Overview of the NTCIR-11 SpokenQuery&Doc Task

Tomoyosi Akiba
Toyohashi University of Technology
1-1 Hibarigaoka,
Tohohashi-shi,
Aichi, 440-8580, Japan
akiba@cs.tut.ac.jp

Hiromitsu Nishizaki
University of Yamanashi
4-3-11 Takeda, Kofu,
Yamanashi, 400-8511, Japan
hnishi@yamanashi.ac.jp

Hiroaki Nanjo
Ryukoku University
Yokotani 1-5, Oe-cho
Seta, Otsu, Shiga, 520-2194,
Japan
nanjo@rins.ryukoku.ac.jp

Gareth J. F. Jones
Dublin City University
Glasnevin, Dublin 9, Ireland
gjones@computing.dcu.ie

ABSTRACT

This paper presents an overview of the Spoken Query and Spoken Document retrieval (SpokenQuery&Doc) task at the NTCIR-11 Workshop. This task included spoken query driven spoken content retrieval (SQ-SCR) as the main sub-task. With a spoken query driven spoken term detection task (SQ-STD) as an additional sub-task. The paper describes details of each sub-task, the data used, the creation of the speech recognition systems used to create the transcripts, the design of the retrieval test collections, the metrics used to evaluate the sub-tasks and a summary of the results of submissions by the task participants.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Experimentation, Performance

Keywords

NTCIR-11, spoken document retrieval, spoken queries, spoken content retrieval, spoken term detection

1. INTRODUCTION

The NTCIR-11 SpokenQuery&Doc task evaluated information retrieval systems for spoken content retrieval using spoken query input, i.e. speech-driven information retrieval and spoken document retrieval.

Spoken document retrieval (SDR) in the SpokenQuery&Doc task built on the previous NTCIR-9 SpokenDoc [1, 2] and NTCIR-10 SpokenDoc-2 [3] tasks, and evaluated two SDR tasks: spoken term detection (STD) and spoken content retrieval (SCR). Common search topics were used for the STD and SCR tasks which enabled component and whole system evaluations of STD and SCR.

Spoken Term Detection: Within spoken documents, find the occurrence positions of a queried term. STD was evaluated based on both efficiency (search time) and effectiveness (precision and recall).

Spoken Content Retrieval: In the SCR task, participants were asked to find spoken segments which included relevant information related to a search query, where a segment was either a pre-defined speech segment or a arbitrary length segment. This task was similar to an ad-hoc text retrieval task, except that the target documents are speech data.

The emergence of mobile computing devices means that it is increasingly desirable to interact with computing applications via speech input. The SpokenQuery&Doc task provided the first benchmark evaluation using spontaneously spoken queries instead of typed text queries. Here, a spontaneously spoken query means that the query is not carefully arranged before speaking, and is spoken in a natural spontaneous style. Query generated in this way tend to be longer than a typed text query. Note that this spontaneousness contrasts with spoken queries in the form of spoken isolated keywords which are carefully selected in advance, and represent very different situations in terms of speech processing and composition. One of the advantages of such spontaneously spoken queries as input to a retrieval system is that it enables users to easily submit long queries which give systems rich clues for retrieval, although their spontaneous nature means that they are harder to recognise reliably.

Our task design is illustrated in Figure 1. In this figure, the straight black arrow from the spoken query to the retrieved document (shown in the upper side) indicates our main goal called the spoken query driven spoken content retrieval (SQ-SCR) task. To achieve this task, particinants' systems were required, given audio wave data of spontaneously spoken query topic, to find corresponding relevant audio segments from within the audio wave date of target spoken documents. Automatic speech recognition (ASR) is often applied to obtain the textual representations of both the spoken query topic and the spoken documents in order to find matching between them. Baseline ASR results were also provided by the task organizers, so that ASR system development was not required for task participation.

One specific way of achieving this main task is illustrated in the lower side of the figure, indicated by the curved gray arrow. This consists of three sub-tasks; (0) finding meaningful spoken terms from the spontaneously spoken query topic, (1) detecting the occurrences of each spoken term in

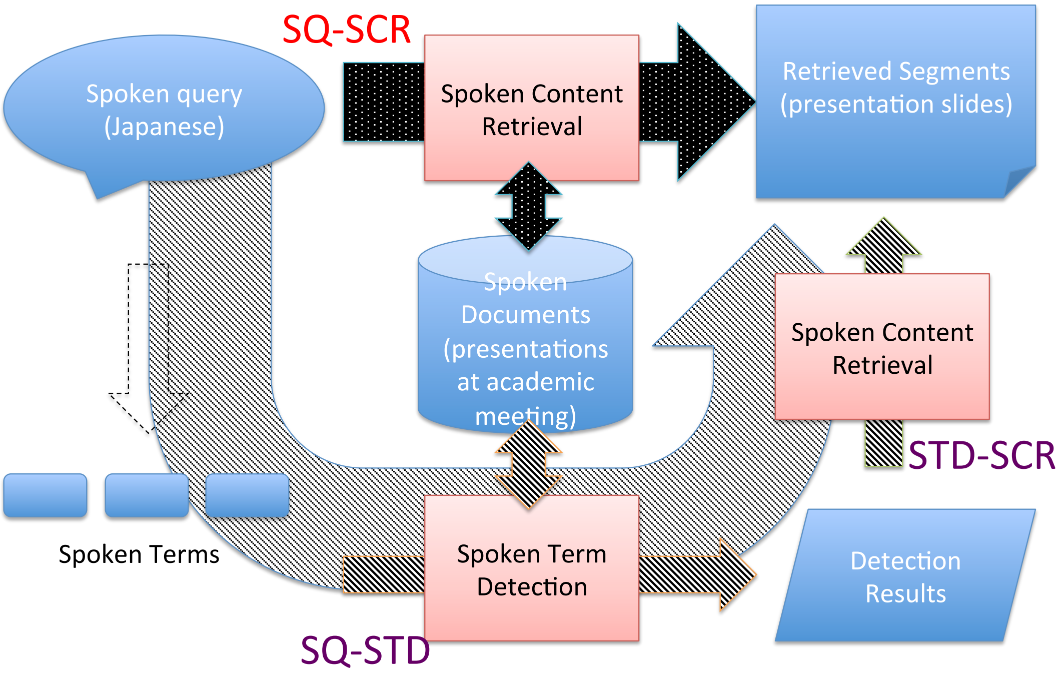

Figure 1: SpokenQuery&Doc task design.

the target spoken documents, and (2) deciding the relevancy of each segment in the spoken documents based on the detected query terms. Assuming that step (0) has been already achieved in some way and that the set of audio segments that represent the query terms are already in hand, steps (1) and (2), which are called spoken query driven spoken term detection (SQ-STD) task and STD results based spoken content retrieval (STD-SCR) task, respectively, were also evaluated in the SpokenQuery&Doc as the two components of total SQ-SCR system.

Our SQ-SCR tasks were defined not to find whole lecture units, but rather to find a shorter relevant speech segments within a complete lecture. For such a speech segment to be searched, we defined two kinds of units which resulted in two different SCR tasks.

The first unit type consists of arbitrary length segments from within the lecture. For these segments we assumes the situation where only the speech data is available. Participants were required to retrieve relevant speech passages. This sub-task continues the one evaluated in the NTCIR-9 SpokenDoc and NTCIR-10 SpokenDoc-2 tasks. Depending on the approach taken, this task may require the retrieval system also to perform topical segmentation of the lecture, and then to find relevant one from the segmented content. This passage retrieval task requires specifically designed evaluation metrics which are described later in the paper.

The other type of search unit investigated is called a slide group segment (SGS). These are naturally defined units based on the speech segment spoken during the display of one or more presentation slides that focus on a single consistent topic. The slide-group-segment (SGS) retrieval task required participants to search for relevant SGS units, and was evaluated using a standard mean average precision (MAP) metrics.

The SQ-STD task is almost same as that conducted in the previous NTCIR SpokenDoc task series, but is different in that spoken query terms are used instead of text query terms. The spoken query term used in the SQ-STD task is also a spontaneous terms that is extracted directly from the spontaneously spoken query topics used for the SQ-SCR task. This makes the SQ-STD task challenging in two ways; i.e. using spontaneous speech and using terms from the spoken information needs instead of artificially selected and balanced STD terms sets. The STD task using textual query terms was also evaluated as in the previous SpokenDoc tasks.

It was also planned to conduct the STD-SCR task as a sub-task in the NTCIR-11 SpokenQuery&Doc task. The task is almost same as the SQ-SCR task except that the search results of the query terms included in a search topic were to be used as search system's input instead of the query topic itself. The search results were provided as the submission for the SQ-STD task from the participants of the task. Unfortunately, there were no result submissions for the STD-SCR task, and we thus removed it from our evaluation of the task.

The rest of this paper is organized as follows. Sec.2 describes the design and our effort for constructing the SpokenQuery&Doc test collection. Sec.3 and Sec.4 describes the task design and the evaluation results of the SQ-SCR main task and the SQ-STD sub-task respectively.

2. TEST COLLECTION

In this section we describe the components of our test collection, including details of the document collection used for the evaluations, construction of the spontaneously spoken query set and transcription of the spoken content.

