Tools for Searching, Annotation and Analysis of Speech, Music, Film and Video—A Survey

Alan Marsden, Adrian Mackenzie and Adam Lindsay
Lancaster Institute for the Contemporary Arts and Institute for Cultural Research, Lancaster University

Harriet Nock, John Coleman and Greg Kochanski
Phonetics Laboratory, University of Oxford

Abstract

This article examines the actual and potential use of software tools in research in the arts and humanities focusing on audiovisual (AV) materials such as recorded speech, music, video and film. The quantity of such materials available to researchers is massive and rapidly expanding. Researchers need to locate the material of interest in the vast quantity available, and to organize and process the material once collected. Locating and organizing often depend on metadata and tags to describe the actual content, but standards for metadata for AV materials are not widely adopted. Content-based search is becoming possible for speech, but is still beyond the horizon for music, and even more distant for video. Copyright protection hampers research with AV materials, and Digital Rights Management (DRM) systems threaten to prevent research altogether. Once material has been located and accessed, much research proceeds by annotation, for which many tools exist. Many researchers make some kind of transcription of materials, and would value tools to automate this process. Such tools exist for speech, though with important limits to their accuracy and applicability. For music and video, researchers can make use of visualizations. A better understanding (in general terms) by researchers of the processes carried out by computer software and of the limitations of its results would lead to more effective use of Information and Communications Technology (ICT).

Correspondence:
Alan Marsden, Lancaster Institute for the Contemporary Arts, Lancaster University, Lancaster LA1 4YW, UK.
E-mail:
a.marsden@lancaster.ac.uk

1 Introduction

Computing for researchers in the arts and humanities has typically been concerned with processing text materials. This is partly because much of the most highly valued cultural forms in the West are stored in print, and partly because computing technology for text reached a highly developed form a couple of decades ago. However, researchers are increasingly interested in non-text materials because of a massive increase in the quantity and availability of audiovisual (AV) materials and rapid developments in technology for handling such materials.

This article examines the prospects for computer-aided research with AV materials, at this convergence of increased availability and emerging software tools. We restrict ourselves to digitized

Literary and Linguistic Computing, Vol. 22, No. 4, 2007.
doi:10.1093/llc/fqm021 Advance Access Published on 21 July 2007

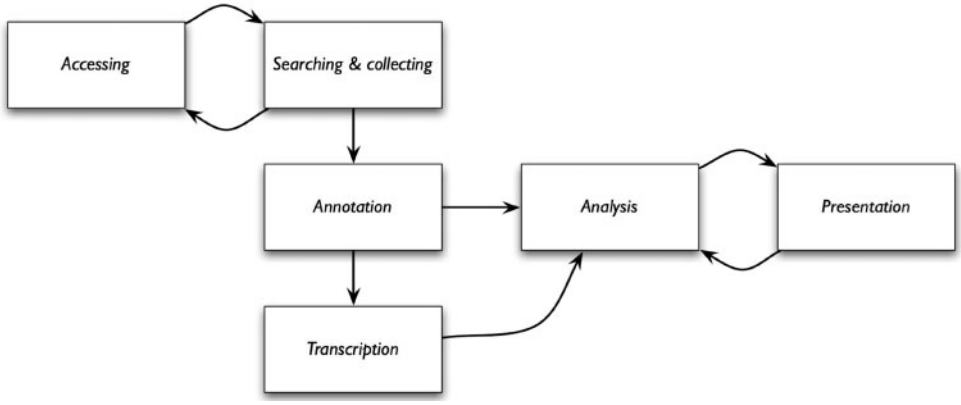

Fig. 1 Schematic model of humanities research with AV materials

time-based audio and visual source material. We exclude primary material based on still images, research material that exists primarily as a visual artefact (such as the image of a musical score), and materials which are primarily symbolic encodings or notations (such as the information in a musical score), except as annotations for search. Although there are common uses and needs, we also exclude materials for teaching and materials used in a creative process (as in the performing, visual, and compositional arts).

We focus on the problems and possibilities of working with primary materials such as recordings, video footage or broadcasts. These issues arise for contemporary scholars in many humanities disciplines. As far as possible, we have avoided addressing technical problems specific to single disciplines. Other recent publications have examined related issues of resources, access, and preservation (British Academy, 2005; BUFVC, 2004), though publications which also deal with issues of technology have generally focused only on speech (Goldman *et al.*, 2005; Koumpis and Renals, 2005; Lee and Chen 2005; Ostendorf *et al.*, 2005; SWAG, 2003).

More details specific to this project can be found in our report for the Arts and Humanities Research Council (Marsden *et al.*, 2006) and the weblog associated with the project: http://mediadescri.be. To discuss the research process, we adopt a simple generic model of humanities research using AV materials (Fig. 1). The model views humanities research as a process of repeatedly accessing, searching, marking up ('annotating'), transcribing, analysing, and presenting materials. The boundaries are of course not as sharp as the boxes imply—the schema is mainly intended to organize our discussion. Section 2 of the article outlines the breadth and abundance of materials becoming available, and some of the difficulties that scholars encounter in making use of them. Section 3 summarizes the major capabilities, possibilities, and lines of future development of digital technologies in humanities research with AV media. Section 4, based on a series of interviews and field visits to humanities researchers in a variety of disciplines, outlines what researchers do and do not do, and what they would like to be able to do in their research.

2 Accessing Audiovisual Materials

Under *access*, we consider issues concerning the location of this material, its quantity, its nature, forms and format, and the problems of right to use this material and its availability in digital form for research.

2.1 Collections of audiovisual materials

The increasing volume of AV materials is obvious. A recent survey estimated that the amount of information recorded on the physical medium of

film is somewhere between 76,000 and 420,000 Terabytes (1 TB = 1,000 Gigabytes). In comparison, the digitized version of the book collections of the US library of Congress would amount to 10 TB of information (19 million books and other printed collections) (Lyman and Varian, 2003).

Only a fraction of this is of interest to arts and humanities researchers, but it is not possible to identify a clear boundary between those materials which are of interest to scholars, and therefore should be preserved and made accessible, and those of no interest. Researchers in the arts and humanities have a massive amount of material available to them, but it is more variable and less organized than the traditional text materials contained in libraries.

While much of the material counted above is currently not in digital form, many archives have digitized some or all of their collections, or plan to do so, and this is often done in conjunction with a programme to make items available online. One such example is the Imperial War Museum's *Collections Online* (Imperial War Museum, 2006a). Another large UK digitization effort is being led by JISC, 'the JISC digitisation programme' (JISC, 2006), funded with a £10 million grant from the Higher Education Funding Council for England, to include such items as archival sound recordings at the British Library (3,900 h) and Newsfilm Online (6,500 h). A similar large-scale project is the effort of Google Video to make 'as much as [...] possible' of the US National Archives public domain video content available online (Google, 2006a; News.com, 2006).

Besides such national projects, there are a huge number of other collections of AV materials, with varying degrees of digitization and accessibility. There are many specialized collections of recorded speech, and those accessible online include poetry readings (e.g. The Poetry Archive, 2005) and oral history (e.g. Black and Ethnic Minority Experience, 2002). Large collections of recorded music exist also, but few collections are accessible without restriction. More commonly, subscription is required, whether for an academic collection such as the Culverhouse Classical Music Collection of recordings (mostly twentieth-century recordings of music from the seventeenth to nineteenth centuries) (Edina, 2006), or a commercial collection such as the Naxos Music Library (the entire CD catalogue of the recording company Naxos, 165,000 tracks from 11,000 CDs) (Naxos, 2006). For film the situation is similar. Collections tend to include material whose copyright is now of little value, such as the 48,000 'ephemeral' (advertising, educational, industrial, and amateur) films held in The Prelinger Archives (Prelinger Archives, 2006), or else payment is required, such as for footage from CNN Image Source (CNN, 2006) or in the increasing number of pay-per-view services.

A particularly significant development in video, however, is the growth of self-publishing through video upload sites such as YouTube (Youtube, 2006) and Google Video (Google, 2006b). This trend does not apply only to 'home movie' material; it can also be seen in efforts by institutions who want to be seen or heard, such as the 'podcasts' of sermons from St George's Church, Leeds (St George's, 2006), and in the growth of educational institutional repositories such as DSpace (MIT, 2006).

