Multimedia Retrieval in MultiMatch: The Impact of Speech Transcript Errors on Search Behaviour

James Carmichael[1], Paul Clough[1], Eamonn Newman[2], Gareth Jones[2]
[1] Dept. of Information Studies, University of Sheffield,
[2] School of Computing, Dublin City University,
{j.carmichael, p.d.clough}@sheffield.ac.uk, {Gareth.Jones, enewman}@computing.dcu.ie

Abstract. This study discusses the findings of an evaluation study on the performance of a *multimedia multimodal information access sub-system (MIAS)*, incorporating automatic speech recognition technology (ASR) to automatically transcribe the speech content of video soundtracks. The study's results indicate that an information-rich but minimalist graphical interface is preferred. It was also discovered that users tend to have a misplaced confidence in the accuracy of ASR-generated speech transcripts, thus they are not inclined to conduct a systematic auditory inspection (their usual search behaviour) of a video's soundtrack if the query term does not appear in the transcript. In order to alert the user to the possibility that a search term may be incorrectly recognised as some other word, a matching algorithm is proposed that searches for word sequences of similar phonemic structure to the query term.

Keywords: automatic speech recognition, multimodal search, user evaluation.

1 Introduction

Multimedia indexing and multilingual information retrieval constitute the main objectives of MultiMatch (http://www.multimatch.org), an EU-funded Specific Targeted Research Project (STREP, contract No. 033104). MultiMatch aims to overcome language boundary, media and distribution problems currently affecting online access to cultural heritage material [3]. Recent years have witnessed a rapid increase in the amounts of multimedia cultural heritage material made available online; a substantial proportion of this information drawing from diverse cultures and languages. The MultiMatch project aims to address some of these issues by developing an information retrieval system able to harvest heterogeneous information from distributed sources, presenting them in a coherent and synthesised manner. The system is designed to exploit such rich and diverse data sources, making content accessible to the widest possible audience. Multilingual searching requires that the document contents be translated into a language known to the user and/or that the actual query term(s) be translated into the language of the target documents. The MultiMatch project has implemented such a multilingual multimodal search engine for a selection of European languages, with the specific objectives of:

- building a corpus of CH-relevant material via an in-depth crawling of online resources originating from various CH organisations;

- automatically classifying the material gathered in a semantic-web compliant fashion, based mainly on the following criteria: (i) **document content**, particularly if CH terms/concepts are present (ii) **metadata descriptors** and (iii) **contextual information** indicating that the found item is likely to be CH-relevant (e.g. images lacking metadata descriptors but which themselves are embedded in a web page featuring CH content);
- optimising the MM corpus data structure to support *focused queries* [2];
- displaying results in a multimodal fashion so that the user is able to simultaneously access and exploit a variety of document types and sources (e.g. audio, video and image files originating from different content providers but displayed on the same web page).

The following section describes such a multilingual multimodal IR engine for inter- and intra-video document searching, an area rich in research possibilities.

1.1 Searching Multilingual Multimedia: What the professional needs

As an example of MM multimodal searching in practice, we consider the requirements of a specific MM client, the *Netherlands Institute for Sound and Vision* (hereafter referred to by the organisation's Dutch name, "Beeld en Geluid" or simply "B&G" [http://www.beeldengeluid.nl]). One of B&G's public services is the provision – upon request – of copies of audiovisual programmes, particularly television documentaries and newscasts, disseminated in the Dutch mass media.

It is usually the case that B&G clients are not interested in entire video documents but only various segments thereof which are relevant to specific search criteria [1]. Accordingly, an online multimedia multimodal document retrieval system was developed allowing users to search within videos for *shot-level* clips relevant to their information needs. Individual episodes of television programmes are displayed in the user interface as a series of representative thumbnail images (*key frames*) that act as a visual summary of the video's contents. Any speech data featured on the video's soundtrack is rendered as a text transcript generated by automatic speech recognition (ASR) technology. Click-and-play functionality allows the user to click on a key frame to initiate video playback from the start of the shot sequence represented by said key frame. Additionally, provision is made for word-level searching of the video's speech transcripts to facilitate the location of relevant shots within the key frame series. This prototypical application, known as the *multimedia multimodal information access sub-system (MIAS)*, was evaluated to determine its usability and efficiency, particularly the accuracy of its ASR search functionality. The results of this evaluation study are presented in the section that follows.

