Using A Priori Information For Speaker Diarization

Daniel Moraru, Laurent Besacier, Eric Castelli

CLIPS-IMAG (UJF & CNRS)
BP 53 - 38041 Grenoble Cedex 9 - France

(daniel.moraru,laurent.besacier,eric.castelli)@imag.fr

Abstract

This paper presents an attempt to use supplementary information for audio data diarization. The approach is based on the use of *a priori* information about the speakers involved in dialogue. Those specific information are the number of speakers involved in conversation, and training data available for one speaker or for all the speakers involved in conversation. The experiments were mainly conducted on the 2003 Rich Transcription Diarization corpus both Dry Run Corpus and Evaluation corpus. The results show that knowing *a priori* the exact number of speakers seems not to be a very useful information. On the other hand, using *a priori* speaker models for one or all speakers involved in the conversation, may improve diarization performance when enough data is available to train reliable speaker models.

1. Introduction

Speaker diarization (or segmentation) is a new speech processing task resulting from the increase in the number of multimedia documents that need to be properly archived and accessed. One key of indexing can be speaker identity.

The goal of speaker diarization is to segment a N-speaker audio document in homogeneous parts containing the voice of only one speaker (also called speaker change detection process) and to associate the resulting segments by matching those belonging to a same speaker (clustering process). In most papers recently published in this new domain [1], [2], [3], an assumption is generally that no *a priori* information is available on the number of speakers involved in the conversation as well as on the identity of the speakers and on the nature of the conversation. A consequence of this is that no speaker reference is supposed to be available before segmenting an audio signal for instance.

This limitation may however not be so rough for some applications and conditions for which we can reasonnably hope to have *a priori* information. For instance, we may be informed of the type of conversation to be segmented (broadcast news, telephone or meeting data) for which speech signal quality and speaker turn length differ. One may also have some speaker reference models known *a priori*. For example, in a telephone meeting room system, one can ask to all the participants to orally present themselves before the meeting starts, in order to get few seconds of speech reference per speaker. In broadcast news documents, one may also have speaker references for a limited number of persons (news presenters for instance). Finally, information about the verbosity of speakers, i.e. wether they are known to speak a lot or not in a conversation, may be also an interesting information.

This paper is an attempt to investigate how helpful could be such kind of *a priori* information for speaker segmentation and how it compares with a reference state of the art segmentation system which uses no *a priori* information. Our state-of-the-art system was developped at CLIPS laboratory for both NIST 2002 and RT (See http://www.nist.gov/speech/tests/rt/rt2003/index.htm for more details) 2003 speaker segmentation evaluations.

Section 2 gives a brief description of the reference CLIPS diarization system used for the experiments. *Section 3* proposes some solutions for including *a priori* information in speaker diarization. The performance of the various propositions are shown and discussed in *Section 4*. Finally, *Section 5* concludes this work and gives some perspectives.

2. The Clips diarization system

This section presents the CLIPS diarization system [4] used during the NIST Rich Transcription Evaluation .

The CLIPS system is a state of the art diarization system based on a BIC (Bayesian Information Criterion) speaker change detector [5], [6], [7] followed by an hierarchical clustering. It uses the ELISA frame-work [8], [9] for the parametrization of the acoustic signal and for the speaker models used for clustering. The clustering stop condition is the estimation of the number of speakers using a penalized BIC criterion.

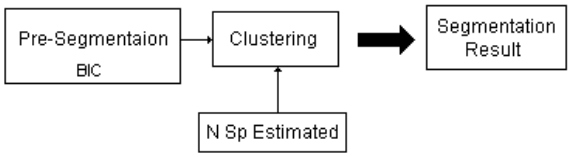


Figure 1: The CLIPS segmentation system

As a preliminary phase we use the LIA hierarchical acoustic pre-segmentation [10] in Speech / Non-Speech, gender and bandwidth. Finally we obtain 4 acoustic classes: male wideband, female wideband, male narrowband, female narrowband. The goal of the acoustic pre-segmentation is to eliminate the false alarm speech errors penalized by the error metric and to improve the performance of the clustering phase (e.g: by not clustering together segments labeled as male and female). The acoustic segmentation is applied individually on each class and the results are merged in the end.

