Martínez-González et al. EURASIP Journal on Audio, Speech, and Music Processing (2021) 2021:12
https://doi.org/10.1186/s13636-021-00196-6

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

Analysis of transition cost and model parameters in speaker diarization for meetings

Beatriz Martínez-González[1†], José M. Pardo[2*†], José A. Vallejo-Pinto[3], Rubén San-Segundo[2] and Javier Ferreiros[2]

Abstract

There has been little work in the literature on the speaker diarization of meetings with multiple distance microphones since the publications in 2012 related to the last National Institute of Standards (NIST) Rich Transcription Evaluation Campaign in 2009 (RT09). Lately, the Second DIHARD Challenge Evaluation has also covered diarization at dinner party meetings that include multiple distant microphones. Dinner party meetings are somehow harder than office meetings because their participants can move freely around the room. In this paper, we studied some of the algorithms on speaker diarization for meetings with multiple distant microphones for the NIST Rich Transcription Evaluation Campaign in 2007 (RT07) and RT09 and provide definite and clear improvements. On the one hand, little or no care has been taken to the problem of penalizing or favoring transitions between speakers other than proposing a minimum duration of a speaker turn or calculating the speakers' probabilities using Variational Bayes (VB). We have studied this issue and determined that a transition penalty term is needed that should be independent both of the number of active speakers and the minimum duration of speaker turns. On the other hand, the determination of a method to automatically select the right number of parameters is crucial in developing good models for speakers. Previous studies have proposed the dynamic selection of the number of parameters based on the duration of the speaker's speech with a mixed performance when tested at one distant microphone meetings or multiple distant microphones meetings. In this paper, we propose a new method that takes into account both the duration of speaker's speech to determine a minimum number of parameters, and the question of overfitting issue to determine a maximum number of them, also taking into account the computation time in order to reduce it.

We have carried out experiments to support our findings, and we have been able to improve our baseline speaker error rate with multiple distant-microphone meetings. Both methods achieve improved performance over the baseline. The first method obtains a 21.6% decrease in relative speaker error for the development set and a 4.6% decrease in relative speaker error for the test set (RT09). The second method obtains a 46.47% decrease in relative speaker error for the development set and a 17.54% decrease in relative speaker error for the test set. Both methods complement each other, and when they are applied in combination, we obtain a 47.2% decrease in

(Continued on next page)

* Correspondence: josemanuel.pardom@upm.es
This work was carried out while the author Beatriz Martínez-González was at Universidad Politécnica de Madrid
†Beatriz Martínez-González and José M. Pardo contributed equally to this work.
[2]Universidad Politécnica de Madrid, Avda. Complutense, 30, 28040 Madrid, Spain
Full list of author information is available at the end of the article

Springer Open

© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

(Continued from previous page)
relative speaker error for the development set and a 22.02% decrease in relative speaker error for the test set. The performance obtained with our proposal is outstanding in some subsets of the development test such as the NIST RT07 and among the best for RT09 using our proposed simple modifications. Furthermore, with our algorithm we obtain gains in computation time without jeopardizing performance. Results with a different publicly available database, augmented multiparty interaction (AMI) obtains a 28.44% decrease in relative speaker error confirming the validity of our methods. Preliminary experiments with a single stream (mfcc) endorse the validity of our findings. Comparisons with an x-vector system deliver superior performance of our system on unseen test data.

Keywords: Speaker diarization, Speaker segmentation, Model complexity selection, Speaker modeling

1 Introduction

Speaker diarization consists of transcribing a recording with speaker labels. This task is usually done with no knowledge as to the number or identity of the speakers. Thus, two tasks are necessary; the first one is to identify the number of speakers, and the second one is to identify the specific regions in which every speaker intervenes. The speaker diarization is needed when transcribing a recording with multiple speakers. An overview of automatic speaker diarization systems is given in [1–3].

There have been the National Institute of Standards (NIST) evaluations for speaker diarization for meetings with multiple distant microphones (MDM) in 2005, 2006, 2007, and 2009. No further NIST evaluations have been made since then. Recently, new interests in speaker diarization have appeared with the launch of The First, Second, and Third DIHARD Speech Diarization Challenge which includes diarization in complex acoustic environments such as broadcast interviews, sociolinguistic interviews, meeting speech, speech in restaurants, clinical recordings, extended child language acquisition recordings, and YouTube videos [4]. However, recordings with multiple distant microphones are only available for dinner parties that differ from the meetings of NIST. Also in the Third DIHARD Challenge, no multiple microphone meetings are included.

The components of a typical speaker diarization are (1) the speech activity detector, (2) the feature extractor, and (3) the segmenting and clustering algorithm. The objective of the speech activity detector is to separate speech from other sounds such as silence or others using two models (speech and non-speech) [5] or more models (i.e., speech, non-speech, and silence) [6, 7]. The feature extractor processes the speech and calculates different spectral characteristics such as the Mel-frequency cepstral coefficients (MFCC), [8, 9], fundamental frequency (F0) [10, 11], the combination of neural network features with MFCC features [12, 13], the use of a phoneme background model [14], and other long-term features [15, 16] or energy features in the case of using multiple distant microphones [17].

The segmenting and clustering algorithm can be either bottom-up [6, 18] or top-down [19]. In a work published in [20], a comparison between both methods is made. In another work, the information on the role of speakers is used to adjust the segmentation [21]. Speaker models can be established using Gaussian mixture models (GMM) [22] or more recent I-Vector models [23, 24], CNN-I-Vectors [25], or X-Vectors [26]. X-Vectors have shown good performance; however, they need much more training data than our experiments because we do not use any external data other than that available from the recording session. In this sense, our GMM system is self-contained both for speaker modeling and for speech activity detection and is independent from any external sources [1]. The more generally used distance metrics depend on the speaker models and the most common are the Bayes information criterion BIC [27], T test distance [28], information theoretic approach [29, 30], and cosine distance and probabilistic linear discriminant analysis (PLDA) for I-Vectors [16, 31].

Since the duration of a speaker's turn is not known, a significant problem is how to decide when a speaker's turn is feasible. One way of doing it is through comparisons of acoustic models before and after the turn. Some people use Viterbi segmentation [32] but penalizing transitions dependent of the number of active clusters. Another possible parameter to use is the minimum duration of a speaker turn, which limits the total number of speaker's turns [2] in a recording. The problem of penalizing transitions has also been analysed in [33] and [34] proposing a different alternative although the number of speakers is known in their experiments. Recent research focuses on this topic and proposes the learning of the

[1] At the time that this technology was created, voice activity detection, for instance, was pretty much dependent on the type of background noise, and in this way, the results of an external VAD could generate unstable results. Equally, if we had to use the system in different rooms, different scenarios, different types of backgrounds etc., the use of external sources would deliver spurious results. If we assume that we have a model that is universal enough that could be used as a background and adapted in a second step to our room, certainly the method could be more robust.

speaker turn priors [24, 35]. In this work, we propose to revisit this problem of classic methods and present alternative solutions that produce better and more robust results.

As regards the cluster (speaker) models, an important decision when modeling a speaker with a GMM or other models is the determination of the number of mixtures or parameters needed. In general, it is known that the amount of available data for training plays a crucial role in defining the number of parameters of a model, since with little data, it is impossible to create good models if the model has a lot of parameters. On the other hand, if we have plenty of data and as many parameters as we want, we encounter the problem of overfitting and the model does not generalize well. This topic is addressed in most pattern recognition books; see for instance [36]. In [37], this problem is analyzed and the number of frames needed to create a model is determined using the so called "cluster complexity ratio" which is a parameter that relates the number of frames of data available to the number of mixtures in a GMM that models this data. After each change in the amount of data assigned to each cluster due to segmentation, a new number of mixtures is defined that is related to the number of frames now assigned to the new model. Some positive results have been obtained in single distant microphone (SDM) experiments with a database of 16 meetings in the development set and 10 meetings in the test set (improvements of 2.9% relative in the diarization error (DER) for the development set and 19.39% relative in the DER for the test set). But when new experiments with a bigger development set (24 meetings) and new set of 8 meetings in the test set (meetings from the NIST Rich Transcription Evaluation Campaign in 2006 (RT06)) and testing in both the SDM and MDM scenario, the results only improve by 2.7% relative in the the DER for the development set and no improvement at all in the test set [32]. Contradictory results are again obtained in [38], in which the SDM results in the test set do not improve but degrade performance by 17.5% relative in the DER. Furthermore, their procedure does not take into account the overfitting issue because more frames, even if they do not add new information, are modeled with more parameters and the speaker model may overfit and not generalize sufficiently. Other researchers [39] have demonstrated that the number of Gaussians used to model a speaker is important in the creation of a good segmentation. Their experiments include a consensus based on different models each trained with a different number of Gaussians.

The problem of selecting the number of parameters is also important when mixing acoustic features with delay features in a weighted model [40]. The delay features do not need as many parameters as the spectral features since their dimensionality is usually lower and should not receive the same treatment.

The objective of this paper is to study the complexity of the models in the context of the MDM meetings' diarization, carry out a thorough analysis of it, propose two parameters and its interrelation for solving the problem, and obtain justified conclusions. This study was not done before. Furthermore, we propose a new strategy to prevent overfitting and save computation time without significantly decreasing performance. Preliminary analysis of our methodology applied to single-channel recordings is also presented.

The paper is organized as follows. In Section 2, the baseline system is described. In Section 3, the database used for experiments is explained. In Section 4, we present the analysis of the problem of the transition penalty when segmenting speakers. In Section 5, we introduce the second objective: how to select the right model for a speaker. Section 6 is a section that merges the approaches of Sections 4 and 5. Section 7 presents results for the best systems with a publicly available database and a set of comparisons of our results with other published data. Finally, Section 8 is the discussion and Section 9 ends with our conclusions.

2 Description of the baseline system

2.1 Introduction

The architecture of the system is presented in Fig. 1. Every microphone produces a signal that is filtered to suppress some channel noise. After that, there is a module that calculates the time difference of arrival (TDOA) between two signals. In our case, these signals are the output of the microphones. The method used is the generalized cross-correlation method (GCC) [41]. The cross-correlations between any pair of channels are calculated as well. The channel with the highest cross-correlation is used as a reference [42]. The next step is the creation of a beamformed signal by delaying and summing the signals coming from the different microphones (weighted sum).

The Mel-frequency cepstrum coefficients (MFCC) are extracted from the beamformed signal, every 10 ms using a window width of 30 ms. The MFCC coefficients form what we call the mfcc vector. The beamformed signal is also processed by a voice activity detector (VAD) that classifies speech frames versus non-speech frames using a two-model Gaussian mixture model (GMM) and Viterbi resegmentation [5]. The output of the VAD module is fed into the agglomerative clustering module.

The localization features estimation creates a vector of TDOAs for each 10 ms frame. This vector is obtained by choosing an optimized set of channel pairs and calculating a TDOA for every pair. The concatenation of the TDOAs forms what we call the tdoa vector. Several

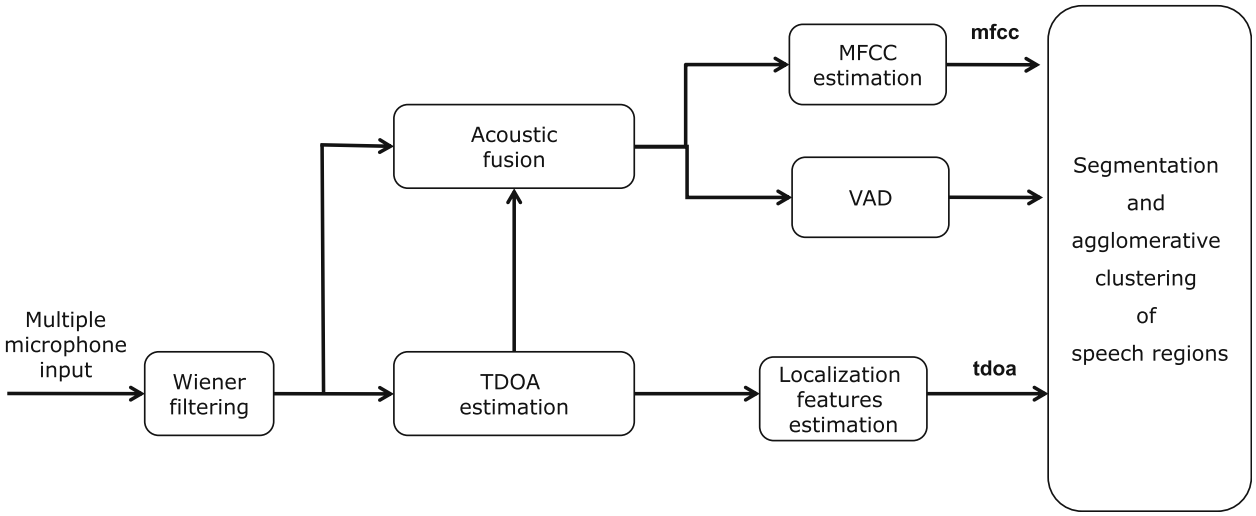

Fig. 1 Baseline system architecture

methods were tried and tested in order to find the optimum representation of the localization features including the principle component analysis (PCA) transformations and using cross-correlation as a measure of quality. The method used was the selection through cross-correlation between channels, see [43].

