New Metrics for Meaningful Evaluation of Informally Structured Speech Retrieval

Maria Eskevich[1], Walid Magdy[2], and Gareth J.F. Jones[1,2]

[1] Centre for Digital Video Processing,
School of Computing, Dublin City University, Dublin 9, Ireland
[2] Centre for Next Generation Localisation,
School of Computing, Dublin City University, Dublin 9, Ireland
{meskevich,wmagdy,gjones}@computing.dcu.ie

Abstract. Search effectiveness for tasks where the retrieval units are clearly defined documents is generally evaluated using standard measures such as mean average precision (MAP). However, many practical speech search tasks focus on content within large spoken files lacking defined structure. These data must be segmented into smaller units for search which may only partially overlap with relevant material. We introduce two new metrics for the evaluation of search effectiveness for informally structured speech data: *mean average segment precision (MASP)* which measures retrieval performance in terms of both content segmentation and ranking with respect to relevance; and *mean average segment distance-weighted precision (MASDWP)* which takes into account the distance between the start of the relevant segment and the retrieved segment. We demonstrate the effectiveness of these new metrics on a retrieval test collection based on the AMI meeting corpus.

Keywords: Speech retrieval, informally structured speech, evaluation metrics

1 Introduction

Increasing amounts of informal spoken content are being collected in digital archives, e.g. recordings of meetings, lectures, internet podcasts and personal data sources. Accessing information from within this content poses challenges beyond retrieval of formally defined documents since this type of informally structured data is not typically divided into topical retrieval units. Existing work on speech retrieval has predominantly focused on retrieval of clearly defined document units in ad hoc search tasks. For these tasks, transcripts generated using automatic speech recognition (ASR) are typically indexed using standard information retrieval (IR) methods with no account being taken of the temporal nature of the spoken content when evaluating retrieval effectiveness. Datasets for speech retrieval tracks at evaluation campaigns such as the TREC Spoken Document Retrieval track (SDR) [4] and CLEF2007 Cross Language Speech Retrieval (CL-SR) track [10] were manually segmented into coherent topical segments considered as documents for IR. Retrieval results were evaluated primarily using mean average precision (MAP). Other tasks within these evaluations focused on a no segmentation condition, but measured only the effectiveness with which an audition drop-in point could be identified in relevant content.

For search of unsegmented speech content, the data must be divided into smaller units for retrieval. These segments can be of fixed size or use an automated content-based segmentation method. The resulting segments typically only partially overlap with relevant material. The temporal nature of speech means that for efficient search users should be accurately directed to relevant content in the audio file so as to avoid spending time auditioning non-relevant material. Standard existing metrics such as mean average precision do not take into account the proportion of a retrieved unit which is relevant to the user.

In this paper we introduce *mean average segment precision (MASP)* a new metric which takes account of both the rank of relevant segments and quality of segmentation with respect to proportion of the retrieved segments which are relevant. We further introduce *mean average segment distance-weighted precision (MASDWP)* which additionally incorporates a component relating to the ideal drop-in point to begin playback. The behaviour and suitability of these metrics for the evaluation of speech retrieval for informally structured data in comparison to existing metrics is illustrated using an IR test collection based on the AMI meeting corpus [2].

The remainder of this paper is structured as follows: Section 2 reviews scores currently used in relevant existing work and outlines their shortcomings, Section 3 introduces the MASP and MASDWP metrics, Section 4 summarises the experiments which we use to demonstrate the use of MASP and MASDWP, Section 5 reports experimental results and analysis, and finally Section 6 gives conclusions and outlines directions for future work.

2 Related work in speech search evaluation

In this section we review the use of MAP for speech search evaluation with defined retrieval units, related work for evaluation of text passage retrieval in the INEX evaluation campaign and the metric used for evaluation of jump in point detection accuracy in the CLEF CL-SR track.

2.1 Mean Average Precision

MAP is one of the most widely used metrics in IR research [1]. Equation 1 shows the definition of the standard average precision (AP) metric for a single query.

$$AP = \frac{1}{n}.\sum_{r=1}^{N} P[r] \cdot rel(r) \tag{1}$$

where n is the number of relevant documents, N is the number of retrieved documents, $P[r]$ is the precision at rank r (the number of relevant retrieved documents divided by the total number of retrieved documents), $rel(r)$ is the relevance of the document ($rel(r) = 1$ if document is relevant, $rel(r) = 0$ if not). MAP is computed by averaging AP across the topic set.

