Speaker Diarization Using Direction of Arrival Estimate and Acoustic Feature Information: The I^{2}R-NTU Submission for the NIST RT 2007 Evaluation

Eugene Chin Wei Koh[1,2], Hanwu Sun[2], Tin Lay Nwe[2], Trung Hieu Nguyen[1], Bin Ma[2], Eng-Siong Chng[1], Haizhou Li[2], Susanto Rahardja[2]

[1] School of Computer Engineering, Nanyang Technological University (NTU), Singapore 639798 {kohc0026, nguy0059, aseschng}@ntu.edu.sg
[2] Human Language Technology Department, Institute for Infocomm Research (I^{2}R), Singapore 119613 {hwsun, tlnma, mabin, hli, rsusanto}@i2r.a-star.edu.sg

Abstract. This paper describes the I^{2}R/NTU system submitted for the NIST Rich Transcription 2007 (RT-07) Meeting Recognition evaluation Multiple Distant Microphone (MDM) task. In our system, speaker turn detection and clustering is done using Direction of Arrival (DOA) information. Purification of the resultant speaker clusters is then done by performing GMM modeling on acoustic features. As a final step, non-speech & silence removal is done. Our system achieved a competitive overall DER of 15.32% for the NIST Rich Transcription 2007 evaluation task.

1 Introduction

Speaker diarization has often been described as the task of identifying "Who Spoke When". When done in the context of the NIST Rich Transcription 2007 (RT-07) Meeting Recognition evaluations [1], this involves indicating the start and end time of every speaker segment present in the continuous audio recording of a meeting. Segments with common speakers have to be identified and annotated with a single speaker identity. This paper describes our system for the RT-07 speaker diarization task for multiple distant microphone (MDM) recordings.

Speaker diarization has traditionally relied on acoustic features such as Mel Frequency Cepstral Coefficient (MFCC) [2] or Perceptual Linear Prediction (PLP) [3] to perform segmentation and clustering. Segmentation is commonly done by employing the Bayesian Information Criteria (BIC) [3, 4]. Over-segmentation typically has to be carried out in order to capture most of the speaker turns. This however poses a problem for subsequent clustering as the resulting segments will usually be of short duration and hence do not offer reliable clustering. Our system mitigates this problem by directly using Direction of Arrival (DOA) [5] information to identify speaker transitions and perform clustering. Cluster purification using acoustic features is then performed. Our results from the RT-07 evaluation have shown that purification using acoustic features helps

to overcome mistakes introduced in DOA estimation. The final Diarization Error Rates (DER) (See Table 1) obtained after subsequent non-speech & silence removal were found to be competitive at 15.32%.

The usage of DOA information in the MDM task is not new. In [3, 6], DOA information was used by delay-and-sum beamformers to generate enhanced signals. It has also been used in [7] to augment acoustic features when performing speaker clustering. The key novelty of our system is in how DOA is used early in the diarization process to perform both segmentation and clustering. Acoustic features are then employed to do what DOA alone cannot handle: cluster purification and non-speech removal.

The rest of this paper is organized as follows. Section 2 describes the modules in our system and how speaker diarization was performed using DOA and acoustic feature information. Section 3 then presents our experimental results for the RT-07 evaluation and offers some discussions on the performance of our system. Section 4 then concludes the paper.

2 System Description

Fig. 1 illustrates our system, which consists of 4 modules. The NIST Rich Transcription Spring 2006 (RT-06s) corpus was used in the development of the system.

1. DOA estimation
2. Bootstrap clustering
3. Cluster purification
4. Non-speech & silence removal

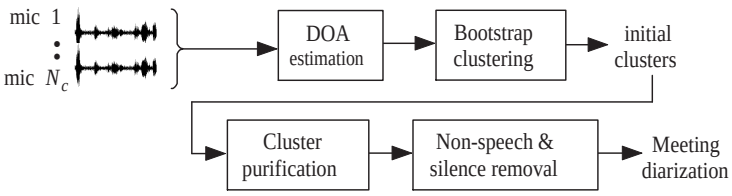


Fig. 1. *Block diagram of the I^2R/NTU RT-07 system.*

2.1 Direction of Arrival (DOA) estimation

A typical MDM task will have two or more distant microphone recordings. As the microphones are usually placed some distance apart, speech originating from a single source will arrive at the different microphones at different times. The time delay between the arrivals can be exploited to estimate the speech's DOA.