A BIC approach is then used to define first potential speaker changes. A BIC curve is extracted by computing a distance between two 1.75s adjacent windows that go along the signal. Mono-Gaussian models with diagonal covariance matrices are used to model the two windows. A threshold is

then applied on the BIC curve to find the most likely speaker change points which correspond to the local maximums of the curve.

Clustering starts by first training a 32 components GMM background model (with diagonal covariance matrices) on the entire test file maximizing a ML criterion using a classical EM algorithm. Segments models are then trained using MAP adaptation [11] of the background model (means only). Next, BIC distances are computed between segment models and the closest segments are merged at each step of the algorithm until N segments are left (corresponding to the N speakers in the conversation).

The number of speakers is estimated (see section 3) using a penalized BIC criterion.

The signal is characterized by 16 mel Cepstral features (MFCC) computed every 10ms on 20ms windows using 56 filter banks. Then the Cepstral features are augmented by energy. No frame removal or any coefficient normalization is applied.

3. Including a priori information

Among the different types of *a priori* information, we are mainly interested in :

- number of speakers involved in the conversation;
- training data available for all speakers involved in conversation;
- training data available for only one or for few of the speakers involved in the conversation.

3.1 Knowing or not the number of speakers

3.1.1 Differences between the types of conversation

The number of speakers involved in dialogue is conversation type dependent: for telephone speech the number is fixed and is 2, for meeting speech the number can be usually known (the list of participants may be known before the meeting) but is greater than two, for broadcast news data the number of speakers is usually unknown and must be estimated.

If we know *a priori* the number of speakers involved in the conversation, since no estimation of the number of speakers is needed, our segmentation system changes as in the next figure:

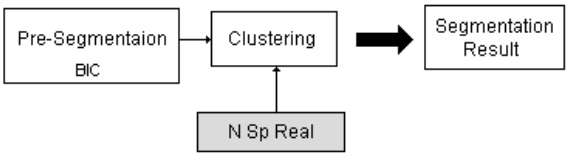


Figure 2: The segmentation system when the number of speakers is known

3.1.2 Estimating the number of speakers

When the number of speakers is unknown, it must be estimated. The estimation of the number of speakers involved in conversation is used as a stop criterion for the clustering phase of the segmentation system. In this case, the number of speakers is estimated using an algorithm based on a penalized BIC criterion.

At first the number of speakers is limited between 1 and 25. The upper limit depends usually on the recording size (e.g.: for 30 minutes audio files the limit is set to 25). Then, we select the number of speakers (NSp) that maximizes:

$$BIC(M)=\log L(X/M)-\lambda\frac{m}{2}NSp\log NX \qquad (1)$$

where M is the model composed of NSp models, Nx is the total number of speech frames involved, m is a parameter that depends on the complexity of the speaker models and λ is a tuning parameter empirically set to 0.6. The first term is the overall log-likelihood of the data while the second term is used to penalize the complexity of the model. We need the second term because the log-likelihood of the data increases with the number of models (speakers) involved in the calculation of $L(X/M)$.

Let X_i and M_i be the speech data and respectively the model of speaker i. Then we have:

$$L(X/M)=L((X_1,X_2,\ldots,X_{NSp})/(M_1,M_2,\ldots,M_{NSp})) \qquad (2)$$

If we make the hypothesis that the data X_i depends only on the speaker model M_i (a speaker data does not depend on other speaker data and models) it can be shown that:

$$L(X/M)=\prod_i L(X_i/M_i) \qquad (3)$$

The number of speakers estimated this way is usually between the real number of speakers and what we call the *optimal* number of speakers namely the number of speakers that minimizes the segmentation error. We will see in an experiment of section 4 that this optimal number is not always the real number of speakers.

3.2 Training data available for all speakers

In this section we assume that we have training data available for all speakers involved in conversation.