2.2 Evaluation of Passage Retrieval at INEX

Since MAP has a binary score relevance, it generally assumes that if the retrieval units are taken from within larger documents, they have been perfectly segmented into coherent topical units. Even for search with multi-topical retrieval units it is assumed that any relevant content is contained completely in the retrieval unit. Therefore MAP is not a good measure when relevant content may have been split between multiple segments.

In order to measure the amount of relevant content contained within a passage the *Mean Average interpolated Precision (MAiP)* metric was introduced for the text passage retrieval task at INEX [7]. Document relevance was not counted in a binary way, but rather it was assumed that the amount of relevant information retrieved should be reflected in the metric. This metric is based on the mean generalised average precision (mGAP) [8] that was introduced to deal with human assessment of partial relevance.

In MAiP, precision at rank r is defined as the fraction of retrieved text that is relevant,

$$P[r] = \left(\sum_{i=1}^{r} rsize(s_i) \right) \quad / \quad \sum_{i=1}^{r} size(s_i) \tag{2}$$

where r is the rank of the document, s_i is the document at rank r, $rsize(s_i)$ is the length of relevant text contained in s_i in characters (if there is no relevant text in s_i, $rsize(s_i)$=0), $size(s_i)$ is the total number of characters in s_i.

Recall at rank r is defined as the fraction of relevant text that is retrieved,

$$R[r] = \left(\sum_{i=1}^{r} rsize(s_i) \right) \quad / \quad Trel(q) \tag{3}$$

where $Trel(q)$ is the total number of relevant characters across all segments, i.e. the sum of the lengths of the (non-overlapping) relevant regions.

The INEX organizers were afraid that $P[r]$ can be biased towards systems returning several shorter segments rather than returning one longer segment that contains them all. This prompted INEX to define MAiP in terms of precision at fixed recall levels rather than ranks. Thus, the measure interpolated precision $iP[x]$ is defined as the maximum precision at the selected recall level x. Retrieval effectiveness is calculated using average interpolated precision A_iP calculated by averaging the interpolated precision scores calculated at 101 recall levels (0.00, 0.01, ..., 1.00),

$$AiP = \frac{1}{101} \cdot \sum_{x=0.00,0.01,\ldots,1.00} iP[x] \tag{4}$$

MAiP is calculated by computing the mean of the AiP values across the topic set. Although MAiP looks to be a suitable metric for evaluating speech search using segments, the way of averaging is inconvenient for speech tasks as discussed later in Section 3.

2.3 CLEF CL-SR Evaluation

Another metric for the evaluation of the results of speech retrieval is an alternative application of mGAP introduced in the CLEF CL-SR task. The implementation of this score for speech data is described in [10]. It measures the errors in finding the start points in time of relevant content by the retrieval system [9]. The calculation of GAP for a single query is shown in Equation 5.

$$GAP = \frac{1}{n}.\sum_{r=1}^{N} P[r] \cdot \left(1 - \frac{Distance}{Granularity} \cdot 0.1\right) \tag{5}$$

where $P[r]$ is the precision at rank r, $Distance$ is the distance between the start of the segment and the beginning of relevant part (the limit was set to 150 seconds by task organizers), $Granularity$ is the step that is used for the penalty function ($Granularity$ = 15 seconds at CLEF). Thus segments that make the user wait for longer than 150 seconds are not considered relevant. This metric seemed reasonable for reflecting the user case scenario, however, it does not reflect the time the user will spend on listening to the relevant content.

3 Time Precision Oriented Metrics

In this section we describe two new metrics for evaluating retrieval effectiveness for searching informally structured spoken content taking into account time information in terms of both precision of the retrieved segments and the distance of the beginning of the retrieved segment to the real start of the relevant content.

3.1 Mean Average Segment Precision (MASP)

MASP is a modification of MAP, inspired by MAiP, but specifically adapted to speech search when no pre-defined segmentation of search units exists. The motivation for MASP is to create a metric that measures both the ranking quality and the segmentation quality with respect to relevance in a single score. Thus, the ideal state for MASP is not only to retrieve the relevant speech segments at the top of the ranked results list, but also to have each segment 100% segmented over relevant speech data without including any non-relevant parts. Unlike MAP, relevance for MASP varies from 0 to 1 according to the amount relevant content present in the segment. This is similar to the measurement of relevance in MAiP, but there are two fundamental differences: the amount of relevant content is measured over time instead of text; and the average segment precision (ASP) is calculated at the ranks of segments containing relevant content rather than fixed recall points as in MAiP.

