Robust Speaker Segmentation for Meetings: The ICSI-SRI Spring 2005 Diarization System

Xavier Anguera[1,2], Chuck Wooters[1], Barbara Peskin[1], and Mateu Aguiló[2,1]

[1] International Computer Science Institute, Berkeley CA 94704, USA,
[2] Technical University of Catalonia, Barcelona, Spain
{xanguera, wooters, barbara, mateu}@icsi.berkeley.edu

Abstract. In this paper we describe the ICSI-SRI entry in the Rich Transcription 2005 Spring Meeting Recognition Evaluation. The current system is based on the ICSI-SRI clustering system for Broadcast News (BN), with extra modules to process the different meetings tasks in which we participated. Our base system uses agglomerative clustering with a BIC-like measure to determine when to stop merging clusters and to decide which pairs of clusters to merge. This approach does not require any pre-trained models, thus increasing robustness and simplifying the port from BN to the meetings domain. For the meetings domain, we have added several features to our baseline clustering system, including a "purification" module that tries to keep the clusters acoustically homogeneous throughout the clustering process, and a delay&sum beamforming algorithm which enhances signal quality for the multiple distant microphones (MDM) sub-task. In post-evaluation work we further improved the delay&sum algorithm, experimented with a new speech/non-speech detector and proposed a new system for the lecture room environment.

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, etc. Within the framework of the Rich Transcription 2005 Spring Meeting Recognition Evaluation, the labels of interest were solely speaker regions. This year's evaluation expands its focus from last year and considers two meeting sub-domains: the conference room, as in previous NIST evals, and the lecture room, with seminar-like meetings. In each sub-domain a test set of about two hours was distributed. Participants' 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. Performance was measured based on the percentage of audio that was incorrectly assigned.

This year is the first time that we participated in the Diarization task for the Meetings environment. The clustering system used is based on our agglomerative clustering system originally developed by Ajmera et al. (see [1] [2] [3] [4]). Its primary advantage is that it requires no pre-trained acoustic models and therefore is robust and easily portable to new tasks. One new feature we have added to the system is a purification step during the agglomerative clustering process. The purification process attempts to split clusters that are not acoustically homogeneous. Another new feature we have added

is multi-channel signal enhancement. For the conditions where multiple microphones are available, we combine these multiple signals into a single enhanced signal using delay&sum beamforming. The resulting system performed well in the meetings environment, achieving official scores of 18.56% and 15.32% error for the MDM and SDM conference room conditions[3], and 10.41%, 10.43% and 9.98% error for the lecture room MDM, SDM and MSLA conditions[4].

In Section 2 we present the detailed description of the different parts in our system. In Section 3 we describe the systems submitted in the evaluation and their performance. In Section 4 we describe some improvements to the system that were made after the evaluation was submitted. Finally, ongoing and future work are presented in Section 5.

2 System Description

The system this year has two parts that are combined to adapt to the different tasks and data available. The first part consists of an acoustic fusion of all the available channels (when they exist) into a single enhanced channel via the delay-and-sum beamforming algorithm. The second part is our basic speaker diarization system, similar to the system submitted for the Fall 2004 Broadcast News evaluation (RT04f) (see [4]). The main differences in this second part are:

1. the use of an un-biased estimator for the variance together with minimum variance thresholding.
2. a purification algorithm to clean the clusters of non acoustically homogeneous data.
3. a major bug-fix in the core clustering system.

The delay&sum beamforming algorithm is used in some tasks where more than one microphone is available (i.e. MDM and MSLA for Diarization). It uses a sliding analysis window of length 500ms, with an overlap of 50%. At each step, a 500ms segment from each of the different channels is aligned to a reference channel producing a delay for that segment. The delay-adjusted segments are then summed to produce an enhanced output, which becomes the input of the basic diarization system. The delays are computed using GCC-PHAT and special care is taken to maintain continuity in the delays given non-speech and multiple speaker areas. For a more detailed description see section 2.1.

