Influence of Transition Cost in the Segmentation Stage of Speaker Diarization

Beatriz Martínez-González, José M. Pardo, Rubén San-Segundo, J.M. Montero

Speech Technology Group
Universidad Politécnica de Madrid, Spain
[beatrizmartinez, pardo, lapiz, juancho]@die.upm.es

Abstract

In any speaker diarization system there is a segmentation phase and a clustering phase. Our system uses them in a single step in which segmentation and clustering are used iteratively until certain condition is met. In this paper we propose an improvement of the segmentation method that cancels a penalization that had been applied in previous works to any transition between speakers. We also study the performance when transitions between speakers are favoured instead of penalized. This last option achieves better results both for the development set (21.65 % relative speaker error improvement-SER) and for the test set (4.60% relative speaker error improvement

IndexTerms— speaker diarization, speech segmentation, speaker recognition

1. Introduction

Speaker diarization is the task of identifying the number of participants in a meeting and creating a list of speech time intervals for each participant. Speaker diarization can be used as a first step in the speech transcription of meetings in which each sentence has to be associated with a specific speaker. The diarization task is carried out without any previous knowledge about the position, number or characteristics of the speakers, the position or quality of the microphones used during the meeting or the characteristics of the room where the recording has taken place. An overview of automatic speaker diarization systems is given in [1], [2] and [3].

When the recording has been done with only one distant microphone we speak of diarization with a Single Distant Microphone (SDM) while if several microphones have been used we speak of diarization with Multiple Distant Microphones (MDM).

Most MDM systems use acoustic features as Mel-Frequency Cepstral Coefficients (MFCC) and localization features as the Time Delay Of Arrival (TDOA) values [4]. This information is extracted from the recordings and then used to analyze them and determine which parts corresponds to which speaker.

As we work in MDM diarization we have more than one recording. As mentioned, there is no previous information about the quality of the microphones, its position in the room or any characteristic of the meeting which could result in recordings with very low signal to noise ratio. One common way to enhance the signal is summing up all the channels, previously filtered and adjusted using the TDOAs as it is shown in [5].

The next step is the Voice Activity Detection module (VAD). VAD algorithms differ, depending on the type of non-speech sounds that appear next to the speech or mixed with it, from the Gaussian mixture models (GMM) to laplacian and gamma probability density functions [6].

The last stage of the diarization task uses all the information previously extracted for segmentation and clustering of speech regions. Some methods use bottom-up agglomerative clustering [7], while others use a top down universal background model (UBM) as a starting point to apply adaptation techniques iteratively to build the speaker models [8].

Clustering algorithms need a distance measure to determine whether two speech clusters belong to the same speaker. The most common used distance is the Bayesian Information Criterion (BIC) distance [9]. Other studies have also presented great improvements using other alternatives based on the t-test distance [10].

The segmentation stage decides, using speech data and the speaker models, which segments of the meeting belong to which speaker. Some works take advantage of more information than MFCC and TDOA features, like prosodic features [11], energy features [12] or even information about the role of the speakers [13] to adjust the segmentation, but MFCC and TDOA are the most common.

As stated in [3], when performing segmentation a minimum duration of speaker turns is usually enforced to avoid the assignment of very short consecutive segments to different speakers. In [14] and [15], authors discuss the convenience of using this minimum duration, setting its optimal value around 3 seconds. As there is no a priori information about the length of a speaker turn in a meeting, the final target was to avoid any other time restriction to the speaker turns apart from the inclusion of the minimum duration parameter. Some algorithms, like [16] or the original algorithm in [14] or [17], included a penalization in the length of the speaker segment, and was modified to make the length of the speaker turns only dependent on the acoustic information. Although the greatest penalization parameter was already cancelled, there is still another weighting parameter dependent on the number of speakers, which actually works as a penalizing factor and varies throughout clustering iterations because the number of speakers is changing after the clustering step. The goal of this work is to study the effect on segmentation stage of this parameter and improve the diarization by making segmentation independent of this varying number of active speakers.

This paper is organized as follows. Section 2 describes the database used. The baseline system is presented in Section 3. The changes proposed are described in Section 4. Finally, results of experiments and conclusions can be found in Section 5 and 6.

