Robust Speaker Diarization for Meetings: ICSI RT06S Meetings Evaluation System

Xavier Anguera[1,2], Chuck Wooters[1], and Jose M. Pardo[1,3]

[1] International Computer Science Institute, Berkeley CA 94704, USA
[2] Technical University of Catalonia, Barcelona, Spain
[3] Universidad Politecnica de Madrid, Madrid, Spain
{xanguera,wooters,jpardo}@icsi.berkeley.edu

Abstract. In this paper we present the ICSI speaker diarization system submitted for the NIST Rich Transcription evaluation (RT06s) [1] conducted on the meetings environment. The presented system is based on the RT05s system, which uses agglomerative clustering with a modified Bayesian Information Criterion (BIC) measure to decide which pairs of clusters to merge and to determine when to stop merging clusters. In this year's system we have eliminated any remaining need for training data, therefore increasing robustness. In our primary system we have introduced several improvements from last year. First, we use a new training-free speech/non-speech detection algorithm. Second, we introduce a new algorithm for system initialization. The third improvement is the use of a frame purification algorithm to increase cluster discriminability. Finally, we describe the use of inter-channel delays as features. We explain each of these improvements and show our system's results on the official evaluation data using hand-aligned references and forced-alignments. We also analyze some of the results and propose improvements.

1 Introduction

The goal of a diarization system is to locate homogeneous regions within an audio segment and consistently label them for speaker, gender, music, noise, silence, etc. Within the framework of the Rich Transcription 2006 Spring Meeting Recognition Evaluation, the labels of interest were solely speaker and silence regions. This year's evaluation continues to focus on two meeting subdomains: the conference room, as in the RT04s and RT02s evaluations, and the lecture room, with seminar-like meetings. In each subdomain, a test set of about two hours was distributed. Participant's systems were asked to answer the question "Who spoke when?". The systems were not required to identify the actual speakers by name, but just to consistently label segments of speech from the same speaker. Prior art in this task can be seen in the different systems participating at RT05s [2], [3], [4]. Performance was measured based on the percentage of audio that was incorrectly assigned. This year was our second participation in the speaker diarization task. The speaker diarization system we used is based on last year's system (see [5]). Our system is based

S. Renals, S. Bengio, and J. Fiscus (Eds.): MLMI 2006, LNCS 4299, pp. 346–358, 2006.

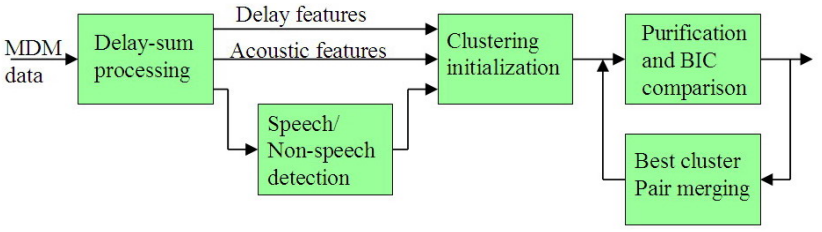

Fig. 1. RT06s Speaker Diarization system blocks diagram

on an agglomerative clustering system developed by Ajmera et al. (see [6]). Its primary advantage is that it requires no pre-trained acoustic models and therefore is robust and easily portable to new tasks.

Some of the improved algorithms are: A new hybrid speech/non-speech detector which combines an energy-based detector with a model based decoder back-to-back in order to avoid the need for outside training data. Also, a new system initialization and an automatic technique for selecting the number of initial clusters. We have also introduced an improved delay&sum algorithm to enhance the signal when multiple acoustic channels are available and a new frame-based purification algorithm that replaces last year's segment-based algorithm and enhances cluster discriminability. Finally, the use of inter-channel time differences as an extra feature stream for the diarization system.

In next section we review the general blocks on which the MDM system is based, sections 3 through 7 introduce the main changes in the system from the last submission in RT05s. Section 8 introduces the use of forced-alignments for this year's development and section 9 presents the main characteristics of the systems submitted. Finally, section 10 shows the systems results and 12 draws some conclusions.

