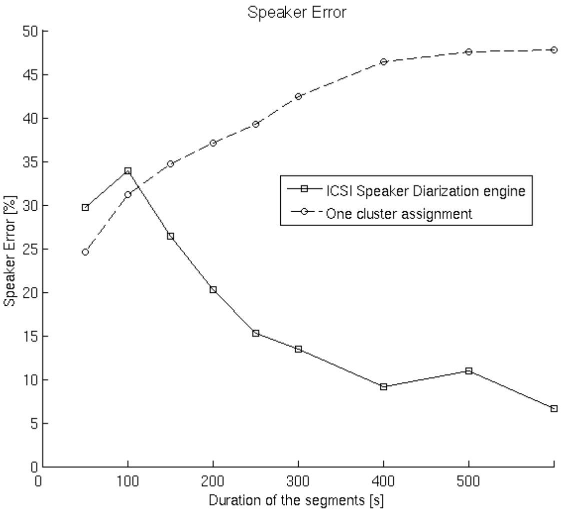

Fig. 2. The performance of the ICSI Speaker Diarization engine on short meetings. For segments of 100 seconds and less, assigning a single speaker to all frames performs best. This underlines the very poor performance of agglomerative hierarchical clustering using fixed initialization parameters for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

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.

IV. Analysis of Short Meetings Diarization

While it has been shown in the past (for example in [11]) that speaker models can be successfully trained on about 50 seconds of speech per speaker for online Diarization, we had anecdotally observed that agglomerative hierarchical clustering methods do not behave very well on short meetings (from 500 seconds down to 100 seconds). In order to systematically study the phenomenon, we randomly split the meetings of the NIST RT-06 development set into smaller pieces of different durations. The meetings were cut into 50, 100, 150, 200, 250, 300, 400, 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 meeting segments with the number of initial clusters $k = 16$ and the amount of Gaussians per initial cluster $g = 5$ (see Section III) and evaluated against the ground truth. This system with the initialization parameters $k = 16$ and $g = 5$ is referred to as the baseline system.

Since the speech activity detector works online, the speech/non-speech error is almost constant even for shorter segments but the Speaker Error is clearly growing as the durations of the meeting segments become shorter (see Figure 2). At first, it seems surprising that the Speaker Error 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 agglomerative hierarchical clustering with the wrong initialization parameters (as will be shown later in this article).

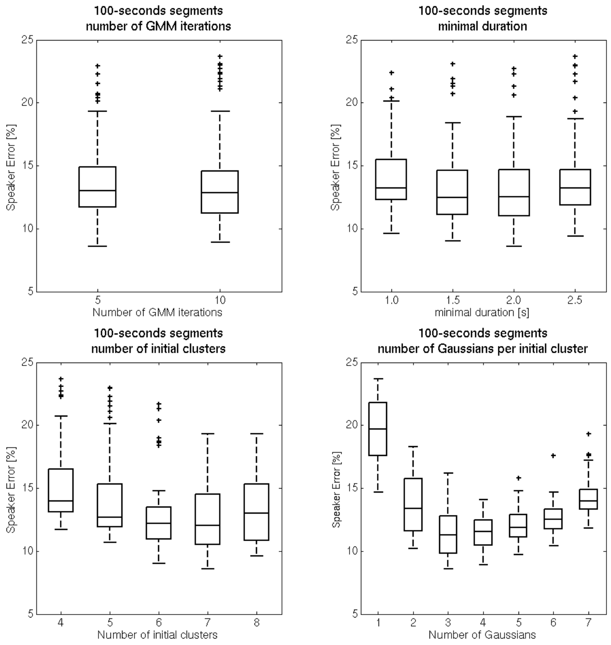

Fig. 3. Boxplots (see [12] for information about boxplots) of the performance of the ICSI Speaker Diarization engine for 100-second-segments. One observes that the variations of the number of initial clusters and the number of Gaussians per initial cluster have most influence on the Speaker Error.

