This is a preprint of an article accepted for publication in *Journal of the Association for Information Science and Technology*

Extracting Audio Summaries to Support Effective Spoken Document Search

Damiano Spina*, Johanne R. Trippas, Lawrence Cavedon, Mark Sanderson
RMIT University, Melbourne, Australia
*Corresponding author.
damiano.spina@rmit.edu.au, johanne.trippas@rmit.edu.au,
lawrence.cavedon@rmit.edu.au, mark.sanderson@rmit.edu.au

Abstract—We address the challenge of extracting *query biased audio summaries* from podcasts to support users in making relevance decisions in spoken document search via an audio-only communication channel. We performed a crowdsourced experiment that demonstrates that transcripts of spoken documents created using Automated Speech Recognition (ASR), even with significant errors, are effective sources of document summaries or "snippets" for supporting users in making relevance judgments against a query. In particular, results show that summaries generated from ASR transcripts are comparable, in utility and user-judged preference, to spoken summaries generated from error-free manual transcripts of the same collection. We also observed that content-based audio summaries are at least as preferred as synthesized summaries obtained from manually curated metadata, such as title and description. We describe a methodology for constructing a new test collection which we have made publicly available.

Index Terms—Spoken Document Retrieval; Query Biased Summarization; Crowdsourcing

INTRODUCTION

This paper investigates the presentation of search results for audio/spoken documents, such as podcasts, via an audio-only communication channel. In particular, we examine what form of *audio* search result summary is preferred by users. Supporting effective retrieval of long audio/spoken documents is increasingly important given the proliferation of such content (Larson and Jones, 2012). While *Spoken Document Retrieval* has a long history of research (Garofolo, Auzanne, and Voorhees, 2000), the main focus has been on retrieval. Result *presentation* via audio is a rather overlooked component.

We are particularly interested in presentation via a speech-only interface, where, "displaying" a summary means *playing* a segment of audio. Information search over audio has a number of significant challenges (Sahib, Tombros, and Stockman, 2012), including appropriate search result presentation. Our prior work (Trippas, Spina, Sanderson, and Cavedon, 2015) investigated aspects of preferred and effective *forms* of spoken/audio summaries, specifically, their length. The current work focuses on preference and effectiveness in relation to the *content* of summaries.

Podcasts and other audio documents often have manually created text metadata associated with them, e.g., title and/or description. Such information can be used both for retrieval and for result presentation. However, it has been shown that using *Automated Speech Recognition (ASR)* to convert spoken documents into text and retrieve against the text leads to better retrieval effectiveness than retrieving using metadata alone. Such improvement occurs even in the presence of significant ASR errors (Besser, Larson, and Hofmann, 2010).

Evidence also exists that for *informational* searches, users prefer search results containing *query biased summaries* (*snippets*) extracted from the content of retrieved documents (Tombros and Sanderson, 1998; Clarke, Agichtein, Dumais, and White, 2007) over a presentation of metadata. Hence, we explore the effectiveness of using summaries generated from automatically transcribed spoken documents, to select snippets of the podcast audio to playback as an audio summary for making relevance judgements. Specifically, we investigate the quality of query biased summaries generated from *noisy* transcripts (i.e., containing significant ASR errors) as compared to those generated from *manually created* full-document transcripts, as well as to summaries generated from metadata.

In particular, the following research questions are addressed:

- *Implicit preference in relevance judgments.* Do text/audio query biased summaries generated from automatic transcripts of podcasts allow users to effectively judge document relevance? Are these judgments as accurate as those made using summaries generated from corresponding manual transcripts?
- *Explicit preference.* Which form of summary do users prefer? In particular, when using an audio-only channel, do users prefer summaries extracted from the original audio or do they prefer speech-synthesized summaries from metadata?
- *Impact of ASR quality.* What is the impact of recognition errors when identifying relevant segments in the content of podcasts to generate audio summaries?

The above research questions are addressed via an experimental methodology that includes both text and audio-only presentation. Our ultimate goal is to determine characteristics

and strategies for effective search over audio-only communication channels.

In the next section, we describe related work followed by the construction of the dataset and our experimental methodology. Next, we present and discuss the results. Then, further detail aspects of the impact of noisy transcripts on search summary generation are presented, before concluding.

Related Work

We organize past work into three categories: (i) retrieval and summarization of spoken documents (ii) podcast search and (iii) relevance perception.

Retrieval and Summarization of Spoken Documents

Retrieving spoken documents has been extensively studied (Larson and Jones, 2012). The TREC Spoken Document Retrieval Track (Garofolo et al., 2000) was a benchmark initiative providing test collections to evaluate the effectiveness of different retrieval systems over spoken content. The effect of recognition error on *known-item* and *ad hoc* retrieval effectiveness was investigated in Garofolo, Voorhees, Auzanne, Stanford, and Lund (1999).

Past research has also explored ways of visually presenting summaries for multimedia, such as lecture recordings (Abdulhamid and Marshall, 2013; Abdulhamid, 2013; Munteanu, Penn, Baecker, and Zhang, 2006b; Stark, Whittaker, and Hirschberg, 2000) and spoken content from the cultural domain (Ordelman, Heeren, Huijbregts, de Jong, and Hiemstra, 2009; Heeren and de Jong, 2008). Typically, these include access via a visual interface to a playback functionality that allows users to search and browse audio content.

Experiments have measured the impact of ASR errors when searching passages in automated transcripts (Ranjan, Balakrishnan, and Chignell, 2006; Munteanu et al., 2006b; Stark et al., 2000). Results show that a large number of errors have a significant impact on relevance assessment and summarization, suggesting that in some cases it is preferable to avoid the use of low accuracy transcripts (Munteanu et al., 2006b).

Jing, Lopresti, and Shih (2003) investigated summarization from documents created by Optical Character Recognition (OCR) output. They also found that the quality of summarization is directly tied to the level of OCR error in a document.

Podcast Search

Podcast search engines typically index only the manually generated metadata to facilitate search and retrieval. However, it has been shown that retrieval using automatically transcribed content and metadata is more effective than using metadata alone (Besser et al., 2010; Goto, Ogata, and Eto, 2007; Ogata and Goto, 2009, 2012; Mizuno, Ogata, and Goto, 2008). For example, Besser et al. (2010) showed that retrieval of (Dutch) podcasts based on (noisy) full transcripts was more effective –via a 55% relative improvement in terms of Mean Average Precision (MAP)– than search using metadata alone. They also performed a user study that revealed that subjects –who were familiar with commercial podcast systems such as iTunes– preferred full transcript search.

The search engine Speechbot (Van Thong, Moreno, Logan, Fidler, Maffey, and Moores, 2002) indexed multimedia content from the Web, including popular radio programs. When transcripts were not available, ASR was performed. Experiments showed that good retrieval performance was achieved even when the transcription was highly inaccurate. However, using inaccurate automatic transcripts for search result summaries was problematic.

Podcastle (Goto et al., 2007) searched English and Japanese podcasts, using ASR transcripts to both retrieve and present search results with the possibility of playing the segments of the corresponding audio. Crowdsourcing was used to manually correct transcript errors (Ogata and Goto, 2009, 2012). The corrections were used to improve retrieval effectiveness and the result presentation, as well as to train better speech recognizers.

Relevance Perception from Summaries

Tombros and Crestani (2000) studied users' perception of relevance in audio summaries for text documents. They measured speed and accuracy of judgments when varying the way in which the search results were presented to users: on-screen display; read by human; read by human over a telephone; or read by synthesizer over telephone. They found that users judge relevance less accurately when listening to audio summaries, in particular when a synthesized voice is used.

The impact of summary quality on the effectiveness of users' ability to locate relevant documents was addressed by Turpin, Scholer, Jarvelin, Wu, and Culpepper (2009) who showed that poor quality summaries (in a web search context) lead users to misjudge the relevance of documents.

Methodology: Evaluation Infrastructure

Here, we describe:

1) The construction of a Test Collection, consisting of a set of podcasts and dual sets of associated text documents (manual and automatic transcripts of each podcast), along with queries and associated relevant documents;
2) The Indexing and Retrieval of podcasts;
3) The Query Biased Summary Generation approach.

Constructing the Test Collection[1]

The purpose of the test collection is to support comparison of effectiveness and preference of different types of query biased document summaries. This requires:

1) A collection of podcasts as well as manual and automatically-constructed transcripts from which to generate different versions of summaries;
2) A set of test queries and associated podcasts judged to be relevant to each query.

1. The collection and scripts to download the audio content are available at http://damiano.github.io/podcastsummaries/.

Figure 1 shows the overall workflow. First, a collection of English language news-related podcasts from the Australian Broadcasting Corporation (ABC) along with their corresponding manual transcripts[2] were gathered from the ABC website. Second, the podcasts were processed by an ASR system to generate a collection of automatic transcripts (see ASR Transcripts Collection).