This could be reasonably achieved for meeting and telephone data. Data for speakers involved in a meeting could easily be obtained by asking each participant to present himself at the beginning of the meeting. For telephone speech since there are only two speakers involved, data could be obtained by manual segmentation of a short part of the conversation.

Using the available data, speaker models are trained for every speaker. The models are derived by MAP (means only) from a background model trained the same way it was trained for clustering (see section 2). Then a decision (see Fig 3) is made for every segment obtained from the BIC-pre-segmentation module, namely the segmentation system becomes in this case:

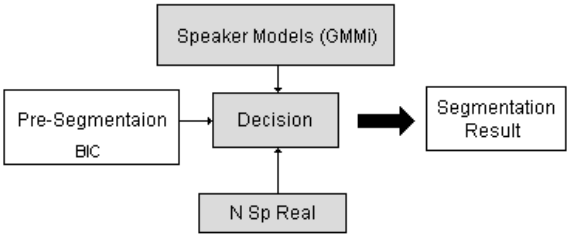


Figure 3: The segmentation system when training data is available for all speakers

The decision is made segment by segment and not frame by frame since experiments have proven that the decision made on an entire segment is more reliable than the decision made on a single frame. A maximum likelihood decision is made for each segment. Given a speech segment X and a speaker model M_i the probability that the segment belongs to that speaker is given by:

$$P(M_i/X)=P(X/M_i)\frac{P(M_i)}{P(X)} \quad (4)$$

Considering that the speakers and the segments are all equal as probability we are actually computing $P(X/M_i)$ for each segment.

However for broadcast news data it is usually impossible to obtain data for all speakers involved in conversation. We can easily obtain data for the news hosts using recordings from the same source (same TV or Radio station) but we could never obtain data for all speakers (e.g: people interviewed that speak for only 30 seconds during a 1 hour news journal).

This is a situation that corresponds to the next case. For the moment we treat only the particular case where training data is available for one of the speakers only.

3.3 Training data available for only one speaker

3.3.1 Pre-Segmenting the audio file

When we have data available for one speaker only, one way to use it is to identify the segments of the known speaker by pre-segmenting the entire recording and segment then the speech not labeled as the known speaker using the conventional segmentation system.

Assuming we have two models available: one for the known speaker M and a universal model for the unknown speakers U and let X be the speech segment data. Then the segment X is labeled to the known speaker if:

$$\log L(X/M)-\log L(X/U)>\alpha \quad (5)$$

where α is a threshold. This pre-segmentation is actually a tracking of the known speakers.

However, this method requires tuning data for the threshold and also the final diarization error may vary with respect to the quantity of training data available for the known speaker.

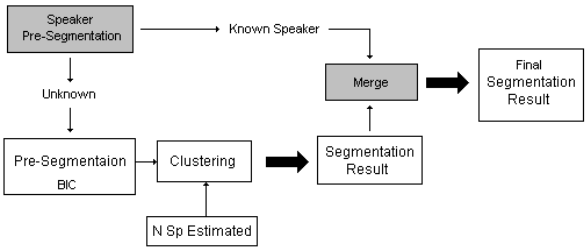


Figure 4 : The segmentation system when data is available for one speaker only (Speaker Pre-Segmentation)

This kind of system (Fig. 4) is actually a cascaded system made of a pre-segmentation and a diarization system. This means that the final system error is the cumulated error of both systems.

3.3.2 Post-Segmentation speaker labeling

There is another way to use the *a priori* known speaker model. We noticed, as stated above, that the estimated number of speakers is usually greater than the optimal number of speakers and smaller than the real number of speakers. The goal in this case would be to decrease the estimated number of speakers in order to get closer to the optimal number of speakers.