In order to find out which initialization parameters are actually 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 iterations to train the Gaussian Mixture Models in each step of the Speaker Diarization algorithm, the minimum duration for a speech region (default: 2.5 seconds, as explained in Section III), the number of initial clusters k, and the amount of Gaussians per initial cluster g. The results are plotted in Figure 3. In each subfigure (a, b, c and d) the same data is presented, and each boxplot (see [12]) shows the Speaker Error when one parameter value is varied. We observe that small changes in the number of GMM training iterations and the minimal duration do not have as much influence on the Speaker Error as the amount of Gaussians per initial cluster and the number of initial clusters.

As mentioned in Section II, in [4] the notion of seconds per Gaussian is 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 per Gaussian is a combination of two parameters (the number of initial clusters k and the amount of Gaussians per initial cluster g). It was claimed that the seconds per Gaussian is a constant of value 4.8 (CSPG). We therefore conducted an exhaustive search on different meeting lengths to verify this claim. Figure 4 presents the results. Two major observations can be made:

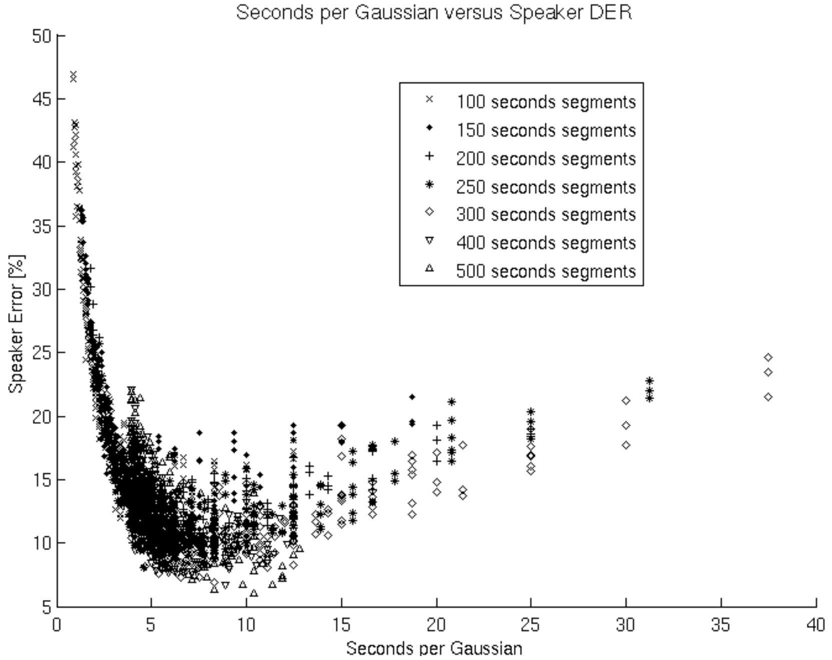

Fig. 4. Speaker Error versus seconds per Gaussian. Each data point corresponds to the average Speaker Error of 12 meetings (2.26 hours of data) for one particular configuration. Configurations for all tested segment durations are shown in the same plot. One can recognize a combination of curves, the minimum seems to be similar for different recording durations.

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 minimum that is very similar for different segment durations and is between 5 and 12.

V. Robust Short Recording Diarization

Our analysis presented in the previous section indicates that it is possible to achieve a low Diarization Error Rate by tuning two parameters, namely the number of initial clusters (k) and the number of Gaussians per initial cluster (g), which can be summarized into a seconds per Gaussian parameter $secpergauss$. Among all tested parameter configurations (presented in Figure 4), the best performing ones for each segment duration were picked and the correlation between the duration of every processed segment versus the corresponding seconds per Gaussian was calculated. The correlation value for the speech duration of the segments versus the Gaussians per second is 0.68. The relatively high correlation value leads to an exploitable linear regression model. Given the definition of the parameter $secpergauss$ and knowing the speech duration after the speech/non-speech 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 (adaptive seconds per Gaussian, ASPG). 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 the number of Gaussians, because for different meeting lengths, there is less variation in the optimal value choice for that parameter than for the number of initial clusters. If the amount of Gaussians is set to four, low variance and mean are attained for 100-second-segment as seen in Figure 3. The boxplots for several different segment lengths look very similar (see [13], p.27) and therefore we decided to fix the number of Gaussians to four. This system is summarized in Equations (1) to (3).