2 Speaker Diarization System

As explained in [5], our speaker diarization system is based on an agglomerative clustering technique. Its main blocks are shown in figure 1 for the case of multiple microphones. It initially splits the data into K clusters (where K must be greater than the number of speakers and is chosen using the algorithm presented in [7]), and then iteratively merges the clusters (according to a metric based on ΔBIC) until a stopping criterion is met. Our clustering algorithm models the acoustic data using an ergodic hidden Markov model (HMM), where the initial number of states is equal to the initial number of clusters (K). Upon completion of the algorithm's execution, each remaining state is taken to represent a different speaker. Each state in the HMM contains a set of MD sub-states, imposing a minimum duration on the model (we use $MD \simeq 3$ seconds). Within the state, each one of the sub-states shares a probability density function (PDF) modelled via a Gaussian mixture model (GMM) for each particular data-stream.

The system works as follows:

1. If more than one recorded channel is available for a given meeting recording, combine them all into a single "enhanced" channel using a delay&sum algorithm further described in [8].
2. Run speech/non-speech detection on the "enhanced" data using the speech/non-speech algorithm presented in [9] and explained in section 3.
3. Extract acoustic and delay features from the data and remove non-speech frames from the agglomerative processing.
4. Estimate the number of initial clusters K using the algorithm presented in [7].
5. Create models for the K initial clusters using the new cluster initialization algorithm explained in section 4 and in [10].
   (a) Run a Viterbi decode to resegment the data.
   (b) Retrain the models using the Expectation-Maximization (EM) algorithm and the segmentation from step (a). Iterate between (a) and (b) until the segmentation stabilizes.
   (c) Select the cluster pair with the largest merge score (based on ΔBIC) that is > 0.0 using the frame purification technique introduced in [11] and section 5.
   (d) If no such pair of clusters is found, stop and output the current clustering.
   (e) Merge the pair of clusters found in step (c). The models for the individual clusters in the pair are replaced by a single, combined model.
   (f) Go to step (a).

As the stopping criterion for clustering and a distance measure for merging, we use a variation of the commonly-used BIC [12]. The variation that we use was introduced by Ajmera et al. [6], and consists of the elimination of the tunable parameter λ by ensuring that, for any given ΔBIC comparison, the difference between the number of free parameters in both models is zero.

One of the main overall changes for this year is that we eliminated all remaining dependency of our system on training data. This was achieved by the creation of a training-free speech/non-speech detector introduced in the next section. Furthermore, this year we introduce the use of data other than acoustic data for clustering by successfully using the delays between channels (in the MDM condition) as a new feature stream in the agglomerative clustering. This is further explained in section 6. Apart from these, a new clustering initialization algorithm and a frame purification algorithm contributed to the increase in the system's robustness and therefore improved its performance. Last year's segment purification algorithm was not used this year. The following sections introduce all these techniques.

3 Speech/Non-speech Detection Algorithm

In speaker diarization it is important to use a speech/non-speech detector as non-speech frames adversely affect the clustering performance. In the RT05s

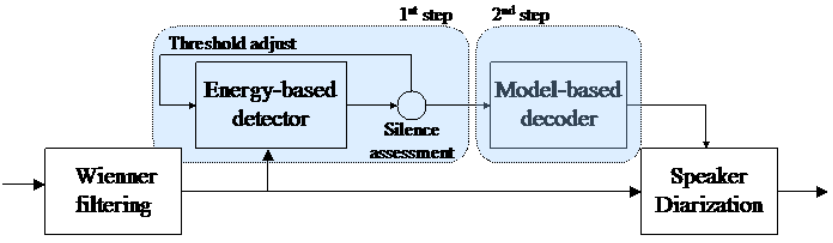

Fig. 2. Speech/non-speech block diagram

evaluation the speech/non-speech system we were using was based on pre-trained acoustic models for both speech and non-speech. This forced the readjustment of the models every time a new environment was to be processed, e.g. "conference room" versus "lecture room" data. For this year's evaluation we have developed a new speech/non-speech detector [9] that is train-free and therefore more robust to unseen data, as long as the main non-speech event in the recording is silence (which is a common trait of meeting data).