Both mfcc vectors and tdoa vectors are fed to the next block which is the segmentation and agglomerative clustering of speech regions. This block has several parts (see Fig. 2). There is an initialization module that creates a first set of segments based on a maximum number of clusters (speakers) L (we use the maximum number of expected speakers). The full recording (only the speech part) is divided uniformly into L parts.

Each cluster is modeled by a Gaussian mixture model (GMM). There is a minimum duration per cluster, typically 2.5 s (see Fig. 3) [22]. The minimum duration per cluster is determined empirically. Due to this minimum duration, short interjections such as "yes," "heah," and "no" will be ignored by the system. However, the scoring mechanism does not ignore such words. They will be considered errors in our system. The problem of short words or affirmation is one of the drawbacks of our method. The GMM consists initially of a minimum number of components, 5 for the mfcc vector and 1 for the tdoa vector. The next module is the segmentation and training module. The sentence is segmented by the Viterbi algorithm using the original cluster models. Then, after segmentation, a new training is carried out followed by a subsequent segmentation. This process is repeated several times (from 3 to 5). The next module is "Cluster pair comparison". Every combination of two

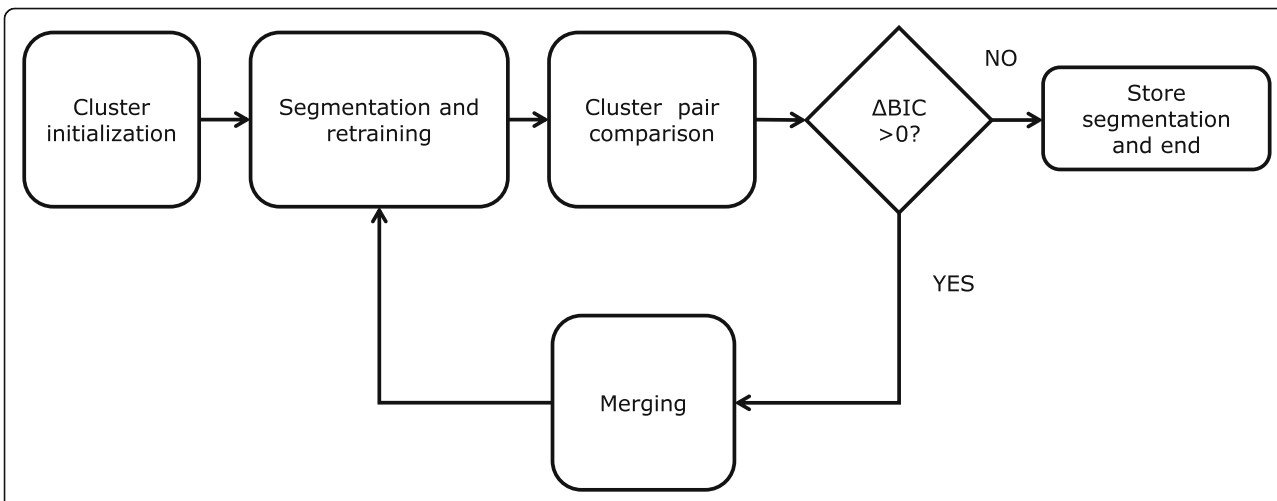

Fig. 2 Block diagram of the baseline segmentation and clustering

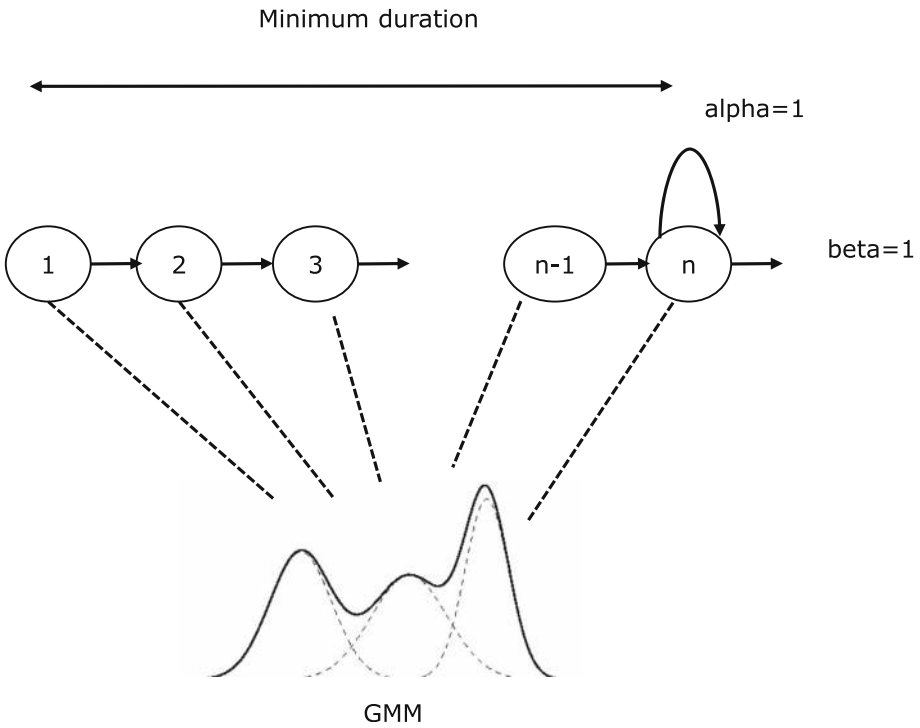

Fig. 3 Cluster models with minimum duration

clusters is compared to determine if they should be merged or not. If the stopping criterion is not met, a pair of clusters is selected to be merged. When this happens, the number of components in the merged cluster is the sum of the components of the individual clusters. When the stopping criterion is met, the process ends. The number of components of any cluster model will depend on the number of times that this cluster has participated in a merging, regardless of the duration of the final cluster once the resegmentation has been carried out.

We use the ΔBIC measure to decide if any merging is still possible (see Eq. 1) [27]. Notice that there is no penalty term λ in the BIC score because there is no difference in the number of parameters from the two modeling hypotheses as shown in [22]. In the following equation, X represents the full recording. X_A represents the part of the recording assigned to speaker A, and X_B represents the part of the recording assigned to speaker B.

$$\begin{aligned} &\Lambda BIC = \log p(X|\xi) - \log p(X_A|\xi_A) - \log p(X_B|\xi_B) \\ &X = X_A \cup X_B \\ &\xi_A \text{is the model created with} X_A \\ &\xi_B \text{is the model created with} X_B \\ &\xi \text{is the model created with} X \end{aligned} \tag{1}$$

The combination of the mfcc vector and the tdoa vector is made using the methodology presented in [40]. We apply a weight factor to the mfcc vector of 0.85 as in [43] since we use the same set of localization features.

2.2 Baseline segmentation method

The model of a cluster consists of a series of Hidden Markov Model (HMM) states that share the same GMM. The number of these states is equal to the minimum number of frames assigned to a speaker turn (in the baseline this is 250 equivalent to 2.5 s). In the last state, following the recommendation in [32, 37], the probability of staying in the last state (alpha) or jumping to another cluster (beta) is set to 1. At this point, neither value can be considered as probabilities anymore since they do not add up to 1. But when calculating the accumulated Viterbi probability, alpha and beta do not add any extra duration model to the last state of a cluster. After the jump to another cluster, the value of beta changes to beta/M, M being the number of remaining active clusters (see Fig. 4).

This value beta/M adds a new penalization factor to a transition. Furthermore, this penalization factor is dependent on the number of active clusters since it changes after every iteration in the clustering and merging process starting from the L initial clusters and decreasing by one at each step. The penalization factor then increases at each iteration (M is lower). This increase is somehow artificial and totally independent of the number of speakers in the recording (since it is not

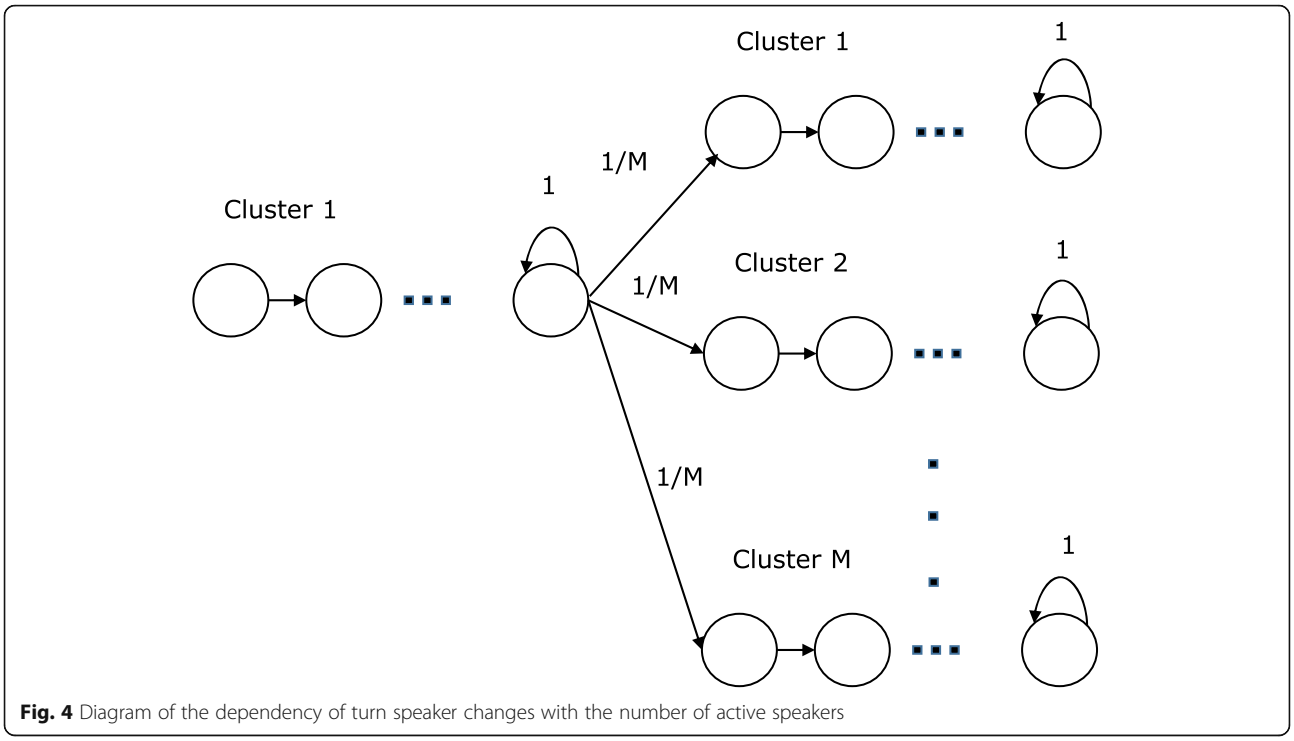

Fig. 4 Diagram of the dependency of turn speaker changes with the number of active speakers

known). This factor is not usually taken into account in classic diarization systems. Some recent research proposes methods to learn this factor [24, 35]. One of the objectives of this paper is to focus on the study of this factor and propose an alternative that improves the baseline system. Preliminary experiments on this topic have been presented in [44].

3 Database and metrics

There has been little work in the literature on speaker diarization of meetings with multiple distance microphones since the last RT09. There is some new work on RT09 but using only one distant microphone and oracle speech/non-speech detector [12] or assuming that the number of speakers is known a priori [45].

We do not have a training set. Our development set used to train hyper-parameters consists of a subset of 12 meetings extracted from NIST Rich Transcription 2002–2005 sets (RT02-05). This set was previously used by us in published work (devel06 in [40]). We add the RT06 set and RT07 set and conform what will be called DEVE LSET from now on, see Table 1. The evaluation set will be RT09. The performance of the systems was calculated using the scored speaker time and the segments of the recordings officially selected by NIST for the annual evaluations. The amount of time is 15,484.34 s or 4.3 h (1,548,434 frames) for the DEVELSET and 5932.88 s or 1.64 h (593,288 frames) for the RT09 set. We did include overlap regions and 0.25 s of forgiveness factor as in the official evaluations. The calculation of the DER and the speaker error (SER) is carried out using the tools provided by NIST [46]. We will focus primarily on the SER since the miss speaker error (MISS) and the false alarm error (FA) are fixed in all our experiments. DER is also presented for comparison purposes with other published works with the same data sets.

4 Segmentation independent of the number of active clusters

4.1 Statement of the problem

We have mentioned above that in the baseline, every change of speaker includes a factor $1/M$. M is the number of current clusters after the previous merging. The factor $1/M$ (always less than 1) decreases the probability of changing the speaker versus staying with the current speaker. An undesirable extra effect is that M is variable at each iteration so the factor $1/M$ is also variable. One would reasonably be tempted to think that if M is bigger, the probability of a speaker change should be higher but this kind of probability is not known neither it is attempted to use in our system. Thus, in the absence of this information, what is not right is to take into consideration the number of "remaining clusters M" in the algorithm. Let us use a penalizing or regularizing factor similar to the penalizing factor that weighs language model versus acoustic model in speech recognition.