Queries and relevance judgments were generated using a variation of the known-item approach from Azzopardi and de Rijke (2006) (see Query Design). The documents from which queries were generated became the known-item for each query (see Assigning Relevance Judgments), and thereby deemed relevant to the query.

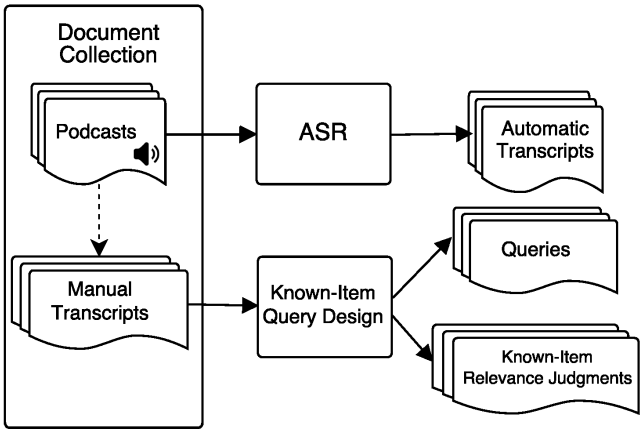


Fig. 1: Methodology's workflow for creating the test collection.

Audio Podcast Collection

The podcast collection consists of 3,012 recordings from the programs: AM, PM, Correspondents Report, and The World Today. The episodes were broadcast between October 1, 2014 and April 1, 2015. Each podcast typically consists of a spoken overview of a news story, followed by a different (often on-site) reporter providing further details and/or an interview.

The collection comprised 217 hours. On average, each podcast was 4 minutes 19 seconds long, with seven words in the textual metadata title and forty words in the description.

Manual Transcripts Collection

The manual transcripts were found to contain meta-comments that were not verbalized in the podcast, e.g., name of interviewer/interviewee or description of background noise (e.g., "Traffic noise in Hong Kong"). Such comments and names were identified and removed using regular expressions. The accuracy of the removal process was checked manually by inspecting a small sample of transcripts, making unlikely, but not impossible, the removal of acronyms within the audio that matched a regular expression. The manual transcripts of the podcasts contained an average of 492 words after this processing.

ASR Transcripts Collection

The automatic transcripts were created using the AT&T WATSON Speech API.[3] Other systems were considered but were found to be unusable or ineffective. For example, the Google Speech API did not allow the volume of ASR processing required for our task.

The original MP3 files of the podcasts were converted to 16-bit PCM WAV format using `SoX`. The WATSON API required the full podcasts to be split into one-minute segments for processing: this was performed using silence detection. The original American English acoustic models were used.

There is little published information on the accuracy of WATSON. Morbini, Audhkhasi, Sagae, Artstein, Can, Georgiou, Narayanan, Leuski, and Traum (2013) tested six datasets involving different dialogue domains where custom language models were used. Five leading ASR systems were tested and WATSON was found to be competitive across the datasets. On average, a Word Error Rate (WER) of under 30% was reported. This was achieved using customized language models.

We obtained an estimated WER of 61.1% (standard deviation of 8.9%) by comparing the manual and automatic transcripts. However we found the manual did not always match the audio content. The error rate was not unexpected as podcasts of speech are known to be prone to more error due to quality of recording and background noise (Ogata and Goto, 2009). The high incidence of Named Entities in the audio also contributed significantly to WER. However, as seen in Table 1, automatic transcripts with such error rates may still be useful for identifying appropriate segments of audio to use as a summary.

Query Design

The *known-item document identification* approach of Azzopardi and de Rijke (2006) has been shown to be effective when used for comparative evaluation of retrieval models (Balog, Azzopardi, Kamps, and de Rijke, 2007; Kim and Croft, 2009; Naji and Savoy, 2011). This approach assumes that a user wishes to retrieve a particular document that they have already previously identified (the *known-item*). This assumption eliminates the need for explicit relevance judgments as the known-item is deemed to be the relevant document. However, a criticism of the approach is that automatically generated queries often look artificial or unrealistic. To address this criticism, we manually created queries, asking human annotators to extract a query from a given known-item document (i.e., a manual transcript). Each such query was designed to correspond to one that a user might submit to a search system to successfully retrieve the item.

Known-Item Document Selection: When using the known-item approach, it is important to minimize likelihood of selecting documents associated with topics that are under-represented or outliers in the collection. Azzopardi, de Rijke, and Balog (2007) modeled the distribution of priors used for selecting representative documents as document importance, measured using inlinks. Given that podcasts in our collection

2. http://www.abc.net.au/transcripts

3. http://developer.att.com/apis/speech/docs

TABLE 1: Example of manual and automatic transcripts (examples are truncated to facilitate legibility).

Manual Transcript	Automatic Transcript
Thousands of artists from around the world have arrived in Adelaide for the annual Fringe Festival The month long event kicks off tonight with a parade through the city It's anticipated it will be visually spectacular, with 80 colourful floats winding through the streets And for the first time, vision impaired and blind people have been able to touch the floats and the costumes beforehand to help them picture what they might look like…	Who is is system around will sit around the lake city annual fringe festival …monthly meetings kicks off to lunch with her right to the city. Its anticipated. It'll be hugely spectacular with icy comma for flights morning through the streets for the first time vision impaired and blind people have been able to touch the flights and costing for hand to help them picture what you look like…

do not contain inlinks, we used the Latent Dirichlet Allocation (Blei, Ng, and Jordan, 2003) version of *topic modeling*[4] to identify representative documents. The model was configured to generate 100 topics against which to distribute the documents in the collection.

Documents were randomly sampled from those with at least 0.33 probability of belonging to one of the identified topics. The output of this step was 476 documents, 15% of the total collection. We then randomly sampled 45 documents which were the known-items selected for use in the construction of test queries.

Query Construction: For each known-item k, four annotators (three authors of this paper plus a non-IR expert) were instructed to build queries q_k that were specific to the main topic of the document, such that k would be definitively relevant to the query. At the same time, the query should not be specifically tailored to precisely retrieve that document in preference to other very similar ones. Queries were to be kept to around 3–7 terms. One of the 45 known-items was not found by the search engine using the automatic transcripts; since the main purpose of this work was to study the impact of transcripts in search summaries, this test case was discarded, leaving 44 known-items.

The final queries were generated by randomly selecting one from the four candidate queries, for each known-item. The full query set is found in the appendix.

Assigning Relevance Judgments

It is reasonable to consider the known-item k as relevant to the query q_k designed to retrieve it (Azzopardi and de Rijke, 2006). While we cannot ensure that other retrieved documents are relevant, this is not problematic given our focus is to judge decision-making and preference of summaries in the context of a given query-document pair.

Indexing and Retrieval

The manual and automatic transcripts as well as the podcast metadata were indexed using the Solr search engine (Lucene 4.6 library).[5] The three collections were indexed into separate fields. In addition, combinations of the collections (i.e., metadata plus manual transcripts, metadata plus automatic transcripts) were indexed in fields. Documents were preprocessed using the Lucene libraries for tokenization, stopword removal, and stemming (Porter, 1980). We used the Solr default ranking function, which is based on the Vector Space Model (VSM) (Salton, Wong, and Yang, 1975). For each of the 44 test queries, Solr retrieved the top 150 documents from each of the collections (i.e. manual, automatic, metadata). Summaries were produced for each of these documents.

Query Biased Search Result Summary Generation

Dynamic *keyword-in-context* (Manning, Raghavan, and Schütze, 2008) query biased summaries (i.e., windows in the content containing one or more query words) were extracted for each retrieved document, using the "Standard Highlighter" included in the Solr tool for summary generation.[6] This tool scores fragments of a document by the number of unique query terms found. The fragment with the highest score is the chosen summary. Fragments/summaries were limited to one hundred characters in length (12.35 words on average). When no summary was able to be generated (this occurred 5.4% of the time), the first one hundred characters of the corresponding indexed content (i.e., automatic transcript, manual transcript or metadata) was used. For each of the 44 test queries, ten summaries were generated for each ranked list –which includes the known-item– for each of the text and audio summary versions.

Figure 2 illustrates the different types of summaries generated for our experiments, in both the text and audio channels. The labels text_auto, text_manual and text_metadata correspond to text summaries generated from ASR transcripts, error-free manual transcripts and manually curated metadata (title and description), respectively.

Text_metadata were extracted automatically using the query biased summary technique. Audio_auto and audio_manual were obtained as follows. First, a text summary was generated from the corresponding text_auto or text_manual. Next, time-stamps in associated with the *start* and *end* words of the summary were obtained.[7] Finally, the corresponding audio segment between the *start* and *end* time-stamps was used as the audio summary.