Let NS_P be the estimated number of speakers, let X_i be the data labeled to speaker i and let M and U be the models corresponding to the known speaker and the unknown speakers, then we compute the likelihood ratio for all NS_P speakers detected:

$$llr(i)=\log L(X_i/M)-\log L(X_i/U) \quad (6)$$

Using a Bayesian decision, the speaker i that maximizes the $llr(i)$ is matched to the known speaker. This simple decision seems to perform well since the right speaker was well matched to the a priori known model in all the test files used in the experiments of section 4.3.2. Now, suppose the known speaker was split in two or more speaker clusters by the segmentation system. The bad speaker clusters should also obtain a likelihood ratio close to the maximum.

So all speakers clusters that have a positive likelihood ratio and that are close enough to the maximum likelihood ratio should be re-labeled as the known speaker. As a measure of closeness we empirically decided that a speaker cluster is labeled as the known speaker only if its likelihood ratio is superior to a decision threshold fixed to 0.5 of the maximum ratio.

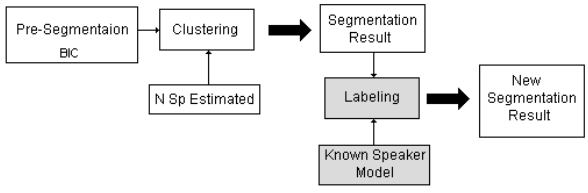


Figure 5: The segmentation system when data is available for one speaker only (Post-Segmentation Labeling)

As shown in Figure 4, some segments of the output segmentation can be re-labeled or not as the known speaker. It is equivalent to a constrained clustering.

4. Experiments And Results

Experiments were conducted on three databases, namely the broadcast news Dry Run and Evaluation database used during the broadcast news NIST RT 2003 Diarization Task and also on the french broadcast news ESTER [12] database. The segmentation error is the official RT 2003 metric used during the spring 2003 campaign.

The RT Dry Run database consisted of 6 news audio files recorded in 1998 from 6 different sources: MNB_NBW, PRI_TWD, NBC_NNW, CNN_HDL, VOA_ENG, ABC_WNT. All files were 10 minutes long, recorded at 16 KHz, 16bit. The number of speakers involved was between 6 and 18.

The RT Evaluation database consisted of 3 news files recorded in 2001 from three different sources: MNB_NBW, PRI_TWD, VOA_ENG. All files were 30 minutes long, recorded at 16 KHz, 16bit. The number of speakers involved was between 10 and 27.

As we can see there are 3 sources that are common to both RT Dry Run and Evaluation databases. Those sources will probably contain speakers that are common to both databases. In experiments of section 4.3, data available from one database will be used as training data for the experiments on the other database.

Finally the ESTER[1] database (the French version of the RT evaluation) see [12] consisted of 40h of audio data recorded from French radio broadcast news shows. It was divided in train data, dev data and test data. A subset was used for the experiment of section 4.3.1. The subset consisted of 21 train files (13h40) from two different sources France Inter, and RFI, 4 dev files (2h40) from the same sources and the 4 test files from the same sources. All files were 1h (RFI) or 20 minutes (France Inter) long, recorded at 16KHz, 16 bit. The number of speakers involved was between 13 and 39.

The performance measure used for the RT 03 segmentation evaluation is an error function based on an optimum one to one mapping of reference speaker IDs to system output speaker IDs. The measure of optimality will be the aggregation, over all reference speakers, of the time that is attributed to both the reference speaker and the corresponding system output speaker to which that reference speaker is mapped. This will always be computed over all speech, including regions of no speech (silence, music, noise, etc).

The measure excludes from scoring the overlapping speech regions. Also some segments, for which transcriptions were not provided (e.g. commercials) using an Unpartitioned Evaluation Map (uem) file, are excluded.

There are two kinds of errors: the speech detection errors and the speaker mapping errors:

- miss speech: a segment was incorrectly labeled as no-speech
- false alarm speech: a segment was incorrectly labeled as speech
- speaker error: a segment was assigned to the wrong speaker

Each audio segment was labeled as a correct diarized segment or as one of the error type segment.

The final score is the fraction of scored time (given by the uem file) that is not correctly labeled. It is the sum of speech error time and of the speaker error time over the scored speech time.