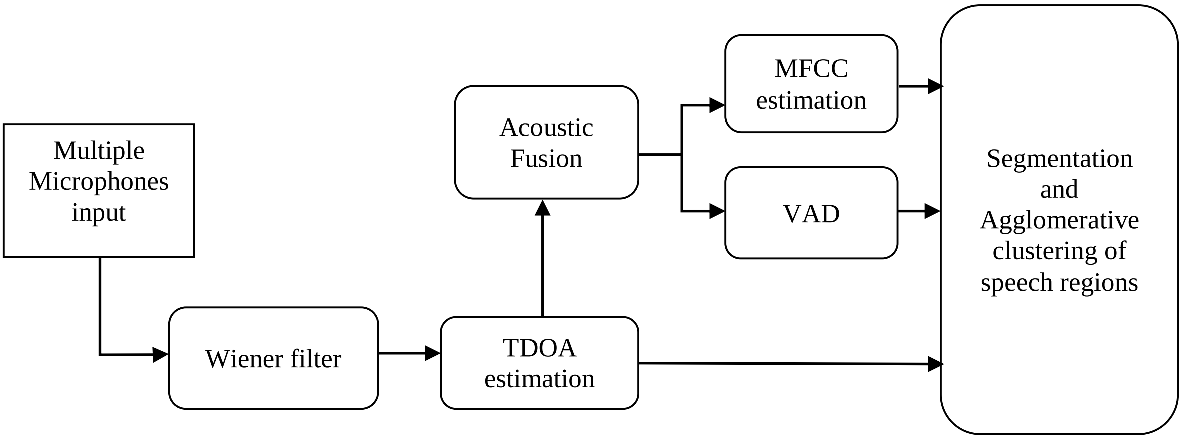
Multiple Microphones input
Wiener filter
TDOA estimation
Acoustic Fusion
MFCC estimation
VAD
Segmentation and Agglomerative clustering of speech regions

Figure1: *Baseline system architecture*

2. Database

In this work we have used a subset of 12 meetings extracted from NIST Rich Transcription 2002-2005 sets (named devel06 in [1]), the set of 8 meetings of NIST Rich Transcription 2006 and the set of 8 meetings of NIST Rich Transcription 2007 to form our development set that will be called ALL06_07 from now on. The evaluation set will be composed of a set of the NIST Rich Transcription, the one from year 2009, which have been called RT09.

The segments defined by NIST for the official evaluations have been used to measure the performance of the systems described in this work. In this paper we use the scored speaker time. These parts consist of 15484.34 seconds (1,548,434 frames) evaluated for the ALL06_07 set, and 5932.88 seconds (593,288 frames) for the RT09 set. Specific meetings included in both databases are listed in Table 1 and Table 3

3. Baseline system

A general diagram of the baseline system is included in Figure1.

The input coming from several different microphones is first Wiener filtered in order to reduce the background noise.

Then, in order to estimate the TDOA between two segments from two microphones, we use the generalized cross correlation with phase transform" (GCC-PHAT). First, we calculate the average cross-correlation between any channel and all the rest of them and select the microphone with highest average cross-correlation to be the reference channel [14]. Finally, a TDOA value will be calculated every 250 ms, between any available microphone and the reference one

The set of TDOAs from each microphone to the reference channel will form what we call the TDOA vector which has a dimension of N-1 being N the number of channels. Once this TDOA vector is calculated, a weighted delay-and-sum algorithm is applied in the acoustic fusion module, where the input signals are delayed and added together to generate a new composed signal. For more detailed information see [5].

The composed signal is then processed by the MFCC estimation module, where MFCC vectors of 19 components mfcc are calculated every 10 ms with a window of 30ms.

The composed signal is also processed by the VAD module. The VAD module is a hybrid energy-based detector and model-based decoder [18].

	Set	Meeting	# mics
1	ALL06_07	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		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		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

Table 1: *List of meetings used for the development set (ALL06_07).*

In the TDOA estimation module the system estimates also another TDOA vector to be used in the segmentation and clustering phase. This new vector will be composed of the TDOA´s from only the 4 pairs of microphones with the highest average cross-correlation. First the system calculates the cross correlation between all the pairs of channels, then selects the four of them with the highest cross correlation and then estimate the delays between those pairs, but this time it is recalculated with a frame rate of 10 ms in order to have the same number of frames as the MFCC vector. The resulting TDOA vector will have a dimension equal to 4. Full detailed information is included in [17].