We also created text summaries corresponding to playing back spoken segments to users by manually correcting ASR

4. taken from the Mallet machine learning toolkit (McCallum, 2002)
5. http://lucene.apache.org/solr/
6. https://cwiki.apache.org/confluence/display/solr/Standard+Highlighter
7. The WATSON API does not provide this information. Therefore, we used an auxiliary ASR tool (Pocketsphinx) to obtain the time-stamps. We then aligned text_auto and text_manual summaries with their corresponding segments in the Pocketsphinx transcripts with highest term overlap. This automatic process accurately obtained 71% of the 880 audio summaries included in audio_auto and audio_manual. Finally, all summaries were manually checked and wrongly aligned summaries fixed.

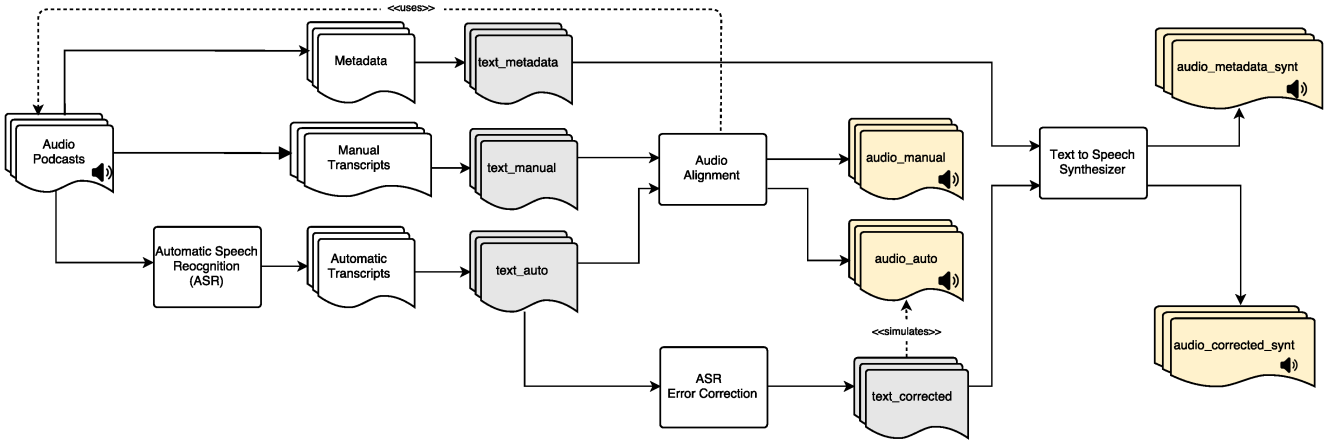


Fig. 2: Workflow for generating text and audio summaries from the different sources.

errors and then using display of the corrected text summary as a surrogate for played-back audio segment (text_corrected). During this correction process, we found that for 40% of the known-items, ASR errors resulted in a different part of the podcast transcript being selected for use as a summary compared to when manual transcripts are used (text_manual).

The audio version of summaries generated from metadata (audio_metadata) was obtained by using a text-to-speech synthesizer. For comparison we also generated synthesized versions of the text_corrected summaries: audio_corrected_synt.

Table 2 shows the different summary versions, both audio and text, generated from the documents associated with that known-item, i.e., the automatic transcripts (both raw and corrected for ASR errors), the manual transcripts and the metadata.

Experiments: Obtaining User Feedback

We used a crowdsourcing platform (CrowdFlower) to gather judgments for the summaries generated for each query.[8] Previous research suggests that non-expert crowdsourced workers produce work of a similar standard to expert workers or to controlled experiments (Munro, Bethard, Kuperman, Lai, Melnick, Potts, Schnoebelen, and Tily, 2010; Sabou, Bontcheva, and Scharl, 2012; Snow, O'Connor, Jurafsky, and a.Y. Ng, 2008). We used *Gold Questions* with clear answers (Buchholz, Latorre, and Yanagisawa, 2013) to ensure quality control.

Details of Tasks

Two tasks were run: relevance-judging, designed to test whether summaries generated from automatic transcripts were as effective as those generated from manual transcripts; and summary preference, designed to measure crowd source worker preferences between two versions of a summary.

8. All experiments were performed under Ethics Application BSEH 10-14 at RMIT University.

Task One: Relevance

As shown in Figure 3, workers were presented with a query and asked to indicate which summaries were relevant, were not relevant, or to indicate that results were unclear in relation to the query. Figure 4 shows the interface for judging relevance from audio summaries. Results returned by the search engine were organized into pages of ten summaries each. The page from which the known-item was retrieved was the one shown to the workers. A summary for the known-document was deemed to be effective if the worker marked that summary as relevant (judgments for the other summaries were not considered).

Task Two: Preference

As shown in Figure 5, workers were asked to read a complete manual transcript, Figure 5a and then proceed to Question 1 (Figure 5b and Figure 6 for text and audio, respectively) where they were presented with two lists of summaries generated from different collections.

Next, workers were asked to choose one summary from each list as the most representative of the given document. Workers were then asked to prefer one summary over another (see Figure 5c). An option of *no preference* was included. Recall that the two summaries were built from different transcripts (e.g., text_manual vs. text_auto).

At the end of both tasks, workers were asked to comment on the task they completed.

Pilot runs were conducted in order to optimize the study design. In addition, per-task payment rates were set via a questionnaire posed to the pilot users. Workers were paid 0.25 American dollar per page with each page containing three tasks.

Selecting Crowdsource Workers

Only workers with an IP address from Australia, Ireland, New Zealand, the United Kingdom and the United States were allowed to participate in order to maximize the likelihood that participants were native English speakers or had high English fluency. Workers could perform up to 33 assignments per job

TABLE 2: Types of search result summaries.

Name	Channel	Document	Description
text_auto	Text	Automatic Transcript	Text summary generated from automatically transcribed content using ASR
text_manual		Manual Transcript	Text summary generated from manually transcribed content
text_metadata		Metadata	Text summary generated from title, and description (i.e., metadata)
text_corrected		Automatic Transcript	text_auto after manually correcting ASR errors. It simulates audio_auto in text
audio_auto	Audio	Automatic Transcript	Audio segment in the original content that corresponds to text_auto
audio_manual		Manual Transcript	Audio segment in the original content that corresponds to text_manual
audio_metadata_synt		Metadata	Synthesized version of text_metadata
audio_corrected_synt		Automatic Transcript	Synthesized version of text_corrected

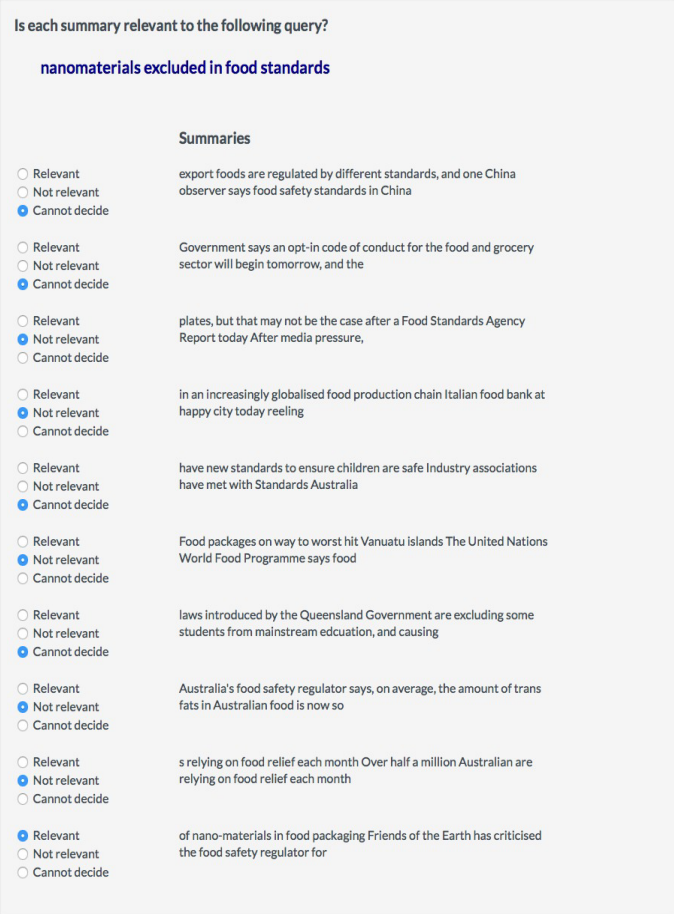


Fig. 3: CrowdFlower setup of the Relevance Task for text summaries.