When a Viterbi segmentation is carried out, there is an accumulated log-likelihood associated with the last sub-state of each cluster, which is the accumulated sum of log-likelihoods corresponding to the previous frames.

Table 1 List of meetings used for the development set (DEVELSET)

		Meeting	# of microphones
1	devel06	AMI_20041210-1052	12
2		AMI_20050204-1206	16
3		CMU_20050228-1615	3
4		CMU_20050301-1415	3
5		ICSI_20000807-1000	6
6		ICSI_20010208-1430	6
7		LDC_20011116-1400	8
8		LDC_20011116-1500	8
9		NIST_20030623-1409	7
10		NIST_20030925-1517	7
11		VT_20050304-1300	2
12		VT_20050318-1430	2
13	RT06	CMU_20050912-0900	2
14		CMU_20050914-0900	2
15		EDI_20050216-1051	16
16		EDI_20050218-0900	16
17		NIST_20051024-0930	7
18		NIST_20051102-1323	7
19		VT_20050623-1400	4
20		VT_20051027-1400	4
21	RT07	CMU_20061115-1030	3
22		CMU_20061115-1530	3
23		EDI_20061113-1500	16
24		EDI_20061114-1500	16
25		NIST_20051104-1515	7
26		NIST_20060216-1347	7
27		VT_20050408-1500	4
28		VT_20050425-1000	7

A speaker's turn takes place when the left-hand part in the formula below is lower than the right-hand part

$$\sum_{i}^{i+MIN\_DUR} log\mathcal{L}(cl_j;fr_i) < log(K) + \sum_{i}^{i+MIN\_DUR} log\mathcal{L}(cl_u;fr_i) \quad (2)$$

in which logL() is the log-likelihood, K the transition weight (in the baseline system this is $1/M$), cl_u the candidate cluster ,cl_j the current cluster, and fr_i the frame being evaluated. The left part represents the sum of the last "minimum duration" log-likelihoods of the frames if they belong to the current cluster. The right-hand part represents the same total of log-likelihoods of the frames if they belong to a different cluster plus the log of a transition weight K. Every increase in the transition weight force a speaker turn since the condition in (2) is easily met. On the other hand, if K is very small (much smaller than one if M is big), it makes the transition more difficult because in the right-hand part, we substract some quantity. In summary, in the current formula, a factor is included that has no relation to the current acoustics and is somehow arbitrary.

In the baseline system, at the beginning of the agglomerative clustering, M is large, K is small, log (K) is negative, and the condition in (2) does not hold, so the transitions are penalized; however, at the end of the iterations, M is much lower, thus favoring transitions. This undesired effect is the one that we want to eliminate.

In order to do so, we propose a set of experiments with variations of this factor, but independently of the number of current active clusters. We will also experiment with high values of K, thus favoring transitions between speakers. The case of K = 1 (the change of speaker determined only by the acoustics) will also be tested.

4.2 Experiments

As explained before, we will be focusing on the SER. The difference between the speaker error rate and the DER is that the DER includes SER plus MISS that is the part in which the system does not find or identify a speaker (in our system, each overlap time will contribute to one or more errors since we deliver just one speaker hypothesis) and FA which is the part in which the system proposes a speaker and there is silence (the VAD module is responsible for this error). The VAD module also contributes to the MISS error (there is a true speaker and the VAD module thinks that it is not speech). Since we are not changing the VAD module neither doing any overlap handling, the MISS error plus the FA error will be 7.44% in our DEVELSET in all the experiments. In our test set (RT09), the MISS error plus the FA error is 8.70% in all cases. We give those two values for comparison purposes to be able to calculate the DER. We use a no-score collar of 0.25 at speaker boundaries as usual in standard Rich Transcription (RT) evaluations. We use a weight factor of 0.85 for the mfcc vector and 0.15 for the tdoa vector.

Figure 5 represents the speaker error versus the transition weight K in formula (2) for different values of the minimum duration of a speaker's turn. The baseline SER is also shown. Analyzing the results, we notice a big dispersion across the K values and across the minimum duration values. For minimum duration = 200 (2 s), the new methodology improves the results of the baseline for an ample range of values of K. At the same time, it is less dependent on the values of K. It is interesting to note that for $K < 1$, the results are less stable than for $K > 1$. We cannot find a good reason to justify it. It is very

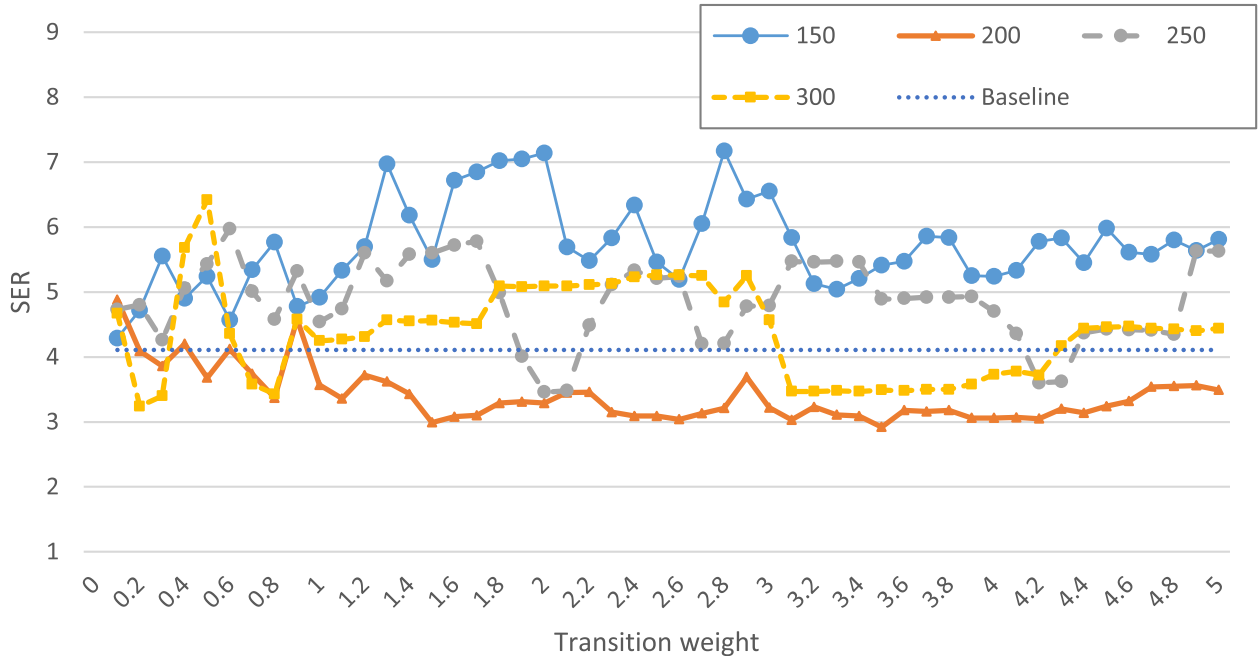

Fig. 5 SER of the development set for minimum duration of speaker turn of 150, 200, 250, and 300 frames against the transition weight

much dependent on the kind of acoustics and the type of meetings and speakers at each meeting. But we have used many different meetings in different rooms, so the experimental results are solid. The fact that for $K < 1$, the results are less stable give us a good reason not to rely on a $K = 1/M$ which is even more unstable since it depends on the iteration of the algorithm. Remember that in the baseline, K is variable at each iteration and less than one. Two proofs are shown in the picture, the first one is that K should not follow the previous strategy (changing it depending on the number of active speakers) but it should be independent of it. At the same time, the parameter "minimum duration" is dependent on K, so both parameters should be explored to find and optimum.

In Table 2, the performance of both DEVELSET and the test set (RT09) are presented for different values of K. Since in the baseline, the minimum duration is 250, and in the new methodology, the minimum duration is 200; we have included the case for baseline and minimum duration equal to 200 in the table. In the baseline system, there is no significant difference from minimum duration of 250 to the minimum duration 200. It can be observed that for the development set, any value of K

Table 2 SER for all the systems developed, confidence intervals are also included. M is the number of active clusters at each iteration. Weight applied to MFCCs is 0.85 and weight applied to TDOA is 0.15

Transition weight	Minimum duration	DEVELSET	Relative improvement over DEVELSET (%)	RT09	Relative improvement over RT09 (%)
1/M (baseline)	250	4.11 ± 0.03		7.82 ± 0.07	
1/M	200	4.07 ± 0.03	1.94	7.73 ± 0.07	1.15
1.0	200	3.57 ± 0.03	13.14	8.45 ± 0.07	− 8.05
2.0	200	3.29 ± 0.03	19.95	7.72 ± 0.07	1.28
3.0	200	3.22 ± 0.03	21.65	**7.46 ± 0.07**	**4.6**
4.0	200	3.06 ± 0.03	25.55	7.57 ± 0.07	3.20

between 1 and 4 and minimum durantion 200 is better than the baseline with either minimum duration of 250 or 200, demonstrating the validity of our approach. If we analyze the results on the test set, we notice that every value of K with the exception of $K = 1$ improves the baseline. The best result with the DEVELSET, which is $K = 3$ or 4 also improves the baseline results. We can conclude that the new methodology delivers better results than the baseline methodology.

It is interesting to note that with $K > 1$, we are favoring speaker changes while in the baseline, K is always less than 1, thus penalizing speaker changes. At the same time that we have discovered that favoring speaker changes is better in our experiments, we have eliminated the somehow arbitrary variations of K depending on the iteration of the algorithm and the number of active speakers at each iteration (baseline).

One important characteristic of speaker diarization for meetings is that the results across different rooms, different location of microphones, different number of microphones, and different number of speakers etc., are very unstable. Some of them are very good but some others are terrible [47]. Thus, the best way to demonstrate technological improvements is to test the system with as many recordings as possible. We have tried the system with 28 meetings for development and 9 for test so our experimentation is ample. Furthermore, the data that we use belong to a community standard and can be contrasted with results of other researchers.

4.3 Experiments with a single channel

In order to check if the previous method works for single-channel recordings, we have selected the mfcc vector coming from the acoustic fusion (see Fig. 1) and discarded the tdoa feature channel. In this case, the diarization is similar to the use of a single microphone recording. Figure 6 represents the speaker error across different values of the transition weight and different minimum duration values. It can be seen that there are several values below the baseline of 8.98 SER. This picture demonstrates the validity of our proposal. The baseline uses a transition weight dependent on the number of remaining clusters, but a constant transition weight improves the SER performance. However, it can be noticed that in this case, the minimum SER values are located at different working points. Two minimums can be considered, one at the point 350 minimum duration and transition weight of 0.001 with an 8.29% SER which represents an improvement of 8.3% relative and another one at a minimum duration of 400 frames and transition values of 0.01 with an 8.39% SER that represents an improvement of 7.0% relative SER.

Table 3 presents the results obtained for the working points for the test set. Both points improve the baseline of the system by 42% relative SER and 7.0% relative SER, respectively. We can notice in the table that the optimum working point with a single channel differs substantially from the optimum working point obtained previously (minimum duration of 200 and transition weight of 3). The conclusion that we extract from this result is that the method is valid also for a single channel and it can be used, but the parameters should be tuned for each case. The minimum duration and the transition weight interact with each other in the system, and they cannot be universally determined but through an empiric study. But it can be proved that both working points also improve the test set, in a case with noticeable improvement.

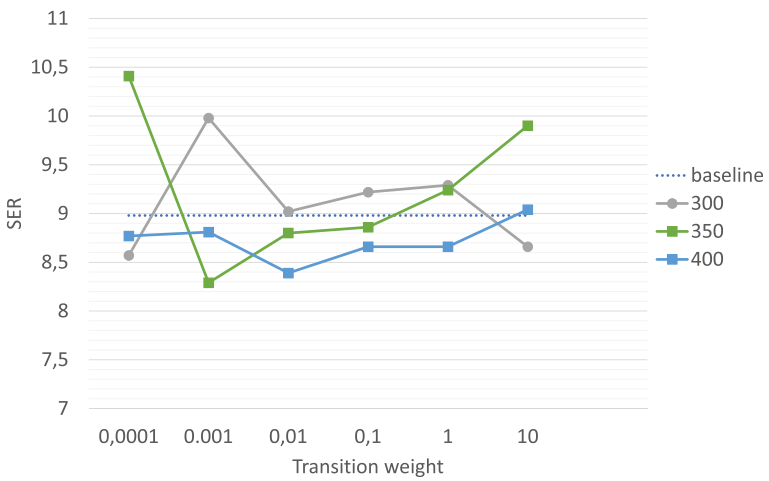

Fig. 6 SER for the DEVELSET for a single mfcc stream across different values of the transition weight and different minimum durations of a speaker turn of 300, 350, and 400 frames

Table 3 SER for all the systems developed using a single channel. *M* is the number of active clusters at each iteration

Transition weight	Minimum duration	DEVELSET	Relative improvement over DEVELSET	RT09	Relative improvement over RT09
1/M (baseline)	250	8.98		21.21	
0.001	350	8.29	8.3%	**14.87**	**42%**
10	400	8.39	7.0%	19.81	7.0%

5 Model selection

5.1 Introduction to the problem

In the baseline system, when merging two clusters, the ΔBIC distance used to determine whether the clusters should be merged eliminates the need for the adjustable λ parameter by setting the number of Gaussians of the merged cluster as the sum of the Gaussians of the original clusters to be merged. In this way, the merged clusters now have many more Gaussians independently of their duration. But the new number of Gaussians may be too small or too big to model properly the new cluster and the remaining clusters after a segmentation step have been carried out.