Segment precision ($SP[r]$) at rank r in MASP is calculated as follows,

$$SP[r] = \sum_{i=1}^{r} rperiod(s_i) \quad / \quad \sum_{i=1}^{r} length(s_i) \tag{6}$$

where *length*(s_i) is the length of segment s_i in time units (minutes or seconds), and *rperiod*(s_i) is the length of the relevant period in the segment s_i. Unlike MAiP, the average segment precision (ASP) is calculated at the ranks where relevant content is found as follows:

$$ASP = \frac{1}{n}.\sum_{r=1}^{N} SP[r] \cdot rel(s_r) \tag{7}$$

where n is the number of segments that contain relevant content, and $rel(s_r)$ is equal to 1 if s_r contains any relevant content, and 0 otherwise. MASP is defined as the mean of ASP across a query set Q.

$$MASP = \frac{1}{|Q|}.\sum_{q \epsilon Q} ASP_q \tag{8}$$

The motivation behind taking the average of $SP[r]$ over the ranks of relevant content is the same as that for MAP. The assumption is that the position where a user stops checking the ranked list is usually a relevant item, which varies for different users. The stopping position at a relevant rank is assumed to be uniformly distributed, which is why the AP is calculated in this way.

The claim for applying the averaging of MAiP at fixed recall points, as described by the INEX organizers [7], is that the score can be biased towards retrieving shorter segments. However, we hypothesize that this issue is automatically resolved within the implementation of MASP. In MASP retrieving shorter segments of relevant content will increase the number of segments with relevant content (n), therefore the averaging process will be applied on a larger number of ranks and ASP will thus not be biased to the length. MASP is low when the percentage of relevant parts in segments is consistently low, which indicates bad segmentation, or when the ranks of the relevant contents are deep in the results list. Table 1 shows a simple illustrative example of how ASP is measured and compared to AP. The example topic has 4 relevant segments appearing in the top 6 ranks. As shown in Table 1, ASP takes into consideration the length of each segment as well as the percentage of relevant content in each one. It can be seen that long and short segments are not treated the same; the score gets lower values when long periods of irrelevant speech are returned on the top of the list. This factor is not measured when using standard AP.

3.2 Mean Average Segment Distance-weighted Precision (MASDWP)

The ASP metric reflects the amount of relevant content present at different ranks. However it does not show how far the user has to listen into the segment at a certain rank until the relevant part actually begins or whether the segment starts after the beginning of the relevant part and the user will have to rewind in the recorded audio signal beyond the beginning of the segment in order to get to the starting point of the relevant content. In order to take this information

Table 1. Example comparing AP, ASP, and ASDWP. The average values are calculated at ranks in bold, the segment at the first rank starts with the relevant information, the relevant content at the third rank position starts only later within the segment, the relevant content starts long before the segments found at ranks 4 and 6.

Rank	**1**	2	**3**	**4**	5	**6**	Avg
rperiod/length	2/3	0/5	3/4	6/6	0/2	5/10	Value
Prec.[r]	1	1/2	2/3	3/4	3/5	4/6	0.771
SP[r]	2/3	2/8	5/12	11/18	11/20	16/30	0.557
SDWP[r]	2/3 * 1.0	2/8	5/12 * 0.9	11/18 * 0.0	11/20	16/30 * 0.0	0.260

into account we introduce the same style of penalty function that was used at CLEF CL-SR Evaluation (Equation 5):

$$ASDWP = \frac{1}{n}.\sum_{r=1}^{N} SP[r] \cdot rel(s_r) \cdot \left(1 - \frac{Distance}{Granularity} \cdot 0.1\right) \qquad (9)$$

In the illustrative example in Table 1, if we suppose that the first segment starts at the relevant point, that the third has a playback drop-in point inside the segment at one step from the start of the segment and that the fourth and sixth segments are actually far beyond the limit set for the distance to be of practical relevance to the user, only the first and third results are included in the calculation of ASDWP, and thus the metric reflects whether while listening to the results in the ranked list the user will start the playback close to the beginning of the actual relevant data.

