# An Overview of Audio Information Retrieval

Jonathan Foote
Institute of Systems Science,
National University of Singapore
Heng Mui Keng Terrace, Singapore 119597
(Now at FX Palo Alto Laboratory, Inc., 3400 Hillview Ave, Building 4, Palo Alto, California 94304)
jtfoote@bigfoot.com

December 18, 1997

**Abstract**

The problem of audio information retrieval is familiar to anyone who has returned from vacation to find an answering machine full of messages. While there is not yet an "AltaVista" for the audio data type, many workers are finding ways to automatically locate, index, and browse audio using recent advances in speech recognition and machine listening. This paper reviews the state-of-the-art in audio information retrieval, and presents recent advances in automatic speech recognition, word spotting, speaker and music identification, and audio similarity with a view towards making audio less "opaque." A special section addresses intelligent interfaces for navigating and browsing audio and multimedia documents, using automatically-derived information to go beyond the tape recorder metaphor.

## 1 Introduction

The problem of audio information retrieval is familiar to anyone who has returned from vacation to find an answering machine full of messages. If you are not fortunate, you may have to listen to the entire tape to find the urgent one from your boss. Determining there is no such message is guaranteed to be time-consuming.

While there is not yet an "AltaVista" for the audio data type, many workers are finding ways to automatically locate, manipulate, skim, browse, and index audio using recent advances in speech recognition and machine listening. Such methods will be indispensable to cope with the burgeoning amounts of audio available both on the Internet and elsewhere. (One German site[1] advertises more than three days' worth of available audio).

This paper will attempt to link a variety of research efforts across many fields, at a level of detail suitable for the non-specialist. A full list of references and URL pointers is given for the interested reader. The focus here will be on systems that can automatically extract information from an audio signal, rather than approaches that depend on human-generated

[1] http://www.icf.de/RIS/

annotations or decisions. (See, for example, the music recommendation service of Firefly[2] as an example of the latter). Because of the wide variety of disciplines involved, from speech recognition, information retrieval, audio analysis, signal processing, psychoacoustics, and machine learning, it is difficult to include all pertinent work, for which I must apologize in advance[3].

## 1.1 Technologies

A variety of technological advances can be used to make audio less "opaque," that is, provide some insight into the content of an audio file, and perhaps ways of using it as other than a monolithic block of digital data. The available methods can be roughly divided into those that assume some speech content in the audio, and those that don't. The general outline of this paper will follow this division; Section 2 will consider approaches based on automatic speech recognition (ASR), while Section 3 will consider more general audio analysis suitable for a wider range of audio such as music or sound effects (while not excluding speech). This parallels Smoliar *et alia*'s classification of multimedia retrieval into "expression" and "semantic" approaches [1]. In the first, objects are retrieved by some *physical* description, or an example from a similar medium. Thus most content-based image retrieval systems require an image or sketch as a search query. The second approach requires some *semantic* analysis or knowledge, for example annotations or picture captions. While semantic retrieval is perhaps the ideal, as it would hopefully retrieve what an intelligent human would, it requires either intensive human-assisted annotations or impractically sophisticated automatic content analysis (imagine the difficulty of automatically finding a photo of, say, Martin Luther King in a large database of uncaptioned images). Recent advances in ASR, however, have yielded audio retrieval systems that may approach the semantic ideal, as they can actually recognize—if not understand—the words uttered in an audio stream. Thus it is entirely possible to automatically locate audio of someone *talking* about Martin Luther King, or indeed one of his speeches.

**Information Retrieval** Conventional Information Retrieval (IR) research has been mainly based on (computer-readable) text ([2, 3]), and is familiar to many through the popular web search engines such as Lycos or AltaVista. The classic IR problem is to locate desired text documents using a search query consisting of a number of keywords. Typically, matching documents are found by locating query keywords within them. If a document has a high number of query terms, it is regarded as being more "relevant" to the query than other documents with fewer or no query terms. Documents can then be ranked by relevance and presented to the user for further exploration, as the web search engines do. Though powerful IR algorithms are available for text, it is clear that for audio, or multimedia in general, common term-matching approaches are useless due to the simple lack of identifiable words (or comparable entities) in audio documents. The problem becomes even more open-ended when one considers audio, such as music, which may have no speech.

Even when a desired audio document can be located in a large archive, another problem to overcome is the linearity of audio files. To ensure that nothing important is missed, the entire audio file must be auditioned from start to finish, which takes significant time. In contrast, the transcription of a minute-long message is typically a paragraph of text,

[2] http://www.firefly.com

[3] In particular, though there is much interesting work in the fields of psychoacoustics and Auditory Scene Analysis, I must reluctantly consider them as beyond the scope of this overview.

which may be scanned by eye in a matter of seconds. Even if there is a "fast-forward" button it is generally a hit-or-miss operation to find a desired section in a lengthy file. A typical interface treats audio as an undifferentiated stream: the "tape recorder" metaphor (with stop, play, rewind and fast-forward buttons) is ubiquitous. In contrast, most text-processing software has a "find" command that does simple string-match word searching to locate desired information in large files. Besides a scroll bar, many word processors can scroll by paragraph, page, function, or chapter. As above, an audio interface of similar flexibility is impossible unless suitable indexing entities, analogous to "words" or "pages," can be located in the audio. Section 4 discusses approaches to precisely this problem.

**Automatic Speech Recognition** Automatic Speech Recognition (ASR) is a technology rapidly coming out of the research laboratories into everyday use. While there have been decades of hard effort on the task[4], recent advances both in search algorithms and commonly-available computing power are rapidly making ASR practical. A perfect ASR system that could quickly transcribe spoken audio documents would be an ideal solution to most audio indexing and retrieval tasks (at least for speech). Such a system would would essentially reduce the audio retrieval problem to the straightforward text retrieval problem described above.

# 2 Speech Recognition

Practically all ASR systems in use are based on hidden Markov models (HMMs) [4]. A hidden Markov model is a statistical representation of a speech event like a word; model parameters are typically trained on a large corpus of labeled speech data. Given a trained set of HMMs, there exists an efficient algorithm for finding the most likely model sequence (the recognized words), given unknown speech data. This approach has proved successful not only for large-vocabulary recognition systems, but for "keyword-spotting" systems where the location of only a few words or phrases is desired. Typically, this is done by training HMMs for both the desired keywords and a "filler" model that attempts to match everything not a keyword [5, 6, 7]. Such systems can be both accurate and computationally far less expensive than a large-vocabulary recognition system, while being flexible enough to handle unconstrained real-world speech [8].