The second part of the system is our basic speaker diarization system. This system uses agglomerative clustering and begins by segmenting the data into small pieces. Initially, each piece of data is assigned to a separate cluster. The system then iteratively merges clusters and resegments, stopping when there are no clusters that can be merged. This procedure requires two measures: one to determine which pair of clusters to merge, and a second measure to determine when to terminate the merging process. In our baseline system, we use a modified version of BIC [5] for both of these measures. The modified BIC equation is defined as:

[3] After the evaluation we made some simple changes to the delay&sum algorithm that considerably changed these results.

[4] Although these are not the primary submission results, as explained below, these are obtained using the clustering system just described.

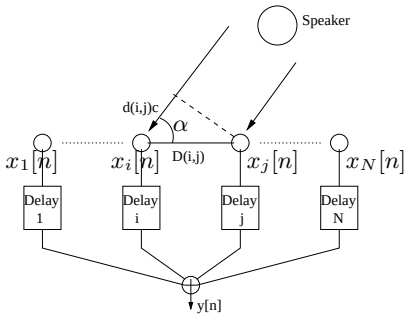

Fig. 1. *Delay-and-sum system*

$$\log p(D|\theta) \geq \log p(D_a|\theta_a) + \log p(D_b|\theta_b) \tag{1}$$

where:

- D_a and D_b represent the data in two clusters and θ_a and θ_b represent the models trained on the data assigned to the two clusters.
- D is the data from $D_a \cup D_b$ and θ represents the model trained on D.

Eq. 1 is similar to BIC, except that the model θ is constructed such that the number of parameters is equal to the sum of the number of parameters in θ_a and θ_b. By keeping the number of parameters constant on both sides of the equation, we have eliminated the traditional BIC penalty term. This increases the robustness of the system as there is no need to tune this parameter.

We can compute a merging score for θ_a and θ_b by combining the right and left-hand sides of Eq. 1:

$$\begin{aligned} \text{MergeScore}(\theta_a, \theta_b) = \\ \log p(D|\theta) - (\log p(D_a|\theta_a) + \log p(D_b|\theta_b)) \end{aligned} \tag{2}$$

2.1 Delay-and-Sum Beamforming

The delay&sum (D&S) beamforming technique [6] is a simple yet effective way to enhance an input signal when it has been recorded on more than one microphone. It doesn't assume any information about the position of the microphones or their placement. The principle of operation of D&S can be seen in Figure 1.

Given the signals captured by N microphones, $x_i[n]$ with $i = 0 \ldots N-1$ (where n indicates time steps) if we know their individual relative delays $d(0, i)$ (called Time Delay of Arrival, TDOA) with respect to a common reference microphone x_0, we can obtain the enhanced signal using equation 3.

$$y(n) = x_0[n] + \sum_{i=1}^{N-1} x_i[n - d(0, i)] \tag{3}$$

By adding together the aligned signals the usable speech adds together and the ambient noise (assuming it is random and has a similar probability function) will be reduced. Using D&S, according to [6], we can obtain up to a 3db SNR improvement each time that we double the number of microphones. We were able to obtain a 15.62% DER

using D&S over multiple microphones compared to 21.32% on SDM for the RT04s development set.

In order to estimate the TDOA between two segments from two microphones we used the generalized cross correlation with phase transform (GCC-PHAT) method (see [7]). Given two signals $x_i(n)$ and $x_j(n)$ the GCC-PHAT is defined as:

$$G_{PHAT}(f) = \frac{X_i(f)[X_j(f)]^*}{|X_i(f)[X_j(f)]^*|} \tag{4}$$

where $X_i(f)$ and $X_j(f)$ are the Fourier transforms of the two signals and $[\,]^*$ denotes the complex conjugate. The TDOA for these two microphones is estimated as:

$$\hat{d}_{PHAT}(i,j) = \underset{d}{argmax}\left(\hat{R}_{PHAT}(d)\right) \tag{5}$$

where $\hat{R}_{PHAT}(d)$ is the inverse Fourier transform of $G_{PHAT}(f)$. Although the maximum value of $\hat{R}_{PHAT}(d)$ corresponds to the estimated TDOA, we have found it useful to keep the top N values for further processing.