In the proposal of Anguera [37], an attempt to solve the problem was addressed. Instead of keeping the number of Gaussians dependent on the number of times that a cluster has been merged with another one (because the total number of parameters is kept constant after merging), the number of Gaussians is always recalculated depending on the duration of the clusters after merging and resegmenting. In this way, a small cluster could be modeled with a single Gaussian. But the proposal by Anguera does not address the problem of using many Gaussians for a long cluster—thus expending a lot of resources—or the risk of overfitting the model. In this paper, we have shown that there is very little improvement by increasing the number of Gaussians after a certain limit because even if more data were available, this data would not add new information to the model.

In Fig. 7, the normalized log likelihood of a speaker extracted from a session from the development set using the true references is plotted versus the number of Gaussians used to train it. The speaker has 55,114 frames (551 s). We normalize the log likelihood by dividing the total log likelihood by the number of frames. It can be observed that the likelihood has a long tail, and it does not improve substantially when the number of Gaussians is over 100 indicating that there is no need to use so many Gaussians to model the speaker. This fact is better illustrated when we plot the derivative of the normalized log-likelihood (see Fig. 8). We notice that after a certain number (i.e., 100) of Gaussians, the derivative remains approximately constant. On the one hand, we need a minimum number of Gaussians to model a speaker of a certain number of frames adequately (duration). On the other hand, we do not gain a lot by augmenting the number of Gaussians after a certain value and we could save computation by limiting the maximum number of them. Figure 9 illustrates the same concept, this time the number of Gaussians is kept constant at 5 and the normalized log-likelihood is plotted against the number of frames. This picture clearly demonstrates that when few frames are available 5 Gaussians is not a good parameter to use in this case and it distorts the model. The picture shows a minimum of log-likelihood at 3674 frames (36 s) having a value at that point which is comparable to values in the previous picture for the same number of Gaussians. This figure

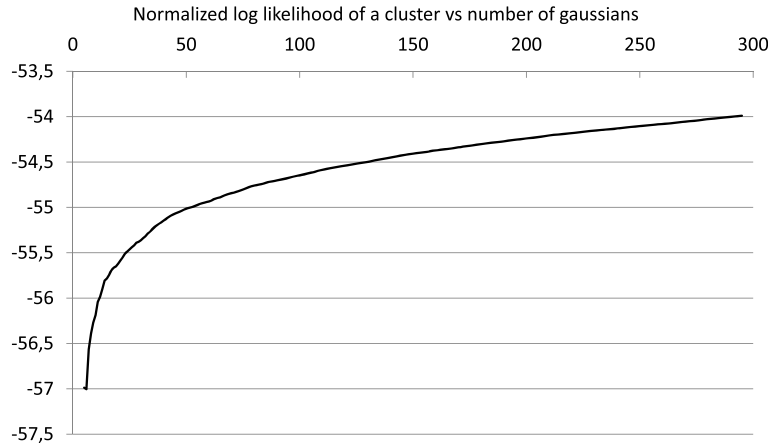

Fig. 7 Normalized log-likelihood of a cluster versus the number of Gaussians used to train it. The likelihood improves asymptotically and very slightly to a maximum number independent of the number of Gaussians

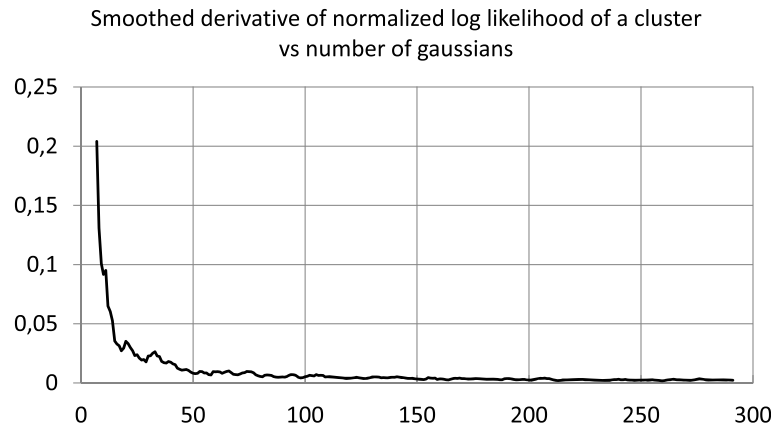

Fig. 8 Smoothed derivative of normalized log-likelihood of a cluster versus the number of Gaussians used to train it. The derivative changes abruptly for a low number of Gaussians and remains constant after 50 Gaussians

shows that having 3674 frames generates a model that can be compared to other clusters. But if we have many fewer frames, using 5 Gaussians is not appropriate and the comparison would have been biased favoring the cluster with a smaller duration (note also that we are using logarithms so the dynamic range of the arithmetic is lower). On the other hand, adding more frames to the model does not significantly change its log likelihood.

Our proposal is to modify this strategy and use two new parameters to determine the number of Gaussians per model, one is the number of frames, and the other is the maximum number of Gaussians per cluster, as will be presented in the next section.

5.2 Proposed method

We propose a method to take into account the problems mentioned in Section 1. The algorithm is presented in Fig. 10 in which the modules that change with respect to the baseline algorithm are marked with "NEW". In the new proposed algorithm, both at the initialization step and after any new segmentation a recalculation of the number of Gaussians used to model each cluster is implemented according to formula (3). Two parameters are used, A = the minimum duration to train any single Gaussian and B = the maximum number of Gaussians used to model a cluster.

$$n = \min\begin{cases} n.\text{of seconds of the cluster}/A \\ B \end{cases} \tag{3}$$

In Fig. 11, the DER values for the DEVELSET for different parameters of the minimum number of seconds per Gaussian (A) and the maximum number of Gaussians (B) are represented together with the average of all of the values (marked "AVE"). We can see that with this

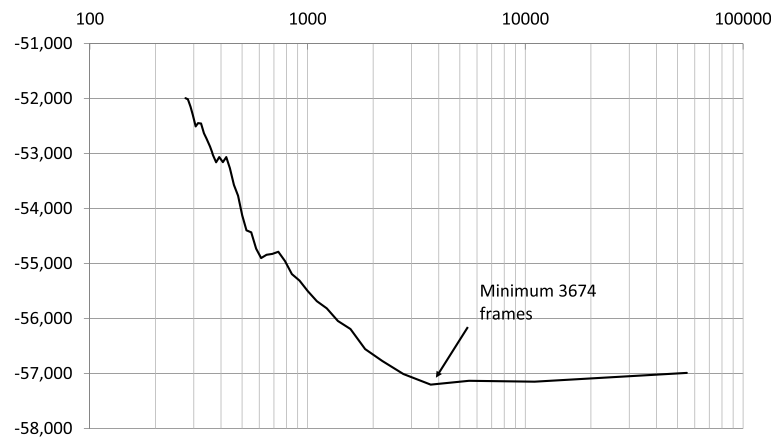

Fig. 9 Normalized log-likelihood of a cluster trained with 5 Gaussians as a function of the number of frames used to train it

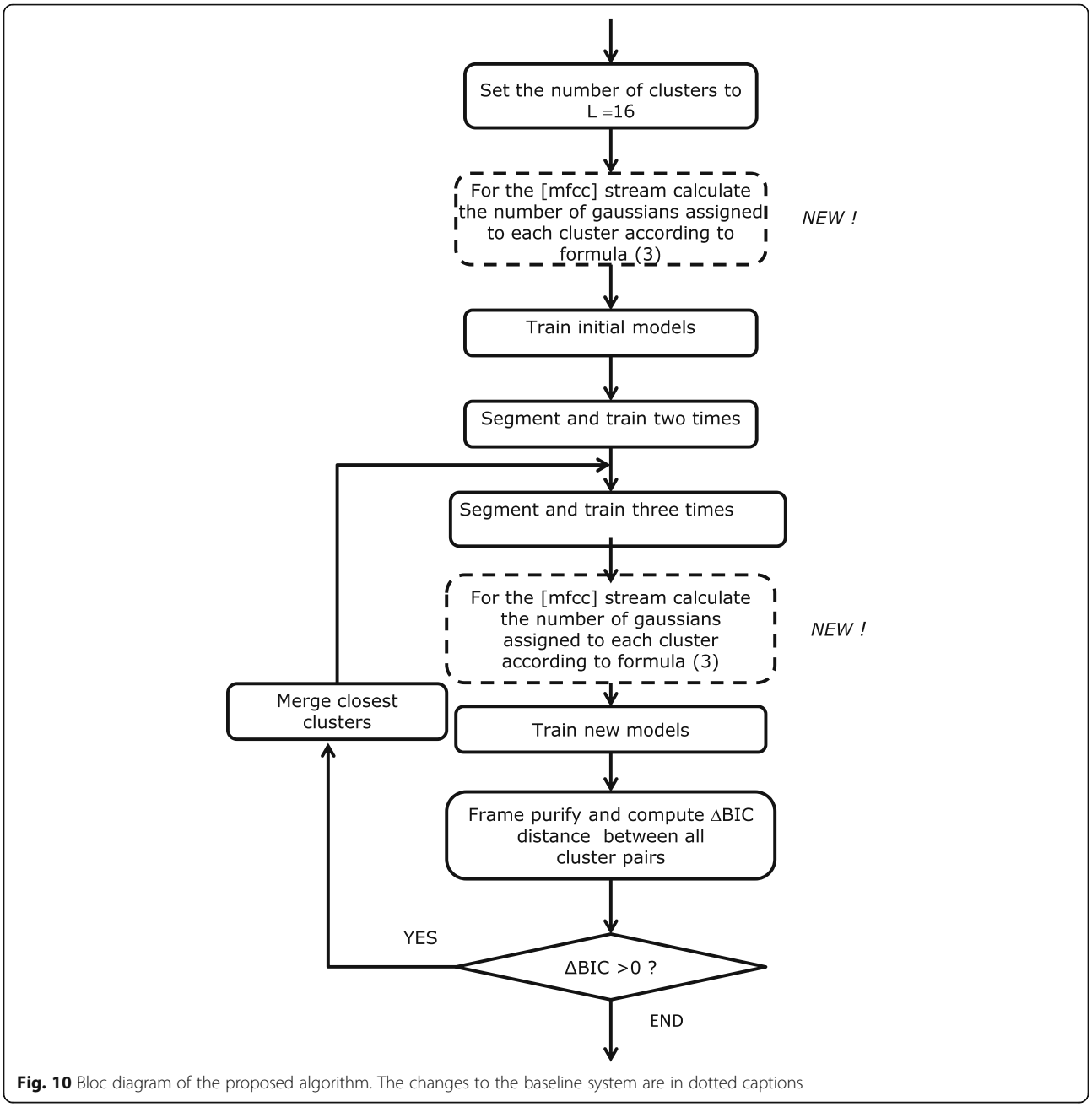

Fig. 10 Bloc diagram of the proposed algorithm. The changes to the baseline system are in dotted captions

algorithm after using 50 or more Gaussians as the maximum, there are improvements in the DEVELSET. We can also observe that the minimum values are obtained with 7 s per Gaussian which also corresponds to the minimum number of frames found in Fig. 9 for 5 Gaussians. The absolute minimum is found with 100 Gaussians which corresponds to the turning point in Fig. 8 in which increasing the number of Gaussians does not add information in the log-likelihood. One hundred Gaussians are also the minimum of the average line (AVE) in this picture. In Fig. 12, the DER for different values of parameter "*A*" are represented for "*B*" = 100. It can be clearly seen that below 7 s per Gaussian, the results are worse than those above it although the evolution of the DER values across parameter A is not descending monotonically. It is important to highlight that the standard DER and SER measure for speaker diarization for meetings is very sensitive to errors in the final number of speakers detected. This occurs because SER is a frame-based measure and one error in its calculation and depending on the duration of the speaker's speech may change the SER significantly. The best way to obtain good conclusions in this area of research is to experiment with as many diverse meetings as possible as mentioned before.