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

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

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

The performance improvement that occurred when we used the linear regression on the NIST RT-06 development data is shown in Figure 5 (baseline: $k = 16$ and $g = 5$, see Section III). In order to test how general the model was, we applied it to other datasets that we split up in the same way. Table I shows the results. The applicability of the same exact linear regression formula to other data sets encourages us to say that the linear regression model is not very dependent on other intrinsic parameters in the ICSI Speaker Diarization engine but might be easily used with any agglomerative hierarchical clustering approach based on Gaussian Mixture Models trained with MFCCs as well.

To compare our new system to related ideas (“Cluster Com-

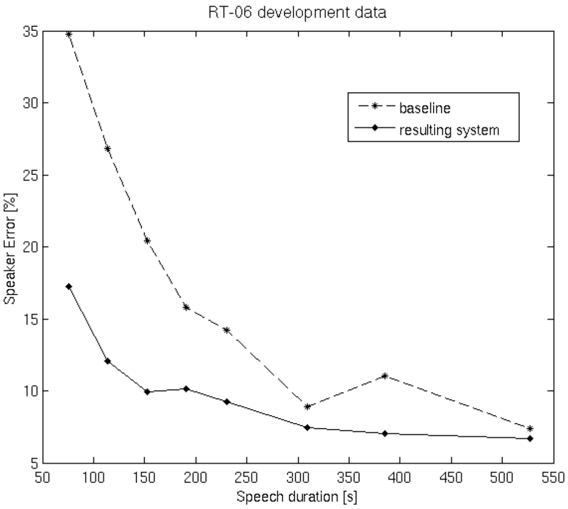

Fig. 5. Performance of the linear regression model vs the baseline on the RT-06 development data.

plexity Ratio", CCR, [9] and "constant seconds per Gaussian", CSPG, [4]) we implemented these methods and tested them experimentally for the first time. For it, the ICSI baseline system described in [10] was used and the two initialization parameters k and g for the different system configurations were chosen, as it is shown in Table II. In [9], $g = 5$ was used during tuning of the Cluster Complexity Ratio. For the CCR, the value that results in the lowest Diarization Error Rate (DER) on the test set (see [9], Table 6.13 on page 163) was used (optimal CCR = 8). For the approach proposed in [4] (CSPG), the seconds per Gaussian were fixed to 4.8 and we decided to use $g = 5$. Originally, it was proposed in [4] to fix the number of initial clusters to 16 and to use the CSPG relation to estimate the amount of Gaussians. The estimations of k for the different systems are shown in Equations (3), (4) and (5).

$$k = \frac{\text{speech in seconds}}{CCR \cdot g} \tag{4}$$

$$k = \frac{\text{speech in seconds}}{CSPG \cdot g} \tag{5}$$

TABLE I
COMPARISON OF THE SPEAKER ERROR OF THE BASELINE SYSTEM VS THE NEW SYSTEM DESCRIBED IN THIS ARTICLE WITH 100-SECOND-SEGMENTS ON DIFFERENT DATA SETS.