There are two cases where the GCC-PHAT computation can provide inaccurate estimates for speaker clustering. These are:

- The analysis window is mainly analyzing a non-speech portion of the signal. As we don't eliminate the regions of non-speech from the signal prior to delay&sum and due to the small size of the analysis window (500ms), when trying to estimate the TDOA from a non-speech region it returns a random delay value with a very small correlation. To avoid this we consider only TDOA estimates with GCC-PHAT values greater than 0.1 (of a normalized maximum value of 1), and carry over the previous estimates to the current segment otherwise.
- There are two or more people talking at the same time. In such cases the estimated TDOA will focus on one or another of the sources, producing an instability and diminishing the quality of the output. To solve this problem we compute the 8 biggest peaks of the GCC-PHAT in each analysis window and select the TDOA by magnitude but favoring TDOA continuity between consecutive analysis windows.

2.2 Speech/Non-Speech Detection

In this year's system we continue to use the SRI STT system's speech/non-speech (SNS) detector to eliminate the non-speech frames from the input to the clustering algorithm. Its use in our speaker diarization system was introduced in last year's RT04f evaluation. The SRI SNS system is a two-class decoder with a minimum duration of 30ms (three frames) enforced with a three-state HMM structure. The features used in the SNS detector (MFCC12) are different from the features used for the clustering. The resulting speech segments are merged to bridge short non-speech regions and padded. The speech/non-speech detector used in RT05s has been trained on meetings data (RT-02 devset data and RT-04s training data). The parameters of the detector were tuned on the RT05s meetings development data to minimize the combination of Misses and False Alarms as reported by the NIST mdeval scoring tool.

2.3 Signal Processing and System Initialization

For our system this year, we used 19 MFCC parameters, with no deltas. The MFCCs were computed over a 30 millisecond analysis window, stepping at 10 millisecond intervals. Before computing the features for each meeting, we extracted just the region of audio specified in the NIST input UEM files. The features are then calculated over this extracted region.

The first step in our clustering process is to initialize the models. This requires a "guess" at the maximum number of speakers (K) that are likely to occur in the data. We used K=10 for the conference room data and K=5 for the lecture room data. The data is then divided into K equal-length segments and each segment is assigned to one model. Each model's parameters are then trained using its assigned data. To model each cluster we use mixtures of gaussians with diagonal covariance matrix starting with 5 gaussians per model. These are the models that seed the clustering and segmentation processes described next.

2.4 Clustering Process

The procedure for segmenting the data consists of the following steps:

1. Run the SRI Meetings SNS detector.
2. Extract 19 MFCCs every 10ms.
3. Discard the non-speech frames.
4. Create the initial models as described above in Section 2.3.
5. The iterative merging process consists of the following steps:
   (a) Run a Viterbi decode to re-segment the data.
   (b) Retrain the models using the segmentation from (a).
   (c) Select the pair of clusters with the largest merge score (Eq. 2) that is > 0.0. (Since Eq. 2 produces positive scores for models that are similar, and negative scores for models that are different, a natural threshold for the system is 0.0.)
   (d) If no pair of clusters is found, stop.
   (e) Merge the pair of clusters found in (c). The models for the individual clusters in the pair are replaced by a single, combined model.
   (f) Run the purification algorithm (see section 2.5 for details) if the number of merging iterations is less than the initial number of clusters.
   (g) Go to (a).

2.5 Purification Algorithm

We have observed that the performance of our system is significantly affected by the way the models get initialized. Even though the initial models are re-segmented and retrained a few times during the clustering process, there are "impure" segments of audio that remain in a model in which they don't belong and negatively affect the final performance of the system. Such segments are either non-speech regions not detected by the SNS detector, or actual speech.

A particular segment of the audio that is quite dissimilar to the other segments in that model may not get assigned to any other model due to: a) the current model overfitting that data, or b) there is not another model that provides a better match.

The purification algorithm is a post-merging step designed to find these segments and extract them, thus "purifying" the cluster. The segments considered are continuous intervals as found in the Viterbi segmentation step. The algorithm that we use to do the purification is applied after each cluster merge as follows:

1. For each cluster, we compute the normalized likelihood (dividing the total likelihood by the number of frames) of each segment in the cluster given the cluster's model. The segment with the highest likelihood is selected as the one that best fits the model.
2. For each cluster, we compute the modified BIC score (as seen in eq. 2) between the best fitting segment (as found in the previous step) and each of the other segments. If all comparisons give a positive value, the cluster is assumed to be pure, and is not considered a candidate for purification.
3. The segment with the lowest score below a certain threshold (-50 in our system) is extracted from the cluster and is re-assigned to its own cluster.

