# Creating a Data Collection for Evaluating Rich Speech Retrieval

**Maria Eskevich[1], Gareth J. F. Jones[1], Martha Larson[2], Roeland Ordelman[3]**

[1] Centre for Digital Video Processing, School of Computing, Dublin City University, Dublin 9, Ireland
[2] Delft University of Technology, Delft, The Netherlands
[3] University of Twente, The Netherlands
{meskevich, gjones}@computing.dcu.ie, m.a.larson@tudelft.nl, roeland.ordelman@utwente.nl

**Abstract**

We describe the development of a test collection for the investigation of speech retrieval beyond identification of relevant content. This collection focuses on satisfying user information needs for queries associated with specific types of speech acts. The collection is based on an archive of the Internet video from Internet video sharing platform (blip.tv), and was provided by the MediaEval benchmarking initiative. A crowdsourcing approach was used to identify segments in the video data which contain speech acts, to create a description of the video containing the act and to generate search queries designed to refind this speech act. We describe and reflect on our experiences with crowdsourcing this test collection using the Amazon Mechanical Turk platform. We highlight the challenges of constructing this dataset, including the selection of the data source, design of the crowdsouring task and the specification of queries and relevant items.

**Keywords:** Speech Search, Speech Collection Creation, Speech Retrieval, Crowdsourcing

## 1. Introduction

The increasing capacity of digital storage media and advances in networking technologies for content delivery are resulting in an ever increasing expansion in the volume of audio and video data accumulated on the web and in offline archives. The nature of this data can vary significantly, from broadcast news and lectures to informal videos recorded for Internet TV channels by semi- and non-professionals. Whatever the form of the data, its potential can only be realised if users can locate relevant content in a timely and efficient manner. The diversity in online archives content potentially gives rise to multiple possible information needs and resulting formulations of search queries. Conventional research on speech retrieval has focused on locating content containing information relevant to a specific information need expressed in a text search query (Garofolo et al., 2000) (Pecina et al., 2008).

In this paper, we describe the design and construction of a speech search benchmarking collection that extends this goal to one when users are considered to be interested in not only informational content, but also in the speaker's intention while uttering their speech. The focus on speaker intention is motivated by the observation that the same words pronounced in different ways can have have different illocutionary meaning - one can promise or warn the listener. We describe our methodology for developing our test collection for investigating search involving speaker intention and our approach to preparing our ground truth. The test collection includes audio-visual spoken content, search queries and corresponding identified relevant data,

Developing meaningful queries and determining relevant segments in videos expressing speaker intention in an unbiased way is particularly challenging. One possible solution for this lies in crowdsourcing, the dissemination of a task to large numbers of human annotators, referred to as workers through an Internet platform (Surowiecki, 2004). This technique permits researchers to gather data from a diverse community of workers in return for micro-payments for their contributions. The potential of crowdsourcing has already been explored in a number of other speech and language applications, e.g. (Snow et al., 2008), (Callison-Burch and Dredze, 2010). Creating a test collection of the type we planned to develop is though muc more demanding of the crowdsource workers than the tasks undertaken in these previous studies.

The test collection described in this paper was developed for use in the Rich Speech Retrieval (RSR) task which formed part of the MediaEval 2011 Benckmark[1]. This RSR task is based on the observation that utterances are actually 'illocutionary speech acts'[2] carried out by speakers (Larson et al., 2011a). The queries are thus designed to search for instances of these speech acts which have been identified in our selected corpus of Internet video.

The paper is structured as follows: Section 2. describes the development of the RSR task we used as an example of dataset construction, Section 3. overviews relevant previous work in crowdsouring for collection of speech and language resources, Section 4. describes the preparation of the data and search collection creation, Section 5. describes task reward issues, Section 6. gives details of the results of the collection exercise, and Section 7. concludes and outlines directions for our future work.