Dataset	Baseline	New system	Rel. improve.
Eval06 SDM	53.00 %	23.50 %	48.3 %
Eval06 MDM	48.90 %	25.30 %	50.9 %
Eval07 SDM	47.90 %	20.40 %	55.7 %
Eval07 MDM	40.30 %	19.80 %	57.4 %
Eval09 SDM	41.40 %	20.70 %	50.0 %
Eval09 MDM	41.10 %	19.50 %	52.6 %

TABLE II
DIFFERENT SYSTEMS USED FOR THE COMPARISON EXPERIMENT. ADAPTIVE SECONDS PER GAUSSIAN (ASPG), CLUSTER COMPLEXITY RATIO (CCR) AND CONSTANT SECONDS PER GAUSSIAN (CSPG).

Configuration	k	g	$parameter$
Baseline	16	5	-
ASPG (new system)	Equation (3)	4	adaptive, see Equation (1)
CCR	Equation (4)	5	$CCR = 8$ (see [9])
CSPG	Equation (5)	5	$CSPG = 4.8$ (see [4])

The results of the comparison between the different systems are shown in Table III and IV respectively. All three datasets (RT-06, RT-07 and RT-09 Evaluation set) were not used for any training or tuning. The new approach (ASPG) presented in this paper outperforms all other approaches on very short segments (100-second-segments). Even though, it was not an initial goal, the adaptive seconds per Gaussian approach also performs very well on longer segment durations. On the RT-06 and RT-07 Evaluation sets, in the MDM case, the method performs best for longer segment durations as well. Only if the complete meetings are processed (800-1100 seconds), the baseline system is very slightly better than ASPG. In the SDM case, ASPG outperforms all other approaches when complete meetings are processed, but it performs slightly worse than some related work for 300 and 500 second-segments. In spite of the fact that the 2009 Evaluation set was considered much more difficult than the previous ones because it contains more speakers (up to 11) and more overlap (up to 37% per meeting), the new system behaves robustly. It does not always perform best, but considering all different segment lengths it shows the most robust behavior. Over all sets, there is a tendency that the CSPG of 4.8 ([4]) performs second for very short meetings (100-second-segments), whereas the CCR of 8 ([9]) performs better for longer meeting durations. The new adaptive approach is always best for very short segments and establishes itself well when compared to the baseline and related work on longer segment durations. The consistently better DER of our approach confirms that an adaptive seconds per Gaussian value determined with the help of a linear regression is a better choice than a constant value.

TABLE III
COMPARISON OF THE BASELINE, RELATED WORK AND ASPG ON THE EVALUATION SETS OF THE RT-06 AND RT-07. THE PARAMETER CHOICES FOR THE RELATED WORK CAN BE FOUND IN [9] (CCR) AND [4] (CSPG) RESPECTIVELY. THE BASELINE SYSTEM IS DESCRIBED IN SECTION III.

NIST RT-06 and RT-07 Evaluation sets - MDM				
Duration	Baseline	New system (ASPG)	CCR	CSPG
100	44.00 %	22.10 %	29.00 %	24.30 %
300	23.80 %	15.40 %	16.20 %	16.70 %
500	16.40 %	14.20 %	15.50 %	15.60 %
complete	12.80 %	14.50 %	14.60 %	22.70 %

NIST RT-06 and RT-07 Evaluation sets - SDM				
Duration	Baseline	New system (ASPG)	CCR	CSPG
100	50.10 %	21.70 %	28.90 %	24.60 %
300	27.40 %	17.10 %	16.10 %	18.70 %
500	20.40 %	16.90 %	16.30 %	20.40 %
complete	16.40 %	13.00 %	21.10 %	26.80 %

TABLE IV
COMPARISON OF THE BASELINE, RELATED WORK AND ASPG ON THE EVALUATION SET OF THE RT-09. THE PARAMETER CHOICES FOR THE RELATED WORK CAN BE FOUND IN [9] (CCR) AND [4] (CSPG) RESPECTIVELY. THE BASELINE SYSTEM IS DESCRIBED IN SECTION III.