The next module is the segmentation and agglomerative clustering process which consists of an initialization part and an iterative segmentation and merging process.

The initialization process segments the speech into NClass blocks (equivalent to an initial hypothesis of NClass speakers or clusters) uniformly distributed. NClass has been set to 16 empirically. Every cluster is modelled using a Gaussian mixture model (GMM) initially containing a number of components that has to be specified (we use 5 for mfcc and 1 for tdoa streams). After the initial segmentation a set of training and re-segmenting steps is carried out using Viterbi decoding. Then the merging step takes place.

When a merging takes place, the GMM for the new cluster is retrained with the data now assigned to it and the number of parameters (mixtures) of the merged model is the sum of the number of mixtures of the component models. The segmentation and clustering steps are repeated until a stopping criterion is reached.

The ΔBIC criterion has been used to decide which clusters to merge, and when to stop the merging. When all possible merge pairs give a negative ΔBIC, the process is stopped. A frame purification algorithm is also applied before computing the BIC distance [14]. A diagram of the segmentation and clustering process is shown in Figure 2.

The features used in the diarization task are the MFCC features combined with the TDOA features. More information about the baseline system can be consulted in [17].

3.1. Baseline segmentation method

The diarization system starts the segmentation and clustering stage with a division of the recording into several parts (cluster initialization). Each part is assigned to a different cluster (speaker) and used to train the corresponding GMM. The next stage, named "Segmentation and Training" in Figure 2, uses these GMMs to decide which frames belong to which cluster. The algorithm has been designed to force a minimum number of consecutive frames assigned to one cluster, in the baseline 250 frames.

The system uses an ergodic Hidden Markov Model (HMM) where each state corresponds to each cluster and then performs a Viterbi search. Each state is composed of a number of sub-states which imposes a minimum duration. As seen in Figure 3 the system has to go through all the sub-states before being allowed to change to the first sub-state of other state (a different cluster). Probabilities of changing or remaining in the same state at the last sub-state have been set to 1 following the recommendation in ([14], [19]) who proposed to set alpha=beta =1 (instead of the previous 0.9, 0.1) even if the sum is not 1.

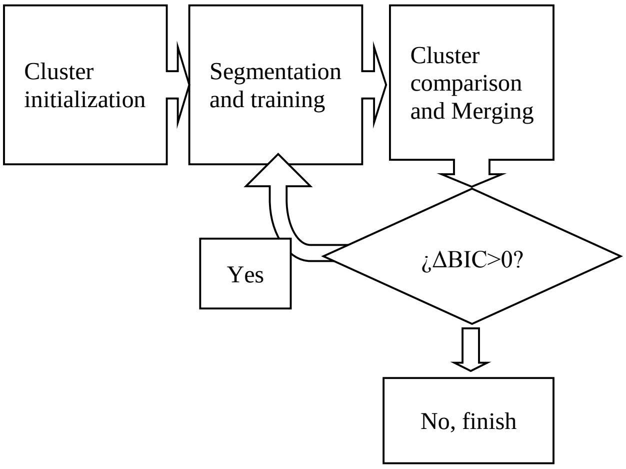

Figure 2: *Diagram of the training and segmentation process of the baseline system.*