4 Experimental Setup

4.1 Test Collection Based on AMI Corpus

Since creating a speech corpus is a very large task, to demonstrate the behaviour of MASP and MASDWP we make use of the existing AMI Corpus[3], collected as part of the AMI project and made publicly available for research purposes. This corpus contains 100 hours of annotated recordings of planned meetings [2]. Meetings last about 30 minutes each, 70% of them simulate a project meeting on product design and usually involve 4 participants. For the majority of the meetings, both manual and automatic transcripts are provided, for the latter the developer of the corpus created an ASR system which makes use of a standard ASR framework employing hidden Markov model (HMM) based acoustic modeling and n-gram based language models (LMs) [12]. The dataset also includes additional materials including the slides projected during the meetings. For this study we use the AMI release 1.4. The provided meeting transcripts were pre-processed to generate a single transcript for each meeting, omitting incompletely transcribed meetings. This gave us a total of 160 meetings.

[3] http://www.amiproject.org/

For our investigation of meeting search we assume the scenario of a meeting participant wanting to find locations in meetings where the topic of a PowerPoint slide projected in one or more of the meetings was being discussed, regardless of whether the slide was being projected at the time that the topic was being discussed. We took a subset of 25 of PowerPoint slides provided with the AMI corpus as a topic set. Corresponding relevance assessments for these topics were manually generated as part of our study using a pooling procedure.

4.2 Segmentation of the AMI Corpus

In order to exploit the AMI Corpus in a search investigation it must be preprocessed to segment it into suitable topical search units. We segmented the manual transcript using simple time- or length-based methods, and content-based algorithms. For the content-based segmentation we used Choi's popular C99 algorithm [3] and Hearst's TextTiling algorithm [5], both performing linear segmentation on the level of sentences. Average length of the segments for C99 and TextTiling is 129 and 136 seconds respectively, segments varied from minimum lengths of 8.11 and 14.53 to maximum lengths of 1352.57 and 818.07 seconds respectively.

In order to compare runs with segments of different, though constant, length, we used time-based segmentation of 60/180 seconds (being the average length of 3 short sentences and the longer segments) (time_60, time_180). Although an unlikely scenario, the non-segmented dataset, i.e. a whole unsegmented meeting as one retrieval unit, was also used in our experiments (one_doc) as an extreme segmentation (average length of these segments is 1998 seconds, with a range between 306.4 and 5297.84 seconds).

Another set of runs was produced by segmenting the transcript into chunks of length that is comparable to the average lengths of the content-based segmentation results (both in time and number of words). Since the average length of content-based segments is 346 and 363 words for C99 and TextTiling, we prepared the following segmentation runs: every 300/400 words (len_300, len_400). Time values of 120 and 150 seconds (time_120, time_150) are close to the average both in time and in the average word-length (313 and 389 respectively). The time boundary points were applied with flexibility to prevent words at the boundaries being split between segments.

Tracking the influence of the ASR performance on retrieval behaviour of the segments is not possible if the two segmented collections have different segment boundary points. Thus we projected the segment borders calculated for the manual transcript onto the ASR transcript by using the word timing information of the transcripts. This resulted in a second collection of segments (asr_man) where the only difference between this and the manual transcript segment collection is that the content of the segments is taken from the ASR transcripts[4]. Sometimes

[4] In practice of course in the absence of manual transcripts segmentation would be carried out on ASR transcripts. We do not include experimental results for ASR transcript derived segmentation for reasons of space.

Table 2. Scores for 1000 retrieved documents for different runs. The scores that position the runs both on man and asr_man transcripts at the same overall rank for the same metric are shown in bold.