Fig. 2 shows how a pair of microphone inputs are used in our DOA estimation system. $r[n]$ denotes the reference microphone signal at time n and $s[n]$ denotes the source microphone. The reference channel is delayed by $L/2$ samples, and the adaptive filter's weights $\mathbf{w}[n]$ is of length L. During the presence of speech, the Normalized Least-Mean Square (NLMS) [8] algorithm is used to adapt the filter. DOA estimation is performed by detecting the peak position of the adaptive filter weights when voice activity is present.

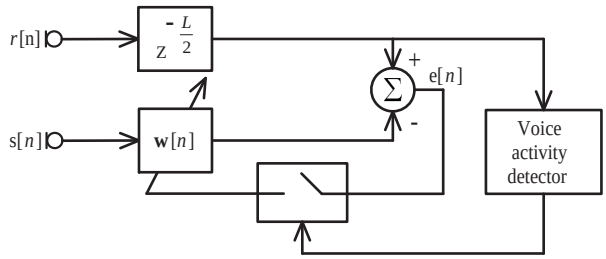


Fig. 2. *DOA estimation module for a microphone pair.*

Our DOA system operates on a frame-by-frame basis with the number of samples for each frame set to 512. For each frame, voice activity detection (VAD) is first performed by checking if the Teager energy [9] of the current frame is greater than an adaptive threshold. If speech activity is detected, the adaptive filter is allowed to adapt. Otherwise weights decay is applied to $\mathbf{w}[n]$. Our experience suggest that weights decay during silence frames improve the turn detection of short segments.

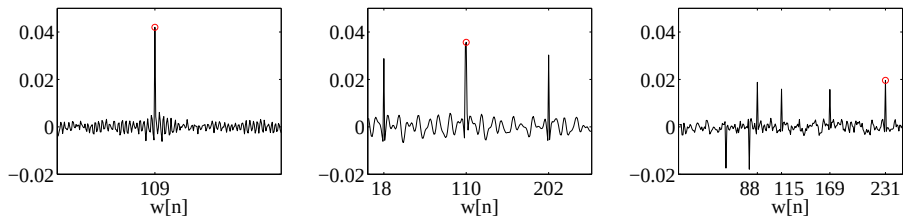
For the RT-06s task, $L = 250$ was found to be suitable. Given that the sampling rate is 16000 samples/sec, and assuming the speed of sound to be 330m/sec, the length $L = 250$ is able to accommodate a maximum microphone pair separation of 2.7 metres. If microphone pairs are separated by distances larger than 2.7 metres, L should be increased accordingly.

Fig. 3a shows a positive example of how peak detection on $\mathbf{w}[n]$ can be used to find the DOA. In this example, the source and reference signals have good SNR and the recordings are not significantly affected by reverberations. After adaptation, the filter's weights reflects an impulse with peak at index 109. This can be interpreted to mean that the reference signal is a delayed version of the source and that the delay is 16 samples (125-109=16).

Given that K pairs of microphones are used to estimate the direction of arrival, the matrix $DOA[n, k]$ stores the DOA values as

$$DOA[n, k] = \underset{j=1..L}{\arg\max}\{w_j[n, k]\} \tag{1}$$

where $n = 1..N$ and $k = 1..K$, N being the number of frames and $w_j[n, k]$ the jth filter coefficient for the kth microphone pair at time n.



(a) (b) (c)

Fig. 3. *Plot of values for filter coefficients* $\mathbf{w}[n]$ *at frame instance* n*, (a) showing a clear peak at index* 109*. (b) where there are 3 simultaneous speakers. (c) showing multiple peaks due to reverberation effects. Index 169 is the true speaker position.*

Microphone pair selection In cases where many microphone recordings are available, the number of microphone pair permutations may be large. Under such circumstances, we can judiciously choose pairs that exhibit characteristics typical of good DOA estimation. Specifically, we choose microphone pairs that have:

- Large highest-peak to next-highest-peak ratio on $\mathbf{w}[n]$.
- High SNR.
- Large DOA dynamic range.