The system shown in figure 2 is a hybrid energy-based detector and model-based decoder. In the first stage, an energy-based detector finds all segments with low energy, while applying a minimum segment duration. An energy threshold is set automatically to obtain enough non-speech segments. In the first pass it takes a very low value and it increases incrementally while the number of non-speech frames falls under 100 and bigger than 10 (chosen empirically). At that point the segmentation is used to train speech and non-speech models in the second module and then several iterations of Viterbi segmentation and model retraining take place, finally outputting the speech/non-speech segmentation when the likelihood converges. In the system we need to define three parameters: the minimum durations for speech/non-speech in the energy module, minimum duration for speech/non-speech in the cluster module and the number of components used to model speech and non-speech in the cluster module. The parameters were tuned using the forced-alignment segmentations on the development set. As shown in [9] even though the miss and false alarm errors are equivalent to those obtained using the pre-trained system, the new system is more robust to changes in the data and appears to be a better fit with the following diarization module.

4 Cluster Initialization Algorithm

In order for the agglomerative clustering to work properly in obtaining the optimum number of speakers for a particular recording, we need to initialize the system with K (where $K > K_{true}$ the true number of speakers) clusters containing acoustically homogeneous data.

Past experiments, using k-means initialization and other techniques, have indicated that one very good option was to do a linear initialization of the data, where K clusters are generated by evenly splitting the acoustic data and then

performing several iterations of model training and resegmentation to allow for homogeneous acoustic data to come together. Although a very simple technique that works extremely well for some cases, in many other cases the resulting clusters contain more than one speaker which affects the (5c-d above) stopping criterion causing the final DER to increase.

The new initialization algorithm, explained in [10], consists of three stages of processing. First, speaker change detection using the Bayesian Information Criterion (BIC) metric (modified not to use a penalty term, as in out clustering system) is used to define acoustically similar segments by finding speaker change points via a scrolling window composed of two one-second regions. The second stage performs a bottom-up clustering by iteratively choosing speaker segments close to an initial segment (friends) to form one cluster and then selecting the segment most dissimilar to all existent clusters (enemy) to initialize the next cluster. Once K clusters are defined, their models are created and a segmentation is performed to assign all remaining segments to either model (third step). Using this technique, we obtain an increase in cluster purity right after the initialization process and a general improvement of the overall DER.

5 Frame Purification for Cluster Comparison

By using an agglomerative clustering technique, the system's performance heavily relies on the metric used to compare the similarity between cluster pairs as well as the clustering stopping criterion. Non-speech data is one of the main causal factors of anomalous behavior, which is one of the reasons a speech/non-speech detector is being used prior to the clustering process. After filtering the non-speech, the data considered to be speech still contains small non-speech segments (normally silence segments in the meeting environment) and other unvoiced speech which affects the cluster's modelling and degrades discriminability between clusters.

The frame purification algorithm (explained in [11]) detects and prevents such acoustic frames from affecting the models during the BIC comparison. To do so, it uses a metric related to the likelihood of the frames given the acoustic model. It is shown that when the cluster model's complexity is greater than two gaussian mixtures, most non-speech frames obtain the highest likelihoods, indicating that these are modelled with a narrower variance. A nice improvement in the model's discriminability is obtained by removing all frames with scores in the top 20% of the likelihood when training models for BIC comparison. This method is demonstrated to work better than filtering based on average frame energy [11].

6 Use of Inter-channel Delays in Clustering

Possibly this year's most effective improvement is the inclusion of inter-channel delays for the tasks where more than one microphone is available (see [13]). The delays are a byproduct of the delay&sum processing. For inclusion in the clustering, the delays are computed between a reference channel and all other

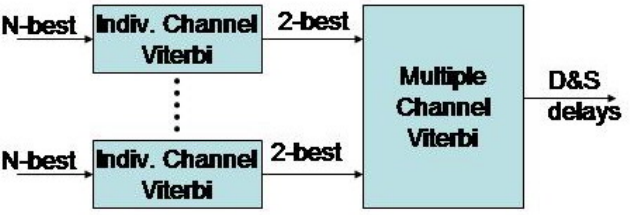

Fig. 3. Delay and Sum double-viterbi delays selection

channels at the same rate as the acoustic features and then post-processed in the same way as in the delay&sum presented below. The delays are initially modelled using single gaussian mixtures, with the same minimum duration as the acoustic features and share the speaker segmentation with the acoustic models. When two clusters are set to be merged their delay models are combined in the same way as the acoustic models.

