Automatic Detection, Indexing, and Retrieval of Multiple Attributes from Cross-lingual Multimedia Data

Qian Hu, Fred J. Goodman, Stanley M. Boykin, Randall K. Fish, Warren R. Greiff, Stephen R. Jones, Stephen R. Moore

The MITRE Corporation
202 Burlington Road, Bedford, MA. USA

Email: {qian, fgoodman, sboykin, fishr, greiff, srjones, srmoore} @mitre.org

Abstract

The availability of large volumes of multimedia data presents many challenges to content retrieval. Sophisticated modern systems must efficiently process, index, and retrieve terabytes of multimedia data, determining what is relevant based on the user's query criteria and the system's domain specific knowledge. This paper reports our approach to information extraction from cross-lingual multimedia data by automatically detecting, indexing, and retrieving multiple attributes from the audio track. The multiple time-stamped attributes the Audio Hot Spotting system automatically extracts from multimedia include speech transcripts and keyword indices, phonemes, speaker identity (if possible), spoken language ID and automatically identified non-lexical audio cues. The non-lexical audio cues include both non-speech attributes and background noise. Non-speech attributes include speech rate, vocal effort (e.g. shouting and whispering), which are indicative of the speaker's emotional state, especially when combined with adjacent keywords. Background noise detection (such as laughter and applause) is suggestive of audience response to the speaker.

In this paper, we describe how the Audio Hot Spotting prototype system detects these multiple attributes and how the system uses them to discover information, locate passages of interest within a large multi-media and cross-lingual data collection, and refine query results.

1. Introduction

The availability of large volumes of multimedia data presents many challenges to content retrieval. Sophisticated modern systems must efficiently process, index, and retrieve terabytes of multimedia data, determining what is relevant based on the user's query criteria and the system's domain specific knowledge. Research systems for spoken document retrieval (SDR) have been developed and evaluated against radio and broadcast news data through TREC SDR 6-9. These systems primarily rely on the automatic speech recognition (ASR) transcribed text for retrieval purposes and return whole documents or stories [4, 12, 17]. Our objective is to return passages within documents rather than entire documents, based upon an extension of attributes beyond the ASR transcript [10, 11]. In addition to the transcribed speech, we have added phoneme-based recognition, speaker and language identification, and less obvious non-lexical cues such as speech rate, shouting and whispering, laughter and applause [7, 8, 9]. Our intent is to begin to explore the rich non-lexical information available in the audio track. In high transcription error rate domains, reading the ASR output is problematic. When users wish to listen to the original audio, we believe that optimum productivity depends upon the system's ability to retrieve the desired portion of the audio file at its precise location rather than just identifying the entire document or story segment in which the event occurred [26]. As described later, we locate and return short passages of text and original media conducive to rapid review and evaluation. Finally, we have experimented with task conditions more challenging than the broadcast news quality of the TREC evaluations including:

- Different language models and genres: lectures, spontaneous speech, meetings etc.
- Multimedia sources
- Multimedia in different languages
- Uncontrolled acoustic environment with ambient background noise and multiple speakers

Early audio information retrieval systems applied straightforward text-based queries to transcripts produced by automatic speech recognition [1,3,4,12,17]. However, audio contains more information than is conveyed by the text transcript produced by an automatic speech recognizer. Information such as: a) who is speaking, b) the vocal effort used by each speaker, and c) the presence of certain non-speech background sounds, are lost in a simple speech transcript. Current research systems make use of speaker identification and prosodic cues for topic segmentation and sentence or discourse boundary detection [5,12,17,19,20,21]. Some also include these non-lexical cues in their search; the most common being the combination of keyword and speaker ID. While there has been no formal evaluation of retrieval performance using these non-lexical cues, we hypothesize that expanding the list of non-lexical cues will assist multimedia content access and retrieval.

Relying on the text alone for retrieval is also problematic when the variability of noise conditions, speaker variance and other limitations of current automatic speech recognizers results in errorful speech transcripts. Deletion errors can prevent the users from finding what they are looking for

from audio or video data, while insertion and substitution errors can be misleading and/or confusing. In order to discover more and better information from multimedia data in the presence of imperfect speech transcripts, we have incorporated multiple speech technologies and natural language processing techniques to develop our research Audio Hot Spotting prototype. By utilising the multiple attributes detected from the audio, Audio Hot Spotting technology allows a user to automatically locate regions of interest in an audio/video file that meet his/her specified criteria. These regions of interest or "hot spots" can be found through keywords or phrases, speakers, keywords in combination with speaker ID, non-verbal speech characteristics, or non-speech signals of interest.

In Section 2, we describe how multiple attributes are detected and used to discover information and refine query results.

In Section 3, we describe keyword retrieval using word-based and phoneme-based speech recognition engines, our indexing algorithms that automatically identify potential search keywords that are information rich and provide a quick clue to the document content. Since speech-to-text speech recognition engines are limited by the language model and lexicon, they will not find words missing from the speech recognizer's lexicon. Furthermore, speech-to-text engines as other audio processing engines are sensitive to the audio quality; thus the word recognition accuracy varies. This affects the Audio Hot Spotting retrieval rate. To mitigate this, a phoneme-based audio retrieval engine is combined with the speech-to-text engine to improve the AHS retrieval rate. We explain our fusing algorithms to merge and rank-order two speech engine results to aid Audio Hot Spotting.

In Section 4, we discuss our query expansion mechanism to improve retrieval rate. We also present different approaches to cross-lingual multimedia audio hot spotting based on the characteristics of the source and target languages and use of the speech recognition engine for a particular language.

In Section 5, we describe the Audio Hot Spotting system architecture and work flow that allow both interactive and batch processing of multimedia data and query requests. We will discuss some current and future work in automatic multimedia monitoring and alert generation when an attribute of interest is detected by the system.

2. Detecting and Using Multiple Attributes from the Audio

Automatic speech recognition has been used extensively in audio and multimedia information retrieval systems. However, high speech Word Error Rate (WER) [5] in the speech transcript, especially in less-trained domains such as spontaneous and non-broadcast quality data, greatly reduces the effectiveness of navigation and retrieval using the speech transcripts alone. Even in applications where WER is low, our approach recognizes that there is more information in the audio file than just the words and that other attributes such as speaker identification and the type of background noise may be helpful in the retrieval of information that words alone fail to provide. One of the challenges facing researchers is the need to identify "which" non-lexical cues are helpful. Since these cues have not been available to users in the past, they don't know enough to ask for them. We have chosen to implement a variety of non-lexical cues with the intent of stimulating feedback from our user community.

2.1 Spoken Language ID

By extending the language model of a commercial spoken language ID engine, we integrated the spoken language ID engine with our custom-built language model aiming to identify a set of specific languages. The benefits of being able to automatically detect the spoken language are multi-fold. It improves human analyst productivity by sorting the media by language, as well as the accuracy of downstream processes that are language-dependent such as automatic speech recognition. It also helps to prioritize the analysis by language.

2.2 Speaker ID and Its Application

Another valuable audio attribute is speaker ID. By extending a research speaker identification algorithm [25], we integrated speaker identification into the Audio Hot Spotting prototype to allow a user to retrieve three kinds of information