Large-vocabulary recognition systems, in contrast, typically use a "sub-word" approach: rather than building an explicit HMM for every one of the tens of thousands of words in the vocabulary, a few hundred sub-word models are used, typically phonetically-based. Given a phonetic dictionary, the appropriate sub-word models can be concatenated to form a word model. For example, the word "right" could be constructed by concatenating the three sub-word models for the phones "R AY T." In addition, a large-vocabulary system requires a statistical "language model" that defines likely word combinations. For example, in English the word bigram "of the" would be far more likely than the bigram "oaf the." The language model can thus constrain the recognizer to word combinations that are more likely, and hence more likely correct. To be useful, language models must be trained on example text (typically millions of words) from a similar domain, which may be more practical for certain domains such as news, where large corpora of newspaper text is available, than others such as conversational speech.

[4] Aleksandr Solzhenitsyn's novel *The First Circle* (1968) describes speech recognition research in a Stalin-era Soviet labor camp.

A particular advantage of using ASR for audio information retrieval is that—unlike dictation or voice-command tasks—most or all of the desired audio is already present, and thus ASR can be performed off-line rather than in real time. A disadvantage is that there can be orders of magnitude more data to actually recognize, and an ASR system sufficiently fast for dictation may be far too slow to use on hours of audio at search time.

The primary drawback of ASR systems is their limited accuracy. Though the best continuous speech recognition systems can achieve better than 90% word accuracy on carefully recorded, limited-domain tasks such as the Wall Street Journal corpus [9], similar systems achieve little better than 50% or 60% word accuracy on real-world tasks such as telephone conversations or news broadcasts [10, 11, 12, 13]. Even though word accuracy rates might appear unusably bad, the results of ASR transcription can still be surprisingly helpful for information retrieval. The reason for this is as follows: even with a 50% word error rate, the chance that a recognizer will miss every word in a three-word query (and thus a desired document) is $\left(\frac{1}{2}\right)^3 = .125$, and will be even lower assuming more relevant documents have more keyword occurrences. A similar effect has been termed "semantic co-occurrence filtering" [14]. Other approaches based on keyword spotting or subword units suffer similar problems. A particular drawback of subword- or phonetic-based systems is the extreme difficulty in correctly recognizing subword units such as phones. Information retrieval based on such systems must be robust to recognition errors.

## 2.1 Keyword Spotting

Automatically detecting words or phrases in unconstrained speech is usually termed "word spotting;" this technology is the foundation of several audio indexing efforts from a number of groups. A popular test is the SWITCHBOARD corpus[5], which contains recordings of spontaneous telephone conversations. Because recording subjects were asked to converse about certain topics (e.g. pets, weather, gun control), several research groups attempted to automatically determine the topic of each conversation. Workers at BBN have used both large-vocabulary ASR and keyword spotting to investigate topic identification [15]. Their approach allowed for a variable number of keywords, determined automatically; they report results from using from 100 to 2500 keywords. Their system results in 88.3% topic identification accuracy (from among 10 topics) given both sides of a SWITCHBOARD conversation are used.

A group at Ensigma used statistical models of keyword co-occurrence to discriminate between radio news topics such as sports or weather, with very good accuracy [16, 17]. The Video Mail Retrieval (VMR) group[6] at Cambridge University (which included the author) has investigated using 35 pre-selected keywords for audio information retrieval [18]. They report retrieval accuracy near 90% of that obtainable from a perfect wordspotter. Kate Knill, while at the Cambridge University Engineering Department, has been investigating fast keyword spotting for hand-held PDAs [19].

## 2.2 Sub-word indexing

Large-vocabulary ASR for audio indexing suffers, as we have seen, from several drawbacks: if a word is not present in the phonetic dictionary, it will not be recognized. Also, a language model must be used, and finding sufficient example text may not be possible. Thirdly,

[5] http://morph.ldc.upenn.edu/ldc/news/newsletter/v1.2/Switch.html
[6] http://svr-www.eng.cam.ac.uk/Research/Projects/vmr/

```
S: [um] IN DOCUMENTS, WHICH MIGHT BE video mail messages  say, [pause]
R: AND  IN DOCUMENT   WHICH MIGHT BE   THE DEAL MINISTRY  IS A

S: [um] AND THAT will tell you WHICH ONES are relevant. You  COULD compute
R: HOME AND THAT WOULD  REVIEW WHICH ONES OF REVENUE         COULD INCLUDE

S: SOME relevance score depending ON HOW CERTAIN you are of FINDING     THEM
R: SOME    OF ITS BOARD    RULING ON HOW CERTAIN OF         FINDINGS ON THEM

S: and so on. [pause] [um] [pause] ANOTHER APPROACH WOULD BE RATHER THAN
R:                                 ANOTHER APPROACH WOULD BE RATHER THAN

S: have A SET OF [pause] distinct DOCUMENTS you HAVE might have ONE
R:      A SET OF       THIS THING DOCUMENT  AND HAVE  MUCH OF   ONE

S: ongoing    SOURCE OF [pause] SOUND SUCH AS   [um] [pause] say a
R: ON GROWING SOURCE OF         SOUND SUCH AS A HOME         SO

S: news bulletin, [pause] BBC     nine o'clock news or something on the
R: NEWSPRINT IN           THE RECENT RESULTS AND

S: RADIO. [loud_breath] [um] [pause] AND THEN and YOU COULD [pause]
R: RADIO                FIRM         AND THEN     YOU COULD

S: listen to THE WHOLE OF IT [um] [pause] AND if you suddenly spotted
R: LOSE INTO THE WHOLE OF ITS OWN         AND  A FEW SOUTHERN IS OFF TWO

S: [pause] [um] THE KEY WORD IN THAT [pause] [um] [pause] then then   you
R:         AND  THE KEY WORD IN THAT        FOUND        THAT THE END TO

S: DECIDE that THE SURROUNDING AREA OF text, [pause] [um] I mean MAYBE
R: DECIDE OF   THE SURROUNDING AREA OF TAX            ON    LINE MAY BE
```

Figure 1: Automatic transcription of spontaneous speech. S: spoken words (lower-case are misrecognized words) R: recognized speech