1.2 The Users' demand: "More Multimodal Info with less Visual Components"

The ten users participating in the evaluation study – all B&G employees with considerable experience in performing multimedia document searches – were presented with a series of pre-defined search scenarios. These scenarios required

participants to locate specific points within a video document where certain key words were spoken and/or certain visual sequences occurred (e.g. a short clip featuring some personage discussing a specific topic). The evaluation group used both MIAS and an in-house multimedia IR system known as the Catalogue, which also features key frame click-and-play searching but no ASR-generated soundtrack speech transcripts. When conducting search operations, the participants were videotaped in order to measure the amount of time and mouse clicks taken to execute various tasks and thus gauge the efficiency of their interactions with the two video search systems. Table 1 lists the results of this time-motion analysis for two tasks involving searches for spoken words/phrases uttered by named individuals in a specified context. For Task 1, (location of term "decoratie"), MIAS' ASR processing correctly recognised the term; this was not the case for Task 2 where the search term, "Maria Callas", was incorrectly transcribed. It is to be noted that, for control purposes, the ten-member user group was divided into two five-member sub-groups, with the first group instructed to use MIAS for Task 1 and the Catalogue for Task 2. This task-to-search engine mapping was reversed for the second sub-group.

Table 1: User Time Motion Measurements for Search Tasks
(average times for each 5-member subgroup as collective)

Search Engine used	**Intra-video search Task Description**	**Avg. No. of Mouse clicks**	**Avg. Time Taken (s)**	**Failed Searches**
MIAS	Locate spoken instance of term "decoratie" [**correct ASR**]	7.4	17.6	0
Catalogue	Locate spoken instance of term "Maria Callas" [ASR unavailable]	29.7	58.3	1
MIAS	Locate spoken instance of term "Maria Callas" [**incorrect ASR**]	41.3	77.2	4
Catalogue	Locate spoken instance of term "decoratie" [ASR unavailable]	33.5	43.5	0

Although the users' longer task execution times and increased number of mouse clicks when using the Catalogue suggest that MIAS is substantially the more efficient system when its ASR technology works correctly, it is somewhat disquieting that the majority of these professional users (90%) failed to locate the query terms in the case of incorrect transcription. Such failure is hardly noted for the Catalogue system which does not offer text-to-speech transcriptions and therefore the users are obliged to revert to their usual behaviour of systematic auditory inspection of the video's soundtrack to find the query term. Moreover, the users' verbal observations while manipulating the MIAS system indicate a misplaced confidence in the ASR soundtrack transcription correctness since previous transcription-based searches had proven reliable. In terms of the actual graphical interface, the consensus among the user group was that both the MIAS and Catalogue interfaces were too cluttered. Four of the ten participants suggested that most of the text boxes could be removed and any

information therein presented in the form of tooltip-style pop-up boxes appearing only if the mouse pointer is placed over some other graphical component, such as a key frame image. These user requests to condense available information into a smaller UI footprint are, paradoxically, accompanied by demands for the provision of a greater range of multi-faceted data. The B&G researchers report that it is not unusual for them to receive IR requests involving personalities or events with which they are unfamiliar, thus it is necessary to conduct a preliminary search – often using the more popular search engines such as Google – to procure still images and/or audio clips in order to know what the investigated person or object looks and sounds like. Such multimodal preliminary searching is supported by MIAS, which features a specialist image-search interface; half of the users considered, however, that this image-search interface could be integrated into the principal search page in order to minimise the number of task-related mouse clicks. These concerns are being addressed in the design of the second MIAS prototype, along with methods of sensitising the user to possible ASR transcription inaccuracies, which is the topic of the section that follows.

1.3 Preliminary Implementation of a Phoneme Matching Algorithm

Given a specific word or word sequence, the proposed phoneme matching algorithm (PMA) scans the entire transcript of the multimedia file's soundtrack to locate other word sequences of similar phonemic structure which might be instances of the misrecognised query term. When this algorithm was applied to the video soundtrack's transcript used in Task 2 (see section 1.2), it correctly identified two phrases as instances of misrecognition of the search item "Maria Callas". When tested on a selection of video and audio documents, the PMA recorded 75 false positives (i.e. wrongly declaring a word/phrase to be a misrecognition of the search term) from 2083 possibilities (the total number of words in the documents' soundtrack transcripts). Conversely, it recorded nine false negatives (i.e. failing to locate a word/phrase that is actually a misrecognition of the search term). Although, it must be noted that only a small selection of polysyllabic search terms were used to test the algorithm, this approach appears to offer some potential in terms of assisting the user to cope with imperfect ASR.

References

1. Carmichael, J., Larson, M., Marlow, J., Newman, E., Clough, P., Oomen, J., Sav, S., "Multimodal Indexing of digital Audio-Visual Documents: A case study for Cultural Heritage Data" *Proceedings of the Sixth International Conference on Content Based Multimedia Indexing (CBMI '08)*, Queen Mary University, London (2008)
2. Jones, G., Li, Q., "Providing Topical Feedback for Link Selection in Hypertext Browsing", *30th European Conference on Information Retrieval Research (ECIR '08)*, Glasgow, Scotland (2008)
3. Marlow, J., Clough, P., Ireson, N., Cigarrán Recuero, J., Artiles, J. and Debole, F., The MultiMatch Project: Multilingual/Multimedia Access to Cultural Heritage on the Web, *Museums on the Web Conference (MW2008): Proceedings, J. Trant and D. Bearman (eds). Toronto: Archives & Museum Informatics*. (2008)