This suggests that the location of collections and repositories used by humanities scholars is shifting. Perhaps the most important access sites are no longer primarily institutionally managed. Instead, commercial services and user-produced archives and collections seem more important and relevant to much current scholarship. This situation yields greater certainty of access in some respects. For example, scholars are likely to have ready access to a much larger collection of recorded music through an online music library such as Naxos than most research libraries hold. On the other hand, catalogues, if they exist at all, can be of uncertain quality, and even in institutionally managed collections they often lack rich content descriptions. Sandom and Enser (2003) report that many film archives have large and growing backlogs of items for which there are no content description.

2.2 Technology and formats

While there are many formats for AV materials, the formats are generally well documented and

widely supported. Issues of fidelity are of diminishing importance with audio, though one must be wary of the growing ubiquity of MP3 encoding, which is adequate for research which involves unsophisticated listening but otherwise lacks the fidelity of other formats. However, fidelity remains a critical issue with video, especially since much material is more heavily compressed. The Cylinder Preservation and Digitization Project, a collection of cylinder recordings dating from 1890 to 1930 (UCSB, 2006), takes the useful approach of making its material available in both compressed restored formats and in a high-resolution raw (unrestored) format; one or the other is likely to be suitable for different research projects.

Some material is available only in streamed, compressed formats (e.g. the Naxos Music Library), which carries issues of reliability and fidelity, and can be frustrating for researchers who frequently need to hop around the material, slice it up, and focus on small sections. On the other hand, it is does not form an absolute impediment, since software to record from a stream to a file is readily available.

The power of modern desktop computers means that they are sufficient for viewing, editing, and storing a moderate collection of AV resources. On the other hand, typical current query-by-audio- or video-content algorithms run roughly equivalent to real time. Application to large collections therefore requires a more powerful technology, such as the pooling of resources available in grid technologies. Parallel to the increasing computing power on the desktop is increasing computing power in mobile devices. AV resources have begun to accompany researchers throughout their work and personal lives (e.g. using a personal video player to store film collections).

2.3 Access rights

The use of AV materials in research soon raises questions of copyright—it was a recurring theme of our interviews (see Section 4)—and they are rarely simple. Even in the case of the Cylinder Preservation and Digitization Project (UCSB, 2006), where copyright might be thought not to be an issue because of the age of the original material, copyright does apply to the restored digitizations, though controls are waived for non-commercial use. Strong voices have been raised in defence of access for research, such as that of the British Library (2006) and the British Academy (2006). However, there are strong commercial pressures in the opposite direction, restricting access for research.

The primary problem for research is likely to be a collection of technologies called Digital Rights Management (DRM). These technologies are designed to allow the rights-owner of content to determine how a consumer may use the content. Typically, the method for implementing DRM is to encrypt the file and tie the encryption key to the content-purchaser, the computer, and/or the date. A specialized, trusted application on the computer or portable device has the ability to decrypt the file and play it; no other applications may do so. This causes difficulties. Such content lacks compatibility with data analysis methods. Trusted applications typically do not offer the analyses that researchers need, and even if they did, the algorithms are private and undocumented, so the results are of uncertain value to researchers. A DRM scheme provider will want to know what an application does with the decrypted content before granting it 'trusted' status, thus open-ended research applications cannot be trusted. For DRM to be valuable to the rights-owner, trusted applications must not let the content escape. Data analysis programmes are thus shut out of working with protected content. Cumbersome workarounds may be possible but are often impractical for large amounts of content.

Even if circumventing DRM is possible, its legality is unclear. It is illegal to construct or possess devices and computer programs with the intention of circumventing protection which infringes copyright. [See, for example, the judgment in (1) *Kabushiki Kaisha Sony Computer Entertainment Inc* (2) *Sony Computer Entertainment Europe Ltd* (3) *Sony Computer Entertainment UK Ltd v (1) Gaynor David Ball & 6 Ors*, (2004) EWHC 1738 (Ch), 19 July 2004]. Section 29 (1) of Part I of the UK Copyright, Designs and Patents Act 1988 as amended (2003) states 'Fair dealing with a literary, dramatic, musical or artistic work for the purposes of research for a non-commercial purpose does not

infringe any copyright in the work provided that it is accompanied by a sufficient acknowledgement,' but DRM software is not typically written to allow such 'fair dealing', and it is not clear how far 'fair dealing' legitimizes devices or programs to circumvent copy protection. In practice, individual researchers cannot negotiate access to content either, because rights owners (quite reasonably) do not wish to sign a contract that incurs legal and engineering costs, brings in negligible revenue, and leads to some risk that the content will escape and be distributed without DRM.

One response is in the use of legal instruments such as Creative Commons (2006) which allow authors to reserve some, rather than all, rights (e.g. the right to benefit if the material is reused commercially). This model has spurred the development of sites storing AV data explicitly licenced for certain kinds of reuse (particularly for creative, non-commercial purposes), such as the BBC Creative Archive (BBC, 2003a) and The Freesound Project (FreeSound, 2006).

Nonetheless, access rights remain an important issue, and commercial usage of Creative Commons remains very small. Access can be a stumbling block for researchers who wish to work with AV materials from established providers and it is expected to become a more serious problem.

3 Technologies—State of the Art, Gaps, Obstacles

The general picture of technologies for handling AV materials in research is complex. There is a small number of tools explicitly designed for research, but researchers also make novel use of technologies originally designed for other purposes entirely. Much software is under development—this is a very active field for research in computer science—and there are many experimental systems available for use, but few finished and established products. This survey looks at several loci of Fig. 1.

3.1 Searching and collecting

With vast collections of digital AV material available, actually searching for and *finding* a resource can be a major barrier to research. Nearly all practical, current, multimedia access depends on good-quality metadata for search. However, content-based search for *text* has become an everyday tool on the world-wide web, and similar tools are desired for AV media. In this area, speech-based information retrieval is by far the most mature, and its current performance can be used as a rough indication of where video and music tools are headed in a decade. Established methods of accessing audio and video on the web include portals such as the BUFVC's Moving Image Gateway (BUFVC, 2006), which collects links to websites involving moving images and sound and their use in higher/further education, and HUMBUL, which includes categories such as Modern Languages—General, Sound/Audio (HUMBUL, 2006).

Free-text searching for speech generally relies on a previous annotation, transcription or content analysis (topics covered in later sections) to derive text from, or associate text with, the spoken word. Systems supporting the free-text querying of textual transcripts are now ubiquitous and similar systems exist for searching speech by querying time-aligned transcripts that can be automatically derived via speech-to-text systems. Surprisingly, with a little engineering ingenuity, the errors and lack of punctuation in these automatically derived transcripts seem to have little impact upon the effectiveness of the search, once the transcription error rate falls below about 40%. [For more details and possible explanations, see Allan (2003). Important words tend to occur multiple times, and it is unlikely that all occurrences will be mistranscribed]. Such an error rate is readily achievable by speech-to-text systems, at least for clean speech. However, difficulties arise if background noise or music is present, or if several people speak at once.

Additionally, speech-to-text systems typically depend upon a dictionary, and this leads to trouble when speech contains words which are not in the dictionary. In such cases, the automatically produced transcription can be missing new, important words until they enter the dictionary (e.g. 'e-mail' in the 1990s). This can be a problem with dynamically

evolving collections such as daily news (Hauptmann, 2005). There has been research into techniques for handling this problem; one technique involves searching a secondary phonetic transcription: a query term which falls outside the dictionary is (hopefully) located by searching for its pronunciation in the phonetic transcription (Amir *et al.*, 2002; Logan *et al.*, 2003).

In practice, searching for music is done via metadata, but the metadata available is often restricted and sometimes inaccurate. Searches within the Naxos Music Library (Naxos, 2006), for example, are restricted to fields such as 'composer' and 'title', according to the organization of the Naxos catalogue. The same applies to the Gracenote database (2006), which holds data on CD recordings, often automatically downloaded by media players when playing music from a CD. Problems include titles appearing in different languages from the original composition, and confusion over whether the 'artist' is the composer or the performer.

Audio content-based search for music, often called 'query by humming', has been a topic of considerable research. While there have been a number of experimental systems, some of them available on the web (e.g. NYU, n.d.), none has reached the stage of a usable tool. There are very significant technical issues to be addressed before this can be achieved and questions have been raised about the degree to which it would ever be a simple-to-use and effective tool (Pardo and Birmingham, 2003).