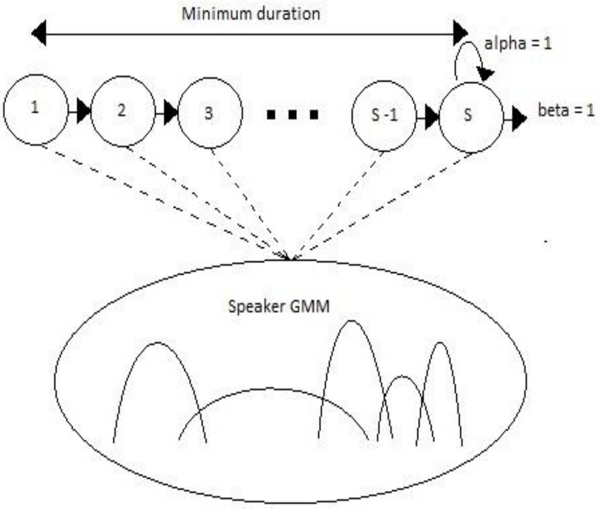

Figure 3: *Cluster models with minimum duration (from [19])*

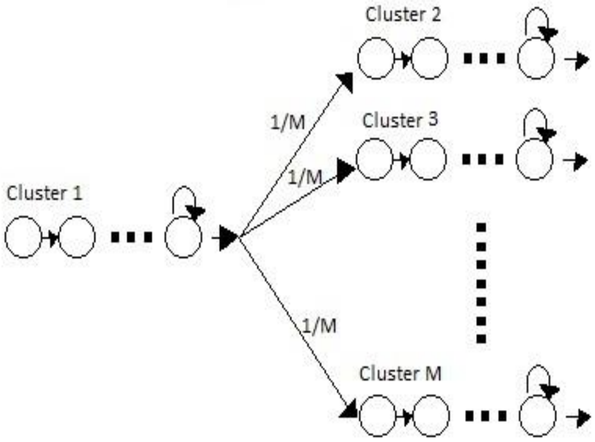

Figure 4: *Diagram of the dependency of turn speaker changes with the number of active speakers.*

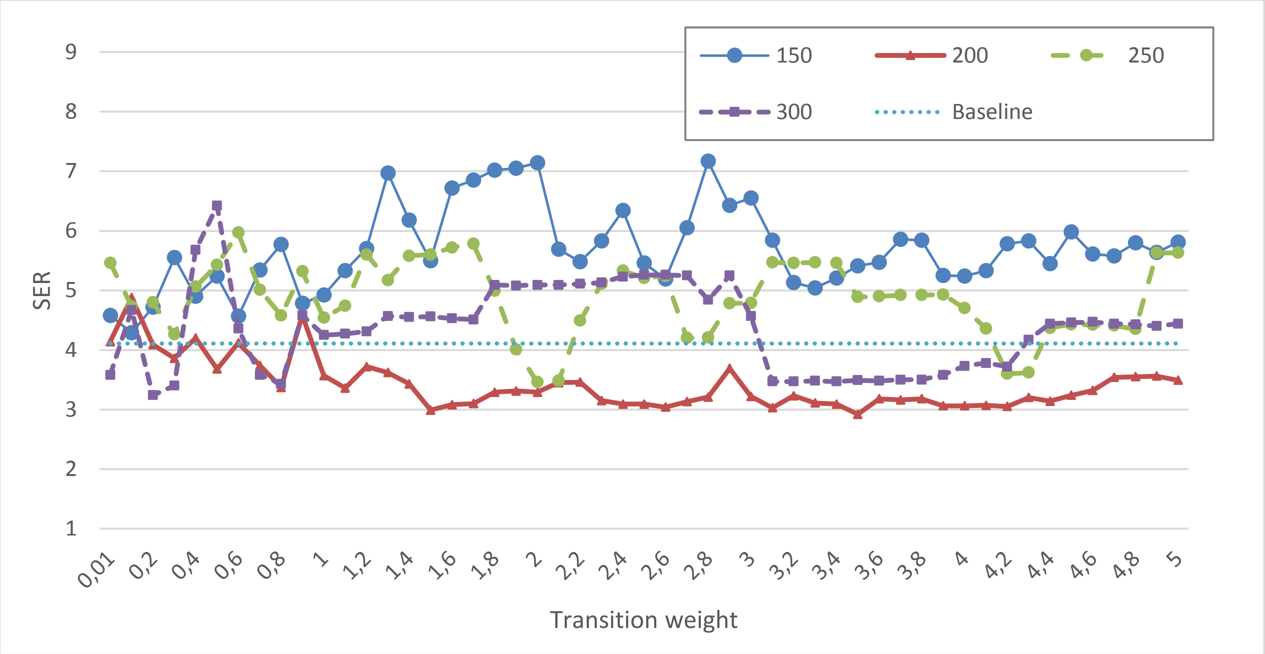

Figure 5: *SER of the Development set for minimum duration of speech of 150, 200, 250 and 300 frames against transition weight.*

This change was implemented to make the length of the speaker turns independent of alpha or beta, so the system will focus only in information from acoustics. However, these values of alpha and beta results in another penalization factor applied to speaker turns of 1/M (beta/M), being M the number of active clusters (see Figure 4).