2.1 Document Collection

The Corpus of 1st to 7th Spoken Document Processing Workshop (SDPWS1to7) was used as the document collection for the NTCIR-11 SpokenQuery&Doc task. It was distributed to the participants by the SpokenQuery&Doc task organisers. It consists of the recordings of the first to seventh annual Spoken Document Processing Workshops with slide-change annotation.

Each lecture in the SDPWS1to7 is segmented using pauses that are no shorter than 200 msec. Each segment forms an Inter-Pausal Unit (IPU). An IPU is short enough to be used to indicate a position in the lecture. Therefore, IPUs are used as the basic unit to be searched in both the STD and SCR tasks.

Unlike "the corpus of Spoken Document Processing Workshop (**SDPWS**)" used in the previous NTCIR-10 SpokenDoc-2 task, **SDPWS1to7** includes an additional 10 lectures from the 7th workshop held in 2013. Furthermore, the time points when a lecture presenter transits her/his presentation slides forward are annotated in the SDPWS1to7. This enables us to divide a lecture into a sequence of speech segments each of which is aligned to a single presentation slide, referred to as *a slide segment.*

Generally, a slide segment can be considered to be a semantically consistent unit with a topic related to its corresponding presentation slide. Actually, most the single slides individually correspond to a semantic topic. However, sometimes a single topic is found to be covered by a series of slides for some technical reason. For example, one may use a series of slides to give an animation effect. In order to deal with such irregularities, we have grouped a series of contiguous slides into *a slide group*, which corresponds to a single presentation topic as a whole. Note that most slide groups in the collection consist of just a single slide, while the other (a few) groups consist of multiple slides. We refer to a speech segment aligned to a slide group as *a slide group segment.* In the SCR-related tasks conducted in the SpokenQuery&Doc, we regard a slide group segment as a search unit, i.e. a document, for retrieval. Therefore, the SCR task is defined as needing to find a set of slide-group-segments that are relevant to a given search topic.

2.1.1 Component Files

The component files of the document collection are grouped into two categories; those provided for each lecture and those provided for each IPU. The former are named using the lecture ID, while the latter are named using its IPU ID, which is the lecture ID followed by a sequential number (starting with 0) for each the IPU connected with a hyphen. Each file has its own extension.

We also refer to slide IDs, which are denoted within some of the files. A slide ID is a number series (starting with 1) of the presentation slides.

VAD file The voice activity detection (VAD) is first applied on an audio file in order to segment it into a sequence of IPUs. The VAD file records the result of the VAD applied on the audio data of the lecture. Its extension is `.seg`. It enables users to know the time stamp of any IPU from the beginning of the lecture.

Each line of a file, which corresponds to an IPU, has two integers formatted as follows:

<start time> <end time>

A unit of the numbers is 1/16000 second from the beginning of the lecture, i.e. 16000 means one second from the beginning.

Slide group file This describes slide groups of the lecture. Its extension is `.grp`. Each line of a file corresponds to a slide group, which is described as a sequence of contiguous slide IDs. Note that, in this file slide IDs are never omitted so that each slide ID appears exactly once in a file.

Time stamps of slide transitions This records the time stamp of the start of each presentation slide. Its extension is `.tmg`. Each line is formatted as follows:

<slide ID> [<minutes> ":"] <second>

The second column denotes the start time of a slide from the beginning of a lecture. Note that the first slide of each slide group must has a corresponding line, but the others are not always a line in this file, i.e. some inner slides in a slide group can be omitted.

Notice that, for most of the lectures in the collection time stamps are recorded at second-level granularity, so that they are not accurate enough to locate the exact position in its corresponding audio file. (This limitation arises from the use of off-the-shell software designed for recording of oral presentations, which was used in most of our recordings.)

Slide-to-IPU alignment file This describes alignments between the starting time of a slide and an IPU. Its extension is `.align`. Each line is formatted as follows:

<slide ID> <IPU ID> ["+"]

The lines without "+" at its end mean that the slide denoted by <slide ID> starts at the beginning of the IPU denoted by <IPU ID>, while those with "+" at its end mean that the slide starts somewhere within the IPU. This file provides an easy way to divide a transcript of a lecture into a set of documents.

Manual transcription file This contains a transcript of a lecture created by a human transcriber. Its extention is `.txt`. Each line is formatted as follows.

<IPU ID> ":" <text>

Several tags, which are explained in another document (the annotation manual), are introduced to describe nonverbal events in the text transcript. Among them, the `(s <slide ID>)` tag is used to indicate the position where the slide denoted by <slide ID> is shown for the first time in the lecture.

Reference automatic transcription The organizers prepared five automatic transcriptions. Three of them, whose file extension is "`_word.jout`", are word-based transcripts created using a large vocabulary continuous speech recognizer using a word-based trigram language model, while the other two, whose file extension is "`_syll.jout`", are subword-based transcripts created using a continuous syllable speech recognizer using a syllable-based trigram language model. The other differences are in their training data used for constructing their language models and the acoustic models.

The five automatic transcriptions are referred to with the following identifiers:

- REF-WORD-MATCH
- REF-SYLLABLE-MATCH

  Their file extension is `.unmatchLM_{word,syll}.jout`. The acoustic model and the language model are trained using the Corpus of Spontaneous Japanese. (the same as ”matched” transcriptions used in the NTCIR-10 SpokenDoc-2)
- REF-WORD-UNMATCH-LM
- REF-SYLLABLE-UNMATCH-LM

  Their file extension is `.unmatchLM_{word,syll}.jout`. The acoustic model is trained by using CSJ, while the language model is trained using newspaper articles. (the same as ”unmatched” transcriptions used in the NTCIR-10 SpokenDoc-2)
- REF-WORD-UNMATCH-AMLM

  New for NTCIR-11. Its file extension is `.unmatchAMLM_word.jout`. Both the acoustic model and the language model are trained in the ”unmatched” condition. These are those distributed as the Julius dictation kit v4.3.1 [1], whose acoustic and language models are trained using the ASJ Continuous Speech Corpus (JNAS) and Balanced Corpus of Contemporary Written Japanese (BCCWJ), respectively.

Audio file The audio files of lectures are stored in WAV format for each IPU. The file names are formatted as follows:

<Lecture ID>_<IPU ID>.wav

2.2 Query Construction

2.2.1 *Collecting Spontaneously Spoken Query Topics*

In order to construct spontaneously spoken query topics that were to be used for SQ-SCR task, subjective experiments were carried out. Before recording spoken query topics, subjects were asked to look over the proceedings of SDPWS1to7, to select papers they were interested in, and, for each paper, to invent a search topic based on its content described within a paragraph. The selected paragraph was preserved for use later in relevance judgment for topic.

In the recording session, subjects were asked to speak their search topics and their speech was recorded using a close microphone and an IC recorder. Throughout the session, they were not allowed to see their selected paper or any other written material. Therefore, we sought to make the subjects try to recall their search topic by themselves. There was no limitation in speaking time; they could even be silent for a while in order to recall what to say and in order to arrange how to say it. Finally, the session was closed when they felt that they had described their search topic as much as they wished to.

We employed 21 graduated students (1 female and 20 males) for the experiment. For each subject, two query topics were recorded through our experiment described above, which resulted in 42 topics. Five topics were selected for our dry-run evaluation. As our dry-run was conducted only for checking the evaluation procedure between the organizers and the participants, we did not conduct relevance judgments on these topics. The remaining 37 topics were used for our formal-run evaluation. The average, maximum, and minimum time duration of the query topics were 44.2, 139.4, 6.1 seconds, respectively. The average, maximum, and minimum word lengths were 97.4, 360, 17, respectively. The distribution of the lengths is shown in Figure 2.

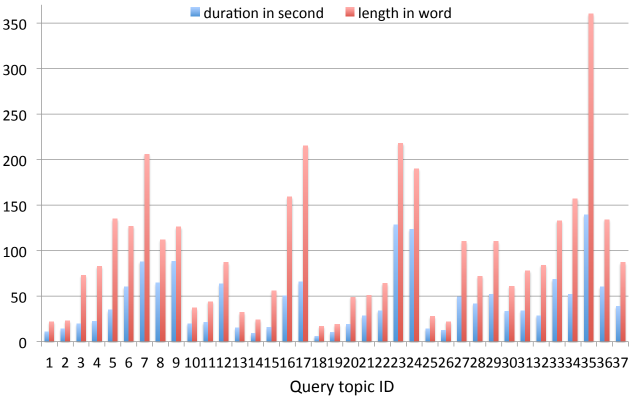

Figure 2: Distribution of the query topic length.

(F えーっと)(D と) 音声認識とかした<息>場合の *(F ま)(F えー)* 場合だとそのテキストが *(F ま)* 話し言葉そのままになるんですけどそれが<息>*(F ま)* 書き言葉の *(D ば)* ものとは *(F ま)(D か)* 書き言葉のものは *(F まー)(F ま)(A ウェブ;Web)* からとってきたりとか *(F ま)* 論文のものだったりとか<息>*(D と)* そういったものは *(F まー)*<息>書き言葉になるんですけどそれとはだいぶ<*H*>*(D かた)(D き)(D き)* 形式が違うというか<息>そのままではあまり一致しないということなので<息>*(F ま)* それを上手く分ける必要があると思うんですけど<息>*(D と)* その書き言葉と話し言葉を上手く分類というかそれを区別する方法<息>についての説明が知りたいです<息>*(F えー)(D と)(F まー)* どういった特徴量使っているとか *(F まー)(D ど)* どういった手法を使っているとかそういうことをですね

Figure 3: An example of a query topic (SpokenQueryDoc-SQSCR-formal-0016).

