NIST RT'05S Evaluation: Pre-processing Techniques and Speaker Diarization on Multiple Microphone Meetings

Dan Istrate[1], Corinne Fredouille[1], Sylvain Meignier[2], Laurent Besacier[3], and Jean François Bonastre[1]

[1] LIA-Avignon - BP1228 - 84911 Avignon Cedex 9 - France
[2] LIUM, Avenue Laënnec, 72085 Le Mans Cedex 9
[3] CLIPS-IMAG (UJF & CNRS & INPG) - BP 53 - 38041 Grenoble Cedex 9 - France

Abstract. This paper presents different pre-processing techniques, coupled with three speaker diarization systems in the framework of the NIST 2005 Spring Rich Transcription campaign (RT'05S).

The pre-processing techniques aim at providing a signal quality index in order to build a unique "virtual" signal obtained from all the microphone recordings available for a meeting. This unique virtual signal relies on a weighted sum of the different microphone signals while the signal quality index is given according to a signal to noise ratio.

Two methods are used in this paper to compute the instantaneous signal to noise ratio: a speech activity detection based approach and a noise spectrum estimate. The speaker diarization task is performed using systems developed by different labs: the LIA, LIUM and CLIPS. Among the different system submissions made by these three labs, the best system obtained 24.5 % speaker diarization error for the conference subdomain and 18.4 % for the lecture subdomain.

1 Introduction

The goal of speaker diarization is to segment an N-speaker audio document in homogeneous parts containing the voice of only one speaker and to associate the resulting segments by matching those belonging to the same speaker. In speaker diarization the intrinsic difficulty of the task increases according to the target data: telephone conversations, broadcast news, meeting data.

This paper is related to speaker diarization on meeting data in the framework of the NIST 2005 Spring Rich Transcription (RT'05S) campaign. Meeting data present three main specificities compared to broadcast news data:

- meeting conversations are recorded with multiple microphones which implies redundancies and different qualities of the same speech recording. The use of information from all channels seems to be an important issue;
- the meeting room recording conditions associated with distant microphones lead to noisy recordings, including background noises, reverberations and distant speakers;

S. Renals and S. Bengio (Eds.): MLMI 2005, LNCS 3869, pp. 428–439, 2006.

- the speech is fully-spontaneous, highly interactive and presents a large number of disfluencies as well as speaker segment overlaps.

This paper is focused on the extraction of pertinent information issued from the different multiple microphone recordings in the particular task of speaker diarization. Indeed, signal processing techniques are applied on the different distant microphone signal recordings in order to determine pertinent portions of signal and to build a unique "virtual" signal. This virtual signal is then used as input for the speaker diarization systems. Basically, the unique virtual signal is based on a weighted sum of the multiple microphone signals. The weights of this sum are estimated according to a signal quality index based on a signal to noise ratio estimate.

Two main factors will be studied in this paper. First, the efficiency of the pre-processing techniques to build a unique virtual signal in the context of speaker diarization will be investigated. Then, the focus will be on how well systems that were tuned with broadcast news data only can handle meeting data. Concerning the last point, different speaker diarization systems will be tested in this study. Developed in three different labs: the LIA, LIUM and CLIPS, these systems have been tuned and evaluated during the French ESTER Rich Transcription evaluation campaign. This campaign, organized in January 2005 and sponsored by the French ministry, was dedicated to Broadcast news data [1]. No particular tuning of the systems was made on meeting data in order to evaluate whether a reliable pre-processing on multi-channel recordings may be sufficient in order to maintain performance.

Finally, the RT'05S evaluation campaign has initiated a new task, based on the Speech Activity Detection (SAD). This processing is classically implemented in both the speech transcription and speaker diarization systems but never scored individually. This paper will present the SAD system proposed by the authors for the RT'05S evaluation campaign and their results.

Section 2 presents the Speech Activity Detection algorithm. Section 3 is dedicated to the pre-processing techniques used in order to obtain a unique signal from the multi-channel recordings. In section 4, the LIA, LIUM and CLIPS speaker diarization systems are presented, followed by a brief description of all the systems submitted for the RT'05S evaluation campaign. Section 5 presents the experimental protocols and results, and finally, section 6 concludes this work.