This last penalization factor (transition weight) has some influence in the number of speaker changes and, as it is related to the number of active hypothesized speakers, it changes at each iteration during the whole diarization process.

The number of active speakers begins at16 and decreases, one by one, every time that the system decides to merge two clusters. The associated transition weight will increase accordingly, as if the probability of having speaker turns would be higher if less speakers are involved in the meeting which is not related to the social activity of the meeting or the number or role of the actors intervening in it.

This factor is not usually tuned in diarization systems and this paper focus on the study of this characteristic and the proposal of a better alternative.

4. Segmentation independent of the number of active clusters

As noted before, the segmentation, particularly the decision of changing from one speaker to another, is dependent of the number of active speakers. The factor 1/M, being M the number of hypothesized speakers at the moment, reduces the probability of moving to another speaker vs remaining in the last one. Also, the factor is variable, because the number of clusters decreases during the process.

In the Viterbi search implemented, the value stored in each final sub-state is the accumulated sum of the previous “Minimum Duration” log-likelihoods for each cluster and frame. The Minimum Duration, as it was previously mentioned, is the minimum number of frames that has to be assigned to one cluster after a turn to avoid unrealistic very fast changes from one speaker to the next one. The system calculates this value for each frame and keeps the cluster with the highest log-likelihood at its final sub-state.

A change of speaker will occur when the sum of the last “Minimum Duration” log-likelihoods from the current cluster is lower than the sum of the last “minimum Duration” log-likelihoods from any other cluster plus a transition weight (log). Therefore, a transition between speakers will take place if the following condition is met:

$$\sum_{i}^{i+MIN\_DUR} log\mathcal{L}(cl_j; fr_i) < log(K) + \sum_{i}^{i+MIN\_DUR} log\mathcal{L}(cl_u; fr_i) \qquad (1)$$

Where $log\mathcal{L}()$ is the log-likelihood, cl_j the current cluster, cl_u the candidate cluster, K the transition weight (in the baseline system this is 1/M) and fr_i the frame being evaluated.

In an extreme situation a very high or very low value of the transition weight could surpass, for every possible value of acoustic data, the difference of the log-likelihoods of the current cluster and the candidate cluster. In practice this would force, or forbid, the transition to other speakers, independently of the information given by the acoustic data and the speaker models.

An extremely low transition weight (negative in logarithm) would make impossible the change between speakers, resulting in a final solution with only one speaker, the first one. The opposite situation would result in a solution where any speaker turn is no longer than the minimum duration established at the beginning.

Transition weight	Minimum duration	All06_07	Relative improvement over All06_07	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%

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.*

In the baseline system, the lowest possible value, and thus the highest opposition to changes, takes place at the beginning of the algorithm, when the number of speakers is 16. From that moment, the transition weight will be increased in every iteration, which would mean that there is higher probability of changes when fewer speakers are present in a meeting.

Although speaker turns could have some dependency on the number of participants, there is no prior information that would make us to think so. This dependency, if it exists, could be related more likely to the role of the speakers or the context or content of the meeting recording.

In this paper we have carried out experiments to eliminate this variability of the transition weight, making it therefore independent of the number of active speakers at any moment.

We also want to study the possibility of using this transition term to actually favour the transition between speakers. As mentioned before a very high value of the transition weight would increase the number of speaker turns drastically, and a too low value would make them nearly impossible. A study of this term is necessary in this case to assure that neither of these situations are encountered.

A value of the transition weight equal to 1 is a special situation where there is no influence of this term, thus it would neither penalise nor favour transitions. Any transition would be determined only by the likelihood of the cluster models given the data.

Experiments have been carried out considering both the transition weight and the minimum duration, because both terms have influence in the duration of the speaker interventions.