The source cluster keeps the same number of gaussians; therefore the purification process increases the total number of gaussians in the system (because a new cluster is created in the last step above). The purification algorithm is executed at most only on the first K iterations of the resegmentation-merging processing. We observed an improvement of approx. 2% absolute using this technique on a development data set built from the RT04s data sets and AMI meetings.

3 Evaluation Performance

For the evaluation we used different combinations of the pieces presented above. Almost all of these combinations share several common attributes:

- 19^{th} order MFCC, no deltas, 30 msec analysis window, 10 msec step size.
- Each initial cluster begins with five gaussians.
- Iterative segmentation/training.
- Cluster purification.

The submitted systems are summarized in table 1.

3.1 Conference Room Systems

For the conference room environment we submitted one primary system in each of the MDM and SDM conditions. The MDM system uses delay&sum to acoustically fuse all the available channels into one enhanced channel. Then it applies the clustering to this enhanced channel. The SDM condition skips the delay&sum processing, as the system's input is already a single channel (from the most centrally located microphone according to NIST).

[1] This system uses a weighted version of delay&sum using correlations, as explained in 4.1.

System ID	room type	Task	Submission	Delay &sum	# Initial clusters	Cluster Min. duration	Mics used
p-dspursys	Conf.	MDM	Primary	YES	10	3 sec	All Available
p-pursys	Conf.	SDM	Primary	NO	10	3 sec	SDM mic.
p-omnione	Lect.	MDM	Primary	NO	n/a	n/a	n/a
c-spnspone	Lect.	MDM	Contrast	NO	n/a	n/a	n/a
c-ttoppur	Lect.	MDM	Contrast	NO	5	5 sec	Tabletop mic.
p-omnione	Lect.	SDM	Primary	NO	n/a	n/a	n/a
c-pur12s	Lect.	SDM	Contrast	NO	5	12 sec	SDM mic.
p-omnione	Lect.	MSLA	Primary	NO	n/a	n/a	n/a
c-nwsdpur12s	Lect.	MSLA	Contrast	YES	5	12 sec	All Available
c-wsdpur12s	Lect.	MSLA	Contrast	YES [1]	5	12 sec	All Available

Table 1. *Distinct configurations of the submitted systems*

3.2 Lecture Room System

In the lecture room environment we submitted primary systems for the tasks MDM, SDM and MSLA, and contrastive systems for MDM (two systems), SDM and MSLA (two systems). Following is a brief description for each of these systems and their motivation:

- MDM, SDM and MSLA primary condition (MDM/SDM/MSLA_p-omnione): We observed in the development data that on many occasions we were able to obtain the best performance by just guessing one speaker for the whole duration of the lecture. This is particularly true when the meeting excerpt consists only of the lecturer speaking, but is often also achieved in the question-and-answer section since many of the excerpts in the development data consisted of very short questions followed by long answers by the lecturer. We therefore presented these systems as our primary submissions, serving also as a baseline score for the lecture room environment. Contrary to what we observed in the development data, our contrastive ("real") systems outperformed our primary ("guess one speaker") submissions on the evaluation data.
- MDM using speech/non-speech detection (mdm_c-spnspone): This differs from the primary submission only on the use of the SNS detector to eliminate the areas of non-speech. On the development data we observed that non-speech regions were only labeled when there was a change of speakers, which never happened for the "all lecturing" sections. This system is meant to complement the previous one by trying to improve performance where between-speech silences are marked.
- MDM using the TableTop microphone (mdm_c-ttoppur): From the available five microphones in the lecture room, the TableTop microphone is clearly of much better quality than all the others. It is located in a different part of the room and is of a different kind, which could be the reason for its better performance. By using an SNR estimator we automatically selected the best microphone (which turned out to always be the TableTop, d05 microphone) and we applied the standard clustering system to it (using models with a five second minimum duration). No SNS detection was used in this system.
- SDM using the SDM channel with a minimum duration of 12 seconds for each cluster (sdm_c-pur12s): This uses our clustering system on the SDM channel. We