2 Speech Activity Detection Task

Considered until now as only a sub-part of speech transcription or speaker diarization systems, Speech Activity Detection has been evaluated in the RT'05S evaluation campaign as an entire task.

Speech Activity Detection is not trivial in a multiple microphone environment. For instance, the portions of silence might be different from one microphone to another. Besides, energy based SAD systems have some difficulties in dealing with background voices.

The Speech Activity Detection (SAD) system, used by most of the systems presented in this paper, was developed by the LIA. It is based on the ALIZE platform [2] and relies on two passes: (1) to apply a speech activity detector process on each individual channel for a given meeting, provided speech and non-speech segments; (2) to keep the non-speech segments, shared over ALL the channels. The speech activity detector process used in the first pass is based on the speech energy modeling and works as follows:

1. the log spectral energy is computed through at a 10ms rate;
2. The energy values are first normalized using a mean removal and a variance normalization in order to fit a 0-mean and 1-variance distribution;
3. They are then used to train a three component GMM, which aims at selecting speech frames. Indeed, $X\%$ of the most energized frames are selected through the GMM, with: $X = w_1 + (\lambda * \alpha * w_2)$ where: w_1 the weight of the highest (energy) gaussian component, w_2 the weight of the middle component, λ an integer ranging from 0 to 1, α a weighting parameter, empirically fixed to 0.6 on the development set. The value of λ is decided according to the likelihood loss when merging the gaussian components 1 and 2 and the components 2 and 3. If the loss is higher for components 1 and 2, λ is set to 0 else to 1;
4. Once all the frames of a signal are labelled as speech or non-speech and concatenated to form segments according to their labels, a final process is applied in order to refine the speech detection. This last process is based on two morphological rules, which consist in constraining the minimum duration of both the speech and non-speech segments (minimum length is 0.3s).

3 Meeting Pre-processing Algorithms

Meeting signals are recorded in smart rooms i.e. a room equipped with several audio and video captors as well as multimedia output devices (video projectors, multi-stereo audio outputs...). According to the distant microphone position in the table, the quality of signal may hugely differ from one microphone to another. For instance, the main speaker utterances may be caught by one or two distant microphones while the other microphones mainly provide background voices, long silence periods, or background noise only. The aim of this pre-processing system is to use redundant channel information in order to extract pertinent information for an enhanced output virtual signal.

This output signal is a weighted mix of all channels available for a given meeting. For each channel a quality measure (signal to noise ratio - SNR) is estimated in order to adapt channel weights. The sum of weights is equal to 1 and the channel weights w_i are computed following equation (1), where N is the number of channels.

$$w_i = SNR_i / \left(\sum_{j=1}^{N} SNR_j \right) \tag{1}$$

To obtain a reliable quality measure, it is necessary to estimate the noise energy, for which two methods have been considered: the use of a speech activity detector (SAD) and the noise spectrum estimate.

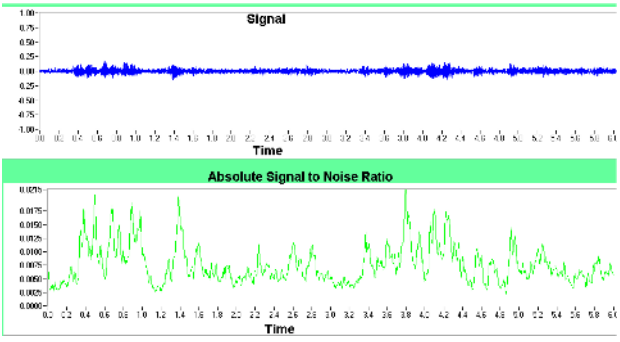

Fig. 1. Example of SNR estimate

If a speech activity detector is used, the labelled *non-speech* segments are used to compute the average noise energy $\overline{E}_{\text{noise}}$ for each channel. The SNR is estimated at each 32ms on frames of 64ms (L=1024 samples) using equation (2).

$$SNR = 10\log_{10}\left(\left(\sum_{i=0}^{L} s_i^2 - \overline{E}_{\text{noise}}\right)/\overline{E}_{\text{noise}}\right) \qquad [dB] \tag{2}$$

where s_i is a signal sample at instant i.