5. Experiments

To measure the error we have used the Speaker Error Rate (SER from now on). This value removes from the typical Diarization Error Rate (DER) the error due to the VAD module and the overlapped speakers. The diarization system classifies each speech segment as a single speaker. When two speakers are speaking simultaneously one of them will be labelled and the other will be added to the Missed Speaker time. Note that the overlapped segments are used to train single speaker models which could degrade the SER of the system. The error due to the VAD module plus the overlapped segments is composed of Missed Speaker plus False Alarm Speaker error which is constant for all the experiments and equal to 7.43 for the development set (All06_07) and 8.70 for the test set (RT09). The no-score collar at speaker boundaries is 0.25

Performance of the baseline system for all the sets used (development and test sets) is shown in Table 2 second line. In previous works the minimum duration parameter was set empirically to 250 sub-states, which means a minimum duration of 2.5 secs. Experiments have been done using MFCC and TDOA values, whose probabilities have been combined using a weight of 0.85 for the MFCCs and 0.15 for TDOAs.

In Figure 5 the results for minimum duration equal to 150, 200, 250 and 300 and a transition weight ranging from 0.01 to 5 are included.

The transition weight in the baseline system is 1/M, being M the number of active speakers. M is reduced in every iteration, going from 16 at the beginning of the process to, at most, 1, which would mean that the system has found only one speaker. As a result this factor will be different in every iteration, while the clusters move from the first 16 to the final hypothesized number.

Note that being this factor equal to 1/M, when the number of speakers is higher than 1 the transition weight will turn lower than 1, resulting in a penalization of the speaker changes. As there is also a forced minimum duration of the speaker turns, penalising further the changes between speakers is not reasonable.

The special case when the transition weight is equal to 1, and therefore, neither penalise nor favour the change of speaker, happens when M=1. This situation is theoretically possible for the baseline system, but with only one active speaker, changes are impossible and the diarization process would end.

The experiments focus in two concepts: making the transition weight constant throughout the diarization process and using values higher than 1 for the transition weight. The first would make the changes independent of the number of cluster at any stage of the segmentation/clustering process. The second would favour the speaker changes instead of penalising them, as it happened in the baseline system.

In Figure 5 we can see that the speaker error rate is very dependent on the transition weight and the minimum duration. The baseline system works only in the region where the transition weight is lower than one, which penalises speaker changes. The results show that favouring these changes instead of penalising them has better results. In the case of minimum duration equal to 200 every single value of transition weight reduce the error of the system, and furthermore, the variability across transition weights is also reduced.

MEETING	# mic.	SPNSP Error	Baseline K=1/M MD=250	K=1/M MD=200	K=1 MD=200	K=2 MD=200	K=3 MD=200
EDI 20071128-1000	24	6.90	0.46	0.82	0.95	0.9	1.21
EDI 20071128-1500	24	12.10	1.64	1.54	1.71	1.46	1.6
IDI 20090128-1600	8	4.80	1.33	1.09	0.72	1.10	1.37
IDI 20090129-1000	8	9.60	4.76	2.14	7.95	4.91	2.06
NIST 20080201-1405	7	19.30	44.68	48.98	43.41	43.09	43.62
NIST 20080227-1501	7	8.80	2.43	2.34	4.99	2.91	3.17
NIST 20080307-0955	7	4.70	13.89	13.44	13.80	13.76	13.7
ALL		8.70	7.82 ±0.07	7.73 ±0.07	8.45 ±0.07	7.72 ±0.07	7.46 ±0.07
Relative improvement over the baseline				**1.15%**	**-8.05%**	**1.28%**	**4.6%**

Table 3*: SER for all meetings of the test set, RT09. MD stands for Minimum Duration. Weight applied to MFCCs is 0.85 and weight applied to TDOA is 0.15*

Meeting	Baseline K=1/M MD=250			K=1/M MD=200			K=1 MD=200			K=2 MD=200			K=3 MD=200		
	SPK	MISS	FA	SPK	MISS	FA	SPK	MISS	FA	SPK	MISS	FA	SPK	MISS	FA
EDI_20071128-1000	4		1	4		1	4		1	4		1	4		1
EDI_20071128-1500	4			4		2	4		2	4		1	4		1
IDI_20090128-1600	4		1	4			4			4			4		1
IDI_20090129-1000	4			4		1	4		1	4		1	4		1
NIST_20080201-1405	3	2		3	2		3	2		3	2		3	2	
NIST_20080227-1501	6			6			6		1	6			6		
NIST_20080307-0955	7	4		7	4		8	3		7	4		8	3	
ALL	32	5	2	32	6	4	33	5	5	32	6	3	33	5	4

Table 4*: number of detected (SPK), missed (MISS) and false alarm (FA) speakers for the test set. MD stands for minimum duration. K stands for transition weight parameter.*

Three points have been chosen to evaluate its performance with the test set, transition term equal to 1, 2 and 3, all with minimum duration of 200 frames.