An example of a SNR estimation algorithm can be found in [10]. To choose the best K pairs, we first rank all the microphone pairs by the above three factors and choose the top K pairs.

2.2 Issues influencing DOA estimation accuracy

Simultaneous speakers In the RT-07 tasks, there are situations when two or more speakers are talking simultaneously. We observed that multiple peaks will be observed for such instances. As can be seen in Fig. 3b, a peak will be registered for each speaker that talks. The filter weights at such an instance appear similar to that when the recordings have reverberations. This thus complicates our selection of the DOA estimate. As our current system is unable to discriminate simultaneous speech instances from reverberant recordings, we only exploit the maximum peak position to identify DOA. To improve system performance, we will develop a simultaneous speech detector to cope with this limitation.

Presence of audio multi-paths Fig. 3c shows the adapted filter weights for the case where the recordings have reverberations. The desired peak, although present is less pronounced and multiple secondary peaks are present. The main path is usually represented by the highest peak while the numerous shorter peaks correspond to the other multi-paths. However as is shown in Fig. 3c, the peaks due to multi-paths can sometimes overwhelm the correct primary location, leading to erroneous DOA estimation. We observed that instances like this occured quite frequently for the VT_20050408-1500 task and suggest that this might be the cause of the poor DOA estimation.

Presence of frequency specific noise The impact of frequency specific noise is also an issue that would warrant further examination. Such noise can be observed in the audio spectrogram as "lines" present at certain frequency bands. As was discussed in [11], such noise may be introduced by way of the microphone system collecting the recording. In [11], frequency specific device noise was found to be present in a NIST MarkIII [12] microphone array setup. That noise was found to have an effect on the cross-correlation of audio channels. This in turn will also affect DOA estimation.

Our own analysis of the RT-07 audio also suggest the presence of such frequency specific noise. Figure 4 shows a spectrogram of the 1st distant microphone of CMU_20061115-1030. In it, a "line" can be observed at the 1.2kHz mark. Similar artifacts can also be found in the recordings for the 2nd and 3rd distant microphones. We believe that this noise has a detrimental effect on our DOA estimations and further exploration will have to be done to confirm this.

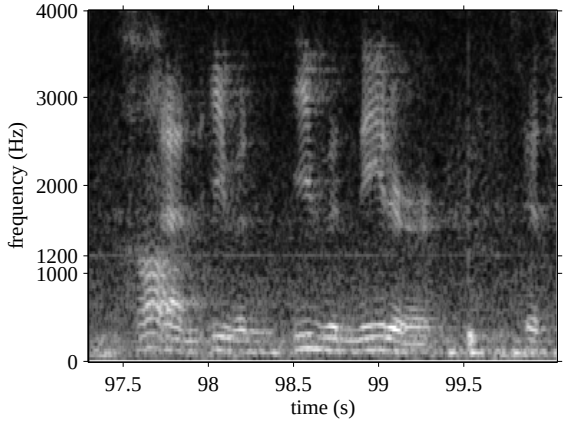


Fig. 4. *Spectrogram of 1st distant microphone for CMU_20061115-1030. Notice a "line" running horizontally at 1.2kHz.*

2.3 Bootstrap clustering

This section describes the 2nd stage of our diarizaton system. Bootstrap clustering uses the frame-wise DOA information from Eqn. 1 to form initial clusters. Since clustering at this stage forms segments using only spatial location information, impure segments containing different speakers will result if speakers move or change places. The presence of these impurities can be mitigated somewhat during subsequent iterative cluster purification where clustering is performed using speaker dependent acoustic features.

The bootstrap clustering process consists of two sub-steps. The first step quantizes DOA estimates for each microphone pair (the columns of $DOA[n, k]$) to commonly occuring locations. The second step merges DOA information across all K microphone pairs (the rows of $DOA[n, k]$), yielding a set of recurring DOA locations. The number of unique clusters formed after this step is taken by our system to be the final number of speakers. These clusters are used as seeds in the cluster purification step. Speech segments will then be formed by grouping together contiguous frames with a common cluster assignment. The details of the two sub-steps follows.