In the second case, an estimate of the noise spectrum is used in order to discard the speech activity detector errors and to have an instantaneous noise energy value instead of an averaged one. The algorithm is based on a minimum statistics tracking method [3]. Assuming the noisy speech power is the summation of clean speech and background noise power, tracking power spectral minima can provide a fairly accurate estimate of the background noise power and then a good estimate of SNR [4]. Also, by tracking minimum statistics, this algorithm can deal with nonstationary background noise with slowly changing statistical characteristics. The noise spectrum is estimated every 2s using signal power spectrum histogram. An example of signal to noise ratio estimate for a part of channel 1 signal from "NIST 20020305-1007" file is presented in Figure 1.

In this case, the SNR is estimated using the signal power spectrum and noise power spectrum, like in equation (3).

$$SNR = 10 * \log_{10}\left(\sum_{i=0}^{M} \widetilde{S}_i / \sum_{j=0}^{M} \widetilde{N}_j\right) \tag{3}$$

where $\widetilde{S}_i$ is signal spectral amplitude at frequency i and $\widetilde{N}_j$ is noise spectral amplitude at frequency j.

In order to evaluate the influence of these pre-processing techniques, an unweighted mix ($w_i = \frac{1}{N}$) has also been computed.

4 Speaker Diarization Systems

Three speaker diarization systems are involved in this work, developed by the LIUM, CLIPS and LIA laboratories individually. Two of them, the LIUM and CLIPS systems, are based on a classical speaker turn detection followed by a clustering phase. For the LIA system, both the speaker turn detection and the speaker clustering are performed simultaneously by using a E-HMM based approach as described in the next section.

No particular tuning on the meeting data has been carried out for these systems to participate at the RT'05S evaluation campaign. Indeed, all these speaker diarization systems have participated at the French Rich Transcription evaluation campaign ESTER. Testing these systems on meeting data without any further tuning will allow the evaluation of their robustness to environment changes, especially if pre-processing techniques are applied beforehand on multiple microphone signals in order to extract pertinent information.

4.1 The LIA System

The LIA speaker diarization system has been entirely developed by using the ALIZE toolkit (freely available thanks to an open software licence), released by the LIA and dedicated to speaker recognition [5]. Its performance has been evaluated firstly during the ESTER evaluation campaign on Broadcast News data. The core of the system is built on a one-step segmentation algorithm implying an E-HMM (Evolutive HMM) [6, 7]. Each E-HMM state characterizes a particular speaker and the transitions represent the speaker changes. All possible changes between speakers are authorized. In this context, the segmentation process has 4 steps:

1. **Initialization:** A first model, named L_0, is estimated on all speech data. The HMM has one state, L_0 state.
2. **New speaker detection:** A new speaker is detected in the segments labelled L_0 as follows: a segment is selected among all the L_0 segments according to the likelihood maximization criterion. This selected segment is then used to estimate the model of the new speaker, named L_x, which is added to the HMM.
3. **Adaptation/Decoding loop:** The objective is to detect all segments belonging to the new speaker L_x. All speaker models are re-estimated through an adaptation process according to the actual segmentation. A Viterbi decoding pass is done in order to obtain a new segmentation. This adaptation/decoding loop is re-iterated while the segmentation is not stable.
4. **Speaker model validation and stop criterion:** The current segmentation is analyzed in order to decide if the new added speaker, L_x, is relevant. In this case the decision is made according to heuristical rules on speaker L_x segment duration. The stop criterion is reached if there is no more segment available in L_0. However, if the contrary, the process goes back to the step 2.

Finally, a resegmentation process is applied, which aims at refining the boundaries and at deleting irrelevant speakers (e.g. speakers with too short speech

segments). This stage is based only on the third step of the segmentation process. A HMM is generated from the segmentation and the adaptation/decoding loop is launched. At the end of each iteration, speakers with too short duration are deleted.

Concerning the front end-processing, the signal is characterized by 20 linear cepstral features (LFCC), computed every 10ms using a 20ms window. The cepstral features are augmented by energy. No frame removal or any coefficient normalization is applied at this stage.

The entire speaker segmentation process is largely described in [8].