Both the delay models and the acoustic models are used to classify the data into the different clusters via a Viterbi segmentation and for cluster comparison using BIC. We make the assumption that delay and acoustic information is uncorrelated and therefore can be modelled with separate models. The joint log-likelihood for any given frame is computed as:

$$lkld(x_{aco}[n], x_{del}[n]|\Theta_{aco}, \Theta_{del}) = \alpha \cdot lkld(x_{aco}[n]|\Theta_{aco}) + (1-\alpha) \cdot lkld(x_{del}[n]|\Theta_{del}) \quad (1)$$

Where Θ_{aco} is the acoustic model, Θ_{del} is the delay model and α weights the effect of each model in the system. The value for α needs to be tuned using development data. In our work, we found that a good value for $\alpha \sim 0.9$.

7 Delay&Sum Improvements

Whenever more than one channel is available for processing, a delay&sum beamforming is applied in order to obtain one single "enhanced" channel. The system used is based on last year's (see [5]), with four added improvements. The first improvement affects the noise filtering. Last year's submission filtered out any delay with cross-correlation value smaller than 0.1 since very low signal correlations indicate less reliability. This caused noisy meetings (or recordings with the lowest quality microphones) to have more frames filtered than in "cleaner" meetings. This year's submission computes a global histogram of all delays, in all channels, and determines the threshold at 10%. As in last year's system, any frame labelled as noisy is replaced by the delay from the previous usable frame, ensuring continuity of the delays.

Another improvement this year involves the delays selection among the N-best GCC-PHAT. As seen in figure 3 we apply a 2-level Viterbi decoding. The first

level consists of a local individual-channel decoding where the 2-best delays are chosen from the N-best delays computed for that channel at every frame. Each possible state has an emission probability equal to the GCC-PHAT value for each delay, and the transition probability between two nodes is inversely proportional to the distance between its delays, ensuring that the N-best probabilities in a particular instant sum up to 1. The second-level viterbi decoding finds the best possible path given all combinations of delays from the 2-best delays in each channel. The emission probabilities are the product of the individual GCC-PHAT values of each considered delay, and the transition probabilities are computed as in the first step, summing all delays distances from all considered delays, and normalized to sum to 1. In both cases the transition probabilities are weighted to emphasize its effect in the decision of the best path (we use a weight equal to 25 in both steps). This newly-introduced technique aims at finding the optimum tradeoff between reliability (cross-correlation) and stability (distance between contiguous delays). We value the second the most as our aim is to obtain an improved signal, avoiding quick changes of the beamforming between acoustic events.

The other two improvements affect the way that channels are summed after their relative delays are obtained. One of last year's post-eval improvements included an adaptive weighting for each individual channel (see [5]). This year we enhanced this concept by using the average cross-correlation between all channels (given the selected delays) to find the relative weights between the channels at each point. This value is also used to eliminate summing any channels with a relative weight smaller than $\frac{1}{4N}$ where N is the number of channels.

The delay&sum beamforming is used to enhance the signals to be used in this year's Speaker diarization systems as well as in the automatic speech recognition (ASR) submissions [14] for both conference and lecture tasks.

8 Use of Forced-Alignments

During this year's development period we experienced difficulties when using hand-made reference files, mostly when scoring on speaker overlap regions. By comparing the hand-made references with the acoustic data we observed that varying amounts of extra padding were inserted around each speaker overlap region, making its duration much longer than the actual acoustic event. We also observed some speaker overlap regions not labelled as such and some speaker overlap labels on non-speaker-overlap regions (although some speaker overlap might be noticed on the IHM channels, its volume is too low to be perceived in the MDM channels). All these artifacts create an extra amount of missed-speech error and of speaker error, which is not consistent over the different evaluation datasets (possibly as the transcription team changes their transcription guidelines). In general, we believe that the hand-made speaker segmentation references show too much transcriber dependency to be able to compare results from different years or to create a consistent and robust speaker diarization system.

For this year's system development we have taken the initiative to use references derives from forced-alignments. We generated the forced-alignments from the hand-transcribed spoken text with the individual IHM acoustic data. This was done at ICSI using the ICSI-SRI speech-to-text system presented for the RT05s evaluation ([15]). The use of forced aligned references was initially proposed by NIST for this year's meetings evaluation, although it was finally not applied.