## 2. Rich Speech Retrieval Task Data Preparation using Crowdsourcing

The focus of the MediaEval 2011 RSR task was to explore the effectiveness with which different types of 'illocutionary speech acts' can be located within audio-visual spoken data. For this we used the ME10WWW archive of semi-professional user generated video downloaded from the Internet video sharing platform blip.tv. The videos for this collection were collected for shows for which the link to

[1] http://www.multimediaeval.org/
[2] http://en.wikipedia.org/wiki/Speech acts

one of their episodes had been tweeted on the Twitter social network. Their licenses were checked to conform that they were Creative Commons. The dataset contains 1974 episodes (247 development and 1727 test) comprising a total of ca. 350 hours of data. The development set is small with respect to the test set and is not intended for training, but rather for parameter tuning. The episodes were chosen from 460 different shows, shows with less than four episodes were not considered for inclusion in the dataset. The ME10WWW dataset for the RSR 2011 task is accompanied by automatic speech recognition (ASR) transcripts (Lamel and Gauvain, 2008), which were generously provided by LIMSI (http://www.limsi.fr/) and Vocapia Research (http://www.vocapia.com/). In order to be included in the ME10WWW set, a video needed to have been transcribed by the ASR-system with an average word-level confidence score of $> 0.7$. The set is predominantly English with approximate 6 hours of non-English content divided over French, Spanish and Dutch. Further details of this video collection can be found in (Larson et al., 2011b).

The richness of the RSR task arises from the specific types of the queries that we are interested in. The information to be found has to be a combination of required audio and visual content and the speaker's intention. Five examples of basic speech acts types were chosen for this task: 'apology', 'opinion' from 'expressives' (speech acts that express on the speaker's attitudes and emotions towards the proposition, e.g. congratulations, excuses and thanks), 'definition' from 'assertives' (speech acts that commit a speaker to the truth of the expressed proposition) 'warning' from 'directives' (speech acts that are to cause the hearer to take a particular action, e.g. requests, commands and advice), 'promise' from 'commissives' (speech acts that commit a speaker to some future action, e.g. promises and oaths).

## 3. Crowdsouring in Development of Speech and Language Resources

### 3.1. Background and Relevant Existing Work

Crowdsourcing is a form of micro-outsourcing that allows tasks to be assigned to remote workers who receive a small financial compensation for their work. The importance of manually developed resources to support speech and language research and the cost of developing them means that crowdsourcing has become a topic of great interest in the development of resources for language technology research.

Looking at the general suitability of the use of untrained crowdsource workers in natural language tasks, Snow et al. (Snow et al., 2008) compared the work of domain experts and with that of non-experts recruited in a general crowdsourcing environment for a range of natural language labelling tasks, including recognising textual entailment and word sense disambiguation. The tasks were restricted to selection of multiple choice response or numeric input within a fixe range. Their results demonstrated that the non-expert crowdsource workers can produce work of a similar standard to expert workers. Callison-Burch and Dredze (Callison-Burch and Dredze, 2010) survey contributions to the NAACL-2010 workshop on crowdsourcing for speech and language resource, and highlight a number of important factors which should be taken into account when designing effective crowdsource tasks in this setting. These include issues of how to attract sufficient suitable workers to undertake the task, the level of payment that should be offered to a worker for undertaking a task, careful design of the task, that the instructions should be clear for the expected participants, and how to deal with the problem of workers who try to cheat on the task to earn payment without undertaking the work properly. It should be noted that here too the tasks examined were relatively straightforward such as involving the selection of appropriate labels from among a number offered or undertaking translation into a language in which the worker is fluent.

In the area of speech resource development Marge et al. (Marge et al., 2010) found that crowdsource workers are able to transcribe speech of varied qualities with reasonable accuracy. Evanini et al. (Evanini et al., 2010) investigated the more challenging task of transcribing non-native read-aloud and spontaneous speech, they found that even merging the results multiple workers produced errorful transcriptions particularly in the case of the spontaneous speech. Thus even in a clearly defined and apparently obvious task, crowdsourcing does not provide a simple solution to challenging tasks. Lane et al. (Lane et al., 2010) found some success exploring the related speech task of collecting spoken corpora using crowdsourcing, but identified issues in relation to the training of the speakers to undertake the task.

In the field of information retrieval, one of the common challenges in the development of test collections for system testing is establishing the relevance of available documents to a user search query. Crowdsourcing provides an intuitively appealing solution to this problem. In this scenario workers can be shown a query and asked whether specific documents are relevant to the query. An early study exploring this topic is described in (Alonso et al., 2008). This examined the important topics of establishing whether workers are actually qualified to carry out the task for which they are volunteering, and seeking to identify those not undertaking the work properly. This is a particular problem in relevance assessment of this nature since clearly the person requesting the work cannot manually check the accuracy of all submitted work. A further study on this topic by Grady and Lease (Grady and Lease, 2010) examined the topic of reward for work done. They examined the issue of worker pay, particularly considering the impact of offering a bonus to the worker for good work, where bonuses were manually assigned when checking the quality of work carried out. Interesting they observed that workers appeared to be attracted to do more work where a bonus was offered, and that they completed the work with greater accuracy on average.