4.2 The LIUM System

The LIUM speaker diarization system is based upon a BIC framework similar to [9, 10], composed of three modules:

1. **signal split into small homogeneous segments:** the initial segment boundaries are determined according to a Generalized Likelihood Ratio (GLR) computed over two consecutive windows of 2s sliding over the signal. No threshold is employed, except for the minimal segment length which is set to 2.5s. The signal is over-segmented in order to minimize miss detection of boundaries, but the minimum segment length is set long enough for a correct estimate of a speaker model using a diagonal Gaussian;
2. **speaker clustering without changing the boundaries:** The clustering is based upon a bottom-up hierarchical clustering. In the initial set of clusters, each segment is a cluster. The two closest clusters are then merged at each iteration until the BIC stop criterion is met. The speaker, ie the cluster, is modeled by a full covariance Gaussian as in the segmentation process. The BIC penalty factor is computed over the length of the two candidate clusters instead of the standard penalty computed over the length of the whole signal [11]. To minimize the clustering time, a first pass of clustering is performed only over adjacent clusters. The λ parameter (eq. 4) is fixed to 2 for the first pass and to 7.5 for the second pass;
3. **boundaries adjustment:** a Viterbi decoding is performed to adjust segment boundaries. A speaker is modeled by a one-state HMM containing a diagonal covariance GMM of 8 components learned by EM-ML over the set of speaker segments. The log-penalty of switching between two speakers is fixed experimentally to 100.

Concerning the front end-processing, the signal is characterized by 12 mel cepstral features (MFCC), computed every 10ms using a 20ms window. The cepstral features are augmented by energy. No frame removal or any coefficient normalization is applied at this stage.

4.3 The CLIPS System

The CLIPS system is based on a BIC [12] (Bayesian Information Criterion) speaker change detector followed by a hierarchical clustering. The clustering

stop condition is the estimate of the number of speakers using a penalized BIC criterion. Whereas the LIUM system clusters homogeneous segments, the CLIPS system clusters segments, which result from a first speaker change detection pass as follows:

1. **speaker change detection:** a BIC curve is extracted by computing a distance between two 1.75s adjacent windows that go along the signal. Mono-component 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 maxima of the curve;
2. **speaker clustering:** Clustering starts by first training a 32 component GMM background model (with diagonal covariance matrices) on the entire test file maximizing a ML criterion thanks to a classical EM algorithm. Segment models are then trained using MAP adaptation 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 N speakers in the conversation).
3. **clustering stop criterion:** the number of speakers (N_{Sp}) is estimated using a penalized BIC. The number of speakers is constrained between 1 and 15. The upper limit is related to the recording duration. The number of speakers (N_{Sp}) is selected to maximize equation (4).

$$BIC(M) = \log L(X; M) - \lambda(m/2)N_{Sp} * \log(N_X) \tag{4}$$

where M is the model composed of the N_{Sp} speaker models, N_X 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 equal to 0.6.

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 here.

The entire speaker segmentation process is largely described in [8].

4.4 Proposed Systems

Different systems have been submitted for the RT'05S campaign. All rely on the following scheme - composition of 3 modules - as summarized in table 1:

1. The pre-processing module can consist in applying:
   - either the *weighted* mix of the multiple microphone signals in which channel weights depend on SNR, estimated either using the speech activity detector (Weighted Mix - SAD) or by applying the noise spectrum algorithm (Weighted Mix - Noise spectrum).
   - or a simple unweighted mix of the multiple microphone signals (Mix).
2. a speaker diarization module, which can be based on the LIA, LIUM or CLIPS systems.

3. the LIA resegmentation process (described in section 4.1) since different studies have shown that a resegmentation phase leads to performance improvement [13, 14, 15, 5].

Table 1. Proposed diarization systems

Systems	Pre-processing	Seg/Re-Seg
WMixSpectrum	Weighted Mix - Noise spectrum	LIA/LIA
WMix	Weighted Mix - SAD	LIA/LIA
MixLIA	Mix	LIA/LIA
MixCLIPS	Mix	CLIPS/LIA
MixLIUM	Mix	LIUM/LIA

5 Experiments

This section presents the protocols and results obtained by the different techniques proposed in this paper and submitted to the RT'05S evaluation campaign.