In table 1 we compare the results using the same system output (a similar version to this year's primary MDM system) evaluated using either hand-aligned references or forced-alignments. We observe a change of between 2% and 5% in DER from non-overlap to overlap speech in the forced-alignment results, while there is a change from 6% to 15% in the hand-alignments, indicating the higher variability in the transcription of speaker overlaps. Additionally, in the evaluations up to RT05s, the non-overlap results are very similar between the hand-aligned and the FA, but in RT06s the difference is very large.

Table 1. Comparison of the DER for all meetings evaluation campaigns using hand-alignments or forced-alignments

Evaluation	MDM Hand-align		MDM Force-align	
campaign	non ovl.	ovl.	non ovl.	ovl.
RT02s	20.79%	26.95%	19.93%	21.89%
RT04s	15.44%	30.55%	13.98%	17.01%
RT05s	10.41%	18.73%	12.52%	15.06%
RT06s	23.06%	36.99%	16.46%	21.19%

Due to the fact that we performed our development experiments using force-aligned references while the eval was scored using hand-alignments, we observed a large increase in our missed-speech error. In most cases this is due to the difference in the extra padding applied to the speaker overlap regions and to the difference in the non-speech labelling criteria (the rule of 0.3sec minimum is applied to the forced-alignments).

9 Evaluation System Descriptions

This year we presented a total of 23 systems in the multiple tasks and subtasks of the evaluation. Each system uses one or more of the improvements presented above. Across tasks, systems with the same ID are equal or very similar, just differing on a few parameters. Their characteristics are:

p-wdels: This is the primary system presented this year for all multi-microphone conditions. It uses all proposed techniques in this paper, and all changes in the diarization code from last year's evaluation.

c-newspnspdelay: This system is presented for the multi-microphone cases and is composed of RT05s evaluation code using this year's delay&sum algorithm, this year's hybrid speech/non-speech detector and taking advantage of the delays for clustering. It uses a minimum duration of 3 seconds, 1/5 initial gaussian mixtures for delays/acoustics and a split weight of 0.1/0.9 between the streams. It is intended to measure the improvements of using the delay features and the new speech/non-speech detector.

c-wdelsfix: This system is identical to p-wdels in all parts except the decision of the initial number of clusters, which is fixed to 16 and 10 clusters for conference and lecture rooms, respectively. It intends to compare the robustness of the initial number of clusters selection.

c/p-nodels: This system contains all of this year's improvements with respect to delay&sum (when available, in MDM), speech/non-speech detection and other diarization algorithms except the inclusion of the delays as an extra feature stream.

c-oldbase: This system uses all improvements in delay&sum (when available, in MDM) and speech/non-speech detection while using the RT05s core speaker diarization system. It is meant to serve as a baseline result for systems this year.

c-guessone: This system guesses one speaker for the entire show. In RT05s we presented this system as our primary system for lecture room data. Since the lecture room data is primarily composed of a single speaker, we believe that this is a reasonable baseline. This year we again present this system as a baseline lecture-room system to be compared with our other lecture-room systems.

10 Evaluation Results

In this section we present the scores for all of the ICSI systems presented in the RT06s evaluation in the speaker diarization (SPKR) task and the speech activity detection (SAD) task. In tables 2 and 3, we show the SPKR results both for conference and lecture room data, and in table 4 we show results for SAD. In all cases we use both the official hand-made references and the forced-alignment references computed as explained previously. In general this year's results using hand-alignments are much worse than in previous years for conference room, which is not so pronounced when evaluating using the forced alignments. This might be due to the increased complexity of the data and of a decrease in the quality of the hand-generated transcriptions for this year's evaluation.