Databases that list information about films and television shows are now common on the web. For instance, the Internet Movie Database (IMDb, 2006) provides reviews, plot summaries, much technical production information and sometimes trailers for over 800,000 films and television series (July 2006). Because volunteers have added so much information about plot summaries and characters to the database, it can be used to find films and television programmes by subject, genre, etc. For current television content, new content alert systems based on programme schedules provide automatic notification of broadcasts that fit certain criteria [e.g. MeeVee (2006) or Radio Times (2006)]. The BBC has announced its commitment to making 1 million hours of television and radio searchable and available online and the BBC Programme Catalogue (BBC, 2006) allows 75 years of broadcasting to be searched. However, these systems do not actually search the content of film, video or broadcast; as in the case of the spoken word resources, they still rely on previous cataloguing, annotation or transcription.

One of the earliest projects to provide a search engine for AV media similar to those for text was Speechbot, a general web-deployed tool for audio indexing speech recognition transcriptions. While Speechbot is now unavailable due to the closure of the Compaq Cambridge Research Lab (US), in the past couple of years a number of similar services have emerged. Some tools crawl the web for audio and video made openly available on websites. For example, Podscope offers the ability to search audio blogs and pod casts, as does Blinkx (2006). Other tools support the search of video or audio submitted by users, for example, Podscope (Price, 2006) and Google Video Upload (Google, 2006c). Yet other tools index content legitimately provided by media companies and archives. For example, Blinkx has major deals with ITN and Fox News Channel (net imperative, 2006).

Finding relevant sources is only one step in collecting research materials. Many scholars collect large amounts of this material on their own computers and on portable storage media. While professionally curated online archives usually have extensive catalogues and indexes, personal collections of AV materials sometimes suffer from lack of organization. Software to facilitate organization of collections of AV material does exist, but it was not found to be commonly used in our interviews with humanities researchers.

3.2 Annotation

In the context of time-based media, annotation associates extra information, often textual but not necessarily so, with particular time points in an AV document or media file. In humanities research, annotation has long been important, but in the context of sound and image, it takes on greater importance. There is no margin on a recording

where a researcher can make notes, and, while pencil-and-paper annotation is still common, software to record and organize annotations of AV media are often useful in research.

Well-known metadata standards do not satisfy the requirements of annotation as described above: the standards do not generally have robust models for marking points *within* the content. One standard which does accommodate such annotations is MPEG-7, though generally available implementations have so far been inconsistent at best. The representation format offered by MPEG-7, however, seems to be one that would serve arts and humanities research very well. It is agnostic to media type and format. It is very general, and can be adapted to serve a variety of different applications. Despite its flexibility, it is far more than a 'guideline' standard: it has very specific rules for ensuring compatibility and interoperability. Annodex (2006) is an open standard for annotating and indexing networked media, which draws to some extent upon experience gained from MPEG-7. Annodex tries to do for video what URL links have done for text and images on the web.

3.2.1 *Manual annotation*

There are numerous tools (and formats) for creating linguistic annotations, many catalogued by the Linguistic Data Consortium (2001). Some speech-analysis tools also support annotation. [See for example, Gonet and Święciński (2002) or the long catalogue of tools listed by Llisterri (2006)]. There is also the open source Transcriber tool (2006) and numerous other commercial solutions for more general transcription of digital speech recordings, such as NCHSwiftSound (2006).

A typical video annotation tool is Transana, developed by WCER, University of Wisconsin (2006), which allows researchers to 'identify analytically interesting clips, assign keywords to clips, arrange and rearrange clips, create complex collections of interrelated clips, explore relationships between applied keywords, and share [their] analysis with colleagues.'

Though not explicitly intended for annotation, music-editing or music-composition software can have annotation capabilities or be used to perform annotation tasks. One example is Tanghe *et al.* (2005), who used *Cakewalk Sonar* (sequencer software by Twelve Tone Systems) to annotate the drum beats in extracts of sound recordings. An advantage of this software was that it allowed the user to check by ear whether or not the percussion strokes had been correctly identified and correctly timed. Perhaps the most highly developed specific music annotation tool is the CLAM Music Annotator (MTG, 2006; Amatriain *et al.*, 2005). This software allows the definition of different annotation types using an XML schema, and software elements can be added to automate some annotation processes.

3.2.2 *Collaborative annotation*

Simple collaborative annotation of AV materials is now common on the web. Sites such as Google Video (Google, 2006b) or Youtube (2006) primarily rely on tags supplied by contributors. Producers and consumers of AV material tag them with keywords which then become searchable via web search engines. While people often choose very generic keywords, and the keywords often apply to large video files, the tags and keywords produce a clearly useful synergy between the descriptions supplied by different users. For example, one user may annotate the style of the image, and another marks the presence of a street sign. Combinations of the annotations supplied by users allow database-driven websites such as flickr.com and youtube.com to provide reasonably powerful and selective search capabilities, more informative than one would expect from any single user's annotations.

A number of projects have attempted to design and construct collaborative software environments for video annotation, allowing a number of people to work on the same video footage. Efficient Video Annotation (EVA) (Volkmer, 2006) is a novel web tool designed to support distributed collaborative indexing of semantic concepts in large image and video collections. Some video annotation tools such as Transana (WCER, 2006) already exist in multi-user versions. In the social sciences, MixedMediaGrid (NCeSS, 2005), an ESRC e-Science funded project, aims to generate tools and techniques for social scientists to collaboratively

analyse AV qualitative data and related materials over the grid. Certainly, these tools and techniques could be used in the humanities too.

3.2.3 *Automatic annotation*

An alternative response to the time-consuming nature of manual annotation is to automate part of the process. Clearly, different kinds of annotations present different levels of difficulty in automation, and it is in the simple and explicit partitioning of audio, in particular, that automatic annotation has had the greatest success. The challenges of semantic annotation are much greater, though some projects in this area have had a degree of success, particularly with respect to music.

The goal of audio partitioning systems is to divide up the input audio into homogeneous segments and (typically) to determine their type. The class types considered may vary by application but a typical partitioning might distinguish pure music, pure speech, noise, combined speech and music, and combined speech and noise (Tranter and Reynolds, 2006). The resultant partitioning may provide useful metadata for the purpose of flexible access, but such partitioning is also an important prerequisite for speech-to-text transcription systems (e.g. it enables the removal of audio that might otherwise generate transcription errors) (Gauvain and Lamel, 2003).

The past decade has seen the birth and rapid growth of the field of Music Information Retrieval (MIR), fed in part by the interest of music businesses in technologies to facilitate user interaction with large databases of downloadable music. While 'query by humming' was an initial impetus to this field, more research has recently been directed at what are effectively various kinds of annotations of music. Some of these are concerned with partitioning (e.g. note onset detection or segmentation into broad sections) and some concerned with richer information such as tempo, beat, harmony and tonality, and various kinds of similarity or classification. Two well-developed tools for MIR are Marsyas, by George Tzanetakis (Tzanetakis, n.d.; Tzanetakis and Cook, 1999), and M2K, by Stephen Downie and others (Information Systems Research Laboratory, 2005), which functions within the D2K 'Data to Knowledge' framework of the US National Centre for Supercomputing Application.

The achievements of recent MIR research are best shown in the results of the MIREX competition (MIREX, n.d.) associated with the international conferences on Music Information Retrieval (ISMIR, n.d.). The competition has categories such as 'Audio Melody Extraction' and 'Symbolic Melodic Similarity'. The best audio systems typically perform with accuracies of 70–80% (but approaching 90% accuracy for key-finding systems), but this is well below the level at which such software would produce reliable results with real saving of effort if details of individual cases are important. The best symbolic systems (e.g. using MIDI data instead of audio) interestingly performed at similar levels of accuracy, despite the much lower complexity of the input data. On the other hand, other tasks on symbolic data such as 'pitch spelling' (i.e. determining a note name and accidental for each note such as 'C sharp' or 'D flat') can be performed with levels of accuracy of greater than 98% (Meredith, 2006), thus promising useful research tools.

For the last decade, many research projects have been working on automated video partition of footage into shots, topics, and face recognition (particularly in news video processing). Some of these systems use manual annotation to start with, and then automatically annotate and index any related video materials. For instance, the Marvel video annotation system (IBM, 2006) demonstrates the ability to generate semantic and formal labels from television news footage. Marvel builds statistical models from visual features using training examples and applies the models to automatically annotate large repositories. Other projects seek to generate topic structures for TV content using TV viewers' comments on live web chat rooms (Miyamori *et al.*, 2006).

3.3 Transcription

Music and speech are commonly transcribed in the course of research, but in the visual domain it is only dance which is regularly transcribed. Transcription requires a canonical means of 'writing down' the sequence of events in a transitory medium, depending on prior tacit agreement as to

what are the basic events and what is significant about them. This does not, and perhaps cannot, exist for generic videos, but transcription within constrained and structured domains such as sports is conceivable.