5.1 Protocols

For RT'05S, the speaker diarization task was evaluated on two subdomains: recordings issued from conference rooms (similar to RT'04S) and from lecture rooms. As for any evaluation campaign, two corpora were available:

- a development corpus: composed of RT'04S development and evaluation corpora (12 meetings of about 10mn each), plus some additional meetings including new recording sites.
- two evaluation corpora, one for each subdomain, composed of 10 meetings of about 10mn each for the conference subdomain and 29 meetings of about 3mn each for the lecture subdomain.

In this paper, only the RT'04S data (development and evaluation) is used as the development corpus, and will be referred to as *dev* corpus in the next sections. On the other hand, the RT'05S evaluation data will be referred to as $eva-conf$ for conference data and $eva-lect$ for lecture data in the next sections.

Analysis of the different corpora leads to the following observations. Regarding the *dev* corpus, we may note the presence of short silence periods, which implies some difficulties to estimate the noise spectrum or the noise energy, low SNRs (minimum average SNR -5.4 dB; 23.75% of files with SNR < 0 dB and 65% of files with SNR < 3 dB), a variable recording level and a bad use of the input scale (a file with a maximum level of 2% of scale and 58.75% of files with a maximum level <50% of scale), and finally several speakers with overlapped speaking segments.

Some similar observations can be made on the $eva - conf$ corpus (same subdomains as the *dev* corpus): short silence periods, with similar consequences, low SNRs (minimum average SNR -1.95 dB; 7.5% of files with SNR < 0 dB and

6.2% of files with SNR < 3 dB), a variable recording level and a bad use of the input scale (a file with a maximum level of 11% of scale and 35% of files with a maximum level <50% of scale), and several speakers with overlapped speaking segments.

Finally, the *eva−lect* corpus reveals some marginal characteristics, enforced by the shortness of the utterances: low SNR (minimum average SNR -2.1 dB; 6.2% of files with SNR < 0 dB and 15.17% of files with SNR < 3 dB), predominantly one speaker per record, and better use of input signal scale.

5.2 Results and Discussion

SAD task. Table 2 shows the performance of the Speech Activity Detection system on the $eva - conf$ and $eva - lect$ corpora in terms of Missed Speaker Error (MiE) and False Alarm Speaker Error (FaE) rates.

Table 2. Results of SAD on RT'05S

Task	MiE	FaE
$eva - conf$	5.3	2.1
$eva - lect$	5.4	1.2

We can observe that the SAD obtains comparable performance on the $eva - conf$ and $eva - lect$ test sets but presents, on both, large Missed Speaker Error rates (≈5.4%). For comparison, the best SAD system has obtained about 5% in terms of both Missed and False Alarm Speaker error rates during the RT'05S campaign.

Speaker diarization task. Experiments presented in this section aim at comparing the performance of the pre-processing techniques proposed in this paper when combined with the LIA speaker diarization system (*WMixSpectrum*, *WMix* and *MixLIA*) as well as at evaluating the robustness of broadcast news speaker diarization systems on the Meetings recordings (*MixLIA*, *MixCLIPS* and *MixLIUM*).

First, all the submitted speaker diarization systems have been evaluated on the *dev* corpus as presented in Table 3. Here, the system performance is expressed in terms of Missed speaker Error (MiE), False Alarm speaker Error (FaE) and Speaker Diarization Error (SDE) rates (the latter include both the MiE and FaE rates as well as the speaker error rate). Details on each meeting are provided as well as the global performance on the *dev* corpus.

The use of the multi-channel information (*WMixSpectrum* and *WMix*), extracted thanks to the pre-processing techniques does not improve globally the speaker diarization performance on the *dev* corpus but obtains very close results from the baseline system (simple sum of the multiple microphone signals: MixLIA). Nevertheless, signal analysis shows that the pre-processing algorithms improve the global SNR of resulting virtual signals; for example, in the case of

Table 3. Results on development corpus (*dev*)