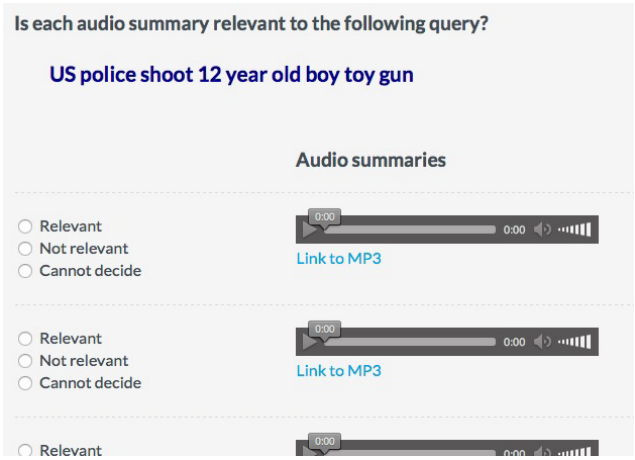


Fig. 4: CrowdFlower setup of the Relevance Task for audio summaries.

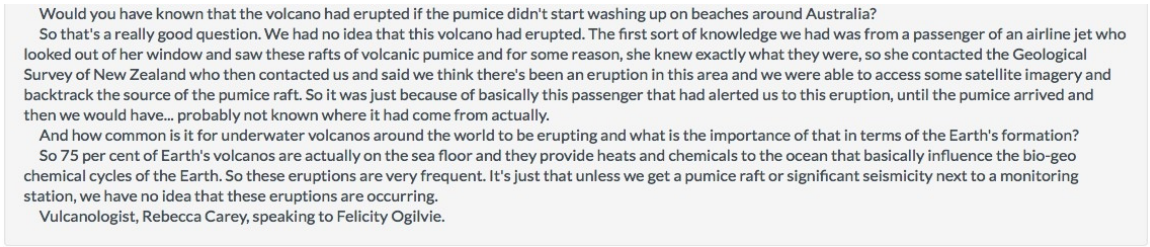
and workers who took less than ninety seconds to complete three assignments were disqualified by CrowdFlower.

Setting Gold Questions

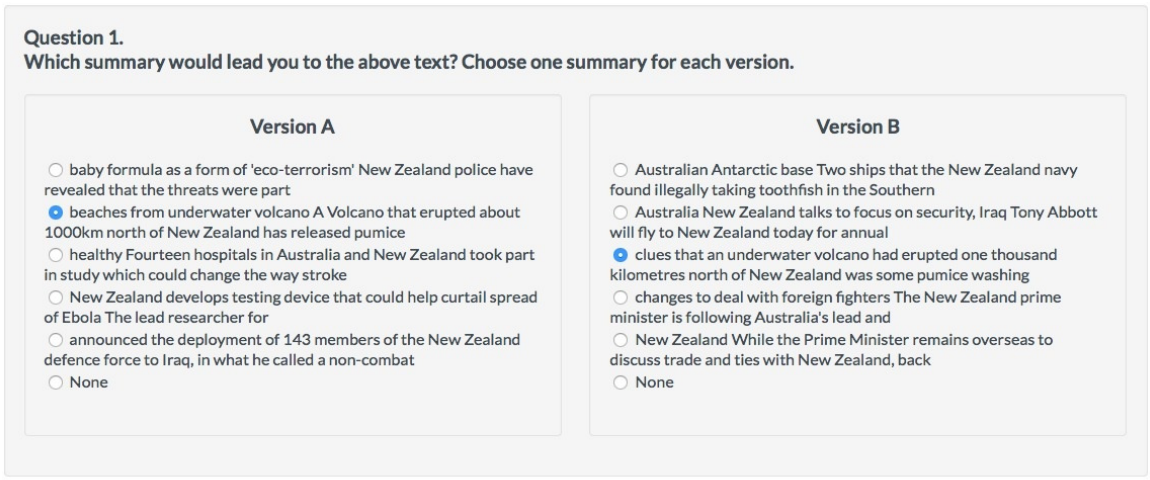
Each CrowdFlower assignment consisted of three tasks, which were randomly selected by the CrowdFlower platform from the set submitted as a job. Nearly 40% of the submitted tasks were Gold Questions. This maximized the likelihood of including a Gold Question in every assignment. Workers were not allowed to perform further assignment if their Gold Question accuracy dropped below 90%. In addition, in order to commence annotating actual tasks, workers had to successfully complete an initial assignment comprised of three Gold Questions.

Gold Questions for both Relevance and Preference Tasks were generated via a semi-automatic process. Artificial rankings were generated by randomly selecting summaries from unrelated queries. Then, a randomly chosen summary was replaced by the summary for the known-item for the query. Finally, the resulting rankings were manually inspected in order to ensure that the randomly-selected summaries were actually not relevant to the query; if they were, they would be manually replaced. This ensured that the known-item was the only relevant result in the list.

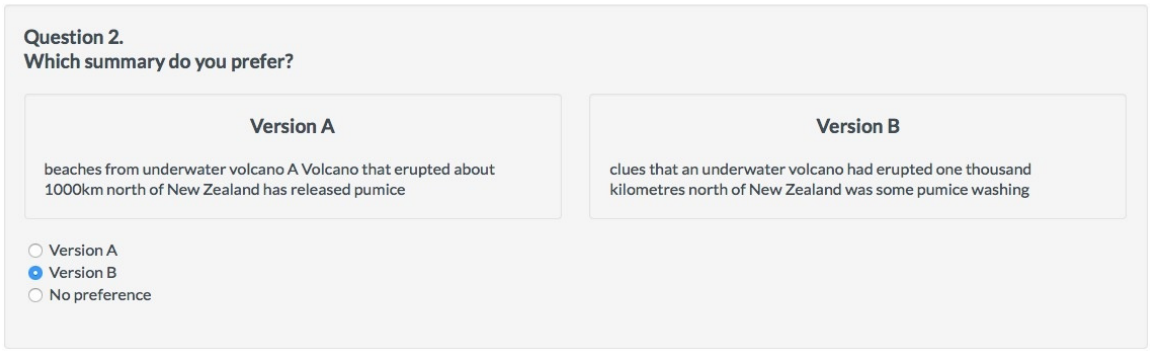
A Gold Question was deemed passed if the worker selected the known-item as relevant. For the Relevance task, this meant other items could also be selected. For the Preference Task, workers were presented with two lists of summaries generated from different collections and only one selection from each list was allowed before moving to the preference comparison (Figures 5 and 6). Here, workers needed to select the summary corresponding to the known-item to successfully pass the Gold Question. The actual summaries of the full document, each in a different list, were generated via different means; e.g., the representative summary in one list was generated from the manual transcript and the other list was generated from the automated transcript (where the order of Versions A and B was rotated). This quality control mechanism aimed to ensure that workers actually read the document before judging preference between the two versions of summaries.



(a) Preference Task - Document



(b) Preference Task - Question 1



(c) Preference Task - Question 2

Fig. 5: CrowdFlower setup of the Preference Task.

Crowdsource Worker Statistics

A total of 122 workers successfully completed the Relevance and Preference Tasks. Workers were allowed to contribute to both tasks, but as the tasks were released at different dates and times only 24% of the workers contributed to both. In total, 86 workers contributed to the Relevance Task and 65 to the Preference Task.

Workers were presented with a contributor satisfaction page from CrowdFlower after they finished the full task and 88 were completed. The overall contributor satisfaction rate was 82%: 77% for the Relevance Task and 84% for the Preference Task.

Workers who successfully passed the Gold Questions, completed between 1 & 138 non-gold question tasks (with a mean of 16.82 and standard deviation of 19.60). A total of 2,052 tasks were completed with 960 judgments for the Relevance Task and 1,092 for the Preference Task. Workers who commenced a task were made aware of the required high accuracy rate of 90% for the Gold Questions.

In total, 525 Gold Questions were answered for the Relevance task, of which 15.43% were not answered correctly. For the Preference Task, workers answered 654 Gold Questions, of which 12.69% were not answered correctly. In total 1,179 Gold Questions were answered with 164 not answered correctly.

Workers could contest a Gold Question if they thought it was an unfair question/answer. When workers contested and made a reasonable argument on why a Gold Question was unfair, it was disabled. Unfair Gold Questions were commonly caused by transcripts with many ASR errors. Workers were not punished for contesting. In total, 3% of the Gold Questions

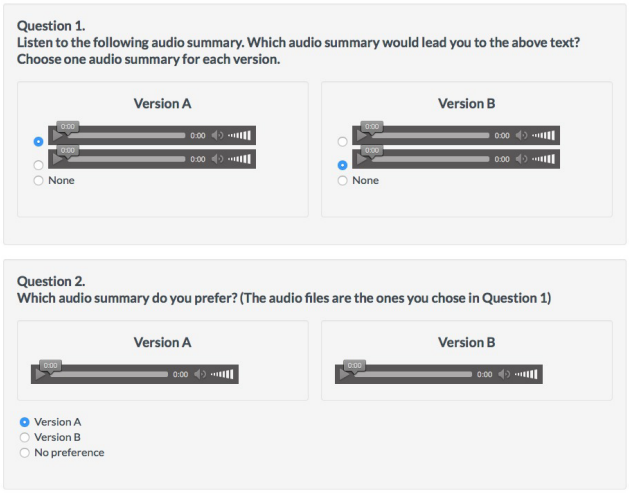


Fig. 6: CrowdFlower setup of the Preference Task for audio summaries.

in the Relevance task and 1.4% in the Preference Task were disabled.