didn't use the SNS detector. We observed that using a minimum duration of 12 seconds, we could bypass the issue of silences marked as speech in the reference files and force the system to end with fewer clusters.
- MSLA with standard delay&sum (msla_c-nwsdpur12s): In order to combine the various available speaker-localization arrays, we included the delay&sum processing, using a random channel from one of the arrays as the reference channel. The enhanced channel that we obtained was then clustered using the 12 second minimum duration system.
- MSLA with weighted delay&sum (msla_c-wsdpur12s): In the time between the conference room and lecture room submissions, we experimented with a weighted version of the delay&sum algorithm with weights based on the correlation between channels (as described in 4.1).

3.3 Scores

The DER scores on non-overlapped speech for this year's evaluation as they were released by NIST are shown in the third column of table 2. The numbers in the fourth column reflect improvements after small bug fixes and serve as the baseline scores used in the remainder of this paper. In the systems using delay&sum, an improvement comes from fixing a small bug in our system that we detected after the eval (the 2% difference in conference room MDM is mainly due to the meeting VT_20050318-1430). In the (non trivial) lecture room systems, the improvement comes from using an improved UEM file for the show CHIL_20050202-0000-E2.

System ID	Room type	DER	post-eval DER
mdm_p-dspursys	Conf.	18.56%	16.33%
sdm_p-pursys	Conf.	15.32%	—
mdm_p-omnione	Lect.	12.21%	—
mdm_c-spnspone	Lect.	12.84%	—
mdm_c-ttoppur	Lect.	10.41%	10.21%
sdm_p-omnione	Lect.	12.21%	—
sdm_c-pur12s	Lect.	10.43%	10.47%
msla_p-omnione	Lect.	12.21%	—
msla_c-nwsdpur12s	Lect.	9.98%	9.66%
msla_c-wsdpur12s	Lect.	9.99%	9.78%

Table 2. *DER on the evaluation set for RT05s*

The use of delay&sum to enhance the signal before doing the clustering turned out to be a bad choice for the conference room systems, as the SDM DER is smaller than the MDM. In section 4.1 we consider what the possible problem could be and propose two solutions.

4 Post-Evaluation Improvements

In this section we present several improvements to the system that were introduced after the evaluation.

4.1 Individual Channel Weighting

After the conference room evaluation, we observed that the straightforward delay&sum processing we had performed using all available distant channels was suboptimal. We found that the quality of the delay&summed output was negatively affected when the channels are of different types or they are located far from each other in the room.

In the formulation of the delay&sum processing, the additive noise components on each of the channels are expected to be random processes with very similar probability distributions. This allows the noise on each channel to be minimized when the delay-adjusted channels are summed. In standard beamforming systems, this noise cancellation is achieved through the use of identical microphones placed only a few inches apart from each other.

In the meetings room we assume that all of the distant microphones form a microphone array. However, having different types of microphones changes the impulse response of the signal being recorded and therefore changes the probability distributions of the additive noise. Also when two microphones are far from each other the speech they record will be affected by noise of a different nature, due to the room's impulse response.

After the conference room evaluation we began working on different ways to individually weight the channels according to the quality of the signal. Here we present two techniques we have tried, plus their combination:

SNR based weighting: A well known measure of the quality of a speech signal is its Signal-to-Noise ratio (SNR). We estimate the SNR value for each channel for all of the evaluated portion of the meeting and we apply a constant weight to each segment of each channel upon summation.

To estimate the SNR value we use a tool provided by Hans-Guenter Hirsch which performs a 2-step process:

1. Detection of stationary segments based on a Mel frequency analysis using the short term subband energies for all subbands. As soon as the subband energy exceeds a certain threshold (defined as the average of the previous energies) this is considered a possible indication for the presence of speech. When a certain number of subbands exceed the threshold it indicates the start of a speech segment. Similar thresholding is used to determine the transition from speech to non-speech.
2. The SNR is computed as $10log_{10}(\frac{S}{N})$ where N is the RMS value of the non-speech parts and S is obtained from the RMS of the speech parts, considering that they are $X = S + N$. Such energy is computed over the "A" filtered data.

More information can be found in [8].