The score tool can be downloaded from the RT web site[2]. The same score is used for the French evaluations ESTER.

4.1 Knowing or not the number of speakers

Our first experiments concerned the estimation of the number of speakers involved in conversation. The following table presents the speaker segmentation error using the optimal, the estimated and respectively the real number of speakers obtained on both speech databases:

Table 1: Estimation of the number of speakers (% diarization error)

Corpus	N Optimal	N Estimated	N Real
Dry Run	14.54	19.65	24.76
Evaluation	14.03	19.25	16.29

The second column is the reference error that we will attempt to decrease using the data available for the speakers.

As stated before, what we call the *optimal* number of speakers is the number of speakers that minimizes the segmentation error and not the real number of speakers. It is usually smaller than the real number of speakers due to the fact that there are a lot of speakers especially in broadcast news data that do not speak enough to train a reliable statistical model (e.g: 4 seconds during a 30 minutes file).

Tests proved that the estimation of the number of speakers generates approximately 5% more absolute segmentation error compared to the optimal number of speakers.

It is however difficult to conclude if it is useful or not to know *a priori* the real number of speakers since the results on Dry Run database show that it is better to estimate the number of speakers and the results on the Evaluation database show that is better to know it *a priori*.

4.2 Training data available for all speakers

For our second experiment we assumed data was available for all speakers.

Different amount of data were available for training starting with 1 second per speaker till 60 seconds per speaker.

The experiment is more suitable for meeting data and certain broadcast news data like a round table discussion but we still used the RT03 Broadcast news data. The speaker data used to build the speaker models was directly extracted from the test data. This is typically what would be done if a human annotator had labeled a part of the conversation in order to have a few seconds of data for each speaker.

For this experiment, a synthesis of results is presented in the Table 2:

[1] http://www.afcp-parole.org/ester/

[2] http://www.nist.gov/speech/tests/rt/rt2003/spring/

Table 2: Diarization error when data is available for all speakers

A priori data	0 sec	1 sec	2 sec	20 sec	60 sec
Dry Run a priori	19.65	29.70	18.92	13.65	15.95
Dry Run Reference	19.65	19.04	18.52	16.85	19.77
Evaluation a priori	19.25	40.51	37.19	20.82	12.67
Evaluation Reference	19.25	18.95	18.69	17.67	21.33

For comparison purposes we present the segmentation error obtained by the stand alone segmentation system (with no *a priori* models). The diarization error presented is the sum of speech detection error and of speaker error but, since we are using the same pre-segmentation in both cases (with and without training data) the gain or the loss in terms of performance is entirely due to the speaker error change. This reference performance for Dry Run and evaluation data changes slightly for each column of the table since the diarization error is always computed on the parts of speech that were not used to train speaker models. In other words the segments used for training are always excluded from scoring the diarization error. It is also important to note that those segments are excluded from scoring but not from system input.

Finally figures 5 and 6 present all results for different amount of data available from 1 second to 60 seconds.

Figure 6: Diarization error for the Dry Run Database according to the amount of training data available a priori

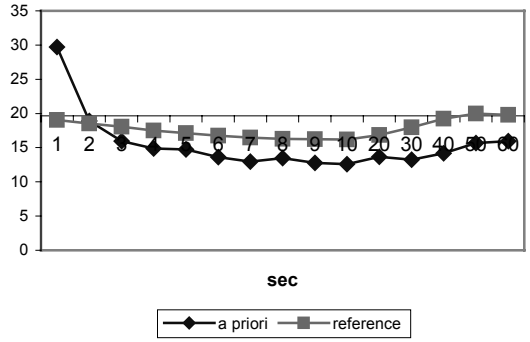
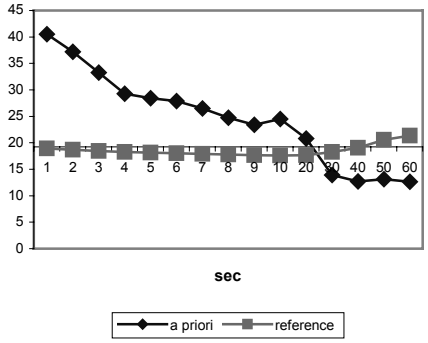


Figure 7: Diarization error for the Evaluation Database according to the amount of training data available a priori



The stand alone reference system segmentation error logically stays constant since it corresponds always to the same reference segmentation output (the slight differences are due to the fact that more and more segments are removed for scoring). On the other hand, the segmentation error decreases as more data is available to learn *a priori* speaker models[3].