Run	man				asr_man			
	MAP	MAiP	MASP	MASDWP	MAP	MAiP	MASP	MASDWP
c99	**0.474**	0.277	**0.235**	**0.187**	**0.438**	0.275	**0.218**	**0.177**
tt	0.467	0.290	**0.241**	**0.179**	0.421	0.275	**0.221**	**0.173**
len_300	0.233	0.337	0.265	0.199	0.416	0.287	0.248	0.181
len_400	**0.491**	0.320	**0.253**	**0.153**	**0.463**	0.286	**0.237**	**0.147**
time_120	0.451	0.287	**0.270**	**0.211**	0.428	0.296	**0.256**	**0.196**
time_150	**0.485**	**0.293**	**0.263**	0.167	**0.448**	**0.283**	**0.243**	0.171
time_180	**0.510**	0.296	0.266	0.176	**0.473**	0.300	0.246	0.163
time_60	0.333	0.259	**0.259**	**0.235**	0.333	0.259	**0.238**	**0.220**
one_doc	**0.712**	**0.103**	**0.088**	**0.008**	**0.686**	**0.109**	**0.085**	**0.009**

manual transcripts do not cover the whole region of the ASR transcripts, since they do not include areas regarded as not relevant to the meetings by the manual transcribers, in these cases the additional words in the ASR transcript were placed in the adjoining manual segment.

4.3 Retrieval Setup

The segments obtained using each segmentation technique from the manual transcripts were indexed for search using a version of the SMART information retrieval system[5] extended to use language modelling (multinomial model with Jelinek-Mercer smoothing) with a uniform document prior probability [6]. Separate retrieval runs were carried out for each topic for each segmentation scheme for the segments containing the manual and ASR transcripts. The retrieval used i = 0:3 for all query terms. Stopwords were removed using the standard SMART stopword list, and words were stemmed using the Porter stemmer [11].

5 Experimental Results

5.1 Baseline Results and Correlations

Table 2 shows results obtained for MAP, MAiP, MASP and MASDWP at rank 1000. We can see that MAP does not properly reflect the fact that the unsegmented transcripts (one_doc) are far from being the best one, being actually the worst in general, because the documents are too long to browse through. All the other metrics consistently give it the worst score for both types of transcripts.

The shortest segments (time_60) are ranked at consistent positions for both MASP and MASDWP, being first in case of the latter (the shorter the segments are, the more highly the ones that are closer to the drop-in point get ranked) and

[5] ftp://ftp.cs.cornell.edu/pub/smart/

at middle rank for the former (these segments are not long enough to contain a lot of relevant information). In general increases in MASDWP correspond to reduced average length of the segment, and the ranking of the runs stays very consistent across different transcripts, suggesting that the user finds the beginning of the relevant content consistently bad or good in both manual and asr_man retrieval outputs. This suggestion also applies to the MASP scores.

MAiP ranking of different runs in man and asr_man transcripts is consistent in putting only the one_doc at the lowest level and keeping time_150 in the middle of the list, suggesting that the segmentation has less effect on the retrieval performance than ASR errors because its calculation does not take segmentation issues into account.

Kendall's Tau Correlation for MAP, MAiP, MASP and MASDWP shows that MAP is not correlated with MAiP, MASP or MASDWP (-0.0043, -0.25 and -0.61 respectively), which reflects our belief that MAP is not really suitable for this task. At the same time MASP and MAiP show higher correlation (0.58) corresponding to the fact that they both consider the amount of relevant content in the segments. MASDWP has low correlation with MAiP (0.17) and much higher correlation with MASP (0.48) reflecting the fact that its calculation, though inspired by MAiP and based on MASP, reflects different parameters of the quality of the system output.

5.2 Detailed Analysis of MASP and MASDWP Behaviour

MASP compensates for certain shortcomings of MAiP because it takes into account not only segment ranking, but also the precision at a particular rank and the time that the user would spend listening to non-relevant content in the list. In this section we provide some examples which illustrate the ability of MASP to capture differences in the ranking between runs for one type of transcript and differences in the ranking for one type of segmentation between different types of transcripts. We use the same examples to show how MASDWP captures more important information from the perspective of the experience of potential users.

Difference Between Segmentations for One Transcript Type Table 3 illustrates cases for runs which have different average segment lengths (from minimum (time_60) to a maximum of the whole document (one_doc)).

For query 13, AiP and ASP differ only in positioning one_doc and time_60 runs. This reflects the trend to favour longer segments by AiP and the coverage of relevant information of the segment at the top rank by ASP (rank (rperiod / length)): (1(553/1323), 2(765/1475), 3(607/1514), 4(277/822), 5(332/1041), ...) - one_doc and 1(56/59), 2(37/65), 3(60/60), 4(38/60), 5(59/59), ...) - time_60.

For this query the no segmentation run (one_doc) has the best AP score, but it is 0.0 for ASDWP, meaning that actually for all of the documents the user would have to listen for a long time in order to get to the relevant part, i.e. for the cases where relevant data does not start from the beginning of the document AP fails to capture this issue.