3.3.1 *Speech-to-text transcription*

Speech-to-text (or automatic speech recognition) systems convert a speech signal into a sequence of words. Progress in the field has been driven by standardized metrics, corpora and benchmark testing through NIST (US National Institute of Standards and Technology) since the mid-1980s, with systems developed for increasingly challenging tasks or 'speech domains': developing from the domain of single person dictation systems to today's research into systems for the meetings and lectures domain. A brief history of speech (and speaker) recognition research can be found in Furui (2005a).

There can be substantial differences between speech domains, and this can create difficulty for automatic systems. For example, speech from the lecture domain has much in common with speech from a more conversational domain including extraneous filler words (like 'okay') and filled pauses ('uh'). Sentences often exhibit false starts or are incomplete. It also displays poor planning at higher structural levels, often digressing from the primary theme.

Generally, speech recognition systems developed for one domain cannot be employed as a black box that can handle any domain: even speech from the same domain that differs from the 'training' data may be problematic (e.g. speech from previously unseen news broadcasts in Le, 2004). Despite significant progress in adaptation techniques to compensate for such differences, the development of systems which are robust to differences in data is a key research goal at present (Le, 2004; Ostendorf *et al.*, 2005).

Consequently, performance measurements are specific to a particular task and data set, and not broadly representative. An evaluation in 2004 reported state-of-the-art transcription systems to achieve a 'word error rate' of 12% for broadcast news in English, but 19% for Arabic. For conversational telephone speech, the figures were 15% for English and 44% for Arabic (Le, 2004). The effect of different recording techniques can be seen in Fiscus *et al.* (2005), where error rates for meetings were 26% when each participant had their own headset microphone, and 38% when the recordings were via multiple distant microphones.

There is a computation time versus accuracy trade-off: a real-time system will typically perform less well than a 10-times-real-time or even unconstrained system, but the degradation will vary with situation. Similarly, memory constraints can affect performance. State-of-the-art systems typically use hardware beyond that of today's average desktop. [The word-error rates for English speech referred to above were achieved in 10 times real time for broadcast news and 20 times real time for conversational telephone speech (Le, 2004)].

Church (2003) shows that speech-to-text transcription researchers have achieved 15 years of continuous error rate reduction: we might wonder what the future holds. At present, the accuracy of current systems lags about an order of magnitude behind the accuracy of human transcribers on the same task (David Nahamoo quoted in Howard-Spink, n.d.; Moore, 2003). Moore has estimated that it would take a minimum of 600,000 h of acoustic training data to approach a zero error rate using current techniques, which he notes is at least four times a typical human's lifetime exposure to speech!

Speech-to-text transcriptions have historically comprised an unpunctuated and unformatted stream of text. There has been considerable recent research into generating 'richer' transcriptions annotated with a variety of information that can be extracted from the audio signal and/or an imperfect transcription. Areas of work include punctuation and structural information (Liu *et al.*, 2005), speaker-related information (tracking utterances from the same speaker, determining who is speaking, verifying a speaker's identity, and determining a speaker's location) (Tranter and Reynolds, 2006; van Leeuwen *et al.*, 2006), named entity extraction (marking references to people, places, organisations, dates, etc.) (BBN Technologies, 2004-06), and various kinds of information extraction, including sometimes the emotional state

of the speaker. While progress has been made in all of these fields, usable tools for arts and humanities researchers involving this kind of technology are still beyond the horizon. Investigations have often used speech from only a small set of domains, such as broadcast news and conversational speech. Emotion-related work in particular is very preliminary.

3.3.2 *Music transcription*

For years, scholars have anticipated a tool which could transcribe musical performances to music notation. A tool which automatically transcribes even a simple musical performance into correct and accurate music notation remains a distant goal, however. Perhaps this should be no surprise, since only highly trained musicians can make any such transcription at all, and even so the process involves a high degree of approximation and guess-work. On the other hand, particularly in the light of recent MIR research, transcription into some form of notation which gives useful information is possible for restricted kinds of musical sound. Such transcriptions have been used in, for example, ethnomusicological research where systems like the melograph (a device which derives a continuous pitch curve from monophonic sound) have been in use for some time. A recent review of the state of the art in music transcription is (Klapuri, 2004).

3.4 Analysis

The location of 'analysis' in Fig. 1 indicates our intended meaning for the term: while many of the tasks and processes of annotation and transcription are in some sense analytical, we mean here that part of research where the results of annotation and transcription are subject to the judgement and intervention of the scholar who seeks to extract useful information, draw lessons, and form conclusions. With respect to AV materials, ICT tools play two distinct but possibly interrelated roles.

The first might be described as 'microscopic analysis', where the tool makes explicit characteristics of or data about the material which are otherwise too small, too fast or otherwise hidden. A prime example is Fourier analysis and other systems which extract time-varying frequency information from an audio signal, important in the analysis of both speech and music. Another example (more important in music than speech) is measurement of the timings of notes or syllables with precision of about 0.01 second.

The second role for ICT tools is to facilitate navigation through AV materials, especially multiple materials, multiple views of materials, or annotations or transcriptions in association with AV materials. Tools make it easy for scholars to jump to specified locations in a source, to align similar materials, to see or hear them aligned, and to view or hear AV material aligned with annotations or visualizations.

3.4.1 *Analysis of audio and music*

Analyses which focus on acoustic properties, in phonetics and music, regularly make use of tools which employ Fourier analysis or other methods such as auto-correlation to determine the component frequencies of a signal and their relative strengths. In the case of non-static sounds, this information is most commonly presented in a sonogram (a 2 1/2-dimensional display with time on the horizontal axis, frequency on the vertical axis, and amplitude or intensity shown by different gradations of colour, or a grey scale). Many such tools exist to effect such analysis: Wavesurfer (Sjölander and Beskow, 2000, 2006) is a good example of software from the research community, while Matlab [with its Signal Processing toolbox (The MathWorks, 1994-2006)] is probably the most commonly used commercial software. Musicians use such tools for many purposes, including the analysis of instrumental tone (e.g. Fitzgerald, 2003) and the analysis of pitch articulations and vibrato in performance (Rapaport, 2004).

Analysing musical performance has become a topic for research, spurred by the two factors of a now substantial history of recorded music and tools to facilitate the analysis of music-as-sound. However, there are still considerable gaps between the information which software can derive from musical sound and the information which researchers want to discover. For example, it is rarely

a simple and straightforward matter to distinguish where notes begin and end in a sonogram, and while the frequency composition of a sound can be derived, that does not always correlate simply with its perceived pitch composition. The most effective use of ICT in this area, therefore, comes when software allows the researcher to bring to play more effectively or more rapidly his or her musical ear and judgement. One example is MATCH (Dixon and Widmer, 2005; Dixon, 2005), which aligns two performances of the same piece, allowing a researcher to quickly and easily compare how two performers treat the same passage of the piece.

3.4.2 *Analysis of film*

Two main avenues of software-augmented analysis of film and video exist. The first seeks to automate analysis of the visual forms and narrative structure of film and television. The second uses databases and presentation software (media players mainly) to facilitate new kinds of manual analysis.

Tools for some limited automated analysis exist already. The Virage VideoLogger software claims to automatically create structured indexes of video content to facilitate search and retrieval (Virage, 2006). Another example is the MoCA Project (Automatic Movie Content Analysis) (Praktische Informatik IV, 2006), which seeks to provide automatic identification of the genre of a film by comparing visual statistics of frames and sequences with genre statistical profiles.

To date, the main software technology used in analysis of film and television has been the database coupled with DVD, which allows easy comparisons in manual analysis. The 'Digital Hitchcock' project, by Stephen Mamber (UCLA, n.d.), represents a well-known early example. It represents all 1,100 shots in the Hitchcock's *The Birds* alongside Hitchcock's storyboard illustrations.

3.5 Presentation

Researchers make use of different ways of visualizing, summarizing or tabulating AV materials. *Presentation* refers to all the different ways in which digital technologies display or render different AV materials apart from simply reproducing them. For instance, the timeline in a video editor or the waveform in a sound editor are presentations of images and sound, respectively. Presentation is closely linked to analysis. In some ways, we could say analysis is nothing but a process of generating increasingly complex, conceptually ordered presentations.