From these existing studies it is clear that crowdsourcing can make a valuable and cost effective contribution to the development of language technology resources. However, workers can find even apparently simple tasks challenging and produce unsatisfactory work. All the tasks examined here are conceptually quite straightforward either relying on workers to use non task specific recognition skills, their own special linguistic knowledge, e.g. being bilingual, or

Table 1: Number of collected queries per speech act for MediaEval 2011 development and test sets

| | Speech act type | | | | | Total |
|---|---|---|---|---|---|---|
| | Apology | Definition | Opinion | Promise | Warning | |
| Development Set | 1 | 8 | 17 | 1 | 3 | 30 |
| Test Set | 1 | 17 | 21 | 5 | 6 | 50 |

transcribing or uttering some speech. There is no personal creativity required to perform any of these tasks. General factors include, the common observation that here is a persistent problem of some workers trying to cheat to receive payment without completing work properly. Also there are interesting questions requiring further exploration relating to the levels of pay offered for tasks, and the potential impact of bonus payments on loyalty and quality of work.

### 3.2. Amazon Mechanical Turk

Currently, the most widely-used platform for crowdsourcing is Amazon Mechanical Turk[3]. In the Mechanical Turk (MTurk) setting tasks are referred to as 'Human Intelligence Tasks' or HITs. To initiate a task, the requester uploads a HIT consisting of relevant instructions, questions, files, etc to be used by the workers while completing the HIT. When the workers have carried out HIT, the requester reviews the completed work and confirms payment to the worker with a previously set payment. If the requester is not satisfied with the work carried out, they can opt not to the worker. Potentially, the requester can also give the worker a bonus, as discussed previously, in order to both express appreciation for the quality of the work and to motivate the worker to continue.

## 4. Development of an Effective HIT

While we had a clear specification of the test collection that we wished to develop using MTurk, as highlighted in the previous section, earlier work using crowdsouring for the development of speech and language resources set workers much less complex tasks than we wished them to undertake. In many cases deliberately designing the task to be as simple as possible to reduce the effort involved for the worker, to maximise the potential number of workers interested and qualified to undertake the task and to minmise the chance of them making mistakes. Thus the creation of our test collection was actually exploring the research question of whether untrained MTurk workers can undertaken extended tasks which require them to be more creative than those examined previously.
For each HIT we required the worker to carry out the following activities:

- View an assigned video to attempt to locate the presence of speech act.
- Label the specific time at which the speech act begins and ends in the video.
- Accurately transcribe the words spoken within the time limits of the labeled speech act.
- Write a full sentence query which they believed would be able to refind this speech act, and write a short web style query to refind the speech act.

The task is thus much longer and more complex than tasks typically offered to workers, since it requires them to carry out multiple activities and also to be creative since they are asked to develop their own search queries as part of the HIT.
While assigning the HIT to a worker we did not require them to have any specific knowledge or experience of work with audio and video data. However we used internal Amazon MTurk platform information about the previous performance of this registered user in order to select those that are familiar with the system itself, and whose previous results satisfied other requesters. This measure is called HIT Approval Rate, and is a simple ratio of how many HITs submitted by each worker have been approved by the requesters. For our HIT we allowed only the workers with HIT Approval Rate greater than or equal to 90 to undertake the task.

### 4.1. Data Management

The videos in the blip TV dataset vary in length. We felt it unrealistic to expect workers to view extended videos while looking for speech acts. Also we observed a bias in our initial crowdsourcing trials with this data that workers tend to identify noteworthy segments in the first few minutes of a video. This may be caused by the fact that they are paid for each HIT, and are therefore interested in completing more HITs in time they have available. Thus for the original 247 and 1727 videos in the ME10WWW for development and test set respectively, we prepared 562 and 3278 starting points for longer videos at a distance of approximately 7 minutes apart. These starting points were then randomly allocated by the Amazon MTurk platform to be presented to the workers in each HIT. Even after presegmenting the videos into shorter parts, the workers rarely found noteworthy content later than the third minute from the start of playback point in the video.
The main technical challenge was that since MTurk does not support playback of multimedia files, the workers needed to watch videos stored on an external server. The path to the remote file was embedded within the html-code of the HIT page. Thus the video player used had to be compatible with different operating systems and browsers. Restrictions on this issue had to be made clear to workers in the general description of the task.