Evaluation Metrics

Mean Reciprocal Rank (MRR) was used to measure retrieval effectiveness of worker judgements of relevance.

$$MRR = \frac{1}{|\mathcal{Q}|} \sum_{i}^{\mathcal{Q}} \frac{1}{\text{rank}_i} \tag{1}$$

Given that we collected from crowd workers several relevance judgments for each known-item, we also measured user agreement on the question of relevance (Alonso and Mizzaro, 2009). Here, we used *Joint Annotator Precision (JAP)*. It is defined as follows. Let $\mathcal{J}_r$ be the set of relevance judgments j_i obtained for a document r being the known-item. The JAP of a group of workers of a known-item being annotated as relevant as:

$$JAP(r = \textit{known-item}) = \frac{1}{|\mathcal{J}_r|} \sum_{i}^{\mathcal{J}_r} rel(j_i) \tag{2}$$

This metric also provided an alternative indication of the quality of summaries, by measuring how well the judges agree that the known-item is relevant. Note, in the equation for JAP *precision* refers to the judgment being made by each worker (annotator) of the (known to be relevant) known-item, reflecting the quality of the summary used to make the judgment.

JAP also can be viewed as a user effectiveness metric (Al-Maskari and Sanderson, 2010). In our experiments, we measured how good users were at identifying a known-relevant document when inspecting different versions of summaries.

When analyzing results, statistical significance was measured using Student's two-tailed t-test. We use $^{\triangle}$ and $^{\blacktriangle}$ throughout to indicate statistical significance for $p < 0.05$ and $p < 0.01$, respectively.

Jayasinghe, Webber, Sanderson, Dharmasena, and Culpepper (2015) showed that it is inappropriate to use t-test to assess statistically significant equivalence, since not rejecting the null hypothesis does not imply accepting it. That is, if no statistical significance is found, it cannot be assumed that M_A and M_B perform equivalently. Therefore, we used Confidence Intervals (CI) to test equivalence. In this case, the null hypothesis is $H_0 = |M_A - M_B| > \delta$. If the intervals are within a given threshold δ, then the measurements are statistically equivalent with a certain grade of confidence. In our analysis, we used a confidence level of 95% and a threshold $\delta = 0.1$ (Sakai, 2014).

Results and Discussion

This section presents and discusses the experimental results.

Ranking Evaluation

We first compare retrieval effectiveness across a range of content types, see Table 3.

TABLE 3: Mean Reciprocal Rank (MRR) for the known-item task varying the indexed content. Statistical significance against *Metadata* is shown.

Indexed Content	MRR
Metadata	0.525
Automatic Transcript	0.561
Manual Transcript	0.598$^{\triangle}$
Metadata + Automatic Transcript	0.566
Metadata + Manual Transcript	0.612$^{\triangle}$

Result 1. *Indexing transcribed content of spoken documents improves retrieval effectiveness compared to indexing metadata only.*

Retrieval over documents represented using only metadata resulted in 0.525 MRR. Retrieval over noisy automatic transcripts resulted in 0.561 MRR, a 7% improvement (not significant). Retrieval over an index built from error-free manual transcripts resulted in MRR of 0.598, 13%$^{\triangle}$and 7% more effective than metadata and automatic transcripts, respectively. Slightly improved results were obtained when transcripts and metadata are indexed together.

Our results were in line with previous findings (Besser et al., 2010): indexing the transcribed content of podcast episodes is more effective and is complementary to using metadata.

Relevance Task

Table 4 shows the results of the crowd sourced relevance judgments obtained for the known-items. For each of the seven summary versions[9] across the 44 queries, between 3 and 7 judgments from different workers were obtained, with an average of 3.6 and a standard deviation of 1.15. Following Turpin et al. (2009), the most frequent relevance judgement across the workers was the one chosen.[10] In the

9. audio_corrected_synt was only used in the Preference Task.

10. A recent study (Davtyan, Eickhoff, and Hofmann, 2015) suggests that considering three judgments with majority voting is a good estimate of the true relevance label.

case of a tie (1.3% of the entire set), the "Cannot decide" judgment was chosen. In order to quantify agreement, we also report macro-averaged JAP of the workers judging the known-item as relevant (Eq. 2).

Result 2. *Recognition errors have significant impact on workers' perception of relevance.*

Regarding text summaries, for over 90% of cases, workers identified the known-item when manual information (metadata or manual transcripts) was used to generate summaries (text_metadata, text_manual). For summaries based on automatic transcripts (text_auto), in 25% of the known-items, workers misclassified the document, either as not relevant (9 known-items) or "Cannot decide" (2 known-items). Moreover, JAP dropped to 0.679 compared to when summaries extracted from the manual transcripts were used, over 30%▼.

Directly showing users summaries generated from noisy automatic transcripts hampered the task of identifying relevant documents. A similar effect was observed in different settings, e.g., using ASR transcripts for completing a quiz after interacting with a lecture webcast (Munteanu, Baecker, Penn, Toms, and James, 2006a) or inspecting the full transcript for relevance judgments (Stark et al., 2000). To our knowledge, the effectiveness of using ASR transcripts to identify relevant passages for generating search result summaries, and its potential application to audio-only communication channels, has not been previously measured.

Result 3. *Corrected summaries are perceived to be as effective as summaries generated from manual transcripts or metadata for judging relevance.*

Using corrected summaries –which simulates a spoken interface scenario where the corresponding fragment of the original audio is played back– we found that the known-item was judged as relevant for all the queries, with a high JAP across the different workers. Results for text_corrected corroborated that the drop in worker accuracy was due to recognition errors.

Result 4. *Audio summaries are in general informative, independent of the source used to obtain their corresponding text segments.*

Users effectively identified more than 95% of the known-items (i.e., at least 42 of the 44 test cases: see Table 4) using any of the audio summaries formats. The relative improvement in terms of JAP over use of text from uncorrected ASR transcripts was 34%▲. This result suggests that even though automatic transcripts had a high estimated WER, they were nevertheless useful for identifying passages to create audio summaries that play back the original audio from the content.

Result 5. *Workers' judgments of relevance using audio summaries from automatic transcripts was equivalent to their judgment using summaries generated from manually curated content.*

So far, we have seen that judgments of summaries containing recognition errors are statistically significantly worse than judgments using error-free text or audio summary. We have also shown that error-free text summaries (text_manual, text_metadata and text_corrected) and audio summaries obtain equivalently high JAP scores. Testing statistical equivalence, Figure 7 shows the confidence intervals for different comparisons.

The first comparison (text_auto vs. audio_auto) highlighted again the effect of showing recognition errors to workers. However, correcting those errors in text (text_corrected) resulted in judgments that were significantly equivalent to audio_auto, if we consider a threshold $\delta = 0.1$. Audio_auto was also significantly equivalent to audio_metadata_synt ($\delta = 0.1$). Audio_manual and audio_auto were not significantly equivalent ($\delta = 0.1$). Finally, comparing text_manual, text_metadata with their analogous audio summary (audio_manual and audio_metadata_synt, respectively) were not significantly equivalent.

The equivalence test analysis suggests that workers performed equivalently at a statistically significance level when listening to audio summaries generated from noisy transcripts (audio_auto) or synthesized summaries from metadata (audio_metadata_synt). Moreover, JAP scores for text_corrected were statistically equivalent to the corresponding audio summaries audio_auto, suggesting that correcting recognition errors was a good proxy to simulating this type of audio summaries in text.

In summary, results for the Relevance task show that, even though automatic transcripts are noisy (with an estimated WER around 60% on average), they can be effectively used to identify audio segments that when played are as effective as those generated from metadata and similar in effectiveness to those generated from manual transcripts.

Preference Task

In the Preference task, workers indicated which of the different summaries they preferred as representatives of a document (see CrowdFlower setup in Figure 5). On average, each known-item received 4.09 preference judgments, with a standard deviation of 1.20. The number of judgments per known-item ranged between 3 and 7. As for the Relevance task, majority vote was taken as the overall judgment.

Tables 5 and 6 show, respectively, the preference results for the comparison of summaries generated from automatic transcripts versus manual transcripts and metadata, for both text and audio channels. For each comparison, the number of the preferred known-items is shown. Confidence intervals are constructed at a significance level of $\alpha = 0.05$ (i.e., 95% confidence level).