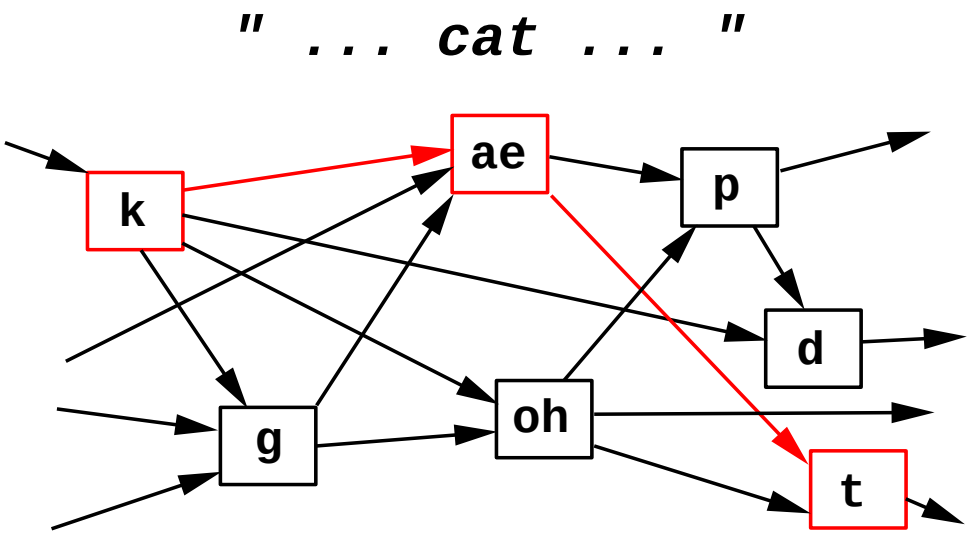


Figure 2: Example phone lattice for utterance "cat"

large-vocabulary ASR may be very expensive in terms of computation and storage (though recent advances in search algorithms are making this much less of a concern). Though this may be acceptable for typical speech recognition applications such as dictation, it is clearly unacceptable to incur several hours of computation when searching an audio corpus of similar length. To avoid some of these drawbacks, several alternatives to large-vocabulary ASR have been pursued. A common feature is the use of sub-word indexing units, such as phones or phone clusters. Typically these are smaller than words, hence there are fewer possible units, dramatically reducing the search space. An unfortunate drawback is that as units get smaller the recognition accuracy will typically decrease as well. Saving the cost of building a detailed language model will unfortunately impact recognition accuracy.

A group at ETH Zürich[7], has reported speech retrieval using a large set of automatically-chosen vowel-consonant-vowel subword units as indexes [20, 21]. A recent paper from MIT's Spoken Language Systems Group[8] investigates a number of possible subword units, from sequences of phones or broader phone classes, automatically-discovered phone "multigrams," to multiple-phone syllable units [22]. Retrieval performance is again related to recognition accuracy; results show no clear benefit of any particular unit, though some might be better than others for certain domains.

Another promising approach is "lattice-based" word spotting. A lattice is a compact representation of multiple best hypothesis generated by a phone or word recognition system. If the phone lattice is generated before need, it can then be searched extremely rapidly to find phone strings corresponding to desired query words. James reports the lattice scanner approach working about 1000 times real time; in other words an hour of audio may be searched in 3.6 seconds [23]. Figure 2 shows an example lattice for the word "cat." Keeping multiple hypotheses makes the system much more robust to recognition errors. For example, in the figure, even though the phone "T" was not the first choice, it is still in the lattice, thus the phone string K AE T can still be found. If the lattice contains too many hypotheses,

[7] http://www-ir.inf.ethz.ch/
[8] http://sls-www.lcs.mit.edu/


Figure 3: PARC audio browser with speaker segmentation

however, recognition accuracy will suffer from too many false alarms, as words that were not uttered can be found in a deep enough lattice. For example, the word "goat" (G OH T) can be found in the lattice of Figure 2, even though the uttered word was presumably "cat." The Video Mail Retrieval group at Cambridge University has successfully used phone lattices for open-vocabulary voice message retrieval [24, 25]; other workers have investigated lattice-based keyword spotting on the TIMIT corpus [26].

## 2.3 Large-vocabulary ASR

Several research groups have used large-vocabulary recognition for audio characterization. Dragon Systems[9] were perhaps the first to use large-vocabulary recognition for speech document characterization; in 1993 they investigated topic identification on the Switchboard corpus [27]. A large effort is the Informedia project[10] at Carnegie Mellon University. There, a combination of verbatim text transcriptions and large-vocabulary recognizer output are used for search indexes of video data such as television broadcasts [28] — Section 4.5 describes the the Informedia project in more detail. The Video Mail Retrieval group at Cambridge University has investigated combining small vocabulary keyword spotting (35 keywords) with large-vocabulary ASR for voice mail retrieval. Their findings suggest that a combination of the two approaches can be superior to either alone, as the keyword spotting allows detection of words not in the large-vocabulary lexicon [10]. Figure 1 shows the results of using a 5,000-word large-vocabulary recognizer on a spontaneously-spoken video mail message. Note that the language model is clearly inappropriate for the domain. Section 4.4 describes more about the VMR project.

## 2.4 Speaker Identification

A somewhat easier proposition than speech recognition is to simply identify differences between voices, rather than determine what the voices are saying. Technology to do this, termed *speaker identification* or *speaker ID*, can be very accurate in the right circumstances [29]. The applications to audio indexing are immediate: work at Xerox PARC allowed recorded meetings to be segmented and analyzed by speaker. A timeline display showed

[9] http://www.dragonsys.com/
[10] http://informedia.cs.cmu.edu/

when particular speakers were talking during the meeting, as well as random-access to play back a desired portion of the recording [30, 31]. Figure 3 shows how various speakers are displayed versus time [32]. Note how non-speech audio events like silence and applause may be located as well, which may be important cues for automatic segmentation.

A novel application of speaker ID was to align multi-hour recordings of the US House of Representatives with the text transcription published as the Congressional Record. Because the published text has been corrected and amended, it is often of interest to determine what a given lawmaker actually said. By using a dynamic-programming approach to align the audio record with the (differing) printed version, different speaker turns can be accurately located, adding much to the utility of the audio transcript [33].