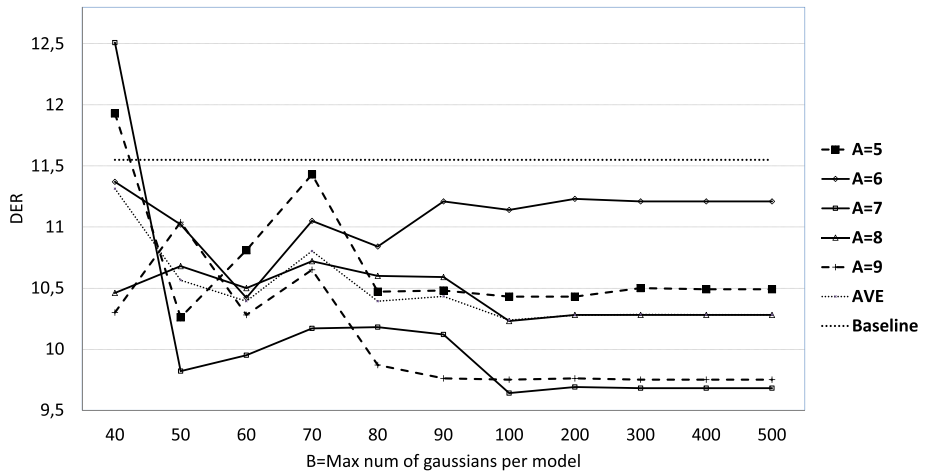

Fig. 11 DER for the DEVELSET versus different values of minimum number of seconds per Gaussian (parameter A) and maximum number of Gaussians (parameter B)

In Fig. 13, the ratio of computation time of the proposed method (using parameter $A = 7$) to the computation time of the baseline for the DEVELSET versus the parameter B (maximum number of Gaussians) is presented. The computation cost increases with the maximum number of Gaussians. When there are more Gaussians to train, the algorithm takes longer. It has a saturation limit at 200 because the maximum is rarely reached at over 200. By observing Fig. 11, we can see that after 100 Gaussians, the error does not diminish. Thus, a good compromising working point would be to use a maximum number of 100 Gaussians. In fact, if we would like to obtain a good working point, we could think of a merit factor that weights 90% the SER and 10% the ratio of computation time over the baseline. If we plot this merit factor against the maximum number of Gaussians (Fig. 14), we can observe this minimum at $B = 100$. There is another minimum at $B = 50$. By using a limit in the number of Gaussians, we can obtain a saving of 25.38% of computational time compared to not using the limit.

The relative improvement in SER over the baseline in the development set is 42.09% for the pair of parameters A-B = (7–50) and 46.47% for the pair of parameters A-B = (7–100) see Table 4. This is a very impressive result. For comparison purposes, we have calculated the SER for a subset of the development set (the RT07 set) obtaining a value of 2.1% which is outstanding performance (remember that the MISS+FA error for RT07 is 6.82). The meetings of this subset is part of our DEVE LSET and has therefore been used for training, still we include the speaker error of this subset separately only for a fast comparison with other works which were using this RT07 set. In Table 4, we also present the results of SER for the test set RT09. Improvements can be obtained for both combinations of parameters provided. Relative improvements in SER range from 15.36 to 17.54% for the two proposed working points.

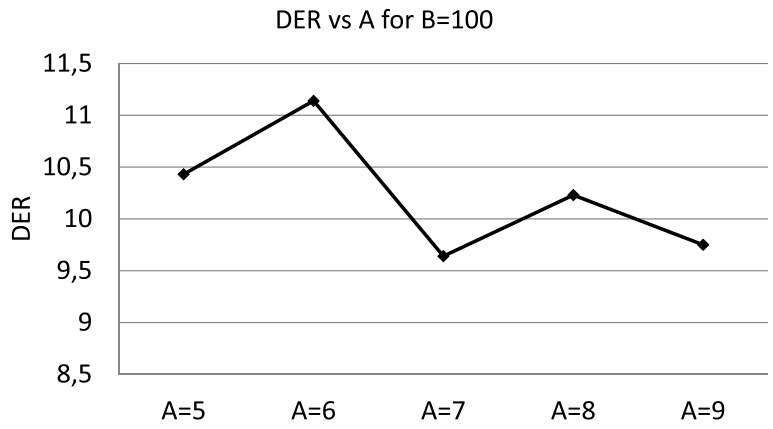

Fig. 12 DER for the DEVELSET across different A values for $B = 100$

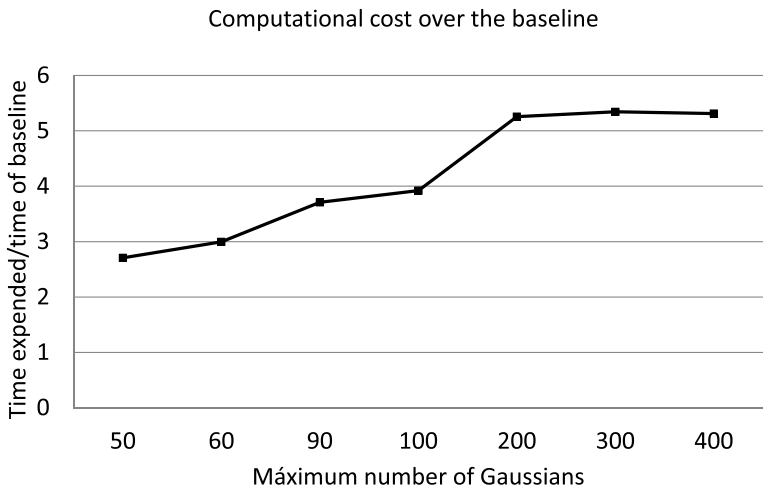

Fig. 13 Number of times that the proposed system is more computationally expensive than the baseline

In Table 5, we present detailed results for the RT09 set meeting by meeting both for the baseline system and for our proposed method using the optimum values (A, B) = (7–100). Four of the meetings present a decrease in SER, two others remain with a similar SER and one of the meetings increases the SER. On average, the SER decreases as we have already mentioned. The results obtained are among the best published to date [10, 39] (you need to add 8.7% of MISS+FA error to obtain the DER).

In Table 6, the number of identified, missed, and false-alarm speakers is presented. The proposed method reveals two more correctly identified speakers than the baseline and two fewer missed speakers although there are three new false-alarm speakers. The influence of new false-alarm speakers in the SER is small. This fact can be easily explained by the fact that the SER is a time-weighted measure, and the new false-alarm speakers possibly intervene for a short period of time and it is not significant in the overall computation. It can also be seen in Table 5 that the meetings that usually have a very high SER also have a very high overlap error, and since we do not propose any solution for the overlaps, we cannot decrease this error with our method.

5.3 Preliminary experiments with a single channel

In order to check how the model selection method behaves for a single channel, we have selected the mfcc vector coming from the acoustic fusion and discarded the tdoa vector. Figure 15 represents the SER for a single mfcc stream and a mixture of parameters (A) and minimum duration across different values of maximum number of Gaussians (B). It can be noticed that the method also improves the baseline results for single-channel recordings although the minimum SER values are obtained at a slightly different parameter values. The first thing to notice is that the optimum minimum duration is now 350 frames compared to 250 frames of the baseline. This change was also noticed in Section 4.3 above with the experiments changing the transition weight. The second change is the number of seconds per Gaussian that in this case is 11 compared to the optimum in previous

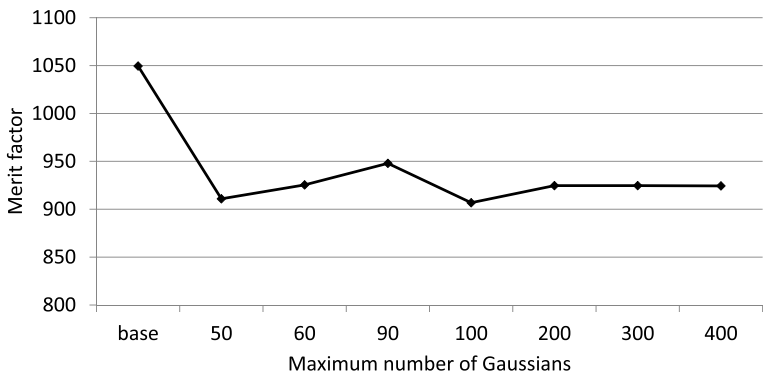

Fig. 14 Merit factor versus number of Gaussians

Table 4 SER for the baseline and for the proposed system with two different parameter settings. The percentage improvement over the baseline is presented in brackets

System	SER DEVELSET (% error improvement)	SER RT09 (% error improvement)
Baseline	4.11 ± 0.03	7.82 ± 0.07
Proposed system I (A, B) = (7–50)	2.38 ± 0.02 (42.09%)	6.61 ± 0.06 (15.36%)
Proposed system II (A, B) = (7–100)	**2.2 ± 0.02 (46.47%)**	**6.44 ± 0.06 (17.54%)**

section that was 7. More seconds per Gaussian are needed in order to create good models. This parameter change may be due to the numeric interaction of the mfcc vector with the tdoa vector that only occurs when using both vector streams. The system that uses both streams has better performance and stability, and it is less sensitive to parameter variations. Both streams complement each other. Looking at Fig. 15, we find several values that improve the baseline. We can choose the value 11-350-40, and the value 12-350-30 that give a SER of 8.3% and 8.14%, respectively, which represents 8.1% and 10.3% relative improvement over the baseline. Table 7 presents the results with the RT09 set. It can be seen that for the point 11,350,40, the relative SER improvement over the baseline is 36.75% which is a very significant improvement. With these results, we demonstrate that the method works not only when using both mfcc and tdoa streams but also for a single mfcc stream.

6 Fusing model selection and speaker segmentation independent on the number of clusters

6.1 Experiments with two streams

After analyzing previous results with transition weight independent of the number of clusters and the results using a method to select an appropriate number of Gaussians per cluster, the obvious next step is to merge both methods. However, the first method as seen in formula (2) uses a tuning parameter K to adjust transition probabilities to penalize or favor speaker changes in the same manner as in a speech recognition system in which the acoustics are appropriately weighted with the linguistic model probabilities in order to insert more or less word hypothesis. If we now consider, our new method of selecting the number of frames per cluster that is dependent on the duration of each cluster and taking into account that there is a maximum number of Gaussians per cluster to model a speaker, the likelihoods calculated in formula (2) may vary. In fact, if the model is a better fit to the speech, the likelihoods should be greater and the transition weight could change. In the same manner, the optimum value for minimum duration may also change.

In Fig. 16, we present the SER across different transition probabilities and for minimum duration 250 (We also did experiments with a minimum duration of 200 with worse results). In Fig. 17, we show the SER for the working point 7-50 compared with the baseline transition probability (1/M). Analyzing Figs. 16 and 17, we notice that now the best K is found at 0.01 with both systems but now the working point 7-50 is slightly better than the 7-100 (although not significant) in contrast with our previous results (SER 2.17% vs 2.2%). What is interesting is that if we compare the results of this experiment for the working point 7-50 with those obtained with the baseline transition weight (1/M), the results

Table 5 Detailed % SER results comparing baseline systems and the improvements for rt09. The last two columns show the overlap speech/non-speech errors common to both of them

Meeting	# Mic.	SER baseline	SER proposed method (system II 7–100)	Overlap	MISS + FA Error
EDI 20071128-1000	24	0.5	0.3	3.06	6.9
EDI 20071128-1500	24	1.6	1.86	7.01	12.1
IDI 20090128-1600	8	1.3	7.09	3.66	4.8
IDI 20090129-1000	8	4.8	2.65	3.54	9.6
NIST 20080201-1405	7	44.7	35.10	14.76	19.3
NIST 20080227-1501	7	2.4	1.84	8.41	8.8
NIST 20080307-0955	7	13.9	4.95	3.53	4.7
All		7.8 ± 0.07	**6.44 ± 0.06**	5.58	8.7
Improvement over the baseline			**17.54%**		

Table 6 Number of identified speakers, missed speakers, and false-alarm speakers for rt09 and all the systems tested

Meeting	Baseline			Proposed method (system II)		
	ID_SPK	MISS	FA	ID_SPK	MISS	FA
EDI_20071128-1000	4		1	4		1
EDI_20071128-1500	4			4		1
IDI_20090128-1600	4		1	4		1
IDI_20090129-1000	4			4		1
NIST_20080201-1405	3	2		3	2	1
NIST_20080227-1501	6			6		
NIST_20080307-0955	7	4		9	2	
ALL	32	6	2	34	4	5

improve significantly 2.17% vs 2.38%. For comparison purposes with other published works, we have calculated the SER for the working point 7-50 for a subset of the development set (the RT07 set) obtaining a value of 1.84% which is again outstanding and even better than the result of the previous section. With these experiments, we can show that both systems can be combined in order to improve the results. With these experiments, we now show that both working points are similarly good ones. Both working points have been tested with the test set giving the results presented in the last two rows of Table 8. The first three rows in Table 8 reproduce the results of the previous section.

If we now analyze the results for the RT09 set, we can see that at the point 7-50, $K = 0.01$ is significantly better than the others and better than those using $K = 1/M$. The results on the test set confirm that both methods contribute to improvements in the system. The fact that the results on the DEVELSET do not change for the 7-100 system may be due to the already very low SER which is quite difficult to decrease. In Table 9, we present the results meeting by meeting for the RT-09 set. We can see that our proposed method improves in all the meetings except one. This single meeting is the one that create the biggest part of the error (SER 46.97), and it is the meeting that has also the biggest overlap error. The average SER for the rest of the meetings is comparable or even better than the results for the baseline and better than the state of the art (see Section 7.2 below).