3.5.1 *Summarization*

AV materials take time to hear and view, and thus there is a strong interest in summarization. There has been some, albeit limited, work on generating summaries of spoken word content. Techniques for audio alone include time compression techniques such as eliminating silence or speeding up the clip (often maintaining pitch for intelligibility) (e.g. Tucker and Whittaker 2005). Spoken word content summarization and usability issues have been considered in some detail by Arons (1997) and by Furui (2005b). Work has also been done using errorful transcripts, by adopting techniques for general text summarization (as is apparently the case in Pickering *et al.*, 2003). Techniques exploiting both audio and transcript include the work by Koumpis and colleagues, who use both lexical- and audio-derived prosodic information to identify elements to include in the summary (Koumpis and Renals, 2001).

The same considerations have motivated research into automatic summarization of music. The common approach is to perform a self-similarity analysis of the audio signal, often by means of a frequency-domain transformation, and then to extract those segments which are similar to other segments. These are likely to correspond to recurring passages such as the chorus of a song, and so to contain music which is salient and typical of the whole. A short segment of audio can then be constructed by stringing together characteristic extracts (see, for example, Peeters *et al.*, 2002).

3.5.2 *Visualization*

Often it is useful to present the information in or derived from AV material in some other graphic form either to enable overall patterns or structure to be seen, or to assist in the identification of

points of particular interest. The topic is particularly common in music research; a discussion of different kinds of music visualization is given in Isaacson (2005).

A common technique is the repurposing of editing software to produce visualizations of the composition or structure of some material. Film scholars use commercial software such as Final Cut Pro and Adobe Premiere not only to edit digital video footage (for example, to extract clips for presentation or personal archives), but also as a way of examining the composition of film at various levels. The editing timeline is a central component in most video editing software, representing the complete set of frames in a film. Using the timeline, scholars can zoom in and out from frames to the overall film, and also view overall structure of the film or analyse transitions between shots. The same applies to audio editing software, where the typical 'waveform' display of the signal provides a quick and easy way of spotting sound and silence, and sometimes allows the beginnings and endings of sound events to be found also.

Specialized software involving visualization exists also. Video editing and mixing tools developed for Vjaying (video DJ'ing, i.e. selecting and mixing found video materials, and setting them to music) have addressed the problem of how to rapidly select and organize quite large collections of film and television footage. Software such as Resolume (n.d.), an instrument for live video performances, allows rapid selection, changing, combining and comparing of video clips on screen. For music, there are examples of projects which attempt to show higher-level or more 'semantic' properties in the audio stream. Examples are provided by aspects of the CLAM Music Annotator, mentioned above, which includes panels to visualize automatically extracted data on harmony and tonality in a time-varying 2D colour display (Gomez and Bonada, 2005).

4 User experience and expectations

In interviews with humanities researchers, we sought to gather information on the 'life cycle' of AV materials gathered for their research purposes. We encountered researchers both who recorded or constructed materials themselves and those who used 'found' materials of various kinds. Examples included the following:

Research resource:
(self-constructed) linguistics corpus, oral history interviews, auditory archaeology recordings;
(found) linguistics corpus, films/television/radio for historical or cultural analysis, poetry readings
Work record:
(self-constructed) archaeological excavation recordings, raw anthropological/documentary footage;
Research outputs and/or dissemination:
(self-constructed) multimedia archives created for use by researchers, technical/scholarly/popular presentations of research results involving multimedia;
Teaching and other purposes:
(self-constructed) phonetics sound examples for class, tutorial exemplars of form;
(found) clip examples for teaching

4.1 Methodology

Our user needs study interviewed 28 humanities researchers and several other technologists who work within the humanities. The research was carried out in three phases. Phase 1 aimed to interview one person per humanities field, using the AHRC Research Subject Coverage for guidance (AHRC 2003). The interviews were loosely structured using an interview questionnaire, supported by screenshots of the following tools:

(1) BLINKX: a live system supporting browsing and free text search for AV on the web (Blinkx, 2006).
(2) ANSES: a demo interface for news summarization including automatically extracted organizations, people, locations, and dates (Pickering, 2006).
(3) FERRET: a meeting browser tool (IDIAP, 2006).
(4) MULTIMODAL ANNOTATION TOOL: a manual annotation tool for video including associated soundtrack (Adams *et al.*, 2002).

Phase 2 aimed to interview modern historians whose web presence suggested AV data might be a potential resource (even if not currently used). Phase 2 interviews were aimed at gaining a more detailed understanding of the work process of researchers in one specific field, extracting information about their use of resources by asking researchers to talk through a typical research project, followed by guided discussion of the screenshots.

Phase 3 interviews concentrated on researchers whose primary interest was in AV material, such as films in popular culture, music within films, video games, or general musicology. The interviews were conducted using a set repertoire of questions on AV media usage and research practice, garnered from experience with the first two phases.

The combined results of all the interviews have been used to generate the following scenarios. They summarize uses of analysis, search and annotation tools that were suggested but also mention some of the associated challenges to deployment (technological or otherwise). The quotations in the scenarios below aim to present the interviewees' comments, but are sometimes abbreviated or paraphrased from notes, and may combine comments from more than one interviewee. Additional scenarios, greater detail, and the provenance of each comment are contained in our full report (Marsden *et al.*, 2006).

4.2 Research scenarios

4.2.1 *Online AV and web AV search tools*

Researcher X is interested in discourse differences across a number of non-Western countries and is currently exploring issues relating to visual grammar and reception. His research begins with a process of data set construction, which requires him to locate sources of moving image data and then filter that data in order to find instances of desired 'events' in the soundtrack or leading imagery, such as clips showing a weapon or alluding to a weapon. These instances form the data set for his research. At present, he primarily obtains data through off-air recordings (e.g. made by colleagues in the region) or from the few academic, area-specific websites online: more online sources of data from the area would benefit his research, particularly if easily locatable through search tools. The filtering process is currently very time-consuming, requiring a full viewing: search tools that could help him identify relevant 'events' within videos would be very helpful in speeding up this filtering process. Such tools would need to support search in his language of interest and perhaps image-related search as well as free text search. Since these envisaged tools do not currently exist in a packaged form, he sees manual tools as a potentially useful and available alternative for the filtering step: a tool such as the IBM annotation tool could support the process of marking up and categorizing soundtrack segments or image regions and this could be combined with a viewer tool which supports the recall of items in the same category (e.g. the category of clips showing a weapon).

4.2.2 *AV Archive Browsing and Search*

Researcher X is a modern historian investigating the social history of an English-speaking country outside the UK who does not currently make much use of time-based media. Very occasionally he will go to archives and read their transcripts of potentially relevant video and he has independently accumulated a few documentaries on prominent political figures in that country, but he finds it 'takes a lot of time to get a little way with video.' He has analysed some propaganda videos in the past, though, and certainly sees uses for AV data in the future if it became more accessible.

The ability to do a free text search within a single archive collection of spoken word data might encourage him to use collections which have not been transcribed, particularly if results are cued up around the query terms and linear scanning of full tapes is not required. Because he investigates mid-century social history from the 'bottom-up', he would be interested in recordings from that era involving 'the people' e.g. 'speeches by the regional mayor or activists.' If this kind of data were to become more accessible, he envisages addressing new questions such as televisual representation or comparative studies of representations of things in television, text and pictures. Most research in his area today cites national newspapers rather than national television stations, even though

more people watch the latter and it is arguably more influential.

4.2.3 *Data preparation*

Researcher X obtains a lot of original material to be edited together later. In order to prepare his material, he relies on a more traditional methodology: he logs tape in his (digital) editing suite. He does not resent this often-laborious process (5–10 h per hour of raw footage), as it gives him a chance to reflect upon the materials he has gathered. What he would welcome, however, is some way of automatically transcribing the speech from the video. As his subjects include non-native speakers and many interviews are conducted in the field (with background noise), his is a wish not likely to be realized in the near term.

4.2.4 *Analysis and interpretation*

Researcher X analyses films and television. He has learned that engineers have developed research tools which can automatically detect shot boundaries and can classify each shot into categories such as cuts or fades. 'With this technology I could explore questions such as the use of "long takes" or statistics about cuts and shot types and so on . . . I could extract statistics such as the number of cuts in the first and last 10 min of a film or historical changes in cutting rates in a TV or film type. It's too time-consuming to manually annotate these things for research on an extensive data set . . . a tool giving this kind of quantitative analysis would be very useful.' Such information may strengthen the empirical foundations of the kinds of research questions currently asked in the field, by providing quantitative evidence.

4.3 Common research issues

Self-recorded research sources were often based upon interviews, oral histories, or as documentary markers. A common issue was the vast amount of material collected and the limited time available to record and sift through it. As such, nearly all such interviewees wanted a way to transcribe the material.