Meetings	SAD MiE	SAD FaE	WMixSpectrum SDE	WMix SDE	MixLIA SDE	MixCLIPS SDE	MixLIUM SDE
CMU 20020319-1400	0.5	5.5	57.9	57.9	57.9	**46.9**	**46.9**
CMU 20020320-1500	0.1	5.3	20.2	20.2	20.2	**18.5**	**18.5**
ICSI 20010208-1430	0.4	3.1	16.5	17.0	19.3	22.5	**13.4**
ICSI 20010322-1450	0.4	1.4	19.6	**13.6**	16.7	17.0	24.6
LDC 20011116-1400	0.4	2.9	**4.5**	15.4	8.0	6.9	7.8
LDC 20011116-1500	0.4	1.6	18.7	12.2	**8.1**	15.8	13.3
NIST 20020214-1148	0.2	8.1	25.4	**16.8**	17.3	22.8	27.2
NIST 20020305-1007	0.0	3.5	33.0	47.8	44.6	**9.4**	19.0
CMU 20030109-1530	0.1	0.7	34.2	34.2	34.2	**27.9**	32.2
CMU 20030109-1600	2.5	1.3	33.5	33.5	33.5	**20.7**	33.5
ICSI 20000807-1000	0.0	3.6	21.2	17.1	**16.2**	17.1	16.3
ICSI 20011030-1030	0.0	3.4	41.4	37.0	**32.3**	51.8	49.4
LDC 20011121-1700	0.0	2.2	32.0	6.7	**3.3**	28.7	39.6
LDC 20011207-1800	0.0	8.6	**26.5**	40.3	44.2	35.7	34.7
NIST 20030623-1409	0.0	1.1	18.9	18.4	24.7	30.5	**11.6**
NIST 20030925-1517	0.4	16.3	64.3	52.0	51.8	70.7	**48.6**
Global performance	0.3	4.1	27.8	26.6	26.2	**25.7**	26.0

LDC 20011121-1700 meeting the unweighted mix leads to a global SNR of -3.88 dB (SNR$\in$ [-10.1;2]dB) to be compared with -0.1 dB (SNR$\in$ [-6.2;5.69]dB) for the *Weighted Mix - SAD* algorithm and with -0.59 dB (SNR$\in$ [-5.0;5.34]dB) for the *Weighted Mix - Noise spectrum* algorithm. The improvement of the SNR on the unique virtual signal does not seem to be helpful for the speaker diarization systems.

Table 4 presents the official results obtained on the RT'05S evaluation corpus for both the conference ($eva-conf$ corpus) and lecture room ($eva-lect$ corpus) recordings (the *WMixSpectrum* system has not been applied on the conference subdomain test set for the evaluation campaign). The best results have been obtained using the two proposed pre-processing techniques as opposed to the results reached on the RT'04S Meeting data (*dev* corpus). The comparison between the simple unweighted sum method and the weighted ones shows a gain of 15% relative on the $eva-conf$ corpus and of 56% on the $eva-lect$. The better quality, in terms of SNR, of RT'05S data can explain the better performance of the systems based on weighted sums. In fact the same SNR gain observed on both RT'04S and RT'05S does not have the same influence in terms of speaker

Table 4. Official results reached for the RT'05S (on *eva-conf* and *eva-lect* corpora)

Show	SAD		WMixSpectrum	WMix	MixLIA	MixCLIPS	MixLIUM
	MiE	FaE	SDE	SDE	SDE	SDE	SDE
eva-conf	4.0	3.0	-	**24.5**	27.7	25.0	30.5
eva-lect	5.6	1.3	**18.4**	21.4	34.2	35.3	20.0

diarization performance according to the initial signal quality. This result tends to demonstrate the relevance of the proposed strategy: designing a virtual signal based on a weighted sum of the multiple microphone recordings.

Concerning the robustness of the different speaker diarization systems against environment changes, it may be observed that their overall performance has significantly decreased on meeting data (about 21% Speaker Error rate) compared with broadcast news (about 12% Speaker Error rate [1] for the French ESTER evaluation broadcast news corpus), even though it is often difficult to compare results obtained on different databases.

Unfortunately, pre-processing techniques applied on multiple microphone signals do not seem to be sufficient to deal with meeting data issues and to avoid specific speaker diarization systems.

6 Conclusions

This paper is concerned with the speaker diarization task in the specific context of meeting data. More precisely, the focus is made on the handling of multiple microphone signals available per meeting. In this framework, a novel approach is experimented based on the rebuilding of a unique and virtual signal, composed of the most pertinent portions of signals issued from the different multiple microphone recordings. The extraction of these pertinent portions is carried out according to a signal quality index using the signal to noise ratio estimate.