6.2 Preliminary experiments with the fusion system and a single stream

In this section, we present experiments fusing the transition weight scheme with the model selection scheme for a single mfcc stream. Figure 18 presents results for the DEVELSET using the fusion scheme. We can find several sets of parameters that have a SER below the baseline. But unfortunately, in our preliminary search, we could not find a minimum better than the minimum that we found using the model selection scheme. Both optimums found are not statistically different (8.22 vs 8.14 SER). The results obtained for the RT09 set for

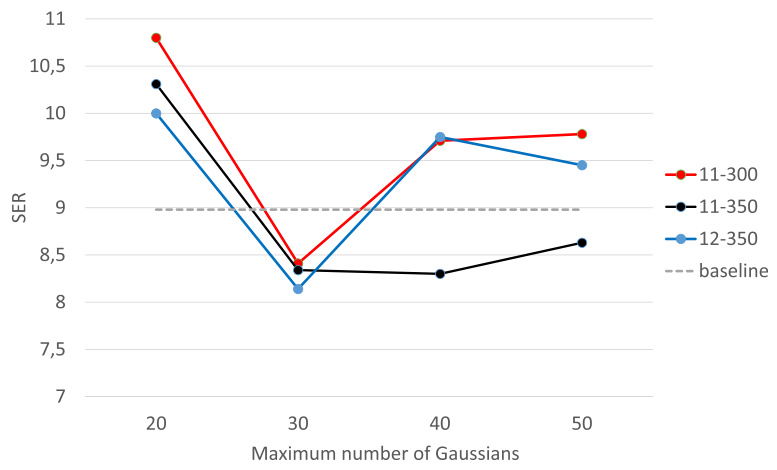

Fig. 15 SER for the DEVELSET for a single mfcc stream across different values maximum number of Gaussians and other parameters such as number of Gaussians per cluster (parameter A) and minimum duration of a speaker turn

Table 7 SER for the DEVELSET and the RT09 set when using the model selection method and a single mfcc stream

Secs per Gaussian	Minimum duration	Maximum number of Gaussians	DEVELSET	Relative improvement over DEVELSET	RT09	Relative improvement over RT09
(Baseline)	250		8.98%		21.21%	
11	350	40	8.3%	8.19%	**15.51%**	**36.75%**
12	350	30	8.14%	10.31%	18.34%	15.64%
11	350	30	8.34%	7.67%	18.37%	15.4%

these optimums are presented in Table 10 and are not as good as the results that we get for a different minimum in the DEVELSET (i.e., the 8.29 DEVELSET SER in Table 10). The search for an optimum using the fusion scheme is more complex since there are many parameters involved. In any case, the results for both the development set and for the evaluation set improve significantly the baseline system, both in the DEVELSET and the RT09 set.

6.3 Comparing results with a single stream versus two streams

Table 11 shows detailed results by meeting comparing the results when using the tdoa features versus the case in which we use a single mfcc stream. It can be noticed a big performance degradation when the tdoa information is not present. There is an exception with the meeting NIST 20080201-1405 whose results are very bad anyhow, but the results with a single mfcc stream are superior. We think that this is due to the speakers moving around the room that corrupts tdoa information.

7 Comparisons with other published results

7.1 Comparison with the AMI dataset

The databases used in this paper devel06, rt06, rt07, and rt09 are not publicly available in full (they are only available to the institutions that participated in the corresponding competitions). In order to test our proposals with other publicly available databases, we have used a subset of the AMI meeting corpus available from the University of Edinburgh [48] just for testing without changing the development set. The set of meetings used (specified in Table 12) includes recordings only from the Idiap Research Institute (IDIAP) site and has been used by other authors [49].

Our results with those databases using the optimum parameters obtained in our development set are presented in Table 13. The MISS+FA error is the same for every experiment and equal to 12.64%.

If we analyse the results, we can see that using the first alternative (changing just the transition cost parameter from the baseline) and using the cost that was obtained in the DEVELSET database does not improve performance. This result may easily be due to the mismatch

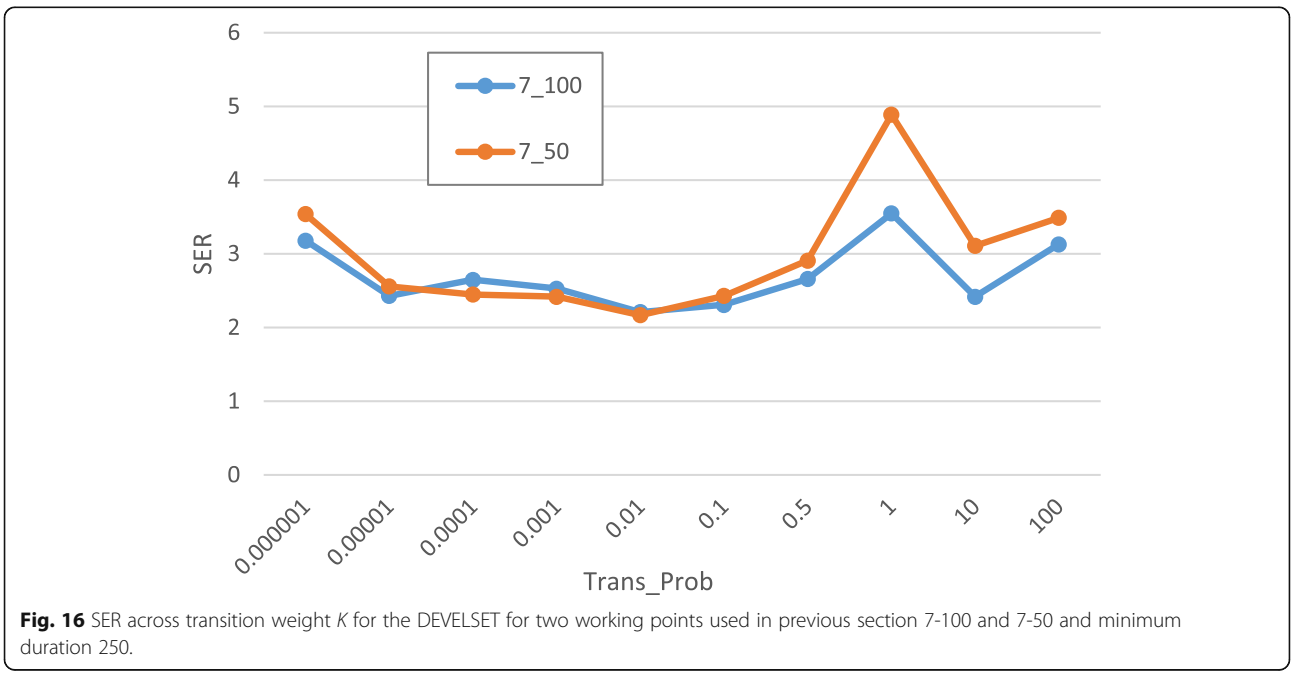

Fig. 16 SER across transition weight *K* for the DEVELSET for two working points used in previous section 7-100 and 7-50 and minimum duration 250.

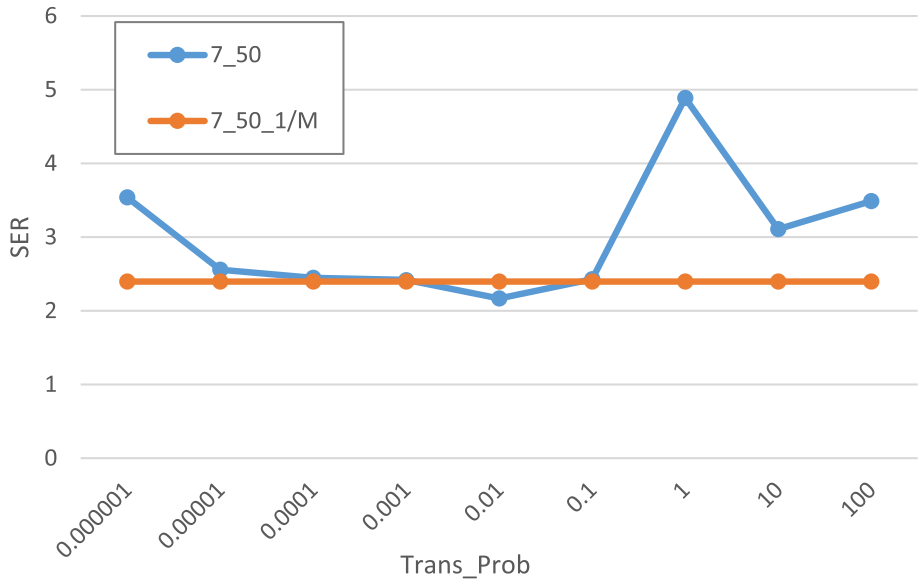

Fig. 17 SER across transition weight K for the DEVELSET for the point 7-50 compared to SER of the baseline (1/M)

between the development set and the test set. We think that the transition cost is an important parameter but that it has to be adjusted with a development set similar to the test set. In contrast, the parameters obtained with the DEVELSET database for the second approach separately (model selection) do improve in both cases (7–50) and (7–100) demonstrating that the second method has somehow obtained robust parameter settings. When the model selection mechanism is joined with the transition cost approach, there are also extra improvements for the (7–100) case. The fact that there is no improvement with the (7–50) set may be due again to the use of a different database for which no development set has been developed and the transition cost may be dependent on the database and on the maximum number of Gaussians per state. In any event, it can be demonstrated again that both the transition cost and the model selection are good strategies that may influence the results in a positive manner.

7.2 Comparison with other RT multiple streams published results

The best results on RT09 published up to now are the ones by Nwe et al. [39], in their Table IV. Table 14 shows their published results compared to ours. It can be noticed that our results are extremely bad for a single meeting (the NIST 20080201-1405 meeting) possibly because the speakers in the meeting move around the room. Our way of using the tdoa vector needs that the speakers stand in one place; otherwise, the tdoa vector corrupts the decision of the system. If we had to report on the SER excluding meeting NIST 20080201-1405 our results would be better than the published results.

Another comparison could be made if we consider the RT07 set (8 meetings). We obtained a SER of 1.84% for this subset, although they were part of our 28 development set meetings. The best RT07 SER published up to now used as a test set is 2.8% (see [39], Table III). In terms of computational complexity, our method is only 2.5 times more

Table 8 Overall results for the develset and the RT09 set for different alternatives

System	SER DEVELSET (% error improvement)	SER RT09 (% error improvement)
Baseline	4.11 ± 0.03	7.81
Proposed system I (A,B) = (7–50) K = 1/M	2.38 ± 0.02 (42.09%)	6.61 ± 0.06 (15.36%)
Proposed system II (A,B) = (7–100) K = 1/M	2.2 ± 0.02 (46.47%)	6.44 ± 0.06 (17.54%)
(A, B) = (7–50) with K = 0.01	**2.17 ± 0.02 (47.20%)**	**6.09 ± 0.06 (22.02%)**
(A,B) = (7–100) with K = 0.01	2.22 ± 0.02 (45.98%)	6.6 ± 0.06 (15.49%)

Table 9 Detailed results for the RT-09 set with the best system ((A-B) = (7-50) with $K = 0.01$)

Meeting	# Mic.	SER baseline	SER best method	Overlap	MISS + FA error
EDI 20071128-1000	24	0.5	0.24	3.06	6.9
EDI 20071128-1500	24	1.6	1.36	7.01	12.1
IDI 20090128-1600	8	1.3	0.59	3.66	4.8
IDI 20090129-1000	8	4.8	1.85	3.54	9.6
NIST 20080201-1405	7	44.7	46.97	14.76	19.3
NIST 20080227-1501	7	2.4	2.12	8.41	8.8
NIST 20080307-0955	7	13.9	4.89	3.53	4.7
All		7.8 ± 0.07	**6.09 ± 0.06**	5.58	8.7
Improvement over the baseline			**22.02%**		

expensive than our baseline. The computational demand of our baseline is just an iterative algorithm of segmenting and merging process AHC. In contrast, the state of the art system in [39] uses several steps one after the other. The first step is by itself an initial clustering process using only tdoa values but with two phases, a previous intra-pair segmentation and clustering and a subsequent inter-pair clustering fusion. The second step is similar to ours with a cluster modelling and cluster merging process. But the third step is quite complex since it includes 10 iterations of training and clustering runs with different number of Gaussians settings (55 different) and MAP adaptation for each run. In total, there are 550 training and clustering runs. Then, there is a process of consensus-based clustering. Although we do not have data to compare absolute computational cost of both systems, we could say that our model is much simpler and easier to reproduce.