Within-pair quantization For the kth microphone pair, the $DOA[n, k]$ values are quantized to locate frequently occurring DOA positions for this pair. The frequently occurring positions are found by constructing a histogram using the kth column of $DOA[n, k]$. Every peak in the histogram indicates that there is a significant amount of speech originating from that particular location. We make the assumption that speech originating from a single location will very likely belong to a single homogenous speaker. As such, the number of peaks can be used as an estimate of the number of speakers present. The peaks in the histogram are taken as centroids and $DOA[n, k]$ values are quantized to these centroids using a nearest neighbor approach, as illustrated in Fig. 5a-c.

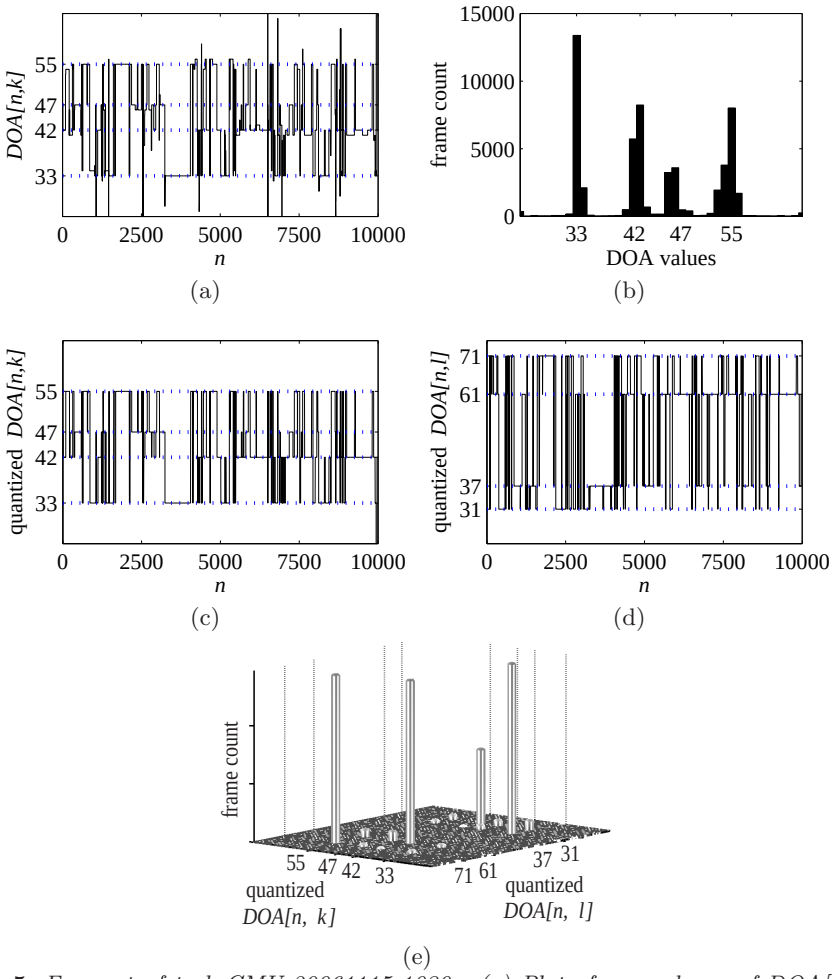


Fig. 5. *For part of task CMU_20061115-1030 : (a) Plot of one column of DOA[n, k]. Horizontal dotted lines correspond to histogram centroids. (b) Histogram of DOA values for selected k^{th} microphone pair. (c) DOA[n, k] values after within-pair quantization. (d) Post-quantization DOA[n, l] values of a different l^{th} microphone pair for the same segment of speech. (e) 2-D histogram of quantized DOA values for 2 microphone pairs (k^{th} and l^{th}). 4 peaks are visible.*

Inter-pair quantization The previous subsection discusses quantization along the columns of $DOA[n, k]$. The second step of this module is to perform quantization along the rows of $DOA[n, k]$, i.e., to identify centroids across K microphones pairs. Using the quantized results from within-pair quantization, a K-dimension histogram is built across all microphone pairs. Centroids can be readily identified within this high dimension histogram by virtue of their relatively high bin counts. An illustration of this is shown in Fig. 5e where 4 centroids can be observed by quantizing across 2 microphone pairs. The remaining histogram bins with low counts will then be clustered into the nearest centroid.