### 4.2. Data Collection Procedure

We used a three-stage approach for our collection procedure. First we prepared and uploaded a pilot version of the HIT and received 55 results, 34 of which we approved.

[3]https://www.mturk.com/mturk/welcome


Figure 1: Amazon MTurk HIT example that was used to gather long and short queries (Questions 6 and 7) associated with certain speech acts (Question 1), time stamps (Questions 3 and 4) and transcript (Question 5) of the relevant content

These answers were not included in the final set, but provided us with valuable feedback to refine our HIT.

The initial HIT was found to contain too many concepts that workers did not understand clearly. The revised HIT thus avoided words such as 'transcripts', 'quote', 'categories' etc. The request to find the segment with a certain speech act was expressed indirectly, and was hard to understand. Initially the description of the task was: "Please watch the video and find a short portion of the video (a segment) that contains an interesting quote. The quote must fall into one of these six categories". Analysis of the results showed that workers were confused with the type of phrases they had to find, the concept of transcription was mixed with the general description of what was said during the video, the workers who were probably not familiar with the video player gave wrong time onsets and offsets for identified relevant speech events.

In our revised HIT, we attempted to make use of a concept with which general workers will be more familiar when working with the videos – sharing. The concept of sharing seemed to us to be part of the everyday experience of people who work with the Internet and would provoke a more natural human response setting. The new phrase of the HIT became: "Imagine that you are watching videos on YouTube. When you come across something interesting you might want to share it on Facebook, Twitter or your favorite social network. Now please watch this video and search for an interesting video segment that you would like to share with others because it is (an apology, a definition, an opinion, a promise, a warning)".

The other way to make the workers more familiar with the task was to provide full examples of how all the questions in one HIT can be answered for each speech act type. At the head of each HIT page we put a link to a webpage with an example video with all fields filled in and a dropout window on the page of the HIT itself with only the textual answers.

All these changes resulted in more appropriate answers to all the questions from a large majority of the workers. An

example of the HIT page is shown in Figure 1.
Additionally the initial trial HIT enabled us to set a suitable worker reward that both we as requesters and workers would be comfortable with. The setting of rewards is discussed in further detail in Section 5.

### 4.3. HIT Refinement

We ran the revised version of the HIT on the development set. An unexpected finding was that the difference in the types of speech acts, together with the limited time that workers are usually prepared to spend on one HIT caused a problem of unbalanced results. We found that it is much easier to assign something that was said by the speaker in the video to be his or her opinion than to find a warning or a promise. We think that this particular feature and the nature of the videos themselves, where we observed that there are not so many incidents of speakers making an apology or promising something in the videos, made our results unbalanced, and that this meant the number of 'opinions' was significantly higher than the number of the four other types. To avoid this situation for the video test set, we decided to run the HIT twice: one HIT with the option of 'rare' speech acts and one only for opinions. At the same time we added new questions about the speaker's appearance and behaviour to help us to detect workers who were not doing their work properly: 'Write one sentence describing the person that you see in this video. If there is more than one person, who is the person who seems to be the most important for the video'; 'Does this person have any particular mannerisms (gestures that they use, particular way of talking, nervous habits)? Please write one sentence to describe anything that you notice'. Our attempt to simply separate the HITs for 'rare' speech acts failed because workers seemed to be more amused by the new questions that had no meaning for our research, the information provided was not useful for our research, and apparently it was harder to find instances of the rare acts in the data.
Thus finally we decided to return to the original single HIT for all types of speech act with the revised wording of the instructions in collecting the queries, assuming that the lack of balance between acts might be just a feature of this dataset, and we used the same HIT questionnaire for the test set as for the development set.
In total we collected 30 queries for the development set and 50 for the test set, Table 1 shows the statistics of the collected speech acts. Examples of long queries that look like a natural sentences, short queries that correspond to the type of queries usually addressed to the Internet search engines, and transcripts of the relevant content are given in Table 2.