7.3 Comparison with an x-vector system for a single channel

With the objective of comparing our system with the recent proposals of neural network-based x-vectors, we have processed our DEVELSET and tested our RT09 set with a system that is available at and that was proposed as a baseline for the Second DIHARD Challenge [4, 50, 51]. We used the same waveform files that we have used in our research for a single-vector stream and the voice activity detector of our system. Table 15 presents our findings. The x-vector approach consists of an x-vector extraction mechanism followed by a PLDA scoring and an adaptation to the development database. The system adjusts its thresholds to the development set. While the SER results for the DEVELSET are better than our results, the results with the test database are much worse. It could be said that the x-vector system overfits its training to the development database but that it has lower prediction power in the test database. Table 16 presents the results of this comparison meeting by meeting. The x-vector system is worse than ours in 5 out of 7 meetings.

7.4 DER comparison with information bottleneck [9] for a single channel

In [9], information bottleneck principle for a single channel is proposed. Our DER results for a subset of the

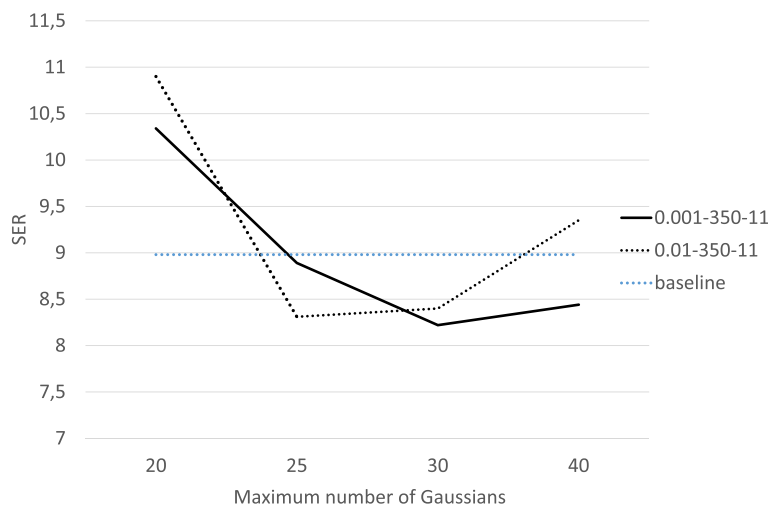

Fig. 18 SER for the DEVELSET using a single mfcc stream across the maximum number of Gaussians and different sets of parameters

Table 10 Results using a single mfcc stream and all the proposed methods

System	SER DEVELSET (± 0.04)	Relative improvement over baseline	SER RT09 (± 0.09)	Relative improvement over baseline
Baseline	8.98%		21.21%	
Transition weight (transition weight, minimum duration) (0.001-350)	8.29	8.3%	**14.87%**	**42%**
Model selection (A, B, minimum duration) (12-30-350)	**8.14%**	**10.31%**	18.34%	15.64%
Model selection (A, B, minimum duration) (11-40-350)	8.3%	8.19%	15.51%	36.75%
Fusion (transition weight, A,B, minimum duration) (0.001-11-30-350)	8.22%	9.2%	20.62%	2.8%
Fusion (transition weight, A,B, minimum duration) (0.01-11-30-350)	8.4%	6.9%	18.70%	13.4%
Fusion (transition weight, A, B, minimum duration) (0.001-11-40-350)	8.44%	6.4%	16.64%	27.46%

DEVELSET, the devel06 set (see Table 17) are 12.24% compared to the weighted published results of [9] 16.53% (see Table 1). For the RT06 set, our DER is 21.56% compared to 22.8% for their system. However, it is fair to mention that the information bottleneck method is faster than our method.

8 Discussion

In the first part of the paper, we have analyzed the effect of the transition cost on the SER by demonstrating that there is a strong influence of the transition cost on the performance of the system both for two streams and for a single stream but the results should be taken carefully. The tuning transition weight K may depends on the quality of the models (as shown in the last part of the paper) and on the minimum duration applied. We have discovered that the method that we were using previously (1/M) is not supported by any solid theory as it varies during the diarization process and that it is better to look for a good match of the transition weight for the problem at hand. In summary, the transition weight K should be adapted to the development data. The adaptation should explore also possible variations on the minimum duration applied to a speaker turn. Both adaptations should be done in tandem. The experiments done with a single stream (mfcc) demonstrate also the validity of our proposal being able to improve the relative SER in the test set by 42%.

Published works on speaker diarization [39] showed some evidence that the number of parameters used to model a speaker is a significant topic. This is also known from other areas of pattern recognition such as speaker verification or speech recognition. The solution in [39] uses a consensus method based on many repetitions of the algorithm and it is very computationally demanding.

We have researched and proposed a simple modification to our previous baseline system that consistently improves the results significantly without dramatically increasing the computation cost. Instead of defining the model only at the initialization step based on empirical data and sticking to it throughout the entire process as

Table 11 Results meeting by meeting comparing the system that uses TDOAS versus the system that does not use it

Meeting	SER two streams	SER single stream
EDI 20071128-1000	0.24	1.8
EDI 20071128-1500	1.36	11.77
IDI 20090128-1600	0.59	15.94
IDI 20090129-1000	1.85	12.47
NIST 20080201-1405	46.97	40.16
NIST 20080227-1501	2.12	10.31
NIST 20080307-0955	4.89	20.06
Overall SER	6.09	14.87

Table 12 List of meetings from AMI meeting corpus

#	Set	Meeting	# of microphones
1	AMI single site	IS1000a	12
2		IS1001a	12
3		IS1001b	12
4		IS1001c	12
5		IS1003d	12
6		IS1006b	12
7		IS1006d	12
8		IS1008a	12
9		IS1008b	12
10		IS1008c	12
11		IS1008d	12

Table 13 Overall results for the AMI single site dataset and the RT09 set for different alternatives

System	SER AMI single site (% error improvement)
Baseline	22.05
Transition cost $K = 3$	22.47 (− 1.90)
Proposed system I (A,B) = (7–50) $K = 1/M$	17.54 (20.45)
Proposed system II (A,B) = (7–100) $K = 1/M$	16.51 (25.12)
(A,B) = (7–50) with $K = 0.01$	19.71 (10.61)
(A,B) = (7–100) with $K = 0.01$	**15.78 (28.44)**

in previously published results or calculating it based only on the number of frames assigned to the speaker, we have chosen it dynamically depending on the duration of the current hypothesis and setting a maximum number of Gaussians. This process has been applied to the mfcc stream leaving the method applied to the tdoa stream unchanged. The strategy used to prevent overfitting has been to do hyper parameter tuning based on a large set of meetings. The set of meetings come from many different places, rooms, speakers, time of recordings, types of microphones, etc.

We have determined two working points of our parameters and have achieved improvements with this method in all sets that we have used. Our algorithm has resulted in astonishing improvements using only 2.5 times the computation time compared to the baseline particularly for the development set (a 42.09% reduction in speaker error). The algorithm also provides an optimum at 4 times the computation time of the baseline (a 46.47% reduction in speaker error). The improvements in SER for the test set with the model selection technique is more modest (17.54%) but still relevant and demonstrates the validity of the approach.

We have tested also the model selection proposal with a single stream meeting obtaining improved performance over the baseline, both in the development set and the test set.

When both methods are combined together, the results go down to 2.17% and 6.09% SER for the development and the test set respectively (a relative improvement of 47.20% and 22.02%). The SER obtained for a subset of the development set, the RT07 set (1.84%) is outstanding without the need for complicated algorithms and using a very simple modification of our baseline.

If we had to report on the DER obtained for the RT09 set, we should notice that it is still high but we should be aware that a large part of it is due to overlap and MISS+FA error (5.58% overlap and a MISS+FA error of 8.7%) and it heavily depends on one single meeting (14.76% overlap and 19,3% total MISS-FA error). One possible reason for it is that our method assumes that the speakers do not move from their places. Another possible reason may be due to overlap. Our algorithm does not take overlap into account because there is only one hypothesized output for every frame. The fact that this meeting has 14.76% of overlap surely corrupts our models. The overlap error is still a problem that remains mostly unsolved [49, 52, 53]. Taking into account that the biggest part of our DER error comes from overlap and speech/non-speech detection, our efforts should go in this direction in the future. If we had to report on SER results for RT09 without taking into account this meeting, our SER results would be better than the state of the art.

We have extrapolated our approach to a new test database (AMI), demonstrating that the proposed methods consistently improve the performance of the baseline method although again a tunning of the K parameter is needed.

Table 14 Comparison of our results for RT09 with other published results [39]

Meeting	# of scored speaker seconds	SER New et al. [39] Table IV	SER our method
EDI 20071128-1000	934.8	1.8	0.24
EDI 20071128-1500	777.93	2.3	1.36
IDI 20090128-1600	1207.97	0.7	0.59
IDI 20090129-1000	957.82	2.6	1.85
NIST 20080201-1405	574.97	7.4	46.97
NIST 20080227-1501	675.17	1.6	2.12
NIST 20080307-0955	804.2	3.6	4.89
Overall SER		2.5	6.09
Weigthed SER excluding NIST 20080201-1405 meeting		2.01	**1.7**

Table 15 Comparison of the results of our system with an x-vector system

System	SER DEVELSET (± 0.04) (%)	SER RT09 (± 0.09) (%)
Transition weight (Transition weight, minimum duration) (0.001-350)	8.29	14.87
Model selection (A,B, minimum duration) (11-40-350)	8.3	15.51
x-vector [50]	5.6	29.04

Finally, we made an effort to compare our system with a more recent x-vector diarization system. While the SER of the x-vector system is lower than ours for the development set, the results for the test set are much higher indicating that the x-vector system is not working well with unseen data. We have compared also the results of our single-stream system with the information bottlenect system in [9] obtaining a superior performance on a subset of the DEVELSET.

9 Conclusions

In this paper, we have demonstrated that a new transition weight and the minimum duration of a cluster are important parameters that should be explored in diarization algorithms. We have also investigated a method to automatically determine the number of GMMs needed to model a speaker. We have established a system that takes into account both the duration of the speaker's speech and the maximum number of Gaussians used. We have added it to our current diarization algorithm and tested it and demonstrated its value. We have obtained improvements in all sets used, both development and test, and reached relative improvement values ranging from 17.54 to 46.47% in speaker error for the test set and development set respectively. When looking for the optimum of these parameters, significant improvements can be made. Our final combined methods obtain 47.2% and 22.02 % relative improvements in SER for the development and test set, respectively. The results obtained are particularly good with a subset of the development set, the RT07 set. Most of the remaining errors of SER for the test set concern a single meeting that has a lot of overlap that corrupts our speaker models. When our methods are applied to a new publicly available database, they show an improvement in performance of 28.44% relative error against the baseline method. Preliminary experiments with a single-stream (mfcc) endorse the validity of our findings. Comparisons with an x-vector system deliver superior performance of our system when tested on unseen data.

Table 16 Detailed comparison for the RT09 of our best system with an x-vector system

Meeting	SER x-vector	SER our best method
EDI 20071128-1000	29.16	1.8
EDI 20071128-1500	43.43	11.77
IDI 20090128-1600	26.51	15.94
IDI 20090129-1000	5.91	12.47
NIST 20080201-1405	39.26	40.16
NIST 20080227-1501	42.31	10.31
NIST 20080307-0955	53.9	20.06
Overall SER	29.04	14.87

10 Methods

The aim of this study is to revise, analyze, and improve some algorithms for speaker diarization of meetings with multiple microphone recordings. The meetings are held in different places and different cities as established in NIST evaluations. The participants in each meeting are variable in number and depend on the meeting place and date of recording. The number of participants is unknown for the algorithms and one of the objectives of the algorithms is to discover it. The characteristics of the participants are detailed in the NIST documents although their identity remains anonymous. All of the participants in the meeting have approved the availability of their recordings for research purposes.

The materials obtained after the recording are the files containing the digitized microphone outputs. The recordings are processed by the algorithms proposed in this paper. The statistical analysis tool to present the results is the standard evaluation script provided by NIST and it is available on their web page [46].

Abbreviations

NIST: National Institute of Standards; RT09: NIST Rich Transcription Evaluation Campaign in 2009; RT07: NIST Rich Transcription Evaluation Campaign in 2007; RT06: NIST Rich Transcription Evaluation Campaign in 2006; RT02-05: NIST Rich Transcription Evaluation Campaign from 2002 to 2005; RT: Rich transcription; VB: Variational Bayes; AMI: Augmented multiparty interaction; MDM: Multiple distant microphones; DIHARD: Diarization is hard; MFCC: Mel frequency cepstral coefficients; F0: Fundamental frequency; GMM: Gaussian mixture models; I-Vector: I-Vector model as defined in the reference [23]; CNN: Convolutional neural network; X-Vector: X-Vector model as defined in the reference [26]; BIC: Bayes information criterion; PLDA: Probabilistic linear discriminant analysis; SDM: Single distant microphone; DER: Diarization error; TDOA: Time difference of arrival; GCC: Generalized cross-correlation; VAD: Voice activity detector; PCA: Principle component analysis; HMM: Hidden Markov Model; SER: Speaker error; MISS: Miss speaker error; FA: False alarm error; AVE: Average line; IDIAP: Idiap Research Institute; DEVELSET: Development set; PLDA: Probabilistic linear discriminant analysis; MAP: Maximum a posteriori

Table 17 DER for a single channel compared with the information bottleneck approach [9] for the devel06 set

Meeting	Scored seconds	Vijayasenan et al [9]	DER our best method
AMI_20041210-1052	474.97	9.6	2.6
AMI_20050204-1206	408.56	14.9	8.33
CMU_20050228-1615	428.87	26.5	17.78
CMU_20050301-1415	418.79	9.6	17.43
ICSI_20000807-1000	443.7	20	7.34
ICSI_20010208-1430	369.66	14.4	6.92
LDC_20011116-1400	411.69	9.2	9.53
LDC_20011116-1500	340.5	21.9	10.54
NIST_20030623-1409	423.42	11.9	2.67
NIST_20030925-1517	336.72	30.6	20.26
VT_20050304-1300	511.54	5.9	18.86
VT_20050318-1430	311.78	34.9	29.73
All		16.53	12.24

Acknowledgements
The authors thank Mark Hallett for the English revision of this paper and all the other members of the Speech Technology Group for the continuous and fruitful discussion on these topics. We also acknowledge the help of Verónica López-Ludeña, Juan M. Montero, and Ricardo de Córdoba with managing computer resources. The authors would like to thank all the reviewers for their valuable comments and suggestions which improved the paper considerably.