Segmenting multimedia streams is a promising area for speaker identification. If the ID technique works well enough on a sub-second time scale, it could be used to detect speaker changes in the soundtrack of a video or multimedia source, allowing it to be indexed and some sort of structure (such as dialogs) to be determined. Wyse and Smoliar report a "novelty measure" based on the cepstral difference between a short (.75 s) and long (3 s) analysis window [34]. Differences are only compared in similar regions of the feature-space to prevent intra-speaker variation (such as different vowels) from generating a high novelty score. When the difference exceeds a threshold, this can signal a new speaker or a significant change in the audio stream. These could be the audio equivalents of scene or camera changes, cuts, fades and wipes. It should be possible to fuse data intelligently extracted from both the video and audio streams, yielding more complete and robust information (about key frames, for example) than is available from either mode alone. A group at the University of Mannheim has been looking at the automatic analysis of film and video soundtracks. They have presented a system for automatic film genre classification based on low-level video and long-term audio frequency and amplitude characteristics [35]. In more recent work, they have attempted to automate violence detection in movie soundtracks by recognizing shots, cries and explosions. This was done by matching characteristics such as amplitude, frequency, and pitch [36].

# 3 Music and Audio Analysis

While ASR can give valuable clues to the content of speech, the universe of possible audio is, of course, much wider than speech alone. Music is a large, and extremely variable (and hence challenging) audio class[11]. Considering the range of sounds that people might want to archive or classify, from the gamut of musical genres through sound effects to animal cries to synthesizer samples, and it is clear that speech-based methods alone are woefully inadequate for general audio discovery. Complicating the task is that any of the above can and will occur in combination; e.g. a narrator speaks over music or natural sounds, a translated version of a speech may be mixed over the original, and widely different sounds may occur sequentially in the same stream.

## 3.1 Music Discrimination

A general problem in audio analysis is to simply discriminate speech from non-vocal music or other sounds. This has immediate applications for speech recognition: in general, there

[11] Few can even agree on a satisfactory definition of "music"; consider John Cage's controversial composition *4' 33"* consisting of four minutes and thirty-three seconds of silent performance.

is no guarantee that a given multimedia audio source contains speech, and it is important not to waste valuable resources attempting to perform speech recognition on music, silence, or other non-speech audio. A straightforward approach to discriminating music from speech was presented by John Saunders [37] based on the statistics of the energy contour and the zero-crossing rate. Saunders reports a 98% classification accuracy on commercial radio broadcasts. Eric Scheirer of the MIT Media Lab and Malcolm Slaney of Interval Research report a speech/music discriminator based on various combinations of 13 features, such as 4 Hz modulation energy, "spectral centroid," and zero-crossing rate [38]. Several classification strategies, including Gaussian mixture models and K-nearest-neighbor classifiers were evaluated. They report a 1.4% error rate on a large and diverse collection of FM radio broadcasts, when looking at relatively long-term (2.4 s) windows. Michelle Spina and Victor Zue of MIT's Spoken Language Systems Group report experiments on two hours of NPR radio news[39]. Using a maximum *a posteriori* approach on mel-frequency cepstral coefficients, the authors report an 80.9% classification accuracy into seven classes, including clean, telephone, and noisy speech, silence, music, and speech plus music. They include a plot of classification performance versus window size, and report that a window size near 0.5 second resulted in optimal classification. Finally, on a much less extensive data set consisting of short samples (0.5 to 5 s) of music or speech, the author reports excellent discrimination accuracy using a classifier based on the statistics derived from a MMI-based vector quantizer [40], albeit on a much smaller sample of audio clips.

# 4 Advanced Audio Interfaces

Arguably more useful than classifying audio would be a system that would let users find audio items of interest, from either a large database of recordings or within a long recording itself. This section presents approaches to audio information retrieval[12]. Many use the "ranked list" user interface familiar from Web search engines. Even when a potentially relevant audio item is found, there are better ways of browsing the audio than the "tape recorder" metaphor, whose controls are typically limited to stop, play, and fast-forward and rewind.

## 4.1 Speech Skimmer

Barry Arons's SpeechSkimmer is an excellent example of pushing audio interaction beyond the tape recorder metaphor [41]. It has long been known that humans can understand speech much quicker than the rate it is typically spoken. Using time-compression processing that alters the audio playback rate without changing the pitch, SpeechSkimmer users can audition spoken documents at several times real-time, as well as backwards. Pauses can be identified (and removed) by detecting speech, using an adaptive algorithm said to be robust to background noise. In addition, the audio may be segmented by analyzing the speaker's pitch to find cues associated with new topics, or by using speaker identification to locate conversational turns. SpeechSkimmer's hierarchy of summarization techniques enable, in Aron's words, a "fish ear" view of an audio document.

[12] Though an interesting topic, we will not consider mere playback interfaces that do not extract information from the audio.

## 4.2 Audio Retrieval-by-Content

Given the proliferation of audio databases on the Internet and elsewhere (some commercial sound effects libraries contain as many as 100 CDs), there is interest in doing for sound what Web search engines do for text. This requires some measure of audio similarity, which is a complicated and subjective matter. Measures of text similarity can be simple as counting the number of words in common. Most approaches to general audio retrieval take a perceptual approach, using measures derived from the audio that reflect perceptual characteristics such as brightness or loudness. A group at Technische Universität Berlin has used a neural net to map a sound clip to a text description, which could be inverted to find sounds by description [42]. An obvious drawback here is the subjective nature of audio descriptions. Sounds that a particular listener describes as "sharp" may be quite different from another's. A later paper used a self-organizing map (SOM) on perceptually-derived spectral features [43]. The net effect was to organize a set of 100 sample-synthesizer sounds into a 2-D matrix such that similar sounds were closer and more disparate sounds were found further away on the grid.

Work by a group at Muscle Fish LLC[13] has resulted in a compelling audio retrieval-by-similarity demonstration[14]. Muscle Fish's approach is to analyze sound files for a specific set of psychoacoustic features. This results in a vector of attributes that include loudness, pitch, bandwidth and harmonicity [44]. Given enough training samples, a Gaussian classifier can be constructed, or for retrieval, a covariance-weighted Euclidean (Mahalonobis) distance is used as a measure of similarity. For retrieval, the distance is computed between a given sound example and all other sound examples (about 400 in the demonstration). Sounds are ranked by distance, with the closer ones being more similar.

Recent work by the author, using an entirely different approach, has resulted in similar retrieval framework. Here distance measures are computed between histograms derived from a discriminatively-trained vector quantizer. Audio is first parameterized into a spectral representation (mel-frequency cepstral coefficients). A learning algorithm then constructs a quantization tree that attempts to put samples from different training classes into different bins. A histogram of an audio file can be made by looking at the relative frequencies of samples in each quantization bin. If histograms are considered vectors, then simple Euclidean or cosine measures can be used to determine similarity between them, and hence their source audio [40]. This approach has been used for speaker identification [45, 46] as well as music and audio retrieval [47].