2.2.2 *Selecting Spontaneously Spoken Query Terms*

In the SpokenQuery&Doc SQ-STD task, we tried to avoid artificial selection of the query term to be detected by selecting them from the actual query topic expression for document retrieval, i.e. the spontaneously spoken queries, described in Sec.2.2.1. Firstly, the audio recordings of the spoken topics were manually transcribed into text. A Japanese morphological analyzer was then applied on the transcribed text, and the maximum contiguous sequences of noun words were extracted to form the candidates for the query terms. Finally, these were manually verified and, if necessary, their boundaries modified in order to make up the appropriate query terms. The selected query terms were annotated on

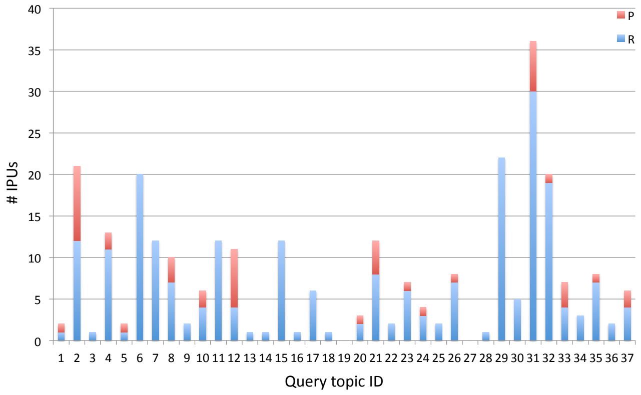

Figure 4: Distribution of relevant SGSs by query topic.

the manual transcription of the query topics.

For the "spoken" query terms in the SQ-STD task, the start and end times for each query term instance (token) were manually annotated (by using an audio editor) on the speech data of the query topic where it uttered. This enabled task participants to locate all the speech segments in the spoken query topics where the query term in question appears. It also enabled them to find the corresponding automatic transcripts of the term by means of the start and end time annotation provided with the query format file.

Through this process, we obtained 63 query terms (types) from the 5 dry-run query topics and 265 from the 37 formal-run topics, respectively.

2.3 Relevance Judgment

Relevance judgment for the SQ-SCR slide-group-segment task was performed against slide-group-segments (SGSs) in the document collection based on two clues: the selected paragraph in the paper used by the topic creator (a subject of the experiment described in Sec.2.2.1) to create the topic, and pooling of SGSs submitted by the task participant's systems. The judgment was performed not only on the SGSs specified in their submissions, but also on all the SGSs included in the same candidate lectures.

Five assessors were employed to carry out the judgments. They annotated three level relevancy, i.e. "R" (relevant), "P" (partially relevant), and "I" (irrelevant), on each SGS in their charge based on both its presentation slide and the manual transcription of its speech segment. The distribution of the relevancy on the formal-run query topics is shown in Figure 4.

Relevance judgment for the SQ-SCR passage retrieval task was performed based on the SGS relevance judgment results. For each SGS judged as either "R" or "P", the assessors tried to find the fine-grained localization of its relevant IPU sequence (arbitrary length passage). Sometimes a relevant SGS might lead to multiple passages, while at other times multiple SGSs might be combined into a single passage.

The relevance judgment for the SQ-STD task was automatically obtained by searching for a query term on the manual transcript of the document collection.

2.4 Transcription for Queries and Documents

Standard SCR methods first transcribe the audio signal into its textual representation by using Automatic Speech Recognition (ASR), followed by text-based retrieval. Additionally, a spoken query also can be transcribed into textual representation by using ASR. The participants can use the following three types of transcripts for both spoken queries and spoken documents.

1. Manual transcripts

   These are mainly used for evaluating the upper-bound performance.

2. Reference automatic transcripts

   The organizers provided five reference automatic transcripts for both spoken queries and spoken documents. These enables participants who are interested in SDR, but not in ASR to participate in our tasks. They also enable comparison of different IR methods based on the same underlying ASR performances. The participants can also use multiple transcripts at the same time to attempt to boost the performance.

   The textual representation are contained in an n-best list of the word or syllable sequence depending on the two background ASR systems, along with corresponding lattice and confusion network representation.

   (a) Word-based transcripts

   There were obtained by using a word-based ASR system. In other words, a word n-gram model was used for the language model of the ASR system. Along with the textual representation, it also provided the vocabulary list used by the ASR. This enabled us to determines the distinction between in-vocabulary (IV) query terms and out-of-vocabulary (OOV) query terms used in our STD subtask.

   (b) Syllable-based transcripts

   These were obtained by using a syllable-based ASR system. A syllable n-gram model was used for the language model, where the vocabulary is all the Japanese syllables. The use of these transcripts can avoid the OOV problem of spoken document retrieval with word-based transcripts. Participants who want to focus on open vocabulary STD and SCR can use this transcription.

3. Participant's own transcription

   The participants could also use their own ASR systems for the transcription. In order to enjoy the same IV and OOV condition, with their word-based ASR systems, they were recommended to use the same vocabulary list as our reference transcript, but this was not a necessary condition.

2.4.1 *Speech Recognition Models for Reference Automatic Transcriptions*

The acoustic models for the ASR system were triphone based, with 48 phonemes. The feature vectors had 38 dimensions: 12-dimensional Mel-frequency cepstrum coefficients (MFCCs); the cepstrum difference coefficients (delta MFCCs); their acceleration (delta delta MFCCs); delta power; and delta delta power. The components were calculated every 10 ms. The distribution of the acoustic features was

Table 1: Speech recognition preformances of reference automatic transcriptions on spoken query topics (%).

	word		syllable	
transcription	Corr.	Acc.	Corr.	Acc.
REF-WORD-MATCH	70.6	63.6	79.7	74.9
REF-SYLLABLE-MATCH	-	-	75.0	70.4
REF-WORD-UNMATCH-LM	50.8	43.9	67.5	59.2
REF-SYLLABLE-UNMATCH-LM	-	-	62.4	52.8
REF-WORD-UNMATCH-AMLM	46.7	42.3	63.5	58.8

Table 2: Speech recognition preformances of reference automatic transcriptions on spoken documents (%).

	word		syllable	
transcription	Corr.	Acc.	Corr.	Acc.
REF-WORD-MATCH	69.6	54.6	85.8	77.0
REF-SYLLABLE-MATCH	-	-	79.6	71.1
REF-WORD-UNMATCH-LM	54.1	41.5	78.6	70.5
REF-SYLLABLE-UNMATCH-LM	-	-	71.1	63.9
REF-WORD-UNMATCH-AMLM	43.5	35.4	69.5	65.8

modeled using 32 mixtures of diagonal covariance Gaussian for the HMMs.

The language models were either word-based or syllable-based trigram language models.

For both spoken queries and spoken documents, the organizers provided five reference automatic transcript with three training conditions on their acoustic and language models, . The three training conditions are referred to as “Match”, “UnmatchLM”, and “UnmatchAMLM”.

“Match” Models

The acoustic model was trained by using the 2,525 lectures (about 600 hours) in the Corpus of Spontaneous Japanese (CSJ). The language models were also trained by using the manual transcripts of the same lectures. They were either word-based trigram or syllable-based trigram, which resulted in word-based transcription and syllable-based transcription, respectively. The resulting two transcripts are referred to as “REF-WORD-MATCH” and “REF-SYLLABLE-MATCH”.

“UnmatchLM” Models

The acoustic model was trained by using the 2,525 lectures (about 600 hours) in the Corpus of Spontaneous Japanese (CSJ), the same as the “Match” models. The language models were trained by using 75 months of newspaper articles. They were either word-based trigram or syllable-based trigram, which result in word-based transcription and syllable-based transcription, respectively. The resulting two transcripts are referred to as “REF-WORD-UNMATCH-LM” and “REF-SYLLABLE-UNMATCH-LM”.

“UnmatchAMLM” Models

Both the acoustic model and the language model were trained in the ”unmatched” condition. These are distributed as the Julius dictation kit v4.3.1, whose acoustic and language models are trained using the ASJ Continuous Speech Corpus (JNAS) and Balanced Corpus of Contemporary Written Japanese (BCCWJ), respectively. Only the word-based transcript, which is referred to as “REF-WORD-UNMATCH-AMLM”, was provided.

Tables 1 and 2 summarize the speech recognition performance of these systems in terms of word correct rate, word accuracy, syllable correct rate, and syllable accuracy, on spoken queries and spoken documents respectively.

3. MAIN TASK: SQ-SCR TASK

3.1 Query

For the task data for our evaluation, the organizers provided two set of files. One was for spoken queries , while the other was for text queries . The query topic IDs are given in the names of these files so that the corresponding files are to be used for searching.

3.1.1 Files for spoken queries

Audio file The audio files of the spoken queries are stored in WAV format. The file names are formatted as follows:

<Query topic ID>.wav

VAD file This records the result of the voice activity detection applied on the audio data of the spoken queries. The file names are formatted as follows:

<Query topic ID>.seg

Each line of a file has two integers formatted as follows:

<start time> <end time>

A unit of the numbers is 1/16000 second from the beginning of the query, i.e. 16000 means one second from the beginning.