Authors' contributions
BM and JP contributed equally in the proposal and test of the algorithms. JV contributed to the analysis of computation costs of the algorithm and provided insight in the results. RS and JF contributed to the refinement of the methods and to the critical analysis of the results. RS helped in the scripts. The authors read and approved the final manuscript.

Funding
The work leading to these results has been funded by AMIC (MINECO, TIN2017-85854-C4-4-R) and CAVIAR (MINECO, TEC2017-84593-C2-1-R) projects.

Availability of data and materials
Some NIST meetings 2002–2005 are available from LDC https://catalog.ldc.upenn.edu/byproject. Other NIST meetings NIST 2006–2009 to the participating teams at the different evaluations https://www.nist.gov/itl/iad/mig/past-hlt-evaluation-projects.The AMI meetings are available from the University of Edingburgh http://groups.inf.ed.ac.uk/ami/download/.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Universidad Tecnológica de Pereira, Pereira, Colombia. [2]Universidad Politécnica de Madrid, Avda. Complutense, 30, 28040 Madrid, Spain. [3]Department of Computer Science, University of Oviedo, Campus of Viesques s/n, 33024 Gijón, Spain.

Received: 8 August 2019 Accepted: 7 January 2021
Published online: 24 February 2021

References

1. S. Tranter, D. Reynolds, An overview of automatic speaker diarization systems. IEEE. Trans. Audio. Speech. Lang. Process. **14**(5), 1557–1565 (2006)
2. X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, O. Vinyals, Speaker diarization: a review of recent research. IEEE. Trans. Audio. Speech. Lang. Process. **20**(2), 356–370 (2012)
3. M.H. Moattar, M.M. Homayounpour, A review on speaker diarization systems and approaches. Speech. Comm. **54**, 1065–1103 (2012)
4. The Third DIHARD Speech Diarization Challenge, [Online]. Available: https://dihardchallenge.github.io/dihard3/index. [Accessed 11 October 2020].
5. X. Anguera, M. Aguiló, C. Wooters, C. Nadeu, J. Hernando, in *Proceedings of Speaker Odyssey*. Hybrid speech/non-speech detector applied to speaker diarization of meetings (2006)
6. C. Wooters, M. Huijbregts, The icsi rt07s speaker diarization system, in Lecture Notes in Computer. Sciences **4625**, 509–519 (2008)
7. M. Huijbregts, F De Jong, Robust speech/non-speech classification in heterogeneous multimedia content, Speech Comm, **53**(2), 143–153 (2011)
8. E. El-Khoury, D. Senac, J. Pinquier, in *in Acoustics, Speech and Signal Processing, 2009*. Improved speaker diarization system for meetings (ICAS SP, Taipei, 2009) IEEE International Conference on, 2009:19–24
9. D. Vijayasenan, F. Valente, H. Bourlard, An information theoretic approach to speaker diarization of meeting data. IEEE. Trans. Audio. Speech. Lang. Process. **17**(7), 1382–1393 (2009)
10. G. Friedland, A. Janin, D. Imseng, X. Anguera, L. Gottlieb, M. Huijbergts, M.T. Knox, O. Vinyals, The ICSI RT-09 speaker diarization system. IEEE. Trans. Audio. Speech. Lang. Process. **20**(2), 371–381 (2012)
11. J.M. Pardo, R. Barra-Chicote, R. San-Segundo, R. Córdoba, B. Martínez-González, Speaker diarization features: the UPM contribution to the RT09 evaluation. IEEE. Trans. Audio. Speech. Lang. Process. **20**(2), 426–435 (2012)
12. S.H. Yella, A. Stolcke, M. Slaney, in *IEEE SLT workshop*. Artificial neural network features for speaker diarization (IEEE, South Lake Tahoe, 2014)
13. S.H. Yella, A. Stolcke, in *Interspeech*. A comparison of neural network feature transforms for speaker diarization (ISCA, Dresden, 2015)
14. S.H. Yella, P. Motlicek, H. Bourlard, in *Interspeech*. Phoneme background model for information bottleneck base speaker diarization (ISCA, Singapore, 2014)
15. G. Friedland, O. Vinyals, Y. Huang, C. Muller, Prosodic and other long-term features for speaker diarization. IEEE. Transact. Audio. Speech. Lang. Process. **17**(5), 985–993 (2009)
16. A. Zewoudie, J. Luque, J. Hernando, The use of long-term features for GMM- and i-vector-based speaker diarization systems. EURASIP J. Audio. Speech. Music. Processing **14** , 1–11 (2018)
17. R. Barra-Chicote, J.M. Pardo, J. Ferreiros, J.M. Montero, Speaker diarization based on intensity channel contribution. IEEE. Transact. Audio. Speech. Language **19**(4), 754–761 (2011)
18. H. Sun, T.L. Nwe, B. Ma, H. Li, in *Proccedings of Interspeech*. Speaker diarization for meeting room audio (2009)
19. C. Fredouille, S. Bozonnet, N. Evans, in *The Rich Transcription 2009 Meeting Recognition Evaluation Workshop*. The lia-eurecom rt 09 speaker diarization system (2009)
20. N. Evans, S. Bozonnet, D. Wang, C. Fredouille, R. Troncy, A comparative study of bottom-up and top-down approaches to speaker diarization. IEEE. Trans. Audio .Speech Lang. Process. **20**(2), 382–392 (2012)

21. A. Sapru, S.H. Yella, H. Bourlard, in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. Improving speaker diarization using social role information (2014)
22. J. Ajmera, C. Wooters, in *ASRU*. A robust speaker clustering algorithm (2003)
23. A. Woubie, J. Luque, J. Hernando, in *Odyssey*. Short- and long-term speech features for hybrid HMM-i-vector based speaker diarization system (2016)
24. M. Diez, L. Burget, P. Matejka, in *Odyssey*. Speaker Diarization based on Bayesian HMM with Eigenvoice Priors (Les Sables d'Olonne, France, 2018)
25. L. Sun, J. Du, C. Jiang, X. Zhang, S. He, B. Yin and C.-H. Lee, Speaker diarization with enhancing speech for the First DIHARD Challenge. [Online]. Available: home.ustc.edu.cn/~sunlei17/pdf/lei_IS2018.pdf. Accessed 17 Jan 2021
26. G. Sell, D. Synder, A. Mc Cree, D. Garcia-Romero, J. Villalba, M. Maciejewski, V. Manohar, N. Dehak, D. Povey, S. Watanabe, S. Khudanpur, in *Interspeech*. Diarization is hard: some experiences and lessons learned for the JHU team in the inaugural DIHARD challenge (2018)
27. S. Chen, P. Gopalakrishnan, in *DARPA Speech Rec. Workshop*. Speaker, environment and channel change detection and clustering via the bayesian information criterion (1998)
28. T.H. Nguyen, E.S. Chng, H. Li, in *Interspeech*. T-test distance and clustering criterion for speaker diarization (2008)
29. D. Vijayasenan, *An information theoretic approach to speaker diarization of meeting recordings* (Ph D. Thesis, École Polytechnique Fédérale de Lausanne, Lausanne, 2010)
30. D. Vijayasenan, F. Valente, H. Bourlard, An information theoretic combination of MFCC and TDOA features for speaker diarization. IEEE. Trans. Audio. Speech. Lang. Process. **19**(2), 433–438 (2011)
31. G. Sell, D. Garcia-Romero, in *IEEE Spoken Language Technology Workshop*. Speaker diarization with PLDAI-vector scoring and unsupervised calibration (2014)
32. X. Anguera, *Robust speaker diarization for meetings, Ph D Thesis* (Universitat Politécnica de Catalunya, Barcelona, 2006)
33. I. Lapidot, J.-F. Bonsatre, in *The Speaker and Language Recognition Workshop*. Generalized Viterbi-based models for time-series segmentation (Odyssey, Singapore, 2012)
34. I. Lapidot, J.-F. Bonastre, in *Interspeech*. On the importance of efficient transition modeling for speaker diarization (ISCA, San Francisco, 2016)
35. G. Sell, D. Garcia-Romero, in *ICASSP*. Diarization resegmentation in the factor analysis subspace (2015)
36. R.O. Duda, P.E. Hart, D.G. Stork, *Pattern classification* (John Willey & Sons, 2001)
37. X. Anguera, C. Wooters, J. Hernando, in *MLMI 2006, Lecture Notes on Computer Science 4299*. Automatic cluster complexity and quantity selection: towards robust speaker diarization (2006)
38. X. Anguera, T. Shinozaki, C. Wooters, J. Hernando, in *International Conference on Acoustics Speech and Signal Processing*. Model complexity selection and cross-validation EM training for robust speaker diarization (ICASSP, Honolulu, 2007)
39. T.L. Nwe, H. Sun, B. Ma, H. Li, Speaker clustering and cluster purification methods for RT07 and RT09 evaluation meeting data. IEEE. Trans. Audio. Speech. Lang. Process. **2**(2), 461–473 (2012)
40. J.M. Pardo, X. Anguera, C. Wooters, Speaker diarization for multiple-distant-microphone meetings using several sources of information. IEEE. Trans. Comput. **56**(9), 1212–1224 (2007)
41. M. Brandstein, H. Silverman, in *International Conference on Acoustics, Speech and Signal Processing*. A robust method for speech signal time-delay estimation in reverberant rooms (1997)
42. X. Anguera, C. Wooters, J. Hernando, Acoustic beamforming for speaker diarization of meetings. IEEE. Trans. Audio Speech. Lang. Process. **15**(7), 2011–2022 (2007)
43. B. Martínez-González, J.M. Pardo, J.D. Echeverry-Correa, R. San-Segundo, *Spatial features selection for unsupervised speaker segmentation and clustering* 73, 27–42 (Expert Systems With Applications, 2017)
44. B. Martínez-González, J.M. Pardo, R. San-Segundo, J.M. Montero, in *Odyssey*. Influence of transtion cost in the segmentation stage of speaker diarization (ISCA, Bilbao, 2016)
45. L. He, X. Chen, C. Xu, Y. Liu, J. Liu, M.T. Johnson, Latent class model with application to speaker diarization. EURASIP. J. Audio. Speech. Music. Process **12** (2019) https://doi.org/10.1186/s13636-019-0154-z
46. Rich Transcription Evaluation Project, National Institute of Technology (NIST), 2002-2009. [Online]. Available: https://www.nist.gov/itl/iad/mig/tools. [Accessed 31 July 2019].
47. N. Mirghafori, C. Wooters, in *International Conference on Acoustics, Speech and Signal Processing*. Nuts and flakes: a study of data characteristics in speaker diarization (ICASSP, Toulouse, 2006)
48. AMI dataset, [Online]. Available: http://groups.inf.ed.ac.uk/ami/download/. Accessed 17 Jan 2021
49. M. Zelenak, C. Segura, J. Luque, J. Hernando, Simultaneous speech detection with spatial features for speaker diarization. IEEE. Trans. Audio. Speech. Lang. Process. **20**(2), 436–446 (2012)
50. T. s. D. Challenge. [Online]. Available: https://arxiv.org/abs/1906.07839. [Accessed 15 October 2020].
51. Kaldi-ASR, [Online]. Available: https://github.com/kaldi-asr/kaldi/tree/master/egs/dihard_2018/v2. [Accessed 11 October 2020].
52. K. Bokye, B. Trueba-Hornero, O. Vinyals, G. Friedland, in *International Conference on Acoustics, Speech and Signal Processing, ICASSP*. Overlapped speech detection for improved speaker diarization in multiparty meetings (2008)
53. M. Huijbregts, D. van Leeuwen, F. de Jong, in *Interspeech*. Speech overlap detection in a two-pass speaker diarization system (ISCA, Brighton, 2009)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen journal and benefit from:

- ► Convenient online submission
- ► Rigorous peer review
- ► Open access: articles freely available online
- ► High visibility within the field
- ► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com