Research with found data can run into several roadblocks. The first is simply knowing where to look. Not all large AV archives are in obvious locations or maintained by the bodies most obvious to those accustomed to more 'traditional' textual scholarship. The second is that there can be access problems: although technical barriers to access are being lifted in the online world, not all of the most relevant archives are digitized or transparent to outsiders. Beyond simple access, access *rights* become terribly important in the digital world: DRM can create difficulties from headaches and inconvenience to completely cutting off a legitimate line of inquiry on AV material (e.g. automated signal processing and analysis on audio).

Once a data store is found and accessed, many find difficulty on the other side of the fence: there can be too much data for a single researcher to work on. A few researchers complained of coming across rich archives of video, but finding that manually searching for things that were interesting to them was too time-consuming for the rewards. Again, transcription was an oft-requested *desideratum*, an implicit demonstration of text's superiority for browsability over AV material. Some researchers give themselves over to serendipity with found media, allowing broadcast media or online sources to open up new avenues for their research.

Once a particular piece of AV content is chosen for deeper analysis after an initial viewing/audition, a common first step is to develop some sort of timeline-based annotation. Although many ICT tools exist for this, many researchers are satisfied with making a table with times of notable events, matched with other relevant notations, on paper. Those who deal with oral histories and other interviews cite the making of transcriptions as a major effort and (often) expense. The later stages of processing may vary from researcher to researcher, but some researchers did show an interest in collaborative annotation (whilst expressing some doubt as to its technical or legal feasibility). Dissemination and other forms of sharing the results of research were similarly up to individual researchers. Those who had made use of ICT in doing so were generally comfortable with the tools available, since the tasks involved are familiar and well documented.

4.4 Technical expectations

Finally, we noted some common misapprehensions about ICT tools and what humanities researchers imagined the tools could achieve:

4.4.1 *Error*

ICT tools cannot generally mimic human perceptions perfectly. While scientists regularly take account of errors in measurement, scholars in the arts and humanities are not generally used to dealing with this. Some make the mistake of assuming that a tool will always do what it is supposed to do or, alternatively, dismiss tools that fail to achieve perfection. To make effective use of these tools, scholars need to allow for error, and have an idea of how much error can be tolerated.

4.4.2 *Robustness*

A lack of robustness is a common source of error. Many of the ICT tools mentioned here (especially in the area of speech recognition) depend on 'training', and can be very susceptible to differences in the nature of material: a system might perform well with data similar to the original training set, but badly with different data. This is particularly problematic in domains where novel data is common (or even an objective), but even in more stable domains, proper use of such tools can rely on a knowledge of how to train and adjust a system with a particular kind of data.

4.4.3 *'I can't do that [with that tool]'*

Software is complex, and researchers are often not aware of a tool's full capabilities. It was common to find researchers who were not aware that a piece of software already performed a task which they thought was not possible.

5 Conclusion

The increasing importance of AV to researchers in the arts and humanities is clear. The quantity available is greater, and its access is easier. It is taking on a role of greater significance for researchers, in line with its increasing role in contemporary society and culture. This will bring with it an increased demand for computing power and network bandwidth on the part of humanities scholars. More importantly, scholars will need somewhat different sets of skills in ICT than have hitherto been regarded as the core of humanities computing.

Commercial interest in digital AV has already grown, seen most dramatically in the rapid expansion of the Google enterprise into the area of AV. We can expect that this will lead to increased access to digital AV for scholars, through greater ease of access via the web, more digitization of materials, and more sophisticated software tools. The needs of scholars, however, will take second place to commercial interests, and we anticipate continuing problems with 'black-box' software whose workings are opaque and whose error and robustness (identified above as important details of software tools) are unknown, limiting their usefulness for research. Metadata might not be reliably accurate or categories important for scholars might simply be missing (as is currently the case for the Gracenote database of CD recordings). Issues of access rights are likely to remain problematic, and adoption of digital rights management systems could cause a serious impediment to research with AV.

Finally, issues have been identified also around those collections of AV materials which scholars make themselves. It is common for a researcher to amass a collection of material sufficiently large to be difficult to organize and keep track of. Systems (both software tools and ways of working) need to be developed to prevent this impeding research. The value of standards to allow the sharing of digital raw materials is already recognized, and bodies such as the Arts and Humanities Data Service in the UK commonly promote their use. Our survey indicates that a vast quantity of the research effort in working with AV is invested in annotation, and mechanisms to facilitate the preservation and sharing of annotated data promise to be of considerable benefit for future research.

Acknowledgements

We gratefully acknowledge the generous amount of time and information given by all of the participants in interviews, and the financial support of the Arts

and Humanities Research Council under its ICT Strategy programme.

References

Adams, B., Lin, C.-Y., and Iyengar, G. (2002). IBM Multimodal Annotation Tool. http://www.alphaworks.ibm.com/tech/multimodalannotation (accessed 4 January 2007).

AHRC, Arts and Humanitites Research Council (2003). Research subject coverage. http://www.ahrc.ac.uk/about/subject_coverage/research_subject_coverage.asp (accessed 4 January 2007).

Allan, J. (2003). Robust techniques for organizing and retrieving spoken documents. *EURASIP Journal on Applied Signal Processing*, 2003, 103–114. doi: 10.1155/S1110865703211070. http://www.hindawi.com/GetArticle.aspx?doi=10.1155/S1110865703211070 (accessed 4 January 2007).

Amatriain, X., Massaguer, J., Garcia, D., and Mosquera, I. (2005). The CLAM Annotator: A Cross-platform Audio Descriptors Editing Tool. In *Proceedings of the Sixth International Conference on Music Information Retrieval*, London, 426–9. http://www.iua.upf.edu/mtg/publications/9317d2-ismir2005-clam-annotator.pdf (accessed 4 January 2007).

Amir, A., Srinivasan, S., and Efrat, A. (2002). Search the Audio, Browse the Video—A Generic Paradigm for Video Collections. http://www.hindawi.com/GetArticle.aspx?Doi=10.1155/S111086570321012X&e=CTA (accessed 4 January 2007).

Annodex (2006). http://www.annodex.net/ (accessed 4 January 2007).

Arons, B. (1997). SpeechSkimmer: A System for Interactively Skimming Recorded Speech. *ACM Transactions on Computer-Human Interaction*, **4**: 3–38, http://xenia.media.mit.edu/~barons/html/tochi97.html (accessed 4 January 2007).

BBC, British Broadcasting Corporation (2003a). Creative Licence Group. http://creativearchive.bbc.co.uk/ (accessed 4 January 2007).

BBC, British Broadcasting Corporation (2006). BBC Programme Catalogue. http://open.bbc.co.uk/catalogue/infax (accessed 4 January 2007).

BBN Technologies (2004–06). IdentiFinder. http://www.bbn.com/Solutions_and_Technologies/Data_Indexing_and_Mining/Identifinder.html (accessed 4 January 2007).

Black and Ethnic Minority Experience (2002). http://www.be-me.org/ (accessed 4 January 2007).

Blinkx (2006). blinkx.tv http://tv.blinkx.com/ (accessed 4 January 2007).

British Academy (2005). E-Resources for Research in the Humanities and Social Sciences—A British Academy Policy Review. http://www.britac.ac.uk/reports/eresources/index.html (accessed 4 January 2007).

British Academy (2006). Copyright and Research in the Humanities and Social Sciences. http://www.britac.ac.uk/reports/copyright/index.html (accessed 4 January 2007).

British Library (2006). Intellectual Property: A balance; The British Library manifesto, http://www.bl.uk/news/pdf/ipmanifesto.pdf (accessed 4 January 2007).

BUFVC, British Universities Film and Video Council (2004). Hidden Treasures: the UK Audiovisual Archive Strategic Framework. http://www.bufvc.ac.uk/faf/HiddenTreasures.pdf (accessed 4 January 2007).

BUFVC, British Universities Film and Video Council (2006). Moving Image Gateway. http://www.bufvc.ac.uk/gateway/ (accessed 4 January 2007).

Church, K. W. (2003). Speech and Language Processing: Where Have We Been and Where Are We Going? In *Proceedings of EUROSPEECH 2003, 8th European Conference on Speech Communication and Technology*, Geneva, 1–4. http://research.microsoft.com/users/church/wwwfiles/papers/Eurospeech/2003/ES032000.pdf (accessed 4 January 2007).

CNN, Cable News Network (2006). Image Source. http://www.cnnimagesource.com/CNIS/index.html (accessed 4 January 2007).