There are 3 main differences between the two databases: the size of the recordings (10 minutes versus 30 minutes), the number of speakers involved in conversation and most important the number of speakers that do not have enough data. That is why the improvement is obtained very fast for the Dry Run (starting with 2 seconds data available) and much later for the Evaluation (starting with 20 seconds data available).

For 1 second per speaker of data available we obtain a high error rate since one second is not really enough to train a speaker model. Thus, the improvement actually depends on the amount of data available and on the number of speakers involved in conversation. For 60 seconds available, the error is mainly due to the false alarm speech error and to the speakers that do not have 60 seconds available (we used the maximum available speech in this case but in Eval03 there are speakers that speaks even less than 7 seconds).

In conclusion, as expected if we do not have *sufficient data* for the speakers involved in conversation it is better to use the stand alone segmentation system. However the *sufficient amount of data* varies with the number of speakers involved in conversation.

For all experiments presented in this section, there is also an important aspect to mention, namely the computation time of the segmentation when *a priori* data is available which was about 0.1 x Real Time on a P III (800 MHz) with 512MB physical memory running Red Hat Linux 7.3, compared to 10 x Real Time for the stand alone segmentation system (with no *a priori*).

4.3 Training data available for only one speaker

4.3.1 Pre-Segmenting the audio file

For this experiment we assumed that data was available only for the news speaker. As described in section 3.3.1 the idea was to pre-segment the audio document in known (news host) speaker speech and unknown speaker speech. For experimental reasons we preferred to use the ESTER database since it provides enough data splitted in training data, dev (tuning) data and test data.

The speaker models were 128 diagonal GMM directly trained on data extracted from the train corpus using the EM algorithm. The data available for every known speaker was between 20 and 70 minutes.

The unknown speaker models were also 128 diagonal GMM and were directly trained on data not labeled as the known speaker extracted from the train corpus. The unknown models were source dependent (1 unknown model per radio source). The data available for the unknown models was between 150 and 250 minutes.

At first the known / unknown speaker pre-segmentation was tuned on the dev part of ESTER. Then it was applied on the test files. The tuning was done only on the decision

[3] When we are saying we have 30 seconds of available data for example it means at most 30 seconds because there are speakers that do not have such an amount of data.

threshold from equation 5 but a finer tuning could also be done on speech parametrization and on speaker modeling.

Table 3 presents the experimental results for the dev files while Table 4 presents the results on the test files. For both dev and test corpora the "Reference Diarization Error" (first line) is the diarization error obtained by the CLIPS segmentation system from section 2 without any *a priori* information used. The "Known speaker error" (second line) is the diarization error obtained from the previous segmentation output (without any *a priori* information) by labeling the hypothesis speakers in either 'known' or 'unknown'. It is always inferior to the "Reference Diarization Error" and corresponds to the diarization errors related to the known speaker. When this error value is high, that indicates the potential of pre-segmenting the signal in known / unknown speaker. For instance, for the first dev signal (FrInter1), the "Reference Diarization Error" is 7.31% while the "Known speaker error" is 7.05%. That means that if we had a perfect known speaker tracking for pre-segmentation, then the overall diarization error would be very low.

For the test files the "Known/Unknown Pre-Seg Error" (third line) scores the pre-segmentation phase in known/unknown speakers. The pre-segmentation output is obtained using the threshold from equation 5 tuned on the dev files.

Finally for the test files the "Final Diarization error" (fourth line) is the error obtained by the overall segmentation system that uses a known/unknown speaker pre-segmentation cascaded with the diarization system applied on the speech part labelled as unknown speaker (Figure 4).