Result 6. *Summaries generated from automatic transcripts are significantly less preferred when summaries shown to workers contain ASR errors, but are no less preferred when those summaries do not include errors as in the case of audio summaries.*

Table 5 shows that text_manual summaries were significantly more preferred than text_auto, i.e., the summaries that included recognition errors. A different behavior was observed when error-free text summaries (text_corrected) were shown

TABLE 4: Relevance judgments (aggregated by voting) and averaged Joint Annotator Precision (JAP) in judgments against the known-items for different versions of generated summaries. Statistical significance against *Automatic (Text)* is shown.

Channel	Summary Version	Relevant	Not relevant	Cannot Decide	JAP(k)
Text	text_metadata	43	0	1	0.972▲
	text_auto	33	9	2	0.679
	text_manual	41	2	1	0.900▲
	text_corrected	44	0	0	0.955▲
Audio	audio_metadata_synt	42	2	0	0.909▲
	audio_auto	43	1	0	0.909▲
	audio_manual	42	1	1	0.962▲

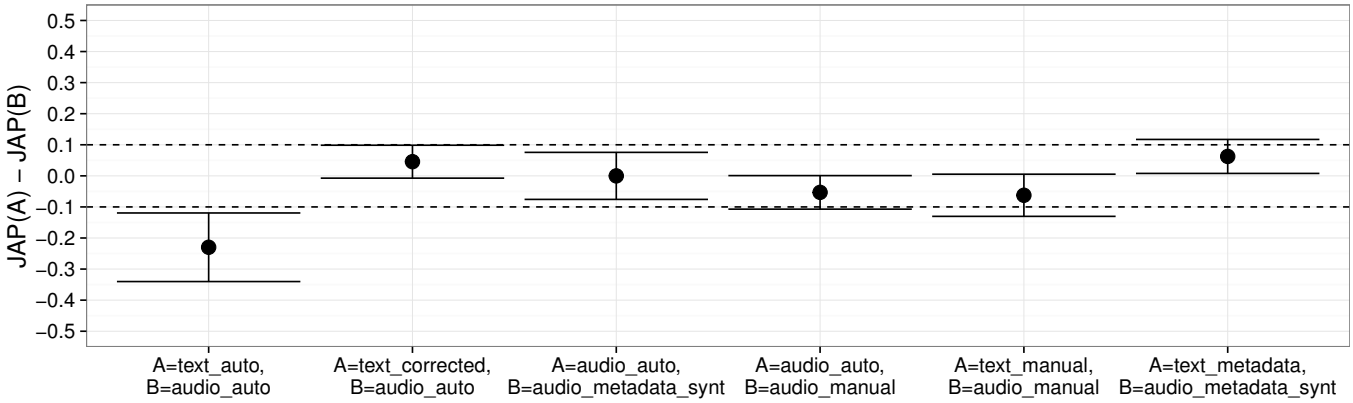


Fig. 7: Confidence intervals ($\alpha = 0.05$) to test significant equivalence in relevance task. Dashed lines correspond to the threshold $\pm\delta = \pm 0.1$.

TABLE 5: Users preference (aggregated by voting) when different versions of summaries based on automatic transcripts were compared against summaries generated from manual transcripts. 95% Confidence Intervals (CI) are shown in brackets.

Comparison		Number of Known-Items [95% CI]		
		Summary Based on Automatic Transcript	No Preference	Summary Based on Manual Transcript
text_auto	vs. text_manual	3 [1,8]	3 [1,8]	38 [32,41]
text_corrected	vs. text_manual	17 [11,23]	15 [9,21]	12 [7,18]
audio_auto	vs. audio_manual	16 [10,22]	15 [9,21]	13 [7,19]

TABLE 6: Users preference (aggregated by voting) when different versions of the summaries based on automatic transcripts were compared against summaries generated from metadata. 95% Confidence Intervals (CI) are shown in brackets.

Comparison		Number of Known-Items [95% CI]		
		Summary Based on Automatic Transcript	No Preference	Summary Based on Metadata
text_corrected	vs. text_metadata	25 [18,30]	2 [0,6]	17 [11,23]
audio_corrected_synt	vs. audio_metadata_synt	25 [18,30]	4 [1,9]	15 [9,21]
audio_auto	vs. audio_metadata_synt	28 [21,33]	6 [2,11]	10 [5,16]

to the workers. The corrected version was most preferred, and there was "no preference" for 15 (34%) of the known-items when compared to text_manual. The same behavior was observed for audio summaries (audio_auto).

Result 7. *Error-free summaries generated from noisy automatic transcripts are no less preferred than summaries generated from metadata.*

When compared to summaries generated from metadata (Table 6), text_corrected and audio_corrected_synt were slightly (but not statistically significantly) preferred to text_metadata and audio_metadata_synt, respectively. This indicates that content-based audio summaries, generated from noisy automatic transcripts, were no less preferred than synthesized summaries obtained from manually constructed metadata.

Result 8. *Audio summaries generated from automatic transcripts were preferred over synthesized summaries generated from metadata.*

When the original audio was played (audio_auto) instead of the synthesized version of corrected summaries, the preference of audio_auto is significantly higher, at a significance level of $\alpha = 0.05$. Even though segments in audio_auto were obtained from noisy transcripts with high WER, workers still preferred human voice summaries against synthesized summaries from manually created metadata. However, this was a preliminary result and the impact of synthesized voice in summaries needs to be further explored.

Tables 5 and 6 show that workers indicated more "No preference" judgments when summaries generated from automatic transcripts were compared to those generated from manual transcripts (Table 5) than when compared to summaries generated from metadata (Table 6). One possible cause of this effect is that summaries generated from automatic transcripts might be closely similar to those generated from manual transcripts – i.e., it is likely that workers would choose the "No preference" option when comparing two summaries that are very similar or identical. In fact, in 27 (61%) out of 44 of the known-items, ASR errors resulted in a segment of the podcast transcript being selected for use as a summary (i.e., the segment used in text_auto, text_corrected and audio_auto summaries) that has some overlap with the one selected when manual transcripts are used (text_manual and audio_manual).

In an effort to shed more light on this, we analyzed the term overlap between three different versions of known-item text summaries: text_corrected, text_manual and text_metadata by computing the Jaccard similarity coefficient (Eq. 3):

$$\text{Jaccard}(s, s') = \frac{W_s \cap W_{s'}}{W_s \cup W_{s'}}. \quad (3)$$

where W_s and W'_s are the set of words obtained after tokenizing and lowercasing the summaries s and s', respectively. The Jaccard similarity between the known-items in text_corrected and text_manual was, on average, $0.58^{\blacktriangle}$, whereas the Jaccard similarity between text_corrected and text_metadata was 0.30, being the difference statistically significant ($\alpha = 0.01$).

Hence, text_corrected being more similar to text_manual than text_metadata likely influenced the fact that workers tended to select "No preference" in comparisons that included more similar summaries (text_corrected vs. text_manual).

Results for the Preference task indicated that audio summaries generated from noisy ASR transcripts were no less preferred than those generated from error-free manual transcripts and curated metadata. They were also preferred over metadata-based summaries generated via synthesized voice.

Further Analysis

We first describe the impact of WER on the Relevance task. We then analyze the relationship between the position of the known-item in retrieval result-lists and worker relevance judgments. We finally describe aspects of the tasks assessed via the crowdsourcing platform.

Impact of Word Error Rate

The WER of the summaries generated from the automatic transcripts was found on average to be 46.4%. Sanderson and Shou (2007) explored the relationship between WER and the position in the ranking of retrieved spoken documents. They found that documents with low WER tended to be ranked higher. Documents containing query words with a high term frequency tended to have a lower WER than the collection average.

We therefore examined whether known-items with low WER in the summaries were better identified as relevant than summaries with high WER. Figure 8 compares JAP (Eq. 2) in the judgments against the WER of text_auto and text_corrected summaries. Each dot represents a known-item in the collection.

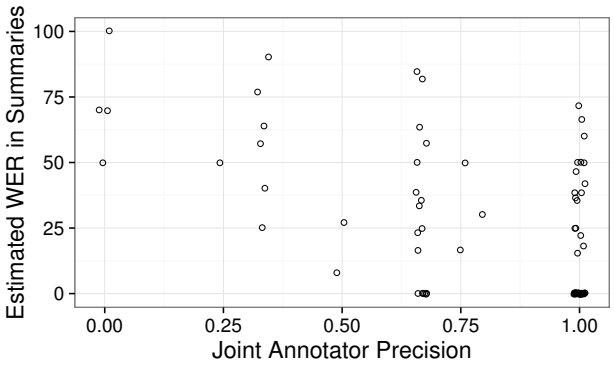


Fig. 8: Joint Annotator Precision (JAP, x-axis) vs. Word Error Rate (WER) in automatic summaries (y-axis).