Note that all the automatic transcripts provided by the task organizers, described below, were obtained by applying ASR on the sequence of the speech segments derived by the VAD process.

Automatic transcription This stores an output of a automatic speech recognition of a spoken query. The file names are formatted as follows:

<Query topic ID>_<recognition condition>.jout

The organizers provided five kinds of recognition results by varying the recognition conditions for each spoken query. The conditions were same as those used to transcribe the target spoken documents as described at *Reference automatic transcriptions* in Section 2.4.

3.1.2 Files for text queries

Manual transcription The manually transcribed text for a spoken query is stored in this file. The file names are formatted as follows:

<Query topic ID>.txt

3.1.3 Query Topic List

A query topic list file summarizes the materials described above into a single XML document. It has a single root level tag "**<QUERY-TOPIC-LIST>**". Under the root tag, there are a sequence of tags "**<QUERY>**", each of which corresponds to a single query topic.

A "**<QUERY>**" has one attribute named "id", where its own query topic id is denoted as its value. Within a "**<QUERY>**" tag, three tags named "**<TXT>**", "**<SPK>**", and "**<STD>**" are specified.

- **<TXT>**

  This has one attribute "file" and its value is the file name of the manual transcript of the query topic.

- **<SPK>**

  This has one attribute "file" and its value is the file name of the audio file of the spoken query topic. Under this tag, a set of **"<TRANSCRIPTION>"** tags are described, each of which refers to an automatic transcription of the spoken query. The recognition condition is described in its "id", "vad", "unit", "acoustic-model", and "language-model" attributes. The "id" attribute denotes the identifier of the recognition condition that is same as that used to identify the condition of the target spoken documents. The "vod" attribute denotes the VAD files on which the ASR is applied. The "unit", "acoustic-model", and "language-model" attributes explain the details of the recognition conditions.

- **<STD>**

  This section is not to be used for the SQ-SCR task, but for the STD-SCR task. Within it, there listed the query terms appeared in the query topic. They are denoted as a set of "**<TERM>**" tags. A **<TERM>** tag has one attribute named "query-term-id", whose value denotes a corresponding query term id.

Figure 5 shows an example of a query topic list file.

3.2 Submission

Each participant was allowed to submit as many search results ("runs") as they wanted. Submitted runs should be prioritized by each group, because a specific number of runs with higher priority would be used for the pooling data for the manual relevance judgments. A priority number should be assigned for each submissions by a participant group, with smaller number having higher priority.

```
<QUERY-TOPIC-LIST>
  <QUERY id="SpokenQD-SQSCR-dry-0001">
    <TXT file="SpokenQD-SQSCR-dry-0001.txt" />
    <SPK file="SpokenQD-SQSCR-dry-0001.wav">
      <TRANSCRIPTION id="REF-WORD-MATCH"
        file="SpokenQD-SQSCR-dry-001_match_word.jout"
        vad="SpokenQD-SQSCR-dry-001.seg"
        unit="word"
        acoustic-model="match"
        language-model="match" />
      ...
    </SPK>
    <STD>
      <TERM query-term-id="SpokenQD-SQSTD-dry-0007" />
      <TERM query-term-id="SpokenQD-SQSTD-dry-0009" />
      ...
    </STD>
  </QUERY>
  <QUERY id="SpokenQD-SQSCR-dry-0002">
  ...
  </QUERY>
...
</QUERY-TOPIC-LIST>
```

Figure 5: An example of a query topic list file.

3.2.1 File Name

A single run is saved in a single file. Each submission file should have an adequate file name following the next format.
SQSCR-*X*-*T*-*I*-*N*.txt

X: System identifier, should be the same as the group ID (e.g., NTC)

T: Target task.

- SGS: Slide-Group-Segment retrieval task.
- PAS: Passage retrieval task.

I: Input modality.

- SPK: Spoken Query.
- TXT: Text Query.

If a run specifies SPK in this field, it is allowed to use only the query files for spoken queries (Sec.3.1.1) but not the files for text queries (Sec.3.1.2.

N: Priority of run (1, 2, 3, ...) for each target document set.

Suppose the group "NTC" submited two files and one file for the slide-group-segment retrieval task by using spoken queries and text queries, respectively, and three files for the passage retrieval task by using text queries. Then, the names of the run files should be "SQSCR-NTC-SGS-SPK-1.txt", "SQSCR-NTC-SGS-SPK-2.txt", "SQSCR-NTC-SGS-TXT-1.txt", "SQSCR-NTC-PAS-TXT-1.txt", and "SQSCR-NTC-PAS-TXT-2.txt".

3.2.2 Submission Format

The submission files are organized with the following tags. Each file must be a well-formed XML document. It has a single root level tag "**<ROOT>**". Under the root tag, it has three main sections, "**<RUN>**", "**<SYSTEM>**", and "**<RESULT>**".

- <**RUN**>

  <**SUBTASK**> “SQ-SCR”, “SQ-STD” or “STD-SCR”. For a SQ-SCR subtask submission, just say “SQ-SCR”.

  <**SYSTEM-ID**> System identifier that is the same as the group ID.

  <**PRIORITY**> Priority of the run.

  <**UNIT**> The retrieval unit to be retrieved. “SLIDE-GROUP” if the unit is a slide group as in the slide-group-segment retrieval task. “PASSAGE” if the unit is a passage as in the passage retrieval.

  <**TRANSCRIPTION**> The transcription used as the text representation of the target document set. “MANUAL” if it is the manual transcription. “REF-WORD-MATCH”, “REF-WORD-UNMATCH-LM”, “REF-WORD-UNMATCH-AMLM”, “REF-SYLLABLE-MATCH”, or “REF-SYLLABLE-UNMATCH-LM”, if it is one of the reference automatic transcription provided from the task organizers. “OWN” if it is obtained by a participant’s own recognition. “NO” if no textual transcription is used. If multiple transcriptions are used, specify all of them by concatenating with the “,” separator.

  <**QUERY-TRANSCRIPTION**> The transcription used as the text representation of the spoken queries. “MANUAL” if text queries are used instead of spoken queries. “REF-*” (“*” should be replaced by a transcription Identifier) if one of the reference transcription provided from the task organizers is used. “NO” if no textual transcription is used. If multiple transcriptions are used, specify all of them by concatenating with the “,” separator.

- <**SYSTEM**>

  <**OFFLINE-MACHINE-SPEC**>

  <**OFFLINE-TIME**>

  <**INDEX-SIZE**>

  <**ONLINE-MACHINE-SPEC**>

  <**ONLINE-TIME**>

  <**SYSTEM-DESCRIPTION**>

- <**RESULT**>

  <**QUERY**> Each query topic has a single “QUERY” tag with an attribute “id” specified in query topic files (Section 3.1). Within this tag, a list of the following “CANDIDATE” tags is described.

  <**CANDIDATE**> Each potential candidate of a retrieval result has a single “CANDIDATE” tag with the following attributes. The CANDIDATE tags should, but do not necessary to, be sorted in descending order of likelihood.

  rank The rank in the result list. “1” for the most likely candidate, incleased one at a time. Required to be totally ordered in a single “QUERY” tag.

  lecture The lecture ID specified in the SDPWS1to7.

```
<ROOT>
  <RUN>
    <SUBTASK>SQ-SCR</SUBTASK>
    <SYSTEM-ID>TUT</SYSTEM-ID>
    <UNIT>SLIDE-GROUP</UNIT>
    <PRIORITY>1</PRIORITY>
    <TRANSCRIPTION>REF-WORD-UNMATCHED,
      REF-SYLLABLE-UNMATCHED</TRANSCRIPTION>
    <QUERY-TRANSCRIPTION>REF-SYLLABLE-UNMATCHED
      </QUERY-TRANSCRIPTION>
  </RUN>
  <SYSTEM>
    <OFFLINE-MACHINE-SPEC>Xeon 3GHz dual CPU, 4GB mem.
    </OFFLINE-MACHINE-SPEC>
    <OFFLINE-TIME>18:35:23</OFFLINE-TIME>
    ...
  </SYSTEM>
  <RESULT>
    <QUERY id="SpokenQueryDoc0-dry-001">
      <CANDIDATE rank="1" lecture="10-09" slide="8" />
      <CANDIDATE rank="2" lecture="12-12" slide="3" />
      ...
    </QUERY>
    <QUERY id="SpokenQueryDoc0-dry-002">
      ...
    </QUERY>
  </RESULT>
</ROOT>
```

Figure 6: An example of a submission file.

  slide *Used for the slide-group-segment retrieval task.* The first slide ID in a slide group (i.e., a document) that is retrieved as a candidate. If the slide ID that is not first, i.e. second or later, in a slide group is specified, its CANDIDATE tag is always marked wrong in evaluation.

  ipu-from *Used for the passage retrieval task.* The Inter Pausal Unit ID, specified in the CSJ, of the first IPU of the retrieved passage (an IPU sequence).

  ipu-to *Used for the passage retrieval task.* The Inter Pausal Unit ID, specified in the CSJ, of the last IPU of the retrieved passage (an IPU sequence).
  NOTE: The IPU sequences specified in a single “QUERY” tag are required to be *exclusive* each other; i.e. no two intervals in a “QUERY”, each of which is specified by “CANDIDATE” tag, are not allowed to have a common IPU.

Figure 6 shows an example of a submission file.