Table 3: Training data available for only one speaker: Experimental results for the *dev* files (pre-seg. Approach)

File	FrInt 1	FrInt 2	RFI0930 1	RFI1130 1
Reference Diarization Err	7.31	8.49	14.55	23.47
Known speaker Err	7.05	7.56	1.31	4.27

Table 4: Training data available for only one speaker: Experimental results for the *test* files (pre-seg. Approach)

File	FrInt 3	FrInt 4	RFI0930 2	RFI1130 2
Reference Diarization Err	13.29	6.89	13.67	25.29
Known speaker Err	6.83	5.48	2.24	6.14
Known / unkown Pre-Seg Err	3.84	3.63	5.16	1.67
Final Diarization Err	12.35	5.64	17.42	21.26

The results obtained on the dev files (Table 3) indicates us the potential of using or not the known/unknown pre-segmentation. It shows that for 3 out of 4 files there is a possible final gain.

Comparing the first line (Reference) and the fourth line (Final) of Table 4 show an improvement of the overall diarization performance on 3 out of 4 files. For the file which did not lead to an improvement (RFI0930 2), the "known speaker error" was very low, which means that the diarization error was mainly due to the unknown speaker signal part. In this case, a known/unknown pre-segmentation cascaded with the diarization system just added errors to the system.

4.3.2 Post-Segmentation speaker labeling

For this experiment we also assumed data was available only for the news host speaker (about 15% of the RT Evaluation database). We used data available from the RT Dry Run database for training since it is anterior to the data available in the RT Evaluation database which was used for testing. For the experiment we started with the reference segmentation output and tried to label the detected speakers as the known speaker. We usually had about 2 minutes data available per known speaker that we used to train a 32 diagonal GMM. For the unknown model we used the rest of the corresponding Dry Run recording. The unknown model was also a diagonal 32 GMM.

The reference segmentation was then re-labelled according to the rule presented in section 3.3.2.

The results are presented bellow:

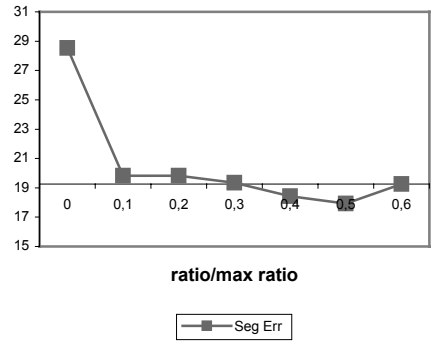
Table 5: Training data available for only one speaker: Experimental results on RT data (post-seg. Approach)

Corpus	Reference	1 known speaker
Evaluation	19.25	17.92

These preliminary results show the potential of using *a priori* information on one speaker. However these results need more experiments in order to have a reliable conclusion. There is one recording for which the gain is 3.4% in absolute decreasing from 10.14% which actually gives an almost perfectly segmented file.

To show how precarious this method is, we need to show Figure 7 which presents the segmentation error with respect to the threshold equals to $llr(i)$ ratio/maximum $llr(i)$ ratio.

Figure 8: Diarization error when data is available for one speaker only, according to the re-labeling threshold



For comparison if all speakers that have a positive likelihood ratio are labeled as known speakers, meaning that we are not using the condition that the ratio should be greater than 0.5 of the maximum, the segmentation error is 28.53% (it is the case ratio/max ratio 0). This proves that most of the positive ratios are concentrated between 0 and 0.1 of the maximum ratio. The ratios of the known speakers are much greater than the ratio obtained by the speakers not labeled as

the known speaker. This way of doing things is equivalent to a constrained clustering.

However the relative small gain, only 1.33% in absolute is due to the fact that the segmentation system already detects well the known speaker (usually the error is about 30 seconds missed speaker time). This means that experiments should be done probably on another database where the possible gain would be greater (more then 30 seconds on a 30 minutes file).