Our experiments conducted on the RT-07 MDM task showed that segments found after applying the bootstrap clustering were mostly of short durations - almost 90% of the segments are less than 3 seconds long and 71% of all the segments are shorter than 1 second.

It is interesting to note that apart from VT_20050408-1500, the initial clusters resulting after this module were observed to yield reasonably low Speaker Errors (See Table 1). The subsequent step of cluster purification only serves to improve the absolute DER by between 0% to 3.64%. This thus shows the effectiveness of our clustering method when the DOA estimations are accurate.

Clustering using only DOA information however will unlikely be robust towards situations where the speakers move significantly during a meeting. In such situations, speech from a single speaker might span across multiple clusters. The resultant clusters may be impure. Cluster purification will help in such situations to move incorrect speech segments to the cluster that they belong.

2.4 Cluster purification

The 3rd module of our system uses acoustic features for iterative re-clustering. This process of re-clustering has the effect of increasing the speaker homogenuity of each cluster. Beamforming is first performed on the available source audio to generate an enhanced recording as described in [6]. MFCC acoustic features are then generated from the enhanced recording. A voice activity detector (VAD) is then applied to retain only the high energy frames.

Initialization step All the MFCC vectors resulting from acoustic feature extraction are then used to train a root Gaussian Mixture Model (GMM), λ_{Root}. The λ_{Root} has 40 Gaussian components with full covariance matrices and was trained using the Expectation Maximization (EM) algorithm as described in [13].

Segments resulting from bootstrap clustering are then pooled together according to their cluster assignments. Individual GMMs are adapted from λ_{Root} for every cluster. Adaptation is performed on the weights, means and variances using the Maximum a Posteriori (MAP) approach [13]. Thus if there are Q speaker clusters resulting from bootstrap clustering, there will be Q GMMs. We denote these Q GMMs as $\lambda_{i,q}$, where i indicates the iteration number and $q = 1..Q$ indicates the qth speaker cluster. For the initialization step, $i = 0$.

Iterated steps The following is then performed:

1. Let $\mathbf{O}_j$ denote the set of feature vectors extracted from the jth segment. Let $s_{i,j} \in \{1..Q\}$ denote the cluster assignment of $\mathbf{O}_j$ for iteration i.
2. Every segment $\mathbf{O}_j$ is scored against models $\lambda_{i,q}$. Each segment is then relabelled by
$$s_{(i+1),j} = \arg\max_{q=1..Q}\{p(\mathbf{O}_j|\lambda_{i,q})\} \tag{2}$$
3. The GMMs $\lambda_{(i+1),q}$ are MAP adapted from $\lambda_{i,q}$ using the segments $\mathbf{O}_j$ corresponding to labels $s_{(i+1),j} = q$.

4. $i = i + 1$; Repeat from Step 2, until the cluster assignments have stabilized and do not vary for successive iterations.

The cluster assignment for segments were found to converge typically within 20 iterations.

2.5 Non-speech and silence removal

Silence and non-speech events such as coughs, laughter or breathing noises are present intermittently within speech segments. These segments are considered to be non-speech and should not be labeled as speaker segments. Hence, they have to be identified and removed.

Non-speech removal We performed non-speech identification on a segment-wise basis as it was found that a frame-wise decision was unreliable. The acoustic features used in this stage are the Log Frequency Power Coefficients (LFPC) [14]. A total of 10 coefficients are extracted from each 20 ms frame with a 10 ms overlap between frames. A model based approach was then used to evaluate every segment. Speech and non-speech were modeled by two separate GMMs, λ_S and λ_N. A classification decision can be made for the jth segment, $\mathbf{O}_j$ as follows.