NIST RT-09 Evaluation set - MDM				
Duration	Baseline	New system (ASPG)	CCR	CSPG
100	41.10 %	19.50 %	27.80 %	22.50 %
300	23.60 %	17.90 %	18.40 %	19.20 %
500	18.30 %	16.70 %	16.10 %	18.10 %
complete	18.20 %	19.50 %	19.50 %	23.80 %

NIST RT-09 Evaluation set - SDM				
Duration	Baseline	New system (ASPG)	CCR	CSPG
100	41.40 %	20.70 %	24.90 %	22.60 %
300	27.30 %	20.60 %	19.80 %	20.30 %
500	23.80 %	19.60 %	17.90 %	21.10 %
complete	24.80 %	19.30 %	20.20 %	24.50 %

VI. CONCLUSION AND FUTURE WORK

The article presented the analysis of the behavior of an agglomerative hierarchical clustering algorithm for Speaker Diarization for short meeting recordings. It was shown that small changes in the two initialization parameters "number of Gaussians" and "number of initial clusters" affect the performance of the system considerably. Based on that analysis, we found a stable correlation between the speech contained in the meeting and the "seconds per Gaussian" parameter by investigating the best performing configurations in a comprehensive series of experiments. Then, a linear regression model was built to balance this relation. The proposed approach is called "adaptive seconds per Gaussian". It elaborates on a statement in [4] that already intuitively indicated the existence of such a relationship but assumed "seconds per Gaussian" to be a constant. We empirically confirmed the relationship and extended the notion of a constant value to an adaptive, linear relation dependent on the amount of speech in the meeting.

A further series of experiments demonstrates that the results presented in this article improve the robustness of short meeting Speaker Diarization in general. The resulting improvement generalizes to different data sets and gives roughly 50 % relative improvement compared to the baseline system for very short meetings (100-second-segments). Until now, the initialization parameters were manually tuned. In this work, it was shown that a simple linear regression model can be used. Further, we implemented two related ideas, namely CSPG and CCR, and evaluated them on different meeting lengths of the RT-06, RT-07 and RT-09 evaluation sets. It was found that in general, even for complete recordings, our linear regression parameter estimation behaves better. We also found that there is potential for our method to be successfully applied to meetings longer than length-standardized NIST benchmark data. Therefore, the investigation of very long meetings (rather than short ones) as well as the generalization to other audio domains, such as broadcast news, constitutes our immediate future work.

ACKNOWLEDGMENT

This work was sponsored by the Swiss NSF through the National Center of Competence in Research (NCCR) on "Interactive Multimodal Information Management" (IM2, www.im2.ch) and the European Integrated Project on "Augmented Multiparty Interaction with Distance Access" (AMIDA, www.amidaproject.org). The authors would like to thank Sherry Zhao from ICSI and Mathew Magimai Doss from Idiap Research Institute for their comments on the paper.

REFERENCES

[1] 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

[2] D. Reynolds and P. Torres-Carrasquillo, "Approaches and Applications of Audio Diarization," *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05).*, vol. 5, pp. 953–956, March 2005.

[3] 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.

[4] 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*, ser. LNCS, vol. 4625/2008. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 475–483.

[5] 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*, ser. LNCS, vol. 4625/2008. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 509–519.

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

[7] 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*, ser. LNCS, vol. 4625/2008. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 543–553.

[8] 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*, ser. LNCS, vol. 4625/2008. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 533–542.

[9] X. A. Miró, "Robust Speaker Diarization for Meetings," Ph.D. dissertation, Universitat Politecnica de Catalonia, 2006.

[10] 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.

[11] G. Friedland and O. Vinyals, "Live speaker identification in conversations," in *MM '08: Proceeding of the 16th ACM international conference on Multimedia*. New York, NY, USA: ACM, 2008, pp. 1017–1018.

[12] 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/ColumnVisual-Presentation-of-Data-by-Means-of-Box-Plots/ArticleStandard/Article/detail/152912

[13] D. Imseng, "Novel initialization methods for speaker diarization," Idiap, Idiap-RR Idiap-RR-07-2009, May 2009, Master thesis.