## 5. Reward Levels

The reward to a worker paid by a requester is generally set in the task description, and the workers take this value into consideration together with the general HIT description when choosing whether to undertake the task. Once some workers submit their work, statistics of the average reward per hour for this HIT are available for viewing by other potential workers to take into account when considering when to take on a HIT.
The availability of the option for the requester to change the reward amount when assessing the data submitted for a HIT by a worker gives the possibility to introduce the notion of a bonus (extra reward) or to decrease the reward if the requester finds that the work was not done correctly.
Our initial trial HIT enabled us to define a suitable reward that both we as requesters and workers would be comfortable with.
Initially we started with a reward of 0.11$ per HIT plus bonus per type of the illocutionary act (the sum varied depending on the rareness of the act). Due to the complicated and confusing formulation of the HIT, we received negative feedback from the workers. Apparently this task was inappropriately time consuming for the reward we set. Thus we worked on reformulation of the HIT to simplify it as described in Section 4.3. Also we added a clear statement in the HIT description that we are a non-profit organization, and we raised the reward to 0.19$ and made the workers themselves suggest their own bonus in the range from 0 to 0.21 $. Our motivation for allowing workers to choose their own bonus level was to demonstrate trust in them and appreciation of their work, which we conjectured would reinforce workers in carrying out work more thoughtfully and carefully. Interestingly, giving workers an opportunity to judge the difficulty of the task themselves resulted in useful answers with little evidence of greed (i.e., people didn't always choose the highest possible bonus). Workers were given a text box in which to provide justification for their requested bonus. Most of them took this opportunity to add a short comment. Sometimes the workers even explained that they were not sure of how well they had done on the task and therefore did not deserve the bonus for completing this HIT. Apart from spam submissions, we found bonus requests always to be reasonable.
In total, the cost of the completing the HITs was the following (10 % of all the rewards paid goes to the Amazon MTurk platform):

- price of the devset: (55*0.11 + 7*0.19 + 46*0.19 ) = 20$ + 16.12 = 36.12 + 10% = 40 $;
- price of the testset: (47.88 + 25.2 (approximately the amount of bonus money)) + 10% = 80.388 $.

## 6. Comments on the HIT Results

Since working with video as required by this HIT is not a common task for crowdsourcing workers, we wanted to support workers that took the effort to undertake our HIT. Thus we accepted reasonably good answers that could not be used in our test collection and even award a small bonus (0.02$) with an explanatory comment to the worker. Another reason for keeping the reward (even a small one), and not rejecting the work carried out by a worker, is that the MTurk platform monitors the level of rejection per worker in order to detect spam or any other inappropriate activity (HIT Approval Rate). Thus we did not want to decrease the HIT Approval Rate of workers who undertook our task and did a substantial amount of work, but could not make it correctly due to misunderstandings due to the nature of the task.

Table 2: Examples of 2 types of queries associated with speech acts and transcripts for the relevant segments

| Speech act | Queries of 2 types and Transcript |
|---|---|
| Apology | Transcript: I'm here now with Terry Denison, who's the President of the Swim Coaches Association in Great Britain. Thanks for joining us on The Morning Swim Show. Oh, well, thank you for inviting me. Actually, I'm Chairman of the Swim Coaches Association.. it's a slightly.. Chairman Denison, I apologize. |
| | Long query: How does Anita Burns, host of the Open Mind Show, save face after the embarrassing comment she made during her interview with Victoria Edwards?<br>Short query: Peter Busch president chairman Denison morning Swim Show apology |
| Definition | Transcript: Equality. How you wanna be equality for his people as far as material possessions and in verse fifteen he compared that to what was said in exodus. as it is written, He that gathered much had nothing over; and he that gathered little had no lack. |
| | Long query: Short video segment defining equality using a segment from a religious book<br>Short query: Equality religious definition |
| Opinion | Transcript: Apple produces this new platform all of a sudden you know within roughly a short period that is twenty five thousand applications. Apple didn't write these applications other people did. You know you look at Twitter A Twitter is as minimal as service it doesn't offer very much and yet thousands of applications are out there adding value |
| | Long query: What makes Twitter more popular than Apple in terms of value<br>Short query: Why Twitter and not Apple? |
| Promise | Transcript: They will launch a new effort to conquer a disease that has touched the life of nearly every American, including me... by seeking a cure for cancer in our time. |
| | Long query: Obama promises to find a cure for cancer!<br>Short query: Obama healthcare promises |
| Warning | Transcript: And there are some here coming for their own purposes and for selfish reasons that are not for your highest good. Not everything out there out there is wonderful and good |
| | Long query: Woman warning that we should be aware of the intentions of the things going on in cosmos<br>Short query: good in cosmos spoiled by selfish reason |

We had several workers who completed several HITs for the development set, but did not participate in the HIT for the collection of the test set. This lack of overlap might cause certain differences in the way people formulated their queries and chose relevant segments.