3.3 Evaluation Measures

3.3.1 Slide-Group-Segment Retrieval

Mean Average Precision (MAP) was used as the official evaluation measure for lecture retrieval For each query topic, the top 1000 documents were evaluated.

Given a question q, suppose the ordered list of documents $d_1 d_2 \cdots d_{|D|} \in D_q$ was submitted as the retrieval result.

Then, $AveP_q$ is calculated as follows.

$$AveP_q = \frac{1}{|R_q|}\sum_{i=1}^{|D_q|} include(d_i, R_q)\frac{\sum_{j=1}^{i} include(d_j, R_q)}{i} \tag{1}$$

where

$$include(a, A) = \begin{cases} 1 & \cdots \quad a \in A \\ 0 & \cdots \quad a \notin A \end{cases} \tag{2}$$

Alternatively, given the ordered list of correctly retrieved documents $r_1 r_2 \cdots r_M (M \leq |R_q|)$, $AveP_q$ is calculated as follows.

$$AveP_q = \frac{1}{|R_q|}\sum_{k=1}^{M}\frac{k}{rank(r_k)} \tag{3}$$

where $rank(r)$ is the rank that the document r is retrieved.

MAP is the mean of the $AveP$ over all query topics Q.

$$\mathbf{MAP} = \frac{1}{|Q|}\sum_{q \in Q} AveP_q \tag{4}$$

3.3.2 Passage Retrieval

In our passage retrieval task, the relevancy of each arbitrary length segment (passage) rather than each whole lecture (document) must be evaluated. Three measures are designed for the task; the one is utterance-based and the other two are passage-based. For each query topic, top 1000 passages are evaluated by these measures.

uMAP

By expanding a passage into a set of utterances (IPUs) and by using an utterance (IPU) as a unit of evaluation like a document, we can use any conventional measures used for evaluating document retrieval.

Suppose the ordered list of passages $P_q = p_1 p_2 \cdots p_{|P_q|}$ is submitted as the retrieval result for a given query q. Suppose we have a mapping function $O(p)$ from a (retrieved) passage p to an ordered list of utterances $u_{p,1} u_{p,2} \cdots u_{p,|p|}$, we can get the ordered list of utterances $U = u_{p_1,1} u_{p_1,2} \cdots u_{p_1,|p_1|} u_{p_2,1} \cdots u_{p_{|P_q|},1} \cdots u_{p_{|P_q|},|p_{|P_q|}|}$. Then $uAveP_q$ is calculated as follows.

$$uAveP_q = \frac{1}{|\tilde{R}_q|}\sum_{i=1}^{|U|} include(u_i, \tilde{R}_q)\frac{\sum_{j=1}^{i} include(u_j, \tilde{R}_q)}{i} \tag{5}$$

where $U = u_1 \cdots u_{|U|} (|U| = \sum_{p \in P} |p|)$ is the renumbered ordered list of U and $\tilde{R}_q = \bigcup_{r \in R_q}\{u | u \in r\}$ is the set of relevant utterances extracted from the set of relevant passages R_q.

For the mapping function $O(p)$, we use the oracle ordering mapping function, which orders the utterances in the given passage p as the relevant utterances come first. For example, given a passage $p = u_1 u_2 u_3 u_4 u_5$ and suppose the relevant utterances are $u_3 u_4$, it returns as $u_3 u_4 u_1 u_2 u_5$.

uMAP (utterance-based MAP) is defined as the mean of the $uAveP$ over all query topics Q.

$$\mathbf{uMAP} = \frac{1}{|Q|}\sum_{q \in Q} uAveP_q \tag{6}$$

pwMAP

For a given query, a system returns an ordered list of passages. For each returned passage, only utterances located in the center of it are considered for relevancy. If the center utterance is included in some relevant passage described in the golden file, basically the returned passage is deemed relevant with respect to the relevant passage and the relevant passage is considered to be retrieved correctly. However, if there exists at least one formerly listed passage that is also deemed relevant with respect to the same relevant passage, the returned passage is deemed not relevant as the relevant passage has been retrieved already. In this way, all the passages in the returned list are labeled by their relevancy. Now, any conventional evaluation metric designed for document retrieval can be applied to the returned list.

Suppose we have the ordered list of correctly retrieved passages $r_1 r_2 \cdots r_M (M \leq |R_q|)$, where their relevancy are judged according to the process mentioned above. $pwAveP_q$ is calculated as follows.

$$pwAveP_q = \frac{1}{|R_q|}\sum_{k=1}^{M}\frac{k}{rank(r_k)} \tag{7}$$

where $rank(r)$ is the rank that the passage r is placed at in the original ordered list of retrieved passages.

pwMAP (pointwise MAP) is defined as the mean of the $pwAveP$ over all query topics Q.

$$\mathbf{pwMAP} = \frac{1}{|Q|}\sum_{q \in Q} pwAveP_q \tag{8}$$

fMAP

This measure evaluates relevancy of a retrieved passage fractionally against the relevant passage in the golden files. Given a retrieved passage $p \in P_q$ for a given query q, its relevance level $rel(p, R_q)$ is defined as the fraction that it covers some relevant passage(s), as follows.

$$rel(p, R_q) = \max_{r \in R_q}\frac{|r \cap p|}{|r|} \tag{9}$$

or

$$rel(p, R_q) = \sum_{r \in R_q}\frac{|r \cap p|}{|r|} \tag{10}$$

Here r and p are regarded as sets of utterances. rel can be seen as measuring the recall of p in utterance level. Accordingly, we can define the precision of p as follows.

$$prec(p, R_q) = \max_{r \in R_q}\frac{|p \cap r|}{|p|} \tag{11}$$

or

$$prec(p, R_q) = \sum_{r \in R_q}\frac{|p \cap r|}{|p|} \tag{12}$$

Then, $fAveP_q$ is calculated as follows.

$$fAveP_q = \frac{1}{|R_q|}\sum_{i=1}^{|P_q|} rel(p_i, R_q)\frac{\sum_{j=1}^{i} prec(p_j, R_q)}{i} \tag{13}$$

fMAP (fractional MAP) is defined as the mean of the $fAveP_q$ over all query topics Q.

$$\mathbf{fMAP} = \frac{1}{|Q|}\sum_{q \in Q} fAveP_q \tag{14}$$

Table 3: SQ-SCR task participants.

Group ID	Group Name, Organization	SGS-SPK	SGS-TXT	PAS-SPK	PAS-TXT
AKBL	Akiba Laboratory, Toyohashi University of Technology	3	7		
CNGL	CNGL, CNGL Center for Global Intelligent Content	24	12		
HYM14	Laboratorie de professeur Chat Noir Gifu University	4			
R531	LabR531, National Taiwan University	4			
RYSDT	RYukoku univ. Spoken Document processing Team, Ryukoku University	8	8	8	8

3.4 Result

Five groups with a total 86 runs submitted their results for the formal-run of the SQ-SCR task. All five groups participated in the slide-group-segment (SGS) task and only one group did in the passage (PAS) task. The group ID and their submitted runs are listed in Table 3. From these submissions, up to nine runs for each combination of the task (SGS or PAS) and transcription type (SPK or TXT) are investigated in this paper because of space limitations.

3.4.1 Baseline

Our baseline runs were implemented by applying conventional methods for IR on the REF-WORD-MATCH transcript. Only verbs, which were transformed into their basic form, and nouns were used for indexing, which were extracted from the transcription by applying the Japanese morphological analysis tool. The retrieval model was a vector space model and the term weighting was TF-IDF with pivoted normalization [5]. From the textual query topics, verbs and nouns were also extracted by applying the same morphological analyzer. For the task using spoken query (SPK), spoken query topic was also transcribed into REF-WORD-MATCH and used as its textual expression.

For the slide-group-segment (SGS) retrieval task, each slide-group-segment was indexed and retrieved. Their run IDs are BASE-SGS-SPK-1 for spoken query topics and BASE-SGS-TXT-1 for text query topics. For the passage retrieval task, we created pseudo-passages by automatically dividing each lecture into a sequence of segments, with N utterances per segment. We set $N = 10$. Their run IDs are BASE-PAS-SPK-1 for spoken query topics and BASE-PAS-TXT-1 for text query topics.

3.4.2 Evaluation Results

Table 4 and Table 5 show the run-by-run evaluation results of the slide-group-segment retrieval task and the passage retrieval task, respectively, where the runs are grouped by their used query transcription and document transcription.

4. SUB-TASK: SQ-STD TASK

4.1 Query

The query terms used for the SQ-STD task are put together into a single file written in an XML format, called query term list. This has a single root level tag "<**QUERY-TERM-LIST**>". Under the root tag, there are a sequence of tag "<**QUERY**>", each of which corresponds to a single query term.

A "<**QUERY**>" tag has one attribute named "id", where its own query term id is denoted as its value. Under the "<**QUERY**>" tag, it has two sections specified by the two tags named "<**SPK**>" and "<**TXT**>".

- **<TXT>**

  This is used to describe the materials used for the STD task from text queries. It has two attributes "text" and "yomi". The value of the "text" tag is the manually transcribed text of the query term, while that of the "yomi" tag is the Japanese pronunciation of the query term written in a Japanese KATAKANA sequence.