Coupled with different speaker diarization systems developed by three different labs: the LIA, LIUM and CLIPS, the proposed approach has been submitted for the NIST 2005 Spring Rich Transcription evaluation campaign (RT'05S). According to the results obtained on the RT'05S evaluation, the use of this pre-processing strategy, which takes advantage of the multi-channel information, seems to have a slight positive influence on the speaker diarization performance.

This study was also focused on the behavior of speaker diarization systems, tuned on broadcast news and tested on meeting data. One assumption was that the application of the pre-processing techniques and the production of the unique and virtual signal would be sufficient to ensure similar performance between broadcast news and meeting corpora. Nevertheless, the level of performance is quite different between both of them. Even though the pre-processing techniques proposed in this paper may still be improved to provide more pertinent virtual signal, further investigation has to be done to study the other particularities

of the meeting data (like spontaneous speech, overlap, ...), which are widely responsible for perturbations of the speaker diarization systems.

References

[1] Galliano, S., Geoffrois, E., Mostefa, D., Choukri, K., Bonastre, J.F., Gravier, G.: The ESTER phase II evaluation campaign for the rich transcription of French broadcast news. In: EuroSpeech'05, Lisboa, Portugal (2005)

[2] Bonastre, J.F., Wils, F., Meignier, S.: ALIZE, a free toolkit for speaker recognition. In: ICASSP'05, Philadelphia, USA (2005)

[3] Cui, X., Bernard, A., Alwan, A.: A noise-robust ASR back-end technique based on weighted viterbi recognition. In: EuroSpeech'03, Geneva, Switzerland (2003)

[4] Hirsh, H.G.: Estimation of noise spectrum and its application to SNR-estimation and speech enhancement. Technical report tr-93-012, ICSI, Berkeley, USA (1993)

[5] Meignier, S., Moraru, D., Fredouille, C., Besacier, L., Bonastre, J.F.: Benefits of prior acoustic segmentation for automatic speaker segmentation. In: ICASSP'04, Montreal, Canada (2004)

[6] Moraru, D., Meignier, S., Fredouille, C., Besacier, L., Bonastre, J.F.: The ELISA consortium approaches in broadcast news speaker segmentation during the NIST 2003 rich transcription evaluation. In: ICASSP'04, Montreal, Canada (2004)

[7] Meignier, S., Bonastre, J.F., Fredouille, C., Merlin, T.: Evolutive HMM for speaker tracking system. In: ICASSP'00, Istanbul, Turkey (2000)

[8] Meignier, S., Moraru, D., Fredouille, C., Bonastre, J.F., Besacier, L.: Step-by-step and integrated approaches in broadcast news speaker diarization. Computer and Speech Language Journal (accepted for publishing in 2005)

[9] Siu, M.H., Rohlicek, R., Gish, H.: An unsupervised, sequential learning algorithm for segmentation of speech waveforms with multi speakers. In: ICASSP'92. Volume 2., San Fransisco, USA (1992) 189–192

[10] Chen, S., Gopalakrishnan, P.: Speaker, environment and channel change detection and clustering via the bayesian information criterion. In: DARPA Broadcast News Transcription and Understanding Workshop, Landsdowne, USA (1998)

[11] Zhu, X., Barras, C., Meignier, S., Gauvain, J.L.: Combining speaker identification and BIC for speaker diarization. In: EuroSpeech'05, Lisboa, Portugal (2005)

[12] Delacourt, P., Wellekens, C.J.: DISTBIC: A speaker based segmentation for audio data indexing. Speech Communication **32** (2000) 111–126

[13] Gauvain, J., Lamel, L., Adda, G.: Audio partitioning and transcription for broadcast data indexation. Multimedia Tools and Applications (2001) 187–200

[14] Reynolds, D.A., Dunn, R.B., Laughlin, J.J.: The Lincoln speaker recognition system: NIST EVAL2000. In: ICSLP'00. Volume 2., Beijing, China (2000)

[15] Adami, A., Kajarekar, S.S., Hermansky, H.: A new speaker change detection method for two-speaker segmentation. In: ICASSP'02. Volume IV., Orlando, USA (2002) 3908–3911