The Muscle Fish and author's retrieval performance have been compared on the same audio corpus of 409 short sound files [47]. When retrieving simple, mono-component sound like isolated instrument samples, the Muscle Fish retrieved sounds of similar timbre but varying pitch, while the author's approach retrieved sounds of similar pitch from instruments of varying timbre. This demonstrates the subjective nature of audio similarity: it is not clear which criterion is more important—the appropriate choice is probably application-dependent. (Similarly for image retrieval, the relative importance of shape vs. color is not clear.)

[13] http://www.musclefish.com

[14] http://www.musclefish.com/cbrdemo.html

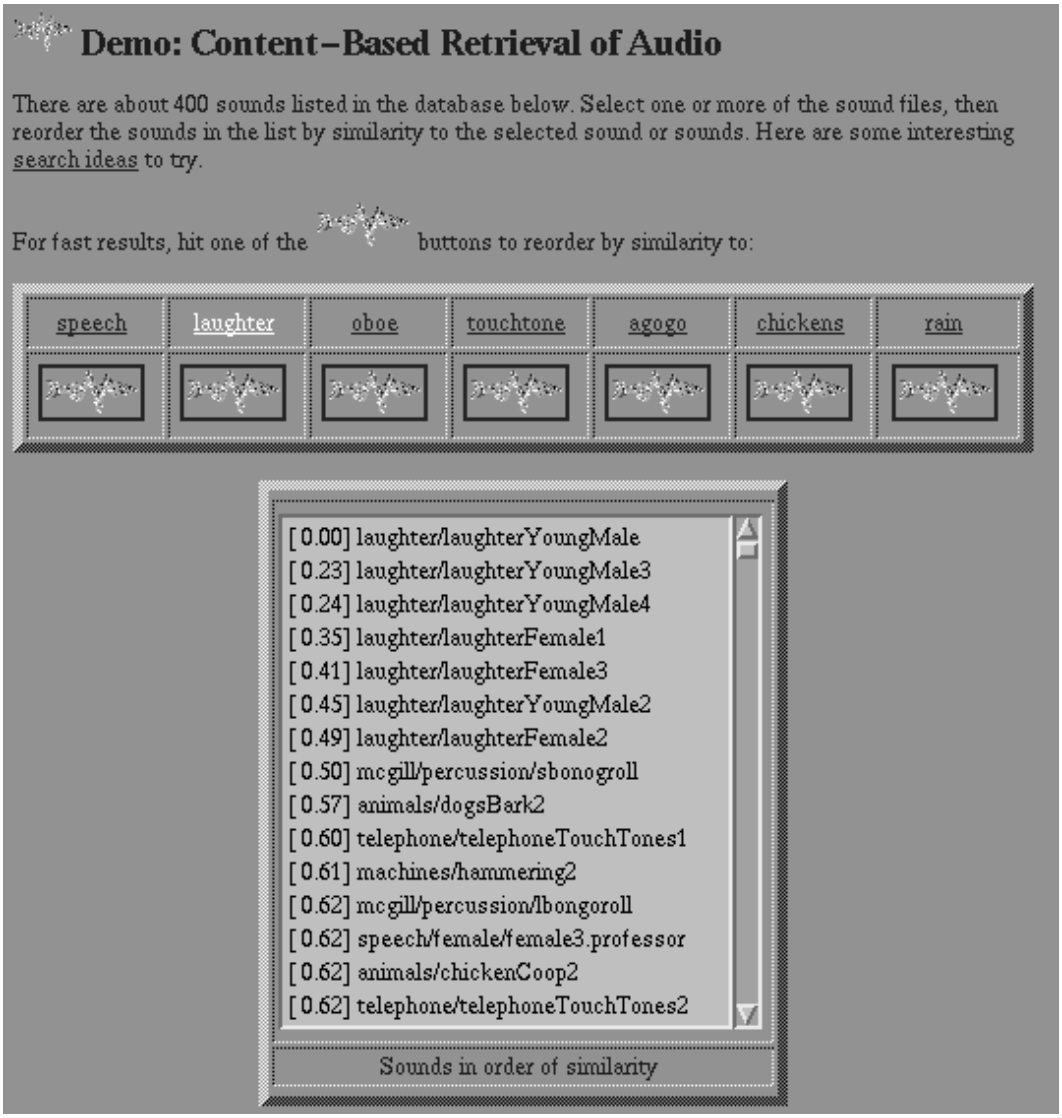


Figure 4: Example Muscle Fish audio retrieval results, searching for audio similar to "laughter" (from http://www.musclefish.com)

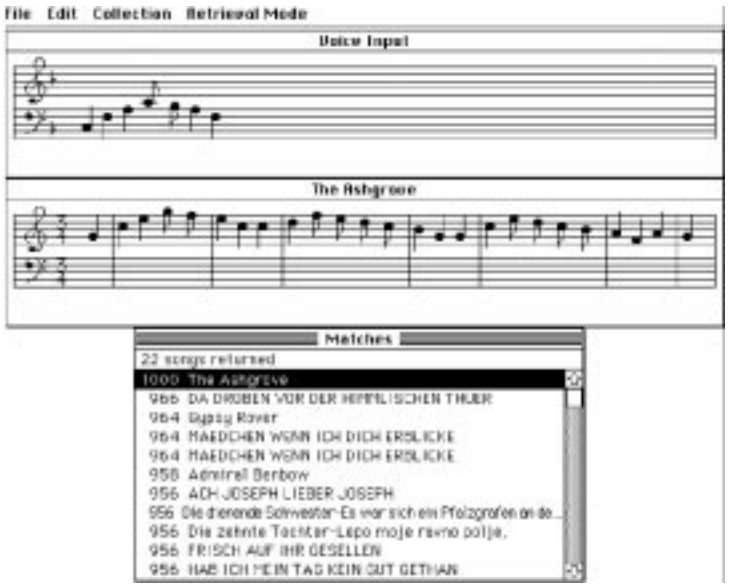


Figure 5: Interface of Waikato tune retrieval application.