In the SPKR task for conference room a substantial improvement can be seen between the first three systems in MDM and the last two due to using delays as features in diarization. In lecture room data (Table 3, third column) the use of

Table 2. Results for Speaker Diarization, conference room environment

Cond.	System ID	%DER MAN	%DER FA
MDM	p-wdels	35.77%	19.16%
	c-newspnspdelay	35.77%	20.03%
	c-wdelsfix	38.26%	23.32%
	c-nodels	41.93%	27.46%
	c-oldbase	42.36%	27.01%
SDM	p-nodels	43.59%	28.25%
	c-oldbase	43.93%	28.21%

Table 3. Results for Speaker Diarization, lecture room environment

Cond.	System ID	%DER MAN	%DER MAN(subset)	%DER FA(subset)
ADM	p-wdels	12.36%	11.54%	10.56%
	c-nodels	10.43%	10.60%	9.71%
	c-wdelsfix	11.96%	12.73%	11.58%
	c-guessone	25.96%	23.36%	24.51%
MDM	p-wdels	13.71%	11.63%	10.97%
	c-nodels	12.97%	13.80%	13.09%
	c-wdelsfix	12.75%	12.95%	12.34%
	c-guessone	25.96%	23.36%	24.51%
SDM	p-nodels	13.06%	12.47%	11.69%
	c-guessone	25.96%	23.36%	24.51%
MSLA	p-guessone	25.96%	23.36%	24.51%

delays affects negatively the performance, possibly due to talkers moving around the room (delays argue for a different speaker for each location).

In general the more microphones available for processing, the better the results. As the diarization system is the same, the improvement is due to the delay&sum processing. This is clear in the conference room data, while in the lecture room data, the results are mixed. We believe this is due to the difference in quality between the microphone used in SDM and all others.

In the lecture room results shown in Table 3 we compare the manual and forced-alignment DER for all systems submitted. The third column shows the results using the latest release of the manual reference segmentations (18 meeting segments). When generating the forced-alignments using the IHM channels from each individual speaker we could not produce them for the meeting segments containing speakers not wearing any headset microphone. The last column shows results using forced-alignment references for a subset of 17 meeting segments containing all speakers who wore a headset microphone. The second to last column shows results using this same subset and using hand-alignments for comparison purposes.

Results using FA references are much better than using hand-alignments in the conference room, while they remain similar in lecture room (with a constant improvement of 0.5% to 1% for FA). We believe the conference room manual

Table 4. Results for Speech Activity Detection (SAD). Results with * are only for a subset of segments.

Env.	Cond.	%DER MAN (%MISS, %FA)	%DER MAN(subset)	%DER FA
Conference	MDM	23.51 (22.76, 0.8)	–	11.10 (7.80, 3.30)
	SDM	24.95 (24.24, 0.8)	–	11.50 (8.80, 2.70)
Lecture	ADM	13.22 (9.3, 3.9)	7.9* (5.0, 2.9)	7.2* (3.7, 3.5)
	MDM	13.83 (9.3, 4.5)	6.5* (5.0, 1.5)	5.6* (3.6, 2.0)
	SDM	14.59 (10.0, 4.6)	7.2* (4.5, 2.7)	6.7* (3.3, 3.4)

references still contain many problems, which have been filtered out in the lecture room references after several redistributions of references.

In table 4 we show the results of our systems on conference and lecture room data for the SAD task, using the new speech/non-speech detector developed for this year's evaluation.

This year's speech/non-speech detector was developed using forced-alignment (FA) data. Therefore the results of the SAD are better as shown in the forced-alignment column. The increase in % MISS in the hand-aligned conference data is probably due to silence regions (greater than 0.3s) that are correctly labelled by the FA transcriptions but are considered speech by the hand-alignments.

As we did for the diarization experiments, we created a subset of meetings to appropriately evaluate the lecture room systems using forced-alignment references, and the counterpart hand-alignments for completeness. One initial observation is that the error rate decreases dramatically when evaluating only a subset of the shows using hand-alignments. a possible explanation for this is transcription errors produced due to the lower quality of the non-headset microphones used in the eliminated set of meetings.

As in the diarization results, in these experiments we also obtain better results with more microphones. When comparing the forced-alignment with the hand-alignment subset, the first group keeps a better balance between misses and false alarms, indicating that parameters defined in development translate robustly to the evaluation data.

Overall, we see an improvement this year with the use of delays between microphones as a feature in the diarization process for conference room data, while mixed results are obtained in lecture room. Also, a general improvement is observed using delay&sum on as many microphone signals as possible.