  Notice that, for the judgment of the term's occurrence in the golden file, "text" is searched against the manual transcriptions, while the "yomi" is never considered for the judgment. Furthermore, the organizers do **not** assure the participants of the correctness of what described in the "yomi" fields, so the participants should take the responsible for using it. Nevertheless, the organizers believes it should help participants to predict the term's pronunciation.

- **<SPK>**

  Under this tag, the materials used for the STD task from spoken queries are described. They consist of a set of "<**SEGMENT**>" tags.

  A "<**SEGMENT**>" specifies a speech segment where a query term is uttered in a spoken query topic. It has three attributes, "query-topic-id", "time-from", and "time-to". A value of a "query-topic-id" attribute is one of the query topic IDs provided from the task organizers. A pair of the attributes "time-from" and "time-to" denotes the time interval that the quey term in question is uttered in the query topic specified by the "query-topic-id". Their values are real numbers denoted in second from the begining of the WAV format file of the spoken query topics.

  Some query term may have several "<**SEGMENT**>" tags, just because it appears several times spread over the query topics. Participants can make use of these segments all together for searching it.

Figure 7 shows an example of a query term list file.

```
<QUERY-TERM-LIST>
  <QUERY id="SpokenQD-SQSTD-dry-0001">
    <TXT text="国立国語研究所"
      yomi="コクリツコクゴケンキュージョ" />
    <SPK>
      <SEGMENT query-topic-id="SpokenQD-SQSCR-dry-0005"
        time-from="3.043042409820067"
        time-to="3.8765430959093545"/>
      <SEGMENT query-topic-id="SpokenQD-SQSCR-dry-0019"
        time-from="29.46664086551418"
        time-to="30.01257631426801"/>
      ...
    </SPK>
  </QUERY>
  <QUERY id="SpokenQD-SQSTD-dry-0002">
  ...
  </QUERY>
...
</QUERY-TERM-LIST>
```

Figure 7: An example of a qyery term list file.

4.2 Submission

Each participant is allowed to submit as many search results ("runs") as they want. Submitted runs should be prioritized by each group. Priority number should be assigned through all submissions of a participant, and smaller number has higher priority.

4.2.1 File Name

A single run is saved in a single file. Each submission file should have an adequate file name following the next format.

SQSTD-*X*-*T*-*I*-*N*.txt

X: System identifier that is the same as the group ID (e.g., NTC)

T: Target task.

- IPU: IPU retrieval task.

For SQ-STD task submission, just say "IPU".

I: Input modality.

- SPK: Spoken Query.
- TXT: Text Query.

N: Priority of run (1, 2, 3, ...) for each target docuemnt set.

For example, if the group "NTC" submits two files and three files by using spoken queries and text queries, respectively, then the names of the run files should be "SQSTD-NTC-IPU-SPK-1.txt", "SQSTD-NTC-IPU-SPK-2.txt", "SQSTD-NTC-IPU-TXT-1.txt", "SQSTD-NTC-IPU-TXT-2.txt", and "SQSTD-NTC-IPU-TXT-3.txt".

4.2.2 Submission Format

The submission files are organized with the following tags. Each file must be a well-formed XML document. It has a single root level tag "<**ROOT**>". It has three main sections, "<**RUN**>", "<**SYSTEM**>", and "<**RESULT**>".

- <**RUN**>

  <**SUBTASK**> "SQ-STD" or "SQ-STD". For a SQ-STD subtask submission, just say "SQ-STD".

  <**SYSTEM-ID**> System identifier that is the same as the group ID.

  <**PRIORITY**> Priority of the run.

  <**TRANSCRIPTION**> The transcription used as the text representation of the target document set. "MANUAL" if it is the manual transcription. "REF-WORD-MATCH", "REF-WORD-UNMATCH-LM", "REF-WORD-UNMATCH-AMLM", "REF-SYLLABLE-MATCH", or "REF-SYLLABLE-UNMATCH-LM", if it is one of the reference automatic transcription provided from the task organizers. "OWN" if it is obtained by a participant's own recognition. "NO" if no textual transcription is used. If multiple transcriptions are used, specify all of them by concatenating with the "," separator.

  <**QUERY-TRANSCRIPTION**> The transcription used as the text representation of the spoken queries. "MANUAL" if text queries are used instead of spoken queries. "REF-*" ("*" should be replaced by a transcription Identifier) if one of the reference transcription provided from the task organizers is used. "NO" if no textual transcription is used. If multiple transcriptions are used, specify all of them by concatenating with the "," separator.

- <**SYSTEM**>

  <**OFFLINE-MACHINE-SPEC**>

  <**OFFLINE-TIME**>

  <**INDEX-SIZE**>

  <**ONLINE-MACHINE-SPEC**>

  <**ONLINE-TIME**>

  <**SYSTEM-DESCRIPTION**>

- <**RESULT**>

  <**QUERY-TERM**> Each query term has a single "QUERY" tag with an attribute "id" specified in a query term list (Section 4.1). Within this tag, a list of the following "TERM" tags is described.

  <**TERM**> Each potential detection of a query term has a single "TERM" tag with the following attributes.

  lecture The searched lecture ID.

  ipu The searched Inter Pausal Unit ID.

  score The detection score indicating the likelihood of the detection. The greater is more likely.

  detection The binary ("YES" or "NO") decision of whether or not the term should be detected to make the optimal evaluation result.

Figure 8 shows an example of a submission file.

```
<ROOT>
  <RUN>
    <SUBTASK>SQ-STD</SUBTASK>
    <SYSTEM-ID>TUT</SYSTEM-ID>
    <PRIORITY>1</PRIORITY>
    <TRANSCRIPTION>REF-WORD-UNMATCHED,
      REF-SYLLABLE-UNMATCHED</TRANSCRIPTION>
    <QUERY-TRANSCRIPTION>REF-SYLLABLE-UNMATCHED
      </QUERY-TRANSCRIPTION>
  </RUN>
  <SYSTEM>
    <OFFLINE-MACHINE-SPEC>Xeon 3GHz dual CPU, 4GB mem.
    </OFFLINE-MACHINE-SPEC>
    <OFFLINE-TIME>18:35:23</OFFLINE-TIME>
    ...
  </SYSTEM>
  <RESULT>
    <QUERY id="SpokenQD0-dry-001">
      <TERM lecture="10-12" ipu="0024" score="0.83"
        detection="YES" />
      <TERM lecture="08-05" ipu="0079" score="0.32"
        detection="NO" />
      ...
    </QUERY>
    <QUERY id="SpokenQD0-dry-002">
      ...
    </QUERY>
    ...
  </RESULT>
</ROOT>
```

Figure 8: An example of a submission file.

4.3 Evaluation Measures

The official evaluation measure for effectiveness is F-measure at the decision point specified by the participant, based on recall and precision averaged over queries. F-measure at the maximum decision point, Recall-Precision curves and mean average precision (MAP) will also be used for analysis purpose.

Mean average precision for the set of queries is the mean value of the average precision values for each query. It can be calculate as follows,

$$MAP = \frac{1}{Q}\sum_{i=1}^{Q} AveP(i) \tag{15}$$

where Q is the number of queries and $AveP(i)$ means the average precision of the i-th query of the query set. The average precision is calculated by averaging of the precision values computed at the point of each of the relevant terms in the list in which retrieved terms are ranked by a relevance measure.

$$AveP(i) = \frac{1}{Rel_i}\sum_{r=1}^{N_i}(\delta_r \cdot Precision_i(r)) \tag{16}$$

where r is the rank, N_i is the rank number at which the all relevance terms of query i are found, and Rel_i is the number of the relevance terms of query i. δ_r is a binary function on the relevance of a given rank r.

4.4 Results

Nine groups with a total 56 runs have submitted their results for the formal-run of the SQ-STD task. All nine groups submitted runs using text query terms, while two groups submitted runs using spoken query terms. The group ID and their submitted runs are listed in Table 6.

4.4.1 Baseline

Five baseline runs for each type (SPK or TXT) of query terms, which resulted in 10 runs in total, were also submitted from the task organizers. These runs are commonly built on the search method that tries to find matchings between the phonetic representation of a query term and target documents in terms of edit distance by using continuous DP matching. The differences among the five runs are only in the transcript used to obtain the phonetic representation. Specifically, either REF-WORD-MATCH, REF-SYLLABLE-MATCH, REF-WORD-UNMATCH-LM, REF-SYLLABLE-UNMATCH-LM, or REF-WORD-UNMATCH-AMLM is used for the priority number 1, 2, 3, 4, or 5 of the baseline run, respectively. The runs using spoken query terms use the REF-WORD-MATCH transcription for phonetic representation for query terms.

4.4.2 Evaluation Results

We found 33 query terms in the formal run query term set did not appear in the target documents at all. We also found 29 query terms appeared more than 500 times in the documents. We excluded these terms and the rest 203 terms were used for our evaluation. Table 7 and Table 8 show the run-by-run evaluation results of the SQ-STD task using spoken query terms and textual query terms, respectively, where the runs are grouped by their used query transcription and document transcription.

5. CONCLUSION

This paper introduced the overview of the Spoken Query and Spoken Document Retrieval Task (SpokenQuery&Doc) in NTCIR-11 Workshop.

6. REFERENCES

[1] T. Akiba et al. Overview of the IR for spoken documents task in NTCIR-9 workshop. In *Proceedings of the Ninth NTCIR Workshop Meeting*, pages 223–235, 2011.