Correlation based weighting: The weighting value is adapted continuously during the duration of the meeting. This is inspired by the fact that the different channels will have different quality depending on their relative distance to the person speaking, which can change constantly during a recording.

The weight for channel i at step n ($\mathcal{W}_i[n]$) is computed in the following way:

$$\mathcal{W}_i[n] = \begin{cases} \frac{1}{\#\text{Channels}} & n = 0 \\ (1-\alpha) \cdot \mathcal{W}_i[n-1] + \alpha \cdot \text{xcorr}(i, \text{ref.}) & \text{otherwise} \end{cases} \quad (6)$$

where $xcorr(i, ref.)$ is the cross-correlation between the delay-adjusted segment for channel i and the reference channel. When i=reference, it is just the power of the reference channel. If the cross-correlation becomes negative it is set to 0.0. By experimenting on the development set we set $\alpha = 0.05$.

Combination of both techniques: We use the SNR to rank the channels and select the best as the reference channel. Then the process is identical to the correlation weighting.

In table 3 we can see the results of running these three proposed techniques on some of the multiple distant microphone conditions.

Submission Desc.	Baseline	SNR Weight	Xcorr Weight	SNR+Xcorr
MDM Conference room	16.33%	17.02%	16.17%	14.81%
MSLA Lecture Room	9.66%	8.94%	9.78%	9.83%

Table 3. *Effect of channel weighting on Eval DER scores*

For Conference room data the correlation technique performs better than the SNR, but when combined together they outperform both individual systems. In Lecture room (on MSLA microphones) the SNR constant weights technique works better that variable weighting. In fact, in the Lecture room environment by having most of the time a single speaker we benefit more from a fixed weight, contrary to when multiple speakers intervene, benefitting from variable weights.

In order to isolate the effect of the weighting techniques, we also ran them using perfect speech/non-speech labels, thus minimizing miss and false alarm errors. In table 4 we can see the resulting DER.

Submission Desc.	chan. Weights	DER
Conference room SDM	n/a	10.95%
Conference Room MDM	equal	11.55%
Conference Room MDM	correlation	10.50%
Conference Room MDM	SNR	10.60%
Conference Room MDM	SNR+corr	10.57%

Table 4. *DER on the evaluation set for RT05s using "perfect" speech/non-speech labels*

4.2 Energy Based Speech/Non-Speech Detector

In our effort to create a robust diarization system that doesn't require any training data and as few "tunable" thresholds as possible, we are experimenting with an alternative to the SRI speech/non-speech(SNS) detector used in this year's evaluation. In this section we present an energy-based detector that performs very well on the test data.

Given an input signal (raw or delay&summed) the processing is done on one minute non-overlapping windows. The signal is first normalized using the average of the largest 50 amplitude values (with outliers removed).

Each normalized segment is then butterworth filtered and also processed with a matched filter (31 points filter, i.e. 2ms) proposed by Li in [9] over the signal to: a) average the signal to round spiky energy regions, and b) create a derivative effect to emphasize the start and end points of the speech/non-speech regions.

The boundary between speech and non-speech regions is given by a double threshold: one to go from non-speech to speech and another to go from speech to non-speech (as implemented in NIST's Speech Quality Assurance Package, see [10]). A finite state machine is implemented to impose minimum durations of the speech and silence segments.

In table 5 we can observe the speech/non-speech error and the DER scores using this speech/non-speech detector on the different tasks. This test was only performed in the conference room domain as we haven't use a speech/non-speech detector in all our lecture room systems.

Submission Desc.	weights	SNS Error		full DER	
		Baseline	Energy-SNS	Baseline	Energy-SNS
SDM Conference room	n/a	4.7%	5.0%	15.32%	14.65%
MDM Conference room	equal	5.30%	3.7%	16.33%	13.93%
MDM Conference room	SNR+corr	5.3%	3.7%	14.81%	13.97%

Table 5. *Energy-based vs. model-based SNS on conference room environment*

4.3 Selective Lecture Room Clustering

On the lecture room data the submitted systems didn't make use of the information regarding the kind of excerpt that was being clustered. As noted by NIST, the excepts ending with E1 and E3 have only the lecturer speaking in them; therefore guessing that only one speaker speaks all the time consistently achieves the best performance. On the other hand, the excerpts ending with E2 belong to the Q&A sections, with more speakers and a structure that more closely resembles the conference room environment.