Creative Commons (2006). Enabling the Legal Sharing and Reuse of Cultural, Educational, and Scientific Works. http://creativecommons.org/ (accessed 4 January 2007).

Dixon, S. and Widmer, G. (2005). MATCH: A Music Alignment Tool Chest. In *Proceedings of the Sixth International Conference on Music Information Retrieval (ISMIR 2005)*, London, 492–497. http://www.ofai.at/cgi-bin/tr-online?number+2005-17 (accessed 4 January 2007).

Dixon, S. (2005). MATCH, Music Alignment Tool Chest. http://www.ofai.at/~simon.dixon/match/index.html (accessed 4 January 2007).

Edina (2006). Film and Sound Online. http://www.filmandsound.ac.uk (accessed 4 January 2007).

Fiscus, J. G., Radde, N., Garofolo, J., Le, A., Ajot, J., and Laprun, C. (2005). The Rich Transcription 2005 Spring Meeting Recognition Evaluation. http://www.nist.gov/speech/publications/papersrc/rt05sresults.pdf (accessed 4 January 2007).

Fitzgerald, R. A. (2003). *Performer-Dependent Dimensions of Timbre: Identifying Acoustic Cues for Oboe Tone Discrimination*. Ph.D. Thesis, School of Music, University of Leeds.

FreeSound (2006). The Freesound Project. http://freesound.iua.upf.edu/ (accessed 4 January 2007).

Furui, S. (2005a). 50 Years of Progress in Speech and Speaker Recognition. In *Proceedings of SPECOM* 2005, Patras, Greece, 1–9. Preprint: http://www.furui.cs.titech.ac.jp/publication/2005/SPCOM05.pdf (accessed 4 January 2007).

Furui, S. (2005b). Spontaneous Speech Recognition and Summarization, The Second Baltic Conference on Human Language Technologies, 39–50. http://www.furui.cs.titech.ac.jp/publication/2005/HLT2005.pdf (accessed 4 January 2007).

Gauvain, J.-L. and Lamel, L. (2003). Structuring Broadcast Audio for Information Access. *EURASIP Journal on Applied Signal Processing*, **2003**(2): 140–50. http://www.hindawi.com/GetArticle.aspx?doi=10.1155/S1110865703211033 (accessed 4 January 2007).

Goldman, J., Renals, S., Bird, S. *et al.* (2005). Accessing the Spoken Word. *International Journal on Digital Libraries*, **5**(4): 287–98. doi: 10.1007/s00799-004-0101-0. Preprint: http://www.cstr.ed.ac.uk/downloads/publications/2005/swag-ijdl05.pdf, full project report: http://www.dcs.shef.ac.uk/spandh/projects/swag/swagReport.pdf (accessed 4 January 2007).

Gomez, E. and Bonada, J. (2005). Tonality Visualization of Polyphonic Audio. In *Proceedings of the International Computer Music Conference*, Barcelona, 57–60.

Gonet, W. and Święciński, R. (2002). Speech Lab @ Work and @ Home. *Speech and Language Technology*, **6**, 57–80, Polish Phonetic Association.

Google (2006a). Nara on Google Video. http://video.google.com/nara.html (accessed 4 January 2007).

Google (2006b). Google Video. http://video.google.co.uk/ (accessed 4 January 2007).

Google (2006c). Google Video Upload Program. http://upload.video.google.com/ (accessed 4 January 2007).

Gracenote (2006). Gracenote Music Fans. http://www.gracenote.com (accessed 4 January 2007).

Hauptmann, A. (2005). Lessons for the Future from a Decade of Informedia Video Analysis Research. http://www.informedia.cs.cmu.edu/documents/CIVR05_Hauptmann.pdf (accessed 4 January 2007).

Howard-Spink, S. (n.d.). You just don't understand! http://domino.watson.ibm.com/comm/wwwr_thinkresearch.nsf/pages/20020918_speech.html (accessed 4 January 2007).

HUMBUL (2006). Intute: Arts and Humanities. http://www.intute.ac.uk/artsandhumanities/langlit-all/ (accessed 4 January 2007).

IBM, International Business Machines (2006). Marvel. http://domino.research.ibm.com/comm/research_projects.nsf/pages/marvel.index.html (accessed 4 January 2007).

IDIAP Research Institute (2006). Ferret Meeting Browser Demo. http://mmm.idiap.ch/demo/ (accessed 4 January 2007).

IMDb (2006). The Internet Movie Database. http://www.imdb.com/ (accessed 4 January 2007).

Imperial War Museum (2006a). IWM Collections Online. http://www.iwmcollections.org.uk/ (accessed 4 January 2007).

Information Systems Research Laboratory, University of Illinois (2005), M2K (Music-to-Knowledge): A Tool Set for MIR/MDL Development and Evaluation. http://www.music-ir.org/evaluation/m2k/ (accessed 4 January 2007).

Isaacson, E. (2005). What You See is What You Get: On Visualizing Music. In *Proceedings of the Sixth International Conference on Music Information Retrieval*, London, 389–95. http://ismir2005.ismir.net/proceedings/1129.pdf (accessed 4 January 2007).

ISMIR (n.d.). The International Conferences on Music Information Retrieval and Related Activities. http://www.ismir.net/ (accessed 4 January 2007).

JISC (2006). JISC Digitisation Program. http://www.jisc.ac.uk/digitisation_home.html (accessed 4 January 2007).

Klapuri, A. (2004). Automatic Music Transcription As We Know it Today. *Journal of New Music Research*, **33**: 269–82.

Koumpis, K. and Renals, S. (2001). The Role of Prosody in Voicemail Summarization Systems. *ISCA Workshop on Prosody in Speech Recognition and Understanding*. NJ: Red Bank(accessed 4 January 2007).

Koumpis, K. and Renals, S. (2005). Content-based Access to Spoken Audio. *IEEE Signal Processing Magazine*,

22(5): 61–90. Preprint: http://www.cstr.ed.ac.uk/downloads/publications/2005/koumpis-spm05.pdf (accessed 4 January 2007).

Le, A. (2004). 2004 Fall Rich Transcription Speech-to-Text Evaluation. http://www.nist.gov/speech/tests/rt/rt2004/fall/rt04f-stt-results-v6b.pdf (accessed 4 January 2007).

Lee, L. and Chen, B. (2005). Spoken Document Understanding and Organization. *IEEE Signal Processing Magazine*, **22**(5): 42–60. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?isnumber=32367&arnumber=1511823&count=17&index=4 (accessed 4 January 2007).

Linguistic Data Consortium (2001). Linguistic Annotation. http://www.ldc.upenn.edu/annotation/ (accessed 4 January 2007).

Liu, Y., Shriberg, E., Stolcke, A., *et. al.* (2005). Structural Metadata Research in the EARS Program. *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. Proceedings* (ICASSP '05), **5**, 957–60. http://ieeexplore.ieee.org/xpl/RecentCon.jsp?punumber=9711, Digital Object Identifier 10.1109/ICASSP.2005.1416464. Preprint: http://www.icsi.berkeley.edu/~yangl/icassp2005-mde.pdf (accessed 4 January 2007).

Llisterri, J. (2006). Speech Analysis and Transcription Software. http://liceu.uab.es/~joaquim/phonetics/fon_anal_acus/herram_anal_acus.html (accessed 4 January 2007).

Logan, B., Moreno, P., and Van Thong, J. M. (2003). Approaches to Reduce the Effects of OOV Queries on Indexed Spoken Audio. *Technical Report HPL-2003-46*, HP Laboratories Cambridge. http://citeseer.ist.psu.edu/logan03approaches.html (accessed 4 January 2007).

Lyman, P. and Varian, H. (2003). How Much Information? http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/ (accessed 4 January 2007).

Marsden, A., Nock, H., Mackenzie, A., Lindsay, A., Coleman, J., and Kochanski, G. (2006). ICT Tools for Searching, Annotation and Analysis of Audiovisual Media. AHRC ICT Strategy Project report. On-line version at http://www.phon.ox.ac.uk/avtools, mirrored at http://ict4av.lancs.ac.uk/report. (accessed 4 January 2007).

MeeVee (2006). http://www.meevee.com/ (accessed 4 January 2007).

Meredith, D. (2006). The *ps13* Pitch Spelling Algorithm. *Journal of New Music Research*, **35**: 121–59.

MIREX (n.d.). http://www.music-ir.org/mirexwiki/index.php/Main_Page (accessed 4 January 2007).

MIT, Massachusets Institute of Technology (2006). Welcome to DSpace. http://www.dspace.org (accessed 4 January 2007).