[2] T. Akiba et al. Designing an evaluation framework for spoken term detection and spoken document retrieval at the NTCIR-9 SpokenDoc task. In *Proceedings of International Conference on Language Resources and Evaluation*, 2012.

[3] T. Akiba et al. Overview of the NTCIR-10 SpokenDoc-2 task. In *Proceedings of the 10th NTCIR Conference*, pages 573–587, 2013.

[4] H. Joho and K. Kishida. Overview of ntcir-11. In *Proceedings of the 11th NTCIR Conference*, Tokyo, Japan, 2014.

[5] A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In *Proceedings of ACM SIGIR*, pages 21–29, 1996.

Table 4: SQ-SCR slide-group-segment retrieval task result (%).

run ID	Query Transcription	Document Transcription	MAP
BASE-SGS-SPK-1	REF-WORD-MATCH	REF-WORD-MATCH	13.1
AKBL-SGS-SPK-1	REF-WORD-MATCH	REF-WORD-MATCH	12.1
CNGL-SGS-SPK-1	REF-WORD-MATCH	REF-WORD-MATCH	6.1
CNGL-SGS-SPK-2	REF-WORD-MATCH	REF-WORD-MATCH	6.3
CNGL-SGS-SPK-3	REF-WORD-MATCH	REF-WORD-MATCH	8.1
CNGL-SGS-SPK-7	REF-WORD-MATCH	REF-WORD-MATCH	6.5
HYM14-SGS-SPK-1	REF-WORD-MATCH	REF-WORD-MATCH	17.2
HYM14-SGS-SPK-2	REF-WORD-MATCH	REF-WORD-MATCH	12.9
HYM14-SGS-SPK-3	REF-WORD-MATCH	REF-WORD-MATCH	12.5
HYM14-SGS-SPK-4	REF-WORD-MATCH	REF-WORD-MATCH	6.0
R531-SGS-SPK-1	REF-WORD-MATCH	REF-WORD-MATCH	4.3
R531-SGS-SPK-2	REF-WORD-MATCH	REF-WORD-MATCH	15.4
R531-SGS-SPK-3	REF-WORD-MATCH	REF-WORD-MATCH	11.9
R531-SGS-SPK-4	REF-WORD-MATCH	REF-WORD-MATCH	12.6
RYSDT-SGS-SPK-1	REF-WORD-MATCH	REF-WORD-MATCH	19.4
RYSDT-SGS-SPK-2	REF-WORD-MATCH	REF-WORD-MATCH	18.8
RYSDT-SGS-SPK-3	REF-WORD-MATCH	REF-WORD-MATCH	18.8
RYSDT-SGS-SPK-4	REF-WORD-MATCH	REF-WORD-MATCH	21.8
RYSDT-SGS-SPK-5	REF-WORD-MATCH	REF-WORD-MATCH	20.7
RYSDT-SGS-SPK-6	REF-WORD-MATCH	REF-WORD-MATCH	21.1
RYSDT-SGS-SPK-7	REF-WORD-MATCH	REF-WORD-MATCH	13.5
RYSDT-SGS-SPK-8	REF-WORD-MATCH	REF-WORD-MATCH	14.3
CNGL-SGS-SPK-8	REF-WORD-MATCH	REF-WORD-UNMATCH-AMLM	1.3
CNGL-SGS-SPK-9	REF-WORD-MATCH	REF-WORD-UNMATCH-AMLM	1.2
CNGL-SGS-SPK-4	REF-WORD-MATCH	MANUAL	9.1
CNGL-SGS-SPK-5	REF-WORD-MATCH	MANUAL	7.2
CNGL-SGS-SPK-6	REF-WORD-MATCH	MANUAL	8.8
AKBL-SGS-SPK-2	REF-WORD-UNMATCH-LM	REF-WORD-UNMATCH-LM	5.2
AKBL-SGS-SPK-3	REF-WORD-UNMATCH-AMLM	REF-WORD-UNMATCH-AMLM	4.6
BASE-SGS-TXT-1	MANUAL	REF-WORD-MATCH	15.9
AKBL-SGS-TXT-1	MANUAL	REF-WORD-MATCH	15.2
AKBL-SGS-TXT-4	MANUAL	REF-WORD-MATCH	16.8
CNGL-SGS-TXT-1	MANUAL	REF-WORD-MATCH	9.0
CNGL-SGS-TXT-2	MANUAL	REF-WORD-MATCH	8.6
CNGL-SGS-TXT-3	MANUAL	REF-WORD-MATCH	8.5
CNGL-SGS-TXT-4	MANUAL	REF-WORD-MATCH	10.2
RYSDT-SGS-TXT-1	MANUAL	REF-WORD-MATCH	21.0
RYSDT-SGS-TXT-2	MANUAL	REF-WORD-MATCH	20.1
RYSDT-SGS-TXT-3	MANUAL	REF-WORD-MATCH	20.1
RYSDT-SGS-TXT-4	MANUAL	REF-WORD-MATCH	20.0
RYSDT-SGS-TXT-5	MANUAL	REF-WORD-MATCH	23.5
RYSDT-SGS-TXT-6	MANUAL	REF-WORD-MATCH	22.1
RYSDT-SGS-TXT-7	MANUAL	REF-WORD-MATCH	15.5
RYSDT-SGS-TXT-8	MANUAL	REF-WORD-MATCH	15.7
AKBL-SGS-TXT-2	MANUAL	REF-WORD-UNMATCH-LM	8.4
AKBL-SGS-TXT-5	MANUAL	REF-WORD-UNMATCH-LM	8.9
AKBL-SGS-TXT-3	MANUAL	REF-WORD-UNMATCH-AMLM	10.1
AKBL-SGS-TXT-6	MANUAL	REF-WORD-UNMATCH-AMLM	10.7
CNGL-SGS-TXT-5	MANUAL	REF-WORD-UNMATCH-AMLM	3.7
CNGL-SGS-TXT-6	MANUAL	REF-WORD-UNMATCH-AMLM	2.2
CNGL-SGS-TXT-7	MANUAL	REF-WORD-UNMATCH-AMLM	4.2
CNGL-SGS-TXT-8	MANUAL	REF-WORD-UNMATCH-AMLM	4.2
AKBL-SGS-TXT-7	MANUAL	MANUAL	17.2
CNGL-SGS-TXT-9	MANUAL	MANUAL	12.1

Table 5: SQ-SCR passage retrieval task result (%).

run ID	Query Transcription	Document Transcription	uMAP	pwMAP	fMAP
BASE-PAS-SPK-1	REF-WORD-MATCH	REF-WORD-MATCH	1.6	5.5	2.4
RYSDT-PAS-SPK-1	REF-WORD-MATCH	REF-WORD-MATCH	2.3	9.8	3.7
RYSDT-PAS-SPK-2	REF-WORD-MATCH	REF-WORD-MATCH	2.3	9.5	3.7
RYSDT-PAS-SPK-3	REF-WORD-MATCH	REF-WORD-MATCH	2.3	9.7	3.8
RYSDT-PAS-SPK-4	REF-WORD-MATCH	REF-WORD-MATCH	2.4	9.8	3.8
RYSDT-PAS-SPK-5	REF-WORD-MATCH	REF-WORD-MATCH	2.4	9.8	3.8
RYSDT-PAS-SPK-6	REF-WORD-MATCH	REF-WORD-MATCH	3.0	9.8	4.1
RYSDT-PAS-SPK-7	REF-WORD-MATCH	REF-WORD-MATCH	1.7	7.0	3.0
RYSDT-PAS-SPK-8	REF-WORD-MATCH	REF-WORD-MATCH	1.7	7.0	3.0
BASE-PAS-TXT-1	MANUAL	REF-WORD-MATCH	2.1	9.0	3.4
RYSDT-PAS-TXT-1	MANUAL	REF-WORD-MATCH	2.8	11.4	4.3
RYSDT-PAS-TXT-2	MANUAL	REF-WORD-MATCH	2.9	11.7	4.4
RYSDT-PAS-TXT-3	MANUAL	REF-WORD-MATCH	2.9	11.5	4.4
RYSDT-PAS-TXT-4	MANUAL	REF-WORD-MATCH	2.9	11.6	4.4
RYSDT-PAS-TXT-5	MANUAL	REF-WORD-MATCH	2.9	11.6	4.4
RYSDT-PAS-TXT-6	MANUAL	REF-WORD-MATCH	3.2	12.5	4.5
RYSDT-PAS-TXT-7	MANUAL	REF-WORD-MATCH	1.8	8.5	3.2
RYSDT-PAS-TXT-8	MANUAL	REF-WORD-MATCH	1.8	8.5	3.2

Table 6: SQ-STD task participants.