In general, for the test set the number of accepted HITs with the speech act chosen was 58.1%, where 39.5% were suitable for use in the dataset and 18.6% were accepted, but not included (either some of the fields were missing or there were some issues with the work of the video). For the remaining 41.9% of HITs the worker indicated that they were unable to find an instance of any of the illocutionary acts, we found that 35% of these responses were reasonable, while the other 6.9% were disputable.

It is worth commenting that the data provided by the workers required an additional manual assessment by the requester because there were a number of spam entries. Some types of spam were easy to capture, even automatically, for example when all fields for the HIT were empty or contained the same word. However some spam workers were more creative and copied field by field the example we have provided in the HIT heading. In our HIT formulation there was the possibility to state that there is no speech segment that can be associated with any of listed speech acts and still get the basic reward. During the collection for the test set, only 16% of these answers seemed to be disputable, and thus could be classified as spam or improper work. In these cases the workers were not paid. This problem was also observed for confirmation of translations in (Callison-Burch and Dredze, 2010), an effective solution was to use images of the text which could not be copied, rather than using text itself. While manual checking by the requestor of the claim that there was no speech act present in the video shown for the HIT was practical for the small scale data collection undertaken here, it would quickly become prohibitive for larger collections. In these cases passing these video playback points to a second round of crowdsourcing might offer a means to check the judgement of the first worker.

Related to this issue, while we mainly only used one assessor to decide on the assign a speech act to each segment and create an appropriate query, this could be done multiple times for each segment. When the dataset was examined by participants in the MediaEval 2011 RSR , in more than 50 % of the cases there was a general consent on the information provided by the worker, however some cases were clearly disputable. Without any information about the workers background, language proficiency and the features in the audio or video that affected the assignment of the speech act to a certain utterance, it is not clear how they made their assignment. To better understand such cases, the same video segment could be given to multiple crowdsourcing workers during the assignment and query generation stage. Segments that all workers agree upon could be chosen for use in the retrieval collection. This would have the additional advantage that there would be multiple queries available for each segment, enabling more extensive RSR experimentation. Additionally, segments, assigned speech acts and associated queries, could be used as

a separate crowdsourcing task in order to get other workers opinions on the reliability of the initial workers judgments in the first round of experiment.

## 7. Conclusions and Future work

This paper has described our successful development of the test collection for the Rich Speech Retrieval task at MediaEval 2011. This work has demonstrated that is possible to use crowdsourcing workers to carry out more extensive and complex tasks in the creation of resources to support speech and language research than has previously been shown. Our experiences in developing the worker task demonstrate the importance of understanding the concepts and vocabulary with which workers are likely to be familiar and to ensure that the required task relates to their general life experiences. Related to the description of the actual task, crowdsourcing workers are currently generally not used to dealing with video and audio and thus tend to be confused by the technical terminology.

The requirement to fully understand the instructions and to successfully complete multiple stages in the HIT, and the somewhat subjective nature of some of the speech acts in the video data means that it may not be possible to reduce the high failure rate of the HIT. In this case while roughly 90 % of the workers were judged to seeking to fulfill the HIT to the best of their ability, and paid accordingly by the requester, with only 10 % not receiving payment, less than 50 % of the paid work was judged suitable for inclusion in the test collection. While the low cost of crowdsouring means that the amount of money wasted is not high, it would be preferable to make the HIT more efficient. Seeking to do this could form the basis of further investigation.

One disadvantage of using this approach is that the crowdsourcing platform is not specifically tuned for video processing, thus we had to use an external video player. Therefore some technical problems that the workers had (the video was not displayed or it was too slow) are hard to control, and it is impossible to detect whether they are caused by the interaction of the platform with external software, or the workers Internet connection affects the video display.

We found that the choice of award level for demanding tasks of the type specified here was very important. Setting the award too low in our initial trial HIT was very unpopular, but this problem was easily addressed when the reward amount was raised, and workers were found to generally be honest in their self assessment of the quality of their work for the HIT and the reward that they deserved.