After the evaluation, we constructed a system to take advantage of this information. The system parses the lecture file name before processing and proceeds accordingly:

- E1 and E3: one speaker all the time
- E2: run the "normal" clustering system

In table 6 we present the results of running this system for the different possible sets of microphones.

Submission Desc.	Baseline DER	Sel. clust. DER
SDM Lecture room	10.47%	9.60%
MDM Lecture room	10.21%	8.75%
MSLA Lecture room	9.66%	9.38%

Table 6. *Selective Lecture room clustering DER*

5 Future Work

Our future work will continue to focus on the use of techniques that require no pre-trained models and as few "tunable" parameters as possible.

- Signal-processing related improvements:
  - Improve SNS without external training data. We will continue work on our energy-based SNS detector, specifically focusing on robustness to different environments including: Broadcast News, Meetings, and Conversational Telephone Speech.
  - Improve delay&sum processing and use extra information extracted from that processing (TDOA values, correlation weights, relative energy between microphones, etc.).
  - Explore the use of alternative front-end signal processing techniques. To date, we have limited our features to MFCC19. We would like to explore alternative front-end features.
- Improvements to the clustering algorithm:
  - Improve the cluster purification algorithm to better deal with SNS errors.
  - Explore the use of techniques from Speaker ID (modified to conform to our philosophy of "no pre-trained models") in the clustering algorithm.
  - Explore the use of alternative stopping and merging criteria.
- General improvements:
  - Bug fixes!
  - Error analysis.

6 Conclusion

The primary advantage of our speaker diarization system is that it requires no pre-trained acoustic models and therefore is robust and easily portable to new tasks. For this year's evaluation, we added a couple of new features to the system. One new feature is the purification step during the agglomerative clustering process. The purification process attempts to split clusters that are not acoustically homogeneous. Another new feature is multi-channel signal enhancement. For the conditions where multiple microphones are available, we combine these multiple signals into a single enhanced signal using delay&sum beamforming. We also experimented with an alternative speech/non-speech detector so that we can eliminate the dependency on the SRI SNS detector, which requires external training data.

The resulting system performed well on the evaluation data. However, there are still many areas for improvement, especially given the large variance in the error rate of individual meetings.

7 Acknowledgments

We would like to acknowledge Hans-Guenter Hirsch for his help with the SNR estimation system. This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-Party Interaction, FP6-506811).

References

1. J. Ajmera, H. Bourlard, and I. Lapidot, “Improved unknown-multiple speaker clustering using HMM,” IDIAP, Tech. Rep., 2002.
2. J. Ajmera, H. Bourlard, I. Lapidot, and I. McCowan, “Unknown-multiple speaker clustering using HMM,” in *ICSLP’02*, Denver, Colorado, USA, Sept. 2002.
3. J. Ajmera and C. Wooters, “A robust speaker clustering algorithm,” in *ASRU’03*, US Virgin Islands, USA, Dec. 2003.
4. C. Wooters, J. Fung, B. Peskin, and X. Anguera, “Towards robust speaker segmentation: The ICSI-SRI fall 2004 diarization system,” in *Rich Transcription Workshop*, New Jersey, USA, 2004.
5. 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.
6. J. Flanagan, J. Johnson, R. Kahn, and G. Elko, “Computer-steered microphone arrays for sound transduction in large rooms,” *Journal of the Acoustic Society of America*, vol. 78, pp. 1508–1518, November 1994.
7. M. S. Brandstein and H. F. Silverman, “A robust method for speech signal time-delay estimation in reverberant rooms,” in *ICASSP-97*, Munich, Germany, 1997.
8. H.-G. Hirsch, “HMM adaptation for applications in telecommunication,” *Speech Communication*, no. 34, pp. 127–139, 2001.
9. Q. Li and A. Tsai, “A matched filter approach to endpoint detection for robust speaker verification,” in *IEEE Workshop on Automatic Identification Advanced Technologies*, New Jersey, USA, october 1999.
10. NIST speech tools and APIs. [Online]. Available: http://www.nist.gov/speech/tools/index.htm