11 Overlapping Speaker Detection

Given that this year's evaluation counts speaker overlap errors in the main metric, we initially spent some time trying to build an overlap detector. In all our experiments we managed to lower the missed speaker error but at the cost of increasing the overall diarization error. We stopped research in this area when we started developing our system using forced-alignments, as the speaker overlap error in this case is less than 5%.

We performed experiments both in the diarization module and in the beamforming module. In diarization we tried a final decoding pass using the resulting speaker models and also all combinations of speaker pairs in order to detect speaker overlap. In the beamforming module we tested several metrics comparing the N-best cross-correlation values under the assumption that two speakers get consistently two main peaks in the correlation function.

12 Conclusions

This paper presents ICSI's submissions to the RT06s speaker diarization and SAD evaluation campaigns. This year's system contains four major improvements from last year. They are: a new training-free speech/non-speech detector, a new initialization algorithm, an improved cluster comparison algorithm, and the use of inter-channel delays as features in the diarization process. In this paper we review the basic system operation and we describe each of the improvements. Results are shown for the submitted systems while comparing the suitability of using hand-alignments versus forced-alignment references. Finally we describe some experiments in detecting speaker overlap.

References

1. NIST rich transcription evaluations, website: http://www.nist.gov/speech/tests/rt.
2. D. Istrate, C. Fredouille, S. Meignier, L. Besacier, and J.-F. Bonastre, "NIST RT05S evaluation: Pre-processing techniques and speaker diarization on multiple microphone meetings," in *NIST 2005 Spring Rich Transcrition Evaluation Workshop*, Edinburgh, UK, July 2005.
3. S. Cassidy, "The macquarie speaker diarization system for RT05S," in *NIST 2005 Spring Rich Transcrition Evaluation Workshop*, Edinburgh, UK, July 2005.
4. D. van Leeuwen, "The TNO speaker diarization system system for NIST RT05s for meeting data," in *NIST 2005 Spring Rich Transcrition Evaluation Workshop*, Edinburgh, UK, July 2005.
5. X. Anguera, C. Wooters, B. Peskin, and M. Aguilo, "Robust speaker segmentation for meetings: The ICSI-SRI spring 2005 diarization system," in *RT05s Meetings Recognition Evaluation*, Edinburgh, Great Brittain, July 2005.
6. J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," in *ASRU'03*, US Virgin Islands, USA, Dec. 2003.
7. X. Anguera, C. Wooters, and J. Hernando, "Automatic cluster complexity and quantity selection: Towards robust speaker diarization," in *MLMI'06*, Washington DC, USA, May 2006.
8. ——, "Speaker diarization for multi-party meetings using acoustic fusion," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Puerto Rico, USA, November 2005.
9. X. Anguera, M. Aguilo, C. Wooters, C. Nadeu, and J. Hernando, "Hybrid speech/non-speech detector applied to speaker diarization of meetings," in *Speaker Odyssey 06*, Puerto Rico, USA, June 2006.
10. X. Anguera, C. Wooters, and J. Hernando, "Friends and enemies: A novel initialization for speaker diarization," in *Proc. ICSLP*, Pittsburgh, USA (to appear), September 2006.

11. ——, "Purity algorithms for speaker diarization of meetings data," in *Proc. ICASSP*, Toulouse, France, May 2006.
12. S. Shaobing Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *Proceedings DARPA Broadcast News Transcription and Understanding Workshop*, Virginia, USA, Feb. 1998.
13. J. M. Pardo, X. Anguera, and C. Wooters, "Speaker diarization for multiple distant microphone meetings: Mixing acoustic features and inter-channel time differences," in *Proc. ICSLP*, September 2006.
14. A. Janin, A. Stolcke, X. Anguera, K. Boakye, O. Cetin, J. Frankel, and J. Zheng, "The ICSI-SRI spring 2006 meeting recognition system," in *Proceedings of the Rich Transcription 2006 Spring Meeting Recognition Evaluation*, Washington, USA, May 2006.
15. A. Stolcke, X. Anguera, K. Boakye, O. Cetin, F. Grezl, A. Janin, A. Mandal, B. Peskin, C. Wooters, and J. Zheng, "Further progress in meeting recognition: The icsi-sri spring 2005 speech-to-text evaluation system," in *RT05s Meetings Recognition Evaluation*, Edinburgh, Great Brittain, July 2005.