Table 3. Comparing scores for the same transcript (asr_manual) and different segmentations.

Query	Segmentation	AP	AiP	ASP	ASDWP
13	c99	0.595 (3)	0.542 (1)	0.505 (1)	0.281 (3)
13	time_180	0.669 (2)	0.482 (2)	0.480 (2)	0.302 (2)
13	one_doc	0.967 (1)	0.385 (3)	0.383 (4)	0.0 (4)
13	time_60	0.477 (4)	0.368 (4)	0.420 (3)	0.349 (1)
21	c99	0.292 (3)	0.382 (1)	0.245 (2)	0.206 (1)
21	time_180	0.375 (2)	0.292 (2)	0.247 (1)	0.145 (2)
21	one_doc	0.511 (1)	0.114 (4)	0.079 (4)	0.008 (4)
21	time_60	0.188 (4)	0.159 (3)	0.160 (3)	0.140 (3)

Although the average numbers in Table 2 might suggest that ASDWP always tends on average to give higher scores to the shorter segments, the example of Query 21 in Table 3 shows that this metric is more subtle, and even though the segmentation at short periods of time (as time_60) has higher chances of accidentally hitting the beginning of the relevant content, segmentation methods that produce longer segments that contain longer parts of the relevant content are rewarded by the score. Indeed, the rankings of these runs are: (4(243/243), 6(105/125), 7(157/204), 8(107/107), 9(350/429), 10(122/122), ...) - c99; (3(179/179), 4(179/179), 5(180/180), 7(179/179), 8(59/179), 9(162/180), 10(143/181), ...) - time_180; and (3(60/60), 4(59/59), 5(60/60), 6(59/59), 7 (59/59), 8(60/60), 9(38/59), ...) - time_60. Whereas the rankings that are close enough to the beginning of the relevant content (i.e. the ones that are calculated for ASDWP) are (rank (distance (a +ve number if the relevant part starts before the beginning of the segment and -ve if it starts within the segment))): (6(20.19), 7(46.98), 8(-44.68), 9(47.2), 10(-16.68), ...) - c99; (5(-68.91), 7(0.0), 9(-4.31), ...) - time_180; and (4(0.42), 6(-9.58), 9(21.05), ...) - time_60.

Although if the user is interested only in finding more relevant information sooner, then time_180 shows better ranking of segments with a greater percentage of relevance and ASP captures this whereas AiP does not. In this case ASP and AiP are both consistent in marking the run without segmentation (one_doc) as the worst one, meaning that the no segmentation condition is not an appropriate approach for this type of query.

Difference Between Different Transcripts for One Segmentation Type When using the same segmentation method, ASR and manual transcripts have different numbers of words in the segments due to recognition errors in ASR transcripts. Assuming ASR and manual runs achieved the same ranking for all relevant segments, AiP will be different since it is based on character matching. However, ASP will be the same for both runs, since it is based on time, which is more important for speech search since the user will be listening to the recording.

Table 4 shows some queries for which ASP is contradictory to AiP. These cases illustrate ASP's advantages, because it reflects not only rank changes, but

Table 4. Comparing scores for the same segmentation and different transcripts

Query	Segmentation	ASP		AiP	
		man >	asr_man	man <	asr_man
1	tt	0.245	0.243	0.272	0.376
1	len_400	0.325	0.322	0.411	0.435
21	tt	0.196	0.182	0.228	0.382
21	time_150	0.242	0.227	0.271	0.272
1	c99	0.287	0.272	0.293	0.377
1	time_120	0.366	0.361	0.367	0.486

also how much relevant content is within the segments that are moved up or down the ranked list. For the cases of query 1 (tt, len_400) and query 21 (tt, time_150), the number of segments ranked higher in the asr_man list in comparison with the man list is lower than the number of segments ranked lower in the asr_man list in comparison with the man list (33 vs 49, 40 vs 57, 38 vs 45, 41 vs 51 respectively), while the amount of relevant content in seconds that is ranked higher in the asr_man list is lower than the amount of relevant content ranked lower in the list for all these cases (2164 vs 2446, 1986 vs 2978, 3009 vs 4705, 3573 vs 4147). Therefore ASP shows that man lists are better than the asr_man, whereas AiP does not capture this difference.