$$p(\lambda_S|\mathbf{O}_j) \geq p(\lambda_N|\mathbf{O}_j) \rightarrow \text{speech} \tag{3}$$

$$p(\lambda_S|\mathbf{O}_j) < p(\lambda_N|\mathbf{O}_j) \rightarrow \text{non} - \text{speech} \tag{4}$$

When expressed as a likelihood ratio, Eqns. 3 and 4 become

$$\frac{p(\mathbf{O}_j|\lambda_S)}{p(\mathbf{O}_j|\lambda_N)} \geq t_1 \rightarrow \text{speech} \tag{5}$$

$$\frac{p(\mathbf{O}_j|\lambda_S)}{p(\mathbf{O}_j|\lambda_N)} < t_1 \rightarrow \text{non} - \text{speech} \tag{6}$$

where $t_1 = \frac{p(\lambda_N)}{p(\lambda_S)}$ is a threshold that was adjusted on the RT-06s corpus.

Silence removal Silence was removed using a "Double-Layer Windowing" method. In the first layer, audio is divided into frames of 20 ms with 10 ms overlapping. The energy for each frame is computed. In order to remove silences that are longer than the 300 ms tolerance specified for the evaluation [1], a second layer window is applied. This 300 ms long window shifts in 10 ms steps. The energy across all the frames in the window is summed.

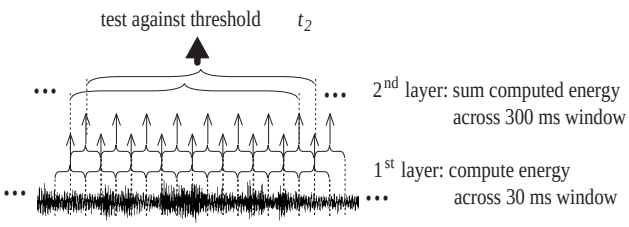


Fig. 6. *The "Double-Layer Windowing" method of removing silence.*

When this energy is found to cross a threshold, t_2, the region covered by the window will be deemed as silence and dropped. The threshold used to make this decision was determined on the RT-06s evaluation data.

3 Experiments

3.1 Results on RT-07

The system described in the previous section was entered as our primary entry for the Rich Transcription 2007 (RT-07) Meeting Recognition MDM evaluation. As listed in Table 1, RT-07 consists of 8 meetings from 4 different meeting rooms (i.e. CMU, EDI, NIST & VT). Each meeting evaluation had a duration of about 23 minutes, giving the the entire evaluation a total duration of 3 hours. The number of microphones channels available for each meeting ranged from 3 for the CMU tasks, to 16 for the EDI tasks. All recordings were done using distance microphones, except those for EDI which were made using 2 microphone arrays (8 channels each). The performance of our system was evaluated by computing the DER (Diarization Error Rate) against the official RTTM released by NIST. The DER is computed from the following error components.

$$DER = \frac{SE + MS + FA}{SPK}(\%) \tag{7}$$

- Speaker Error time (SE): Total time that is attributed to the wrong speaker.
- Missed Speaker time (MS): Total time in which less speakers are detected than what is correct.
- False Alarm Speaker time (FA): Total time in which more speakers are detected than what is correct.
- Scored Speaker time (SPK): Sum of every speaker's utterance time as indicated in the reference.

Table 1. *DER and Speaker Error (SE) time obtained on the RT-07*

RT-07 Task	Actual number of speakers	DER (%) after bootstrap clustering †	DER (%) after cluster purification	DER (%) after non-speech & silence removal	*SE* (s) after bootstrap clustering	*SE* (s) after cluster purification
CMU_20061115-1030	4	22.8 [4]	22.7	19.4	48	48
CMU_20061115-1530	4	17.8 [4]	17.6	12.5	8	7
EDI_20061113-1500	4	24.3 [5]	24.3	20.7	34	34
EDI_20061114-1500	4	30.6 [4]	30.2	15.0	32	29
NIST_20051104-1515	4	23.3 [4]	22.8	12.7	18	14
NIST_20060216-1347	6	22.1 [5]	18.5	13.4	66	39
VT_20050408-1500	5	46.4 [5]	19.8	11.3	234	28
VT_20050425-1000	4	27.4 [4]	27.1	18.5	23	21
Overall	—	27.0	22.8	15.3	463	218

† - Number of resultant clusters for each task is indicated in brackets [ ].