To prove that the improvements obtained are not only due to the reduction of the minimum duration parameter (250 frames in the baseline and 200 frames for the proposed systems) we have checked the performance of the system when the minimum duration is the only parameter modified and the transition weight is 1/M. This experiment and the baseline will differ only in the minimum duration established by the user. Its speaker error, calculated only for comparison purposes, is included in the third line of Table 2. The next rows show performance of the diarization algorithm when transition weight is kept constant and higher than one, and with minimum duration equal to 200.

The good results for the minimum duration equal to 200, in opposition to the baseline, can be easily explained by data. In meetings from our development dataset, short speaker turns are common, and forcing every intervention to go to 250 frames increase the error. Note, however, that results for a minimum duration of 150 are much worse (Figure 5), with very variable values which is what we try to avoid.

As it can be seen in Table 2, cancelling the transition weight (transition weight=1) reduces the SER for the development set but increases it for the test set. On the other hand favouring transitions between speakers (transition terms 2.0 and 3.0) improve the performance of the system for both the development and the test set. The difference is also statistically significant for every value of the development set and for the test set when transition weight term is 3.0.

Though two parameters have been changed from the baseline to the system with transition weight equal to 3.0, results show that the modification of the minimum duration is not the only responsible of the improvement, as the system equal to the baseline, transition weight variable and equal to 1/M, except for the minimum duration of 200, works well but not as much as the system with transition weight fix and equal to 3.0.

In Table 3, we include results for test set meeting per meeting. We can see that although the average performance for the whole set is better for most of the systems, there are some meetings whose performance is actually degraded heavily as in system with transition weight equal to 1/M and minimum duration of 200. Meeting NIST 20080201-1405 has much higher error than the baseline, but is also shorter than other meetings which explains why the high increase in error in that meeting is not degrading the average SER of the system (ninth row. Fifth column in Table 3).

In Table 4, we include the number of detected, missed and false alarm speakers for every meeting in the test set. The number of speakers correctly detected increase when the transition weights is equal to 3.0. This is consistent with the fact that the average SER decreases also for this system. However, the number of wrongly detected speakers (FA) increases also. We have checked that the speech frames assigned to these new FA speakers are very low and they sum the same number of seconds than the FA from the baseline system which explains why its inclusion has no influence in the overall performance of the system. New wrongly detected speakers have its origin in the previous ones but they become splitted.

Our experiments demonstrate that the transition weight should be modified together with the minimum duration to obtain the best results. In contrast to the baseline system, transition weights higher than 1 have shown to obtain less speaker error rate and lower variability as shown in Figure 5.

One feature that is found in speaker diarization of meetings is that there is a high variability of results across different sessions in different rooms and disperse microphone locations and unknown number of speakers so it is very difficult to demonstrate advancements of new methods [20]. Although the improvement of results on the test set are smaller than the ones of the development set, if we consider both sets as an ensemble there is a definite improvement using this new approach. If we take into account that the data that we use is a community standard and that we experimented with an extensive amount of meetings (35 meetings), we can conclude that the new method that we propose has better performance than the previous one, extensively used by different laboratories.

6. Conclusions

In this paper we have proved that the segmentation stage of a speaker diarization algorithm can be improved by not penalizing transitions between different speakers or by making them more probable. The variability of results is reduced when these transitions are not penalised and remain constant throughout the segmentation/clustering iterations. Also, with a transition weight equal to 3.0, we achieved a SER reduction of 21.65% relative for the development set and of 4.6% relative for the test set.

7. Acknowledgment

The work leading to these results has received funding from the European Union under grant agreement nº 287678. It has also been supported by ASLP-MULÁN (TIN2014-54288-C4-1-R), NAVEGABLE (MICINN, DPI2014-53525-C3-2-R) and MA2VICMR (Comunidad Autonoma de Madrid, S2009/TIC-1542) projects.