## 4.3 Music and MIDI retrieval

While information retrieval for text relies on simple text queries, the structure of a query for sound or music is not so obvious. Though textual descriptions can be assigned to sounds, they are not always obvious or indeed well-defined. The content-based retrieval applications of the previous Section have avoided the problem somewhat by using audio examples as the query, in effect saying "look for things that sound like *this*." Some recent work improves on this by allowing the user to sing or whistle a desired tune. Pitch extraction algorithms convert this to a note-like representation, which can be used to query a database of music. Unfortunately, extracting score-like attributes for anything but the simplest pieces has proved extremely difficult. Researchers in the area have finessed this problem by using archives of MIDI (Musical Instrument Digital Interface) files, which are score-like representations of music intended for musical synthesizers or sequencers. Given a melodic query, then, the MIDI files can be searched for similar melodies. Researchers at Cornell report surprisingly effective retrieval using query melodies that have been quantized to three levels, depending on whether each note was higher, lower, or similar pitch as the previous one [48]. Besides simplifying the pitch extraction, this allows for less-than-expert singing ability on the part of the user! A similar, if more advanced, application has been developed the University of Waikato in New Zealand [49]. The Waikato system uses flexible string-matching algorithms to locate similar melodies located anywhere in a piece. Figure 5 (after McNab et al. [50]) shows the recognized tune sung by the user on the top staff, and the retrieved musical tunes[15].

[15]The tune retrieval application (for Macintosh computers) may be downloaded at ftp://ftp.cs.waikato.ac.nz/pub/mac/MRv2.0.1.sea.hqx

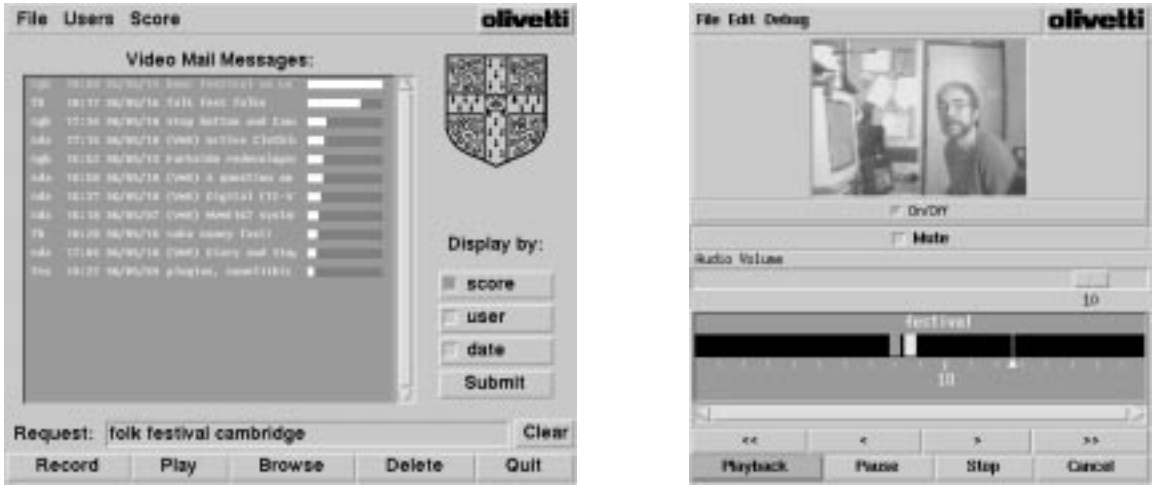


Figure 6: Video Mail Retrieval user interface. Search engine (left) and mail browser(right).

## 4.4 Video Mail Retrieval

As mentioned in Section 2.1, the VMR group at Cambridge University has used open-vocabulary word spotting (based on phone lattices) to retrieve messages from an archive of video mail. In operation, the user types in a text search query, exactly as in a Web search engine. The lattice-based word spotter finds instances of each query word spoken in each message (if any). A matching score for each message is computed by the retrieval engine, and the interface then displays a list of messages ranked by score. Scores are represented by bar graphs, as in the left image of Figure 6 (after Brown et al. [51]).

After the ranked list of messages is displayed, the user must still investigate the listed messages to either find the relevant one(s). The video mail browser (the right image for Figure6) is an attempt to represent a dynamic time-varying process (the audio/video stream) by a static image that can be taken in at a glance. A message is represented as horizontal timeline, and events are displayed graphically along it. Putative keyword hits are displayed along the timeline, as in Figure 6. The timeline is the black bar; the scale indicates time in seconds. When pointed at with the mouse, keyword names are highlighted in white (so it may be read in the presence of many other keyword hits). Clicking on the desired time in the time bar starts message playback at that time; this lets the user selectively play regions of interest, rather than the entire message.

## 4.5 Informedia

The Informedia project[16] at Carnegie Mellon University is an impressive combination of video and audio analysis and text-based information retrieval techniques [52, 28, 12]. Given a video broadcast and a text transcription (from production notes; closed-captions are insufficiently accurate), a HMM-based approach can accurately time-align the spoken words with the transcript. The Informedia project uses this to provide a abstract of the video by extracting both key frames from the video and important words from the text. Word impor-

[16] http://http://www.informedia.cs.cmu.edu/

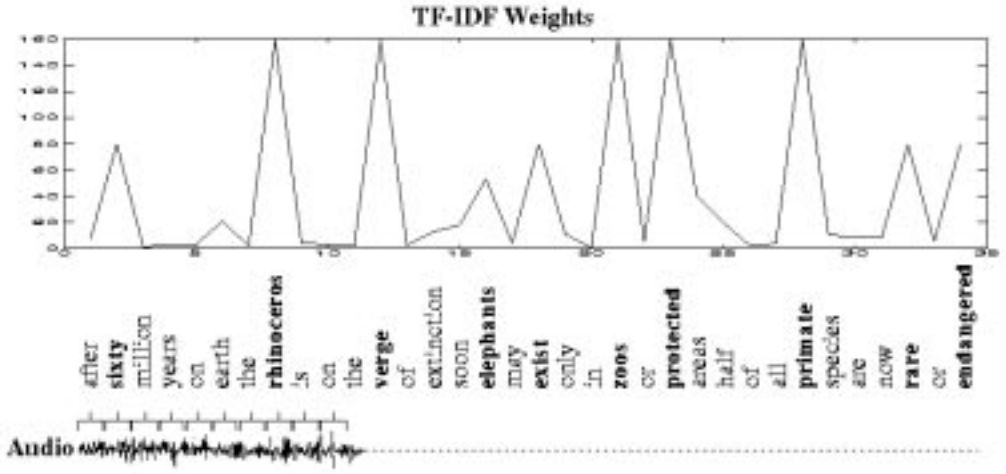


Figure 7: Informedia term weighting approach.

tance is determined from *tf/idf weights*, which stands for "term frequency/inverse document frequency." Basically, terms (words) which appear frequently in a single document have a high "term frequency" (tf) and are thus more important. Similarly, terms which appear across many documents in the collection are less important, and have a lower "document frequency" (df). The product of tf and the inverse of df gives a measure of word importance, as shown in Figure 7. Extracting words with high tf/idf from the audio stream and concatenating them yields a digest that (hopefully) contains the gist of the audio, and which can be auditioned extremely rapidly. Also conventional text-based retrieval techniques can be used to locate desired broadcasts or relevant portions within them.