We presented videos that are longer than 7 minutes to the workers several times, each time starting the playback at a distance of approximately the same length in order to get the queries from all of the data and not only the beginning of the files. However even with this setting, as noted earlier, workers tend to watch only a maximum of the first 3 minutes from the start of the playback which biases our results. Using a smaller window between the playback start points might be a solution to this problem. Although this change is not completely straightforward due to the presence of music and other non-speech sounds that has to be taken into account while assigning the position of playback start points within each file.

In future work we plan to collect more retrieval queries with speech act information for this dataset through crowdsourcing. We assume that the retrieval process might benefit when queries of different speech act types are processed differently. However, the number of queries of different types in the current test collection is not sufficient to draw conclusions in this regard from experiments. We will investigate whether the creation of a set of HITs to collect the query set, and then checking their reliability through crowdsourcing could form a basis for the creation of a large retrieval collection for future investigation in the domain of rich speech retrieval.

## 8. Acknowledgments

This work is funded by a grant under the Science Foundation Ireland Research Frontiers Programme 2008 Grant No: 08/RFP/CMS1677, and funding from the European Commission's 7th Framework Programme (FP7) under grant agreements no. 216444 (EU PetaMedia Network of Excellence) and AXES ICT-269980.

## 9. References

Omar Alonso, Daniel E. Rose, and Benjamin Stewart. 2008. Crowdsourcing for relevance evaluation. *SIGIR Forum*, 42(2):9–15.

Chris Callison-Burch and Mark Dredze. 2010. Creating speech and language data with Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (CSLDAMT 2010)*, pages 1–12, Stroudsburg, PA, USA. Association for Computational Linguistics.

Keelan Evanini, Derrick Higgins, and Klaus Zechner. 2010. Using Amazon Mechanical Turk for transcription of non-native speech. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (CSLDAMT 2010)*, pages 53–56. Association for Computational Linguistics.

John S. Garofolo, Cedric G. P. Auzanne, and Ellen M. Voorhees. 2000. The TREC spoken document retrieval track: A success story. In *Proceedings of RIAO 2000*, pages 1–20.

Catherine Grady and Matthew Lease. 2010. Crowdsourcing document relevance assessment with mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (CSLDAMT 2010)*, pages 172–179. Association for Computational Linguistics.

Lori Lamel and Jean-Luc Gauvain. 2008. Speech processing for audio indexing. In *Advances in Natural LanguageProcessing*, pages 4–15. Springer Berlin / Heidelberg.

Ian Lane, Alex Waibel, Matthias Eck, and Kay Rottmann. 2010. Tools for collecting speech corpora via mechanical-turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (CSLDAMT 2010)*, pages 184–187. Association for Computational Linguistics.

Martha Larson, Maria Eskevich, Roeland Ordelman, Christoph Kofler, Sebastian Schmiedeke, and Gareth J. F. Jones. 2011a. Overview of Mediaeval 2011 Rich Speech Retrieval task and genre tagging task. In Martha Larson, Adam Rae, Claire-Hélène Demarty, Christoph Kofler, Florian Metze, Raphaël Troncy, Vasileios Mezaris, and Gareth J. F. Jones, editors, *Proceedings of the MediaEval 2011 Workshop*, volume 807. CEUR-WS.org.

Martha Larson, Mohammad Soleymani, Pavel Serdyukov, Stevan Rudinac, Christian Wartena, Vanessa Murdock, Gerald Friedland, Roeland Ordelman, and Gareth J. F. Jones. 2011b. Automatic tagging and geotagging in video collections and communities. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval (ICMR 2011)*, pages 51:1–51:8. ACM.

Matthew Marge, Satanjeev Banerjee, and Alexander I. Rudnicky. 2010. Using the Amazon Mechanical Turk for transcription of spoken language. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2010)*, pages 5270–5273. IEEE.

Pavel Pecina, Petra Hoffmannová, Gareth Jones, Ying Zhang, and Douglas Oard. 2008. Overview of the CLEF 2007 Cross-Language Speech Retrieval Track. In Carol Peters, Valentin Jijkoun, Thomas Mandl, Henning Müller, Douglas Oard, Anselmo Peñas, Vivien Petras, and Diana Santos, editors, *Advances in Multilingual and Multimodal Information Retrieval*, Lecture Notes in Computer Science, pages 674–686. Springer Berlin / Heidelberg.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*, pages 254–263, Stroudsburg, PA, USA. Association for Computational Linguistics.

James Surowiecki. 2004. *The Wisdom of Crowds*. New York, Random House Inc.