8. References

[1] M. H. Moattar and M. M. Homayounpour, "A review on speaker diarization systems and approaches," *Speech Communication*, vol. 54, pp. 1065-1103, 2012.

[2] S. Tranter and D. Reynolds, "An overview of automatic speaker diarization systems," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 5, p. 1557–1565, 2006.

[3] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland and O. Vinyals, "Speaker Diarization: A Review of Recent Research," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 2, pp. 356-370, 2012.

[4] J. Pardo, X. Anguera, y C. Wooters, “Speaker diarization for multiple-distant-microphone meetings using several sources of information,” *IEEE Transactions on Computers*, vol. 56, nº. 9, págs. 1212–1224, 2007.

[5] X. Anguera, C. Wooters, y J. Hernando, “Acoustic beamforming for speaker diarization of meetings,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, nº. 7, págs. 2011–2022, 2007.

[6] J.-H. Chang, N. S. Kim, and S. Mitra, “Voice activity detection based on multiple statistical models,” *Signal Processing, IEEE Transactions on*, vol. 54, no. 6, pp. 1965–1976, June 2006.

[7] C. Wooters and M. Huijbregts, “The icsi rt07s speaker diarization system,” in *Lecture Notes in Computer Sciences*, vol. 4625, 2008, pp.509–519.

[8] C. Fredouille, S. Bozonnet, and N. Evans, “The lia-eurecom rt 09 speaker diarization system,” in *The Rich Transcription 2009 Meeting Recognition Evaluation Workshop*, 2009.

[9] S. Chen and P. Gopalakrishnan, “Speaker, environment and channel change detection and clustering via the Bayesian information criterion,” in *DARPA Speech Rec. Workshop*, 1998.

[10] T. H. Nguyen, E.-S. Chng, and H. Li, “T-test distance and clustering criterion for speaker diarization,” in *Proceedings of Interspeech*, September 2008.

[11] G. Friedland, O. Vinyals, Y. Huang and C. Muller, "Prosodic and other long-term features for speaker diarization," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 5, p. 985–993, 2009.

[12] R. Barra-Chicote, J. M. Pardo, J. Ferreiros and J. M. Montero, "Speaker Diarization Based On Intensity Channel Contribution," *IEEE Transactions on Audio, Speech and Language*, vol. 19, no. 4, pp. 754-761, 2011.

[13] Sapru, A.; Yella, S.H.; Bourlard, H., "Improving speaker diarization using social role information," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE*

International Conference on , vol., no., pp.101-105, 4-9 May 2014

[14] X. Anguera. “Robust speaker diarization for meetings”, Ph D Thesis, Universitat Politécnica de Catalunya, October 2006.

[15] J. Ajmera y C. Wooters, «A robust speaker clustering algorithm,» de *Automatic Speech Recognition and Understanding, 2003. ASRU '03. 2003 IEEE Workshop on*, 2003.

[16] C. Barras, X. Zhu, S. Meignier y J. L. Gauvain, «Improving speaker diarization,» de *Proc. DARPA RT04*, Palisades, NY, 2004.

[17] B. Martínez-González, J. M. Pardo, J. D. Echeverry-Correa, J. A. Vallejo-Pinto and R. Barra-Chicote, “Selection of TDOA parameters for MDM speaker diarization,” in *Interspeech* , Portland (OG), 2012.

[18] X. Anguera, M. Aguiló, C. Wooters, C. Nadeu and J. Hernando, "Hybrid speech/non-speech detector applied to speaker diarization of meetings,," in *Speaker and Language Recognition Workshop,* 2006. IEEE Odyssey 2006: The, 2006

[19] X. Anguera, C. Wooters, and J. Hernando, "Automatic Cluster Complexity and Quantity Selection: Towards Robust Speaker Diarization" *Proceedings of the Third Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI 2006)*, Washington DC, pp. 248-256, May 2006.

[20] N.Mirghafori, C. Wooters, “Nuts and flakes: A study of data characteristics in speaker diarization,” de *International Conference on Acoustics, Speech and Signal Processing, ICASSP*, Toulouse, France, 2006.