# 5 Future Directions

This paper has described some novel and powerful ways of extracting information from audio data. Clearly there is a lot of room for further research. One promising direction is the inclusion of Radio Broadcast News into the DARPA speech recognition evaluation effort (CSRIV Hub-4). This will encourage speech recognition researchers to tackle the difficult problem of general audio, rather than just clean speech as has been the case with most research to date. Audio information retrieval is also now a sub-task of the TREC (Text Retrieval Conference), encouraging those in the text retrieval community to consider audio as well. Some interesting work, somewhat beyond the scope of this review, is being done in automatic translation of audio and text, as well as cross-language information retrieval. A particularly exciting area is how to combine information from various modes, such as audio and video. Efforts like that at the Informedia project hint at the power that fusing low-level information from different media can bring to the general problems of multimedia recognition, segmentation, and retrieval. Research at Ryukoku University in Japan, which deserves a wider audience, has pioneered combining video analysis and word spotting to segment articles from TV news [53]. Finally, issues of scale will need to be addressed: how well do various methods cope with vast numbers of large documents? As multimedia archives proliferate on the WWW and elsewhere, technology like that presented here will

be become indispensable to locate, retrieve, and browse audio and multimedia information.

# References


[1] S. W. Smoliar, J. D. Baker, T. Nakayama, and L. Wilcox. Multimedia search: An authoring perspective. In *Proceedings of the First International Workshop on Image Databases and Multimedia Search*, pages 1–8. IAPR, August 1996.

[2] C. J. van Rijsbergen. *Information Retrieval.* Butterworths, London, 2nd edition, 1979.

[3] W. Frakes and R. Baeza-Yates. *Information Retrieval: data structures and algorithms.* Prentice Hall, New Jersey, 1992.

[4] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 77(2):257–286, February 1989.

[5] R. C. Rose and D. B. Paul. A hidden Markov model based keyword recognition system. In *Proc. ICASSP 90*, pages 129–132. IEEE, 1990.

[6] L. D. Wilcox and M. A. Bush. Training and search algorithms for an interactive wordspotting system. In *Proc. ICASSP 92*, volume II, pages 97–100, San Francisco, 1992. IEEE.

[7] P. Jeanrenaud, K. Ng, M. Siu, J. Rohlicek, and H. Gish. Phonetic-based word spotter: Various configurations and application to event spotting. In *Proc. Eurospeech 93*, pages 2145–2148, Berlin, 1993. ESCA.

[8] R. C. Rose. Techniques for information retrieval from speech messages. *Lincoln Laboratory Journal*, 4(1):45–60, 1991.

[9] D. Pallett et. al. Benchmark tests for the ARPA spoken language program. In *Proc. ARPA SLS Technology Workshop*, January 1995.

[10] G. J. F. Jones, J. T. Foote, K. Spärck Jones, and S. J. Young. Robust talker-independent audio document retrieval. In *Proc. ICASSP 96*, volume I, pages 311–314, Atlanta, GA, April 1996. IEEE.

[11] P. Jeanrenaud, E. Eide, U. Chaudhari, J. McDonough, K. Ng, M. Siu, and H. Gish. Reducing word error rate on conversational speech from the Switchboard corpus. In *Proc. ICASSP 95*, pages 53–56, Detroit, May 1995. IEEE.

[12] A. Hauptmann and M. Witbrock. Informedia: News-on-demand — multimedia information acquisition and retrieval. In M. T. Maybury, editor, *Intelligent Multimedia Information Retrieval*, chapter 10. MIT Press, 1997.
http://www.cs.cmu.edu/afs/cs/user/alex/www/.

[13] F. Kubala, H. Jin, S. Matsoukas, L. Nguyen, and R. Schwartz. Broadcast news transcription. In *Proc. ICASSP 97*, volume 1, pages 203–206. IEEE, April 1997.

[14] J. Kupiec, D. Kimber, and V. Balasubramanian. Speech-based retrieval using semantic co-occurrence filtering. In *Proc. ARPA Human Language Technology Workshop*, Plainsboro, NJ, March 1994. ARPA.

[15] J. McDonough, K. Ng, P. Jeanrenaud, H. Gish, and J. R. Rohlicek. Approaches to topic identification on the Switchboard corpus. In *Proc. ICASSP 94*, volume I, pages 385–388, Adelaide, 1994. IEEE.

[16] J. H. Wright, M. J. Carey, and E. S. Parris. Improved topic spotting through statistical modelling of keyword dependencies. In *Proc. ICASSP 95*, pages 313–316, Detroit, MI, May 1995. IEEE.

[17] J. H. Wright, M. J. Carey, and E. S. Parris. Topic discrimination using higher–order statistical models of spotted keywords. *Computer Speech and Language*, 9(4):381–405, Oct 1995.

[18] G. J. F. Jones, J. T. Foote, K. Spärck Jones, and S. J. Young. Video Mail Retrieval: the effect of word spotting accuracy on precision. In *Proc. ICASSP 95*, volume I, pages 309–312, Detroit, May 1995. IEEE.

[19] K. M. Knill and S. J. Young. Speaker dependent keyword spotting for hand-held devices. Technical Report 193, Cambridge University Engineering Department, July 1994.

[20] P. Schäuble and M. Wechsler. First experiences with a system for content based retrieval of information from speech recordings. In *IJCAI Workshop: Intelligent Multimedia Information Retrieval*, August 1995.
ftp://ftp.inf.ethz.ch/pub/publications/papers/is/ir/ijcai95.ps.gz.

[21] M. Wechsler and P. Schäuble. Speech retrieval based on automatic indexing. In C. J. van Rijsbergen, editor, *Proceedings of the MIRO Workshop*, University of Glasgow, September 1995.

[22] K. Ng and V. Zue. Subword unit representations for spoken document retrieval. In *Proc. Eurospeech 97*. ESCA, 1997.
http://www.sls.lcs.mit.edu/~kng/papers/sir-eurospeech97.ps.

[23] D. A. James and S. J. Young. A fast lattice-based approach to vocabulary independent wordspotting. In *Proc. ICASSP 94*, volume I, pages 377–380, Adelaide, 1994. IEEE.

[24] J. T. Foote, G. J. F. Jones, K. Spärck Jones, and S. J. Young. Unconstrained keyword spotting using phone lattices. *Computer, Speech and Language*, 1997. In press.