For query 1 (c99, time_120) the amount of relevant information that moves up the list is higher than the amount of relevant content that moves down the list (2667 vs 2107, 2757 vs 1969), the changes in ranks are less consistent for these two cases (35 vs 42 and 45 vs 43). ASP still shows that man runs are better than the asr_man, because even though there is more relevant content moving up the list, the content that moves down is falling from higher ranks, and even if it is substituted with other segments with relevant information, the ratio between relevant and non-relevant content in these new segments is lower, i.e. it is worse for the user who will have to listen to longer non-relevant segments. To see this effective, compare the ratio of the relevant information for query 1 c99 run (rank (total amount of relevant content in seconds until this rank / total length of the documents in seconds until this rank)): man run - (1(110/125), 2(201/266), 3(308/404), 4(376/530), 5(454/607), 6(492/763), 7(544/945), 8(660/1187), 9(789/1403), 10(816/1659), ...), asr_man - (1(77/77), 2(188/203), 3(279/343), 4(347/469), 5(385/624), 6 (437/807), 7(611/1217), 8 (740/1433), 9(856/1676), 10(927/2180),...).

6 Conclusions and Future work

In this paper we introduced the MASP and MASDWP evaluation metrics for search of informally structured speech. These were inspired by currently existing evaluation scores, but addressed their shortcomings. We have shown that the new metrics are more suitable for the evaluation of retrieval in informally structured speech content. MASP captures the amount of relevant content that appears at different ranks and MASDWP awards runs where segmentation algorithms put boundaries closer to the actual beginning of the relevant parts and these segments are higher in the ranked list.

In future work, this score will be applied for other types of unstructured speech than meetings. The fact that adjacent segments may be present at the adjacent positions in the ranked list may be introduced into the calculations.

7 Acknowledgements

This work was funded under the Science Foundation Ireland Research Frontiers Programme 2008 by Grant No: 08/RFP/CMS1677, and partially by Grant No: 07/CE/I1142 as part of the Centre for Next Generatiion Localisation (CNGL).

References

1. Büttcher, S., Clarke, C. L. A., Cormack, G. V.: Information Retrieval: Implementing and Evaluating Search Engines, MIT Press (2010)
2. Carletta, J.: Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus. Language Resources and Evaluation Journal. 41(2), pp. 181–190 (2007)
3. Choi, F.Y.Y.: Advances in domain independent linear text segmentation. Proc. of the 1st NAACL conference. pp. 26–33. Seattle, Washington, USA (2000)
4. Garofolo, J.S., Auzanne, C. G. P., Voorhees, E. M.: The TREC Spoken Document Retrieval Track: A Success Story. Proc. of RIAO 2000, pp. 1–20. Paris, France (2000)
5. Hearst, M.A.: TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages. Computational Linguistics. 23(1), pp. 33-64. (1997)
6. Hiemstra, D.: Using Language Models for Information Retrieval. Ph.D. thesis, Center of Telematics and Information Technology, AE Enschede, The Netherlands (2000)
7. Kamps, J., Pehcevski, J., Kazai, G., Lalmas, M., Robertson, S.: INEX 2007 Evaluation Measures. In: Fuhr, N., Kamps, J., Lalmas, M., Trotman, A. (eds.) Focused Access to XML Documents. pp.24–33. Springer (2006)
8. Kekalainen, J., Jarvelin, K.: Using graded relevance assessments in ir evaluation. Journal of the American Society for Information Science and Technology. 53(13), pp. 1120–1129. Wiley Subscription Services (2002)
9. Liu, B., Oard, D.W.: One-sided measures for evaluating ranked retrieval effectiveness with spontaneous conversational speech. Proc. of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Seattle, Washington, USA (2006)
10. Pecina, P., Hoffmannova, P., Jones, G. J. F., Zhang, Y., Oard, D.W.: Overview of the CLEF 2007 Cross-Language Speech Retrieval Track. Proc. of the CLEF 2007 Workshop, pp. 674–686. Budapest, Hungary (2007)
11. Porter, M.F.: An algorithm for suffix stripping. Program 14(3), pp. 130-137 (1980)
12. Renals, S., Hain, T., Boulard, H.: Recognition and interpretation of meetings: The AMI and AMIDA projects. Proc. of the IEEE Workshop ASRU (2007)