There is a slight trend of having a greater chance of successfully identifying a known-item as relevant when WER is low. Pearson correlation was measured as $r = -0.59$ (and $r = -0.40$ considering text_auto only). This indicates a weak degree of inverse dependence between the two variables, aligning with the previous work.

We also analyzed the relationship between Keyword Error Rate (keyWER), i.e., the Word Error Rate of query keywords in the snippets, and JAP (Figure 9). The figure shows that users are more likely to identify a known-item as relevant

when Keyword Error Rate is low. The Pearson correlation ($r = -0.38$ and $r = -0.19$ considering text_auto only) is not as high as when considering full WER, but the same trend of inverse dependence between the two variables was observed.

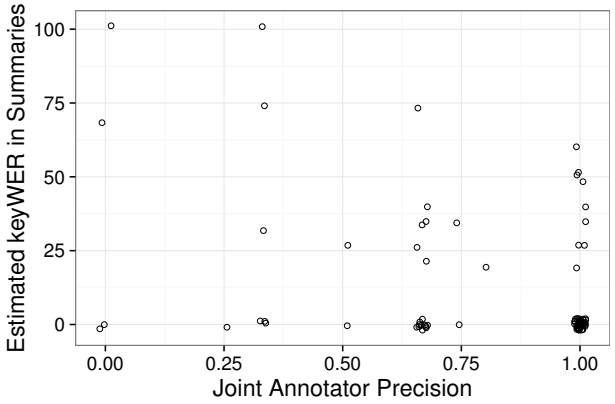


Fig. 9: Joint Annotator Precision (JAP, x-axis) vs. Key-Word Error Rate (keyWER) in summaries (y-axis).

Effect of Ranking Position

Workers were less accurate when judging relevance from more noisy summaries. We examine whether a similar relationship exists with the retrieval system. In order to quantify this relationship, we compared JAP in the relevance judgments to the original ranking position of the known-item (Figure 10).

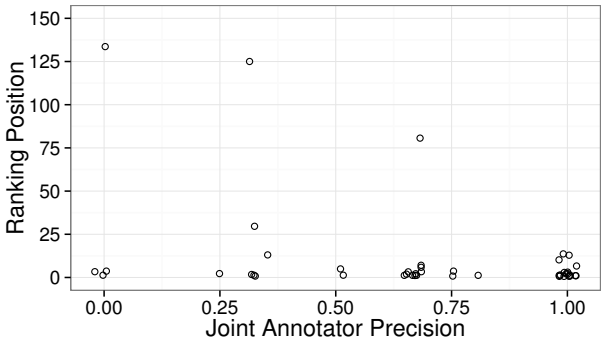


Fig. 10: Joint Annotator Precision (JAP) vs. ranking position of the known-item in summaries generated from automatic transcripts.

A weak inverse correlation ($r = -0.34$ Pearson correlation) was found between JAP and rank. The figure shows that documents with perfect JAP are returned in the first 20 positions of the system's results ranking, while documents ranked lower tend to obtain a lower JAP in the relevance judgments.

Figure 11 depicts the relationship between JAP and the position of all documents in the ranking.

If we consider all the judgments obtained for the ranking (i.e., not only looking at the known-item), the same trend was observed but with a lower correlation ($r = 0.21$).

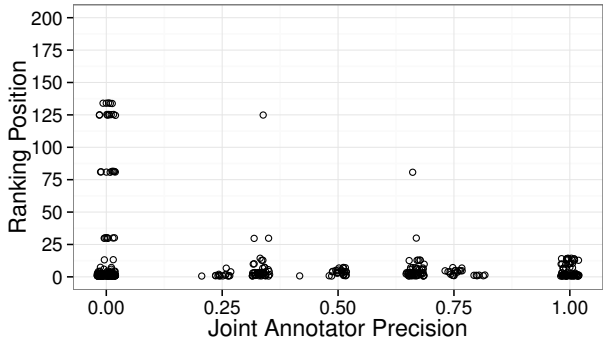


Fig. 11: Joint Annotator Precision (JAP) vs. ranking position of all the summaries generated from automatic transcripts.

In sum, although the correlation is low, documents that appear in first position of the ranking tend to be judged as relevant more consistently by users. This effect was previously observed in summaries (Sanderson, 1998).

Crowdsourcing Tasks

Reflecting on the use of crowdsourcing, the use of Gold Questions was crucial to ensure accurate annotations and to filter out workers who seemed not to pay close attention to the task.

Many workers seemed to require all the terms of a given query to appear in a document before marking that document as relevant. From the Gold Questions we found that such workers tended to annotate as "Cannot decide" summaries containing only a subset of query terms.

Examining the worker optional feedback, some indicated that they found the Relevance task more difficult when judging summaries from automatic transcripts. In particular, they complained about the lack of punctuation and incompleteness of sentences in the summaries (likely truncated due to the fixed summary length).

With regard to the Preference task, some feedback mentioned the problem of ASR errors (e.g., *"There were more errors in version B"*) and emphasized errors in query terms (e.g., "cubani" instead of "Kobane").

CONCLUSIONS

We studied the use of noisy ASR transcripts for generating query biased spoken summaries for podcast search in audio-only interfaces. We compared both text and audio summaries extracted from automatic transcripts to those that would be generated from error-free manual-constructed transcripts or manually curated metadata associated with the same spoken documents.

We found that users accurately judged relevance in a ranked list of query biased audio summaries generated from noisy automatic transcripts. The quality of these judgments is comparable to those obtained using summaries generated from error-free manual transcripts.

This suggests that generating summaries from noisy automated transcripts still results in appropriate document segments being selected as summary segments, which can then be played back for users to hear.

We also found that content-based audio summaries are preferred over synthesized summaries obtained from manually created metadata when the original audio is used, whereas there is no significant preference when synthesized voice is used instead.

These results are important from the standpoint of informing the design of search engine interfaces for retrieving podcast and other spoken-audio content. In particular, the results demonstrate that, in the absence of manual transcriptions, automatic transcripts generated by an ASR engine –even in the context of significant WER– provide a valuable surrogate for generating audio summaries that support users making effective relevance-judgments for spoken document retrieval.

Extensions of the experiments performed here are planned, to generalize the setting and thereby the results. First, our podcast dataset consists of only news-related podcasts, from a single broadcaster, a limitation constrained by the availability of manual transcripts. We plan to extend our experiments to other types of spoken content (e.g., audio books). Second, reducing ASR error –e.g., by customizing acoustic and language models to the collection or using a different ASR system– would allow varying WER to provide a better understanding of the relationship between the level of WER in the automatic transcripts and users' preferences in inspecting summaries for relevance judgment. Note, however, that we would not expect changes in performance/preference results after reducing WER.

ACKNOWLEDGMENTS

This research was partially supported by Australian Research Council Project LP130100563 and Real Thing Entertainment Pty Ltd. The authors wish to thank Falk Scholer, W. Bruce Croft and Douglas W. Oard, who provided valuable feedback.

REFERENCES

Abdulhamid, F. . Spex: A tool for visualising and navigating speech audio. Master's thesis, Victoria University of Wellington, 2013.

Abdulhamid, F. and Marshall, S. . Treemaps to visualise and navigate speech audio. In *Proceedings of the Australian Conference on HCI (OzCHI) '13*, pages 555–564, 2013.

Al-Maskari, A. and Sanderson, M. . A review of factors influencing user satisfaction in information retrieval. *Journal of the American Society for Information Science and Technology*, 61(5):859–868, 2010.

Alonso, O. and Mizzaro, S. . Can we get rid of trec assessors? using mechanical turk for relevance assessment. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, 2009.

Azzopardi, L. and de Rijke, M. . Automatic construction of known-item finding test beds. In *Proceedings of SIGIR'06*, pages 603–604, 2006.

Azzopardi, L. ; de Rijke, M. , and Balog, K. . Building simulated queries for known-item topics: An analysis using six european languages. In *Proceedings of SIGIR'07*, pages 455–462, 2007.

Balog, K. ; Azzopardi, L. ; Kamps, J. , and de Rijke, M. . Overview of webclef 2006. In *CLEF'07*, pages 803–819, 2007.

Besser, J. ; Larson, M. , and Hofmann, K. . Podcast search: user goals and retrieval technologies. *Online Information Review*, 34(3):395–419, 2010.

Blei, D. M. ; Ng, A. Y. , and Jordan, M. I. . Latent dirichlet allocation. *the Journal of Machine Learning Research*, 3: 993–1022, 2003.

Buchholz, S. ; Latorre, J. , and Yanagisawa, K. . Crowdsourced assessment of speech synthesis. *Crowdsourcing for Speech Processing*, pages 173–216, 2013.