[25] S. J. Young, M. G. Brown, J. T. Foote, G. J. F. Jones, and K. Spärck Jones. Acoustic indexing for multimedia retrieval and browsing. In *Proc. ICASSP '97*, volume 1, pages 199–202, Munich, Germany, April 1997.

[26] P. Gelin and C. Wellekens. Keyword spotting for video soundtrack indexing. In *Proc. ICASSP 96*, volume I, pages 299–302, Atlanta, GA, April 1996. IEEE.

[27] L. Gillick, J. Baker, and J. Bridle et al. Application of large vocabulary continuous speech recognition to topic and speaker identification using telephone speech. In *Proc. ICASSP 93*, volume II, pages 471–474, San Francisco, May 1993. IEEE.

[28] M. A. Smith and M. G. Christel. Automating the creation of a digital video library. In *Proc. ACM Multimedia 95*, pages 357–358, San Francisco, November 1995. ACM.

[29] S. Furui. An overview of speaker recognition technology. In *ESCA Workshop on Automatic Speaker Recognition, Identification, and Verification*, pages 1–9. ESCA, April 1994.

[30] L. Wilcox, F. Chen, and V. Balasubramanian. Segmentation of speech using speaker identification. In *Proc. ICASSP 94*, volume S1, pages 161–164, April 1994.

[31] D. Kimber and L. Wilcox. Acoustic segmentation for audio browsers. In *Proc. Interface Conference*, Sydney, Australia, July 1996.
http://www.fxpal.xerox.com/abstracts/kim96.htm.

[32] F. Chen, M. Hearst, D. Kimber, J. Kupiec, J. Pedersen, and L. Wilcox. *Managing Multimedia Data: Using Metadata to Integrate and Apply Digital Data*, chapter Metadata for Mixed Media Access. McGraw Hill, 1997.

[33] D. Roy and C. Malamud. Speaker identification based text to audio alignment for an audio retrieval system. In *Proc. ICASSP 97*, Munich, Germany, April 1997. IEEE.

[34] L. Wyse and S. Smoliar. Toward content-based audio indexing and retrieval and a new speaker discrimination technique. In *Proc. ICJAI '95*, Montreal, 1995.

[35] S. Fischer and W. Effelsberg. Automatic film genre classification. In *Proc. ACM Multimedia '95*, pages 295–304, San Francisco, November 1995.

[36] S. Pfeiffer, S. Fischer, and W. Effelsberg. Automatic audio content analysis. Technical Report TR-96-008, University of Mannheim, D-68131 Mannheim, Germany, April 1996.
ftp://pi4.informatik.uni-mannheim.de/pub/techreports/1996/TR-96-008.ps.gz.

[37] John Saunders. Real-time discrimination of broadcast speech/music. In *Proc. ICASSP '96*, volume II, pages 993–996, Atlanta, May 1996. IEEE.

[38] E. Scheirer and M. Slaney. Construction and evaluation of a robust multifeature music/speech discriminator. In *Proc. ICASSP 97*, volume II, pages 1331–1334. IEEE, April 1997.

[39] M. Spina and V. Zue. Automatic transcription of general audio data: Preliminary analyses. In *Proc. International Conference on Spoken Language Processing*, pages 594–597, Philadelphia, PA, October 1996.

[40] J. T. Foote. A similarity measure for automatic audio classification. In *Proc. AAAI 1997 Spring Symposium on Intelligent Integration and Use of Text, Image, Video, and Audio Corpora*, Stanford, Palo Alto, CA, March 1997.

[41] B. Arons. SpeechSkimmer: A system for interactively skimming recorded speech. *ACM Trans. on Computer Human Interaction*, 4(1):3–38, March 1997.
http://barons.www.media.mit.edu/people/barons/papers/ToCHI97.ps.

[42] B. Feiten and T. Ungvary. Organizing sounds with neural nets. In *Proc. 1991 Int. Computer Music Conf.*, San Francisco, 1991. Internation Computer Music Association.

[43] B. Feiten and S. Günzel. Automatic indexing of a sound database using self-organizing neural nets. *Computer Music Journal*, 18(3):53–65, 1994.

[44] E. Wold, T. Blum, D. Keslar, and J. Wheaton. Content-based classification, search, and retrieval of audio. *IEEE Multimedia*, pages 27–36, Fall 1996.

[45] J. T. Foote and H. F. Silverman. A model distance measure for talker clustering and identification. In *Proc. ICASSP '94*, volume S1, pages 317–32, Adelaide, Australia, April 1994. IEEE.

[46] J. T. Foote. Rapid speaker identification using discrete MMI feature quantisation. *Expert Systems with Applications*, 13(4), 1998.

[47] J. T. Foote. Content-based retrieval of music and audio. In C.-C. J. Kuo et al., editor, *Multimedia Storage and Archiving Systems II, Proc. of SPIE*, volume 3229, pages 138–147, 1997.

[48] A. Ghias et al. Query by humming. In *Proc. ACM Multimedia 95*, San Francisco, November 1995.

[49] R. McNab, L. Smith, I. Witten, C. Henderson, and S. Cunningham. Towards the digital music library: Tune retrieval from acoustic input. In *Proc. Digital Libraries '96*, pages 11–18, 1996.
http://www.cs.waikato.ac.nz/~rjmcnab/papers/mt.ps.gz.

[50] R. McNab, L. Smith, I. Witten, and C. Henderson. Tune retrieval in the multimedia library.
http://www.cs.waikato.ac.nz/~rjmcnab/papers/mmtools.ps.gz, 1997.

[51] M. G. Brown, J. T. Foote, G. J. F. Jones, K. Spärck Jones, and S. J. Young. Open-vocabulary speech indexing for voice and video mail retrieval. In *Proc. ACM Multimedia 96*, pages 35–43, Boston, November 1996. ACM.

[52] A. Hauptmann, M. Witbrock, A. Rudnicky, and S. Reed. Speech for multimedia information retrieval. In *Proc. UIST–95 User Interface Software and Technology*, Pittsburgh, PA, 1995.
http://informedia.cs.cmu.edu/research/uist95.ps.

[53] Sakurai Ariki. Article extraction and classification of TV news using image and speech processing. In *International Symposium on Cooperative Database Systems for Advanced Applications (CODAS-96)*, Kyoto, 1996.
http://banjo.kuis.kyoto-u.ac.jp/~tarumi/juten/J/event/kyoto-camera/L006.ps.