5. Conclusions

In this paper we have investigated the use of *a priori* information for speaker diarization (segmentation).

We started by presenting the stand alone CLIPS diarization system that was used during the NIST Rich Transcription 2003 evaluation.

Then we investigated the possibility of using different *a priori* information in order to improve the stand alone system diarization error. As *a priori* information we were interested in: knowing or not the real number of speakers, using data available for all speakers involved in conversation and using data available for only some of the speakers involved in conversation.

It was difficult to conclude if it is useful or not to know *a priori* the real number of speakers since the results on one of the databases show that it is better to estimate the number of speakers and the results on the other database show that is better to know it *a priori*.

For the case were training data is available for all speakers, experiments have proven as expected that when there is *sufficient data* available for all speakers this data should be used for segmentation. However the amount of *sufficient data* can vary with respect to the number of speakers involved in conversation.

Finally when data was available for only one speaker it was quite difficult to use it and we could only obtain a small diarization error gain. We presented two different methods: one by pre-segmenting the signal in known / unknown speaker and one by post-segmentation speaker labeling.

As a perspective, further experiments should be done on other databases. Experiments on meeting or telephone conversation databases for the case when data is available for all speakers.

6. References

[1] S. Meignier, J.-F. Bonastre, and S. Igounet, “E-HMM approach for learning and adapting sound models for speaker indexing,” In *A Speaker Odyssey*, pp.175-180, Chania, Crete, June 2001.

[2] P. Nguyen, J.-C. Junqua, "PSTL's Speaker Diarization system", *DARPA/NIST Rich Transcription Workshop*, Boston, Massachusetts, May 2003

[3] A. G. Adami, S. Kajarekar, H. Hermansky, "A New Speaker Change Detection Method For Two-Speaker Segmentation", *Proc of ICASSP 2002*, Orlando, Florida May 2002

[4] D. Moraru, S. Meignier, C. Fredouille, L. Besacier, J.-F. Bonastre, "The Elisa Consortium Approaches in Broadcast News Speaker Segmentation During The Nist 2003 Rich Transcription Evaluation", Accepted to *Proc of ICASSP 2004*, Montreal, Canada, May 2004

[5] S.S. Chen, P.S. Gopalakrishnan., "Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion", *Proc. of DARPA Broadcast News Transcription and Understanding Workshop*, Landsdowne, Virginia, 127--132, February 1998.

[6] H. Gish, H-H Siu, R. Rohlicek. "Segregation of Speakers for Speech Recognition and Speaker Identification", *Proc of ICASSP 1991*, Toronto, Canada, May 1991

[7] P. Delacourt and C. Wellekens, "DISTBIC: a Speaker-Based Segmentation for Audio Data Indexing", *Speech Communication,* Vol. 32, No. 1-2, September 2000

[8] I. Magrin-Chagnolleau, G. Gravier, and R. Blouet for the ELISA consortium, "Overview of the 2000-2001 ELISA Consortium Research Activities", In *A Speaker Odyssey*, pp.67–72, Chania, Crete, June 2001.

[9] C. Fredouille, J.-F. Bonastre, and T. Merlin, "AMIRAL: a Block-Segmental Multi-Recognizer Architecture for Automatic Speaker Recognition", *Digital Signal Processing*, Vol. 10, No. 1-3, January/April/July 2000.

[10] S. Meignier, D. Moraru, , C. Fredouille, L. Besacier, J.-F. Bonastre, "Benefits of Prior Acoustic Segmentation for Automatic Speaker Segmentation", Accepted to *Proc of ICASSP 2004*, Montreal, Canada, May 2004

[11] D.A. Reynolds, T.F. Quatieri and R.B. Dunn, “Speaker Verification Using Adaptation Mixture Models”. *Digital Signal Processing*, Vol. 10, No. 1-3, January/April/July 2000.

[12] G. Gravier, J.-F. Bonastre, E. Geoffrois, S. Galliano, K. Mc Tait, K. Choukri, “The ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News”, Accepted to *LREC* 2004, Lisbon, Portugal, May 2004