Clarke, C. L. ; Agichtein, E. ; Dumais, S. , and White, R. W. . The influence of caption features on clickthrough patterns in web search. In *Proceedings of SIGIR'07*, pages 135–142, 2007.

Davtyan, M. ; Eickhoff, C. , and Hofmann, T. . Exploiting document content for efficient aggregation of crowdsourcing votes. In *Proceedings of CIKM'15*, 2015.

Garofolo, J. S. ; Voorhees, E. M. ; Auzanne, C. G. ; Stanford, V. M. , and Lund, B. A. . 1998 trec-7 spoken document retrieval track overview and results. In *Broadcast News Workshop*, volume 99, 1999.

Garofolo, J. S. ; Auzanne, C. G. , and Voorhees, E. M. . The trec spoken document retrieval track: A success story. In *in Text Retrieval Conference (TREC) 8*, 2000.

Goto, M. ; Ogata, J. , and Eto, K. . Podcastle: a web 2.0 approach to speech recognition research. In *Proceedings of INTERSPEECH'07*, 2007.

Heeren, W. and de Jong, F. . Disclosing spoken culture: user interfaces for access to spoken word archives. In *Proceedings of British Computer Society Conference on HCI*, pages 23–32, 2008.

Jayasinghe, G. K. ; Webber, W. ; Sanderson, M. ; Dharmasena, L. S. , and Culpepper, J. S. . Statistical comparisons of non-deterministic ir systems using two dimensional variance. *Information Processing & Management*, 51(5):677–694, 2015.

Jing, H. ; Lopresti, D. , and Shih, C. . Summarization of noisy documents: a pilot study. In *Proceedings of the HLT-NAACL'03 on Workshop on Text Summarization*, pages 25–32, 2003.

Kim, J. and Croft, W. B. . Retrieval experiments using pseudo-desktop collections. In *Proceedings of CIKM'09*, pages 1297–1306, 2009.

Larson, M. and Jones, G. J. . Spoken content retrieval: A survey of techniques and technologies. *Foundations and Trends in Information Retrieval*, 5(4–5):235–422, 2012.

Manning, C. D. ; Raghavan, P. , and Schütze, H. . *Introduction to Information Retrieval*, volume 1. Cambridge university press Cambridge, 2008.

McCallum, A. K. . Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

Mizuno, J. ; Ogata, J. , and Goto, M. . A similar content retrieval method for podcast episodes. In *Spoken Language*

Technology Workshop, pages 297–300, 2008.

Morbini, F. ; Audhkhasi, K. ; Sagae, K. ; Artstein, R. ; Can, D. ; Georgiou, P. ; Narayanan, S. ; Leuski, A. , and Traum, D. . Which asr should i choose for my dialogue system? In *Proceedings of the 14th SIGdial Meeting on Discourse and Dialogue*, pages 394–403. ACM, 2013.

Munro, R. ; Bethard, S. ; Kuperman, V. ; Lai, V. T. ; Melnick, R. ; Potts, C. ; Schnoebelen, T. , and Tily, H. . Crowdsourcing and language studies: The new generation of linguistic data. In *Proceedings of the NAACL-HLT'10 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 122–130, 2010.

Munteanu, C. ; Baecker, R. ; Penn, G. ; Toms, E. , and James, D. . The effect of speech recognition accuracy rates on the usefulness and usability of webcast archives. In *Proceedings of SIGCHI'04*, pages 493–502, 2006a.

Munteanu, C. ; Penn, G. ; Baecker, R. , and Zhang, Y. . Automatic speech recognition for webcasts: how good is good enough and what to do when it isn't. In *Proceedings of the 8th International Conference on Multimodal Interfaces*, pages 39–42, 2006b.

Naji, N. and Savoy, J. . Information retrieval strategies for digitized handwritten medieval documents. In *Proceedings of the Asian Information Retrieval Symposium (AIRS)*, pages 103–114, 2011.

Ogata, J. and Goto, M. . Podcastle: collaborative training of acoustic models on the basis of wisdom of crowds for podcast transcription. In *Proceedings of INTERSPEECH'09*, pages 1491–1494, 2009.

Ogata, J. and Goto, M. . Podcastle: Collaborative training of language models on the basis of wisdom of crowds. In *Proceedings of INTERSPEECH'12*, 2012.

Ordelman, R. ; Heeren, W. ; Huijbregts, M. ; de Jong, F. , and Hiemstra, D. . Towards affordable disclosure of spoken heritage archives. *Journal of Digital Information*, 10(6), 2009.

Porter, M. F. . An algorithm for suffix stripping. *Program*, 14 (3):130–137, 1980.

Ranjan, A. ; Balakrishnan, R. , and Chignell, M. . Searching in audio: the utility of transcripts, dichotic presentation, and time-compression. In *Proceedings of the SIGCHI'06*, pages 721–730, 2006.

Sabou, M. ; Bontcheva, K. , and Scharl, A. . Crowdsourcing research opportunities: Lessons from natural language processing. In *Proceedings of CIKM'12*, pages 1–8, 2012.

Sahib, N. G. ; Tombros, A. , and Stockman, T. . A comparative analysis of the information-seeking behavior of visually impaired and sighted searchers. *Journal of the American Society for Information Science and Technology*, 63(2):377–391, 2012.

Sakai, T. . Designing test collections that provide tight confidence intervals. In *Forum on Information Technology 2014*, volume 2, pages 15–18, 2014.

Salton, G. ; Wong, A. , and Yang, C.-S. . A vector space model for automatic indexing. *Communications of the ACM*, 18 (11):613–620, 1975.

Sanderson, M. . Accurate user directed summarization from existing tools. In *Proceedings of CIKM'98*, pages 45–51, 1998.

Sanderson, M. and Shou, X. . Search of spoken documents retrieves well recognized transcripts. In *Proceedings of ECIR'07*, pages 505–516, 2007.

Snow, R. ; O'Connor, B. ; Jurafsky, D. , and a.Y. Ng, . Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of EMNLP'08*, pages 254–263, 2008.

Stark, L. A. ; Whittaker, S. , and Hirschberg, J. . Asr satisficing: the effects of asr accuracy on speech retrieval. In *Proceedings of INTERSPEECH'00*, pages 1069–1072, 2000.

Tombros, A. and Sanderson, M. . Advantages of query biased summaries in information retrieval. In *Proceedings of SIGIR'98*, pages 2–10, 1998.

Tombros, T. and Crestani, F. . Users' perception of relevance of spoken documents. *Journal of the American Society for Information Science*, 51(10):929–939, 2000.

Trippas, J. ; Spina, D. ; Sanderson, M. , and Cavedon, L. . Towards understanding the impact of length in web search result summaries over a speech-only communication channel. In *Proceedings of SIGIR'15*, 2015.

Turpin, A. ; Scholer, F. ; Jarvelin, K. ; Wu, M. , and Culpepper, J. S. . Including summaries in system evaluation. In *Proceedings of SIGIR'09*, 2009.

Van Thong, J.-M. ; Moreno, P. J. ; Logan, B. ; Fidler, B. ; Maffey, K. , and Moores, M. . Speechbot: an experimental speech-based search engine for multimedia content on the web. *Multimedia, IEEE Transactions on*, 4(1):88–96, 2002.

APPENDIX
QUERY SET

TABLE 7: Full list of queries in the test collection.

aboriginal women died police custody
ABS accepts recommendations
adelaide fringe festival
air attack to islamic state
asylum seekers Immigration
australian arrested in bali to be executed
bangkok bombing
Black Saturday AusNet payout
child sex abuse sydney school
Children in immigration detention centers australia
Clive James poetry new book voice Japanese Maple
cuba america re-establishing diplomatic relations
Dog show Crufts controversy
ebola australia
financial dealing of west australian minister
floral tribute martin place
free trade china australia
Freedom Ride Perkins
Germanwings Andreas Lubitz
heart transplant of stopped heart
hospital bed shortage adelaide
Inflamatory deseases university of queensland treatment breakthrough
interest rate cuts
IS extremists Kobane US airstrikes
Israel Palastine tension worship at Al Asqa Jordan diplomat
Jim Byrnes trade union commision
Kim Jong-un 69th anniversary of the ruling Workers' Party
labour plan to build submarines in australia
malcolm fraser died
nanomaterials excluded in food standards
News Corp Nova Peris leaked email
northern territory chief minister dumped
pro-democracy protests Hong Kong
productivity commission's report childcare system
protest vote against gold company executive pay
Renewable Energy Target
royal commission Hutchins child abuse
royal commission Hutchins child abuse apology Anglican church
Royal commission Mangrove Mountain Yoga Ashram
Sex abuse New Sounth Wales ashram
Tony Abbott offending indigenous people
trade union royal commission report released Gillard CFMEU
US police shoot 12 year old boy toy gun
volcano new zealand