Miyamori, H., Stejic, Z., Araki, T., Minakuchi, M., and Ma, Q. (2006). Proposal of Integrated Search Engine of Web and TV Contents. In *Proceedings of WWW2006, 15th World Wide Web Conference*, Edinburgh. http://www2006.org/programme/files/pdf/p190.pdf (accessed 4 January 2007).

Moore, R. (2003). A Comparison of the Data Requirements of Automatic Speech Recognition Systems and Human Listeners. In *Proceedings of EUROSPEECH 2003, 8th European Conference on Speech Communication and Technology*, Geneva, 2582–4. http://www.dcs.shef.ac.uk/~roger/publications/Eurospeech03%20Comparison%20of%20Data%20Requirements.pdf (accessed 4 January 2007).

MTG, Pompeu Fabra University (2006). http://iua-share.upf.es/wikis/clam/index.php/Music_Annotator (accessed 4 January 2007).

Naxos (2006). Naxos Music Library. http://www.naxosmusiclibrary.com (accessed 4 January 2007).

NCeSS, National Centre for e-Social Science (2005). MixedMediaGrid. http://www.ncess.ac.uk/research/nodes/MiMeG/ (accessed 4 January 2007).

NCHSwiftSound (2006). Express Scribe Transcription Playback Software. http://www.nch.com.au/scribe/index.html (accessed 4 January 2007).

net imperative (2006). Blinkx launches ad-funded video service. http://www.netimperative.com/2006/02/08/Blinkx_ITN (accessed 4 January 2007).

News.com (2006). Google puts National Archives Video Online. http://news.com.com/Google+puts+National+Archives+video+online/2100-1025_3-6043193.html (accessed 4 January 2007).

NYU, New York University (n.d.). Query by Humming. http://querybyhum.cs.nyu.edu/ (accessed 4 January 2007).

Ostendorf, M., Shriberg, E., and Stolcke, A. (2005). Human Language Technology: Opportunities and Challenges. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2005. (ICASSP'05), **5**, 949–52. http://ieeexplore.ieee.org/xpl/RecentCon.jsp?punumber=9711, Digital Object Identifier 10.1109/ICASSP.2005.1416462. Preprint: http://www.speech.sri.com/papers/icassp2005-specialsession.pdf (accessed 4 January 2007).

Pardo, B. and Birmingham, W.P. (2003). Query by Humming: How Good Can it Get? In J.S. Downie (ed.),

The MIR/MDL Evaluation Project White Paper Collection, 3rd edn, 107–9. http://www.music-ir.org/evaluation/wp3/wp3_pardo_query.pdf (accessed 4 January 2007).

Peeters, G., La Burthe, A., and Rodet, X. (2002). Toward Automatic Music Audio Summary Generation from Signal Analysis. In *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR 2002)*, Paris. http://recherche.ircam.fr/equipes/analyse-synthese/peeters/ARTICLES/Peeters_2002_ISMIR_AudioSummary.pdf (accessed 4 January 2007).

Pickering, M. (2006). Automatic News Summarization Extraction System. http://wwwhomes.doc.ic.ac.uk/~mjp3/anses/ (accessed 4 January 2007).

Pickering, M. J., Wong, L., and Rüger, S. M. (2003). ANSES – Summarisation of News Video. In *Proceedings of International Conference on Image and Video Retrieval (CIVR-2003)*. Lecture Notes in Computer Science 2728 (Springer), 425–34. http://www.doc.ic.ac.uk/~mjp3/phd/www-pub/civr2003.pdf (accessed 4 January 2007).

The Poetry Archive (2005). http://www.poetryarchive.org (accessed 4 January 2007).

Praktische Informatik IV, University of Mannheim (2006). Automatic Movie Content Analysis: The MoCA Project. http://www.informatik.uni-mannheim.de/pi4.data/content/projects/moca/ (accessed 4 January 2007).

Prelinger Archives (2006). http://www.archive.org/details/prelinger (accessed 4 January 2007).

Price, G. (2006). Searching for Online Video. http://searchenginewatch.com/searchday/article.php/3576231 (accessed 4 January 2007).

Radio Times (2006). http://www.radiotimes.com/ (accessed 4 January 2007).

Rapaport, E. (2004). Schoenberg-Hartleben's Pierrot Lunaire: Speech – Poem – Melody – Vocal Perforamnce. *Journal of New Music Research*, **33**: 71–111.

Resolume (n.d.). Resolume VJ Software. http://www.resolume.com/features/index.php (accessed 4 January 2007).

Sandom, C. and Enser, P. (2003). Archival Moving Imagery in the Digital Environment. In Anderson, J., Dunning, A., and Fraser, M (eds), *Digital Resources for the Humanities 2001–2002*. London: Office for Humanities Communication, King's College. http://www.cmis.brighton.ac.uk/research/vir/DRH2001.pdf (accessed 4 January 2007).

Sjölander, K. and Beskow, J. (2000). WaveSurfer – An Open Source Speech Tool. In *Proceedings of ICSLP*, Beijing, Oct 16–20, 4:464–467. http://www.speech.kth.se/wavesurfer/wsurf_icslp00.pdf (accessed 4 January 2007).

Sjölander, K. and Beskow, J. (2006). Wavesurfer. http://www.speech.kth.se/wavesurfer/ (accessed 4 January 2007).

St George's, Leeds (2006). Sermons. http://www.stgeorgesleeds.org.uk/church/sermons.htm (accessed 4 January 2007).

SWAG, Spoken Word Archive Group (2003). Report of the EU/US Working Group on Spoken Word Digital Audio. http://www.dcs.shef.ac.uk/spandh/projects/swag/swagReport.pdf (accessed 4 January 2007).

Tanghe, K., Lesaffre, M., Degroeve, S., Leman, M., De Baets, B., and Martens, J.-P. (2005) Collecting Ground Truth Annotations for Drum Detection in Polyphonic Music. In *Proceedings of the Sixth International Conference on Music Information Retrieval*, London, 50–57. http://ismir2005.ismir.net/proceedings/1006.pdf (accessed 4 January 2007).

The MathWorks (1994–2006). MATLAB®. http://www.mathworks.com/products/matlab/ (accessed 4 January 2007).

Transcriber (2006). A Tool for Segmenting, Labeling and Transcribing Speech. http://sourceforge.net/projects/trans/ (accessed 4 January 2007).

Tranter, S. and Reynolds, D. (2006). An Overview of Automatic Speaker Diarisation Systems. *IEEE Transactions on Speech and Audio Processing*, **14: 1557–65.** http://www.ll.mit.edu/IST/pubs/0511_Reynolds1.pdf (accessed 4 January 2007).

Tucker, S. and Whittaker, S. (2005). Novel Techniques for Time-compressing Speech: An Exploratory Study. *International Conference on Acoustics, Speech and Signal Proceessing*, Philadelphia. http://www.dcs.shef.ac.uk/~sat/downloads/ICASSP2005.pdf (accessed 4 January 2007).

Tzanetakis G. (n.d.). MARSYAS: Music Analysis, Retrieval and Synthesis for Audio Signals. http://opihi.cs.uvic.ca/marsyas/ (accessed 4 January 2007).

Tzanetakis, G. and Cook, P. (1999). MARSYAS: a Framework for Audio Analysis. *Organised Sound*, **4**: 169–175. http://www.cs.uvic.ca/~gtzan/work/pubs/organised00gtzan.pdf (accessed 5 July 2007).

UCLA Film & Television Archive (n.d.). Digital Hitchcock, http://www.cinema.ucla.edu/education/dighitch.html (accessed 4 January 2007).

UCSB, University of California Santa Barbara (2006). Cylinder Preservation Project. http://cylinders.library.ucsb.edu (accessed 4 January 2007).

van Leeuwen, D., Martin, A., Przymocki, M., and Bouten, J. (2006). NIST and NFI-TNO Evaluations of Automatic Speaker Recognition. *Computer Speech and Language*, **20**: 128–58.

Virage (2006). Virage Products Overview. http://www.virage.com/content/products/ (accessed 4 January 2007).

Volkmer, T. (2006). Efficient Video Annotation (EVA) System. http://domino.research.ibm.com/comm/research.nsf/pages/r.multimedia.innovation.html?Open&printable (accessed 4 January 2007).

WCER, Wisconsin Center for Education Research, University of Wisconsin (2006). Transana. http://www.transana.org/ (accessed 4 January 2007).

Youtube (2006). Broadcast Yourself. http://www.youtube.com/ (accessed 4 January 2007).