The DER for every task was found to improve after each subseqent processing step. Non-speech & silence removal yielded the largest overall improvement. It was capable of yielding an absolute improvement in the DER of 7.4%. The final DER of 15.3% was found to be competitive versus the other systems submitted for the RT-07 evaluation.

We observed that after bootstrap clustering, the VT_20050408-1500 task had the worst DER of 46.4%. A clue to its poor performance can be seen in the Speaker Error time (SE) component of the DER. It made up almost two-thirds of the diarization errors. As was previously mentioned in Section 2.2, the DOA estimation for this task was highly inaccurate. This resulted in many speech segments being incorrectly attributed to the wrong speaker. The DER for this task however improves considerably to 19.8% after performing cluster purification. The amount of SE fell 88% to 28 seconds. This thus suggests that our cluster purification method is capable of redeeming the speaker assignment errors introduced by inaccurate DOA estimation.

It is noteworthy that after cluster purification, a large DER improvement was observed only for VT_20050408-1500. The improvements for the other tasks were marginal. As can be seen in Table 1, this perhaps is because the Speaker Error time for these tasks were already rather low after bootstrap clustering. Cluster purification thus had little room to improve upon. This thus also suggests that if the meetings recordings are of sufficient quality, DOA estimation alone is capable of producing reasonably pure speaker clusters.

In the course of the review process for this paper*, it was suggested that further gains may be obtained by moving the cluster purification step to be after Non-Speech & Silence removal. This is a possibility that we are currently studying. The merit of doing so would be that cluster purification will be performed on segments that are free from silence and non-speech. The resultant speaker assignments may be more accurate, and this could be a way of further reducing our SE time.

3.2 Performance of Non-Speech & Silence removal

Table 2. *Before and after non-speech & silence removal*

	SAD DER Error (%) ‡	SAD Missed Speaker time (s)	SAD False Alarm Speaker time (s)
Before non-speech & silence removal	14.5	8	1005
After non-speech & silence removal	8.7	326	280

‡ - Speech Activity Detection (SAD) DER error measures the amount of non-speech or silence that is incorrectly attributed to a speaker (SAD Missed Speaker time or $SADMS$ time), and speech that is incorrectly regarded as non-speech or silence (SAD False Alarm Speaker time or $SADFA$ time).

As was mentioned in Section 3.1, the non-speech & silence removal module produced the greatest improvement to the DER scores for RT-07. This module produced an absolute overall DER reduction of 7.4%. Further analysis however suggests that there remains potential for the improvement of this module. The

* We would like to thank the anonymous reviewer for this suggestion.

Speech Activity Detection (SAD) error after this module stands at 8.7%. This value is relatively high when compared to the same for the systems submitted by the other RT-07 entrants.

The further removal of non-speech & silence however cannot be done by simply increasing the thresholds t_1 and t_2. As is reflected in Table 2, the removal of non-speech & silence was done at the expense of increasing the amount of speech misclassified to be non-speech or silence, i.e. the $SADFA$ time. A more aggressive non-speech or silence removal could worsen the system by removing more than it should. Subsequent experiments will thus have to be done to determine the optimum trade-off point where the total error time is lowest.

3.3 Number of detected speakers

It was observed that the number of unique speakers present in each meeting was determined quite accurately by our system. This number was incorrectly anticipated only in the case of EDI_20061113-1500 and NIST_20060216-1347. For EDI_20061113-1500, our system found 5 speakers when there should be 4. The speech of the longest speaker was found to be divided into two clusters, one containing 441 seconds of speech and other 86 seconds.

The numbers for NIST_20060216-1347 and our previous experience on the RT-06s corpus however suggests a shortcoming of our system. In our RT-06s experiments, it was observed that our system tends to under-estimate the number of speakers for those tasks with a large number of speakers. Speakers who spoke for the shortest durations tends to be found clustered together with a longer speaker. Our results for NIST_20060216-1347 reinforced this observation - only 5 speakers were detected when there should be 6. The speaker who spoke the least (59 seconds of speech) was found merged with a speaker who spoke for 128 seconds.