Group ID	Group Name, Organization	SPK	TXT
AKBL	Akiba Laboratory, Toyohashi University of Technology		3
ALPS	ALPS & Utsuro Lab., University of Yamanashi		3
IWAPU	IWAPU-EX3, Iwate Prefectural University	4	15
NKGW	Speech Language Processing Laboratory, Toyohashi University of Technology		1
NKI14	Nitta-Katsurada-Iribe-lab, Toyohashi University of Technology		4
R531	LabR531, National Taiwan University		6
RYSDT	RYukoku univ. Spoken Document processing Team, Ryukoku University		9
SHZU	Kai Laboratory, Shizuoka University	2	1
TBFD	Team Big Four Doragons, Daido University		8

Table 7: Result of SQ-STD using spoken query terms (%).

			micro F.	macro F.		time †
run ID	Query Transcription	Document Transcription	(spec./max.)	(spec./max.)	MAP	[msec]
BASE-SPK-1	REF-WORD-MATCH	REF-WORD-MATCH	34.9/45.5	34.2/43.5	43.4	-
BASE-SPK-2	REF-WORD-MATCH	REF-SYLLABLE-MATCH	24.7/34.6	24.4/32.9	31.4	-
BASE-SPK-3	REF-WORD-MATCH	REF-WORD-UNMATCH-LM	22.1/31.7	20.6/29.2	25.3	-
BASE-SPK-4	REF-WORD-MATCH	REF-SYLLABLE-UNMATCH-LM	13.5/22.6	12.2/21.2	17.2	-
BASE-SPK-5	REF-WORD-MATCH	REF-WORD-UNMATCH-AMLM	19.6/29.6	21.4/29.8	26.7	-
IWAPU-SPK-4	OWN	REF-WORD-MATCH	13.0/50.0	11.6/40.3	42.5	270
SHZU-SPK-1	OWN	REF-WORD-MATCH & REF-SYLLABLE-MATCH	42.3/43.4	36.5/36.9	32.5	792
SHZU-SPK-2	OWN	REF-WORD-MATCH & REF-SYLLABLE-MATCH	33.7/35.7	33.7/34.1	31.7	823
IWAPU-SPK-1	OWN	OWN	14.9/56.1	13.3/50.7	58.6	880
IWAPU-SPK-2	OWN	OWN	14.3/54.3	12.9/50.5	56.1	50
IWAPU-SPK-3	OWN	OWN	14.1/54.3	12.6/45.8	52.2	290

† Search time per query.

Table 8: Result of SQ-STD using text query terms (%).

run ID	Query Transcription	Document Transcription	micro F. (spec./max.)	macro F. (spec./max.)	MAP	time † [msec]
BASE-TXT-1	MANUAL	REF-WORD-MATCH	69.3/69.9	59.4/59.9	54.0	-
IWAPU-TXT-5	MANUAL	REF-WORD-MATCH	15.5/64.5	13.9/60.5	67.6	2420
IWAPU-TXT-11	MANUAL	REF-WORD-MATCH	15.5/71.0	13.9/62.6	59.7	520
R531-TXT-4	MANUAL	REF-WORD-MATCH	13.7/53.4	0.43/32.7	42.6	?
R531-TXT-7	MANUAL	REF-WORD-MATCH	23.7/53.4	19.9/32.7	42.3	?
RYSDT-TXT-1	MANUAL	REF-WORD-MATCH	14.7/69.8	13.1/59.2	56.1	?
RYSDT-TXT-2	MANUAL	REF-WORD-MATCH	14.1/64.5	12.6/53.3	52.4	?
RYSDT-TXT-3	MANUAL	REF-WORD-MATCH	14.0/64.5	12.5/51.1	53.3	?
RYSDT-TXT-4	MANUAL	REF-WORD-MATCH	14.9/70.0	13.3/59.6	57.5	?
RYSDT-TXT-5	MANUAL	REF-WORD-MATCH	14.7/64.9	13.1/54.3	56.1	?
RYSDT-TXT-6	MANUAL	REF-WORD-MATCH	14.7/66.1	13.1/56.0	56.8	?
RYSDT-TXT-7	MANUAL	REF-WORD-MATCH	14.2/69.8	12.7/59.1	54.5	?
RYSDT-TXT-8	MANUAL	REF-WORD-MATCH	14.8/70.4	13.2/60.3	57.3	?
RYSDT-TXT-9	MANUAL	REF-WORD-MATCH	14.5/59.1	12.9/59.1	54.8	?
SHZU-TXT-3	MANUAL	REF-WORD-MATCH	61.3/66.5	54.7/58.3	48.3	445
TBFD-TXT-1	MANUAL	REF-WORD-MATCH	58.3/58.9	54.0/54.9	44.6	180
TBFD-TXT-2	MANUAL	REF-WORD-MATCH	59.2/59.2	55.3/55.3	44.6	171
TBFD-TXT-3	MANUAL	REF-WORD-MATCH	59.0/59.5	54.8/55.6	45.0	144
TBFD-TXT-4	MANUAL	REF-WORD-MATCH	49.5/49.7	42.5/42.9	31.2	71
TBFD-TXT-5	MANUAL	REF-WORD-MATCH	57.6/57.9	53.3/53.8	42.9	97
TBFD-TXT-6	MANUAL	REF-WORD-MATCH	57.8/58.2	53.2/53.9	43.1	120
TBFD-TXT-7	MANUAL	REF-WORD-MATCH	48.9/48.9	42.1/42.1	31.2	90
TBFD-TXT-8	MANUAL	REF-WORD-MATCH	32.0/32.0	30.8/30.8	20.2	9.8
TBFD-TXT-9	MANUAL	REF-WORD-MATCH	32.0/32.0	30.8/30.8	20.2	9.8
BASE-TXT-2	MANUAL	REF-SYLLABLE-MATCH	52.6/52.6	43.3/43.7	40.9	-
AKBL-TXT-1	MANUAL	REF-SYLLABLE-MATCH	45.9/45.9	36.1/36.1	23.5	143
IWAPU-TXT-12	MANUAL	REF-SYLLABLE-MATCH	13.9/60.2	12.4/57.4	61.4	2400
NKGW-TXT-1	MANUAL	REF-SYLLABLE-MATCH	23.5/27.6	21.7/23.6	22.2	6
R531-TXT-8	MANUAL	REF-SYLLABLE-MATCH	8.9/8.9	7.3/7.3	3.6	?
BASE-TXT-3	MANUAL	REF-WORD-UNMATCH-LM	46.6/47.9	37.2/38.2	33.3	-
BASE-TXT-4	MANUAL	REF-SYLLABLE-UNMATCH-LM	28.7/35.8	22.6/29.0	24.4	-
BASE-TXT-5	MANUAL	REF-WORD-UNMATCH-AMLM	46.3/46.4	39.9/40.0	34.2	-
IWAPU-TXT-8	MANUAL	REF-WORD-MATCH & REF-SYLLABLE-MATCH	15.4/68.3	13.8/60.9	61.5	1100
IWAPU-TXT-9	MANUAL	REF-WORD-MATCH & REF-SYLLABLE-MATCH	15.5/68.0	13.8/60.4	61.4	1030
NKI14-TXT-1	MANUAL	REF-WORD-MATCH & REF-SYLLABLE-MATCH	58.3/58.8	54.0/55.5	51.0	0.65
NKI14-TXT-3	MANUAL	REF-WORD-MATCH & REF-SYLLABLE-MATCH	56.3/57.0	53.4/55.4	50.6	11.70
R531-TXT-6	MANUAL	REF-WORD-MATCH & REF-SYLLABLE-MATCH	13.7/13.7	12.3/12.3	39.7	?
R531-TXT-9	MANUAL	REF-WORD-MATCH & REF-SYLLABLE-MATCH	15.2/53.4	13.6/32.7	43.6	?
NKI14-TXT-2	MANUAL	REF-WORD-UNMATCH-LM & REF-SYLLABLE-UNMATCH-LM & REF-WORD-UNMATCH-AMLM	49.6/49.7	46.4/46.8	44.2	0.88
NKI14-TXT-4	MANUAL	REF-WORD-UNMATCH-LM & REF-SYLLABLE-UNMATCH-LM & REF-WORD-UNMATCH-AMLM	44.7/45.1	44.2/45.4	43.0	53.43
IWAPU-TXT-2	MANUAL	OWN	16.7/71.9	14.9/66.9	72.5	290
IWAPU-TXT-3	MANUAL	OWN	15.7/64.8	14.1/61.2	69.8	2410
IWAPU-TXT-4	MANUAL	OWN	13.0/50.0	11.6/40.3	42.5	2460
IWAPU-TXT-7	MANUAL	OWN	14.8/62.7	13.3/59.4	65.8	2400
IWAPU-TXT-10	MANUAL	OWN	14.8/56.7	13.3/50.2	54.0	180
IWAPU-TXT-13	MANUAL	OWN	14.8/56.7	13.3/50.2	54.0	280
IWAPU-TXT-14	MANUAL	OWN	14.7/50.5	13.1/46.5	54.1	720
IWAPU-TXT-15	MANUAL	OWN	12.5/40.1	11.2/36.1	40.1	780
IWAPU-TXT-1	MANUAL	2 OWNs	16.6/70.3	14.9/65.6	73.6	580
IWAPU-TXT-6	MANUAL	2 OWNs	15.9/64.7	14.2/59.1	66.6	1160
ALPS-TXT-1	MANUAL	8 OWNs & REF-WORD-MATCH & REF-SYLLABLE-MATCH	61.4/63.7	56.6/57.2	66.6	8125
ALPS-TXT-2	MANUAL	8 OWNs & REF-WORD-MATCH & REF-SYLLABLE-MATCH	53.6/65.5	50.6/58.5	67.2	6770
ALPS-TXT-3	MANUAL	8 OWNs & REF-WORD-MATCH & REF-SYLLABLE-MATCH	59.9/59.9	52.6/52.9	55.3	887

† Search time per query.