The reason for these missing short speakers would be due to the histogram quantization used in the bootstrap clustering process. When there is a large number of speakers, the resolution of the histogram becomes inadequate. Adjacent speaker distributions tend to overlap. Those speakers who spoke the least had small bin counts and therefore did not register as a clustering centroid. They were thus absorbed into their neighbours.

Our system currently determines the number of unique speakers present using only the DOA estimation values. This could potentially also result in another shortcoming. The number of unique speakers is approximated using those physical locations where most of the speech originates from. In the event where there is a speaker replacement in the middle of a meeting (e.g. in a lecture), this method will mistake both the speaker and its replacement as a common speaker. A possible solution to this problem would be to use acoustic information to test the homogenuity of each cluster prior to purification. Clusters that are found to be highly impure can be split into smaller parts of higher purity. Purification can then proceed on these smaller clusters.

4 Conclusions

The system developed for the I^2R/NTU submission to the RT-07 evaluation was found to yield a competitive overall DER of 15.3%. For 7 of the 8 meetings, the

initial clusters obtained after bootstrap clustering had a relatively low Speaker Error time. This indicates that using DOA alone can serve as an effective way of segmenting and clustering speakers.

It was observed however that the usage of DOA information alone is inadequate for the VT_20050408-1500 task. Clustering mistakes introduced by the unreliable DOA estimation could be redeem by performing cluster purification using acoustic features. We thus conclude that the usage of acoustic features to perform purification serves as an effective complement to segmentation and clustering using DOA information.

References

[1] "Spring 2007 (RT-07) Rich Transcription Meeting Recognition Evaluation Plan," [Online], Available: http://www.nist.gov/speech/tests/rt/rt2007/docs/rt07-meeting-eval-plan-v2.pdf.

[2] X. Anguera, C. Wooters, B. Peskin, and M. Aguilo, "Robust Speaker Segmentation for Meetings: The ICSI-SRI Spring 2005 Diarization System," In Proc. NIST MLMI Meeting Recognition Workshop, Edinburgh, 2005.

[3] D. A. v. Leeuwen and M. Huijbregts, "The AMI speaker diarization system for NIST RT06s meeting data," In Proc. NIST Rich Transcription 2006 Spring Meeting Recognition Evaluation Workshop, Washington DC, pp. 371-384, 2006.

[4] 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 Proc. NIST MLMI Meeting Recognition Workshop, Edinburgh, 2005.

[5] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," In Proc. International Conference on Acoustics, Speech, and Signal Processing, Munich, pp. 375-378, 1997.

[6] X. Anguera, C. Wooters, and J. Hernando, "Speaker Diarization for Multi-Party Meetings Using Acoustic Fusion," In Proc. IEEE Automatic Speech Recognition and Understanding Workshop, San Juan, 2005.

[7] X. Anguera, C. Wooters, and J. Pardo, "Robust Speaker Diarization for Meetings: ICSI RT06s evaluation system" In Proc. Interspeech 2006 ICSLP, Pittsburgh, 2006.

[8] S. Haykin, Adaptive Filter Theory (4th Edition). Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2002.

[9] J. F. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal," In Proc. International Conference on Acoustics, Speech, and Signal Processing, Albuquerque, pp. 381-384, 1990.

[10] H. G. Hirsch, "Estimation of noise spectrum and its application to SNR-estimation and speech enhancement," Technical report tr-93-012, ICSI, Berkeley, 1993.

[11] L. Brayda, C. Bertotti, L. Cristoforetti, M. Omologo and P. Svaizer, "Modifications on NIST MarkIII array to improve coherence properties among input signals," Journal of Audio Engineering Society, 2005.

[12] C. Rochet (Sept 2005), "Technical Documentation of the Microphone Array Mark III," [Online], Available: http://www.nist.gov/smartspace/cmaiii.html

[13] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification using Adapted Gaussian Mixture Models," Digital Signal Processing, vol. 10, pp. 1941, 2000.

[14] T. L. Nwe, S. W. Foo, and L. C. D. Silva, "Stress classification using subband based features," IEICE Trans. Information and Systems, vol. E86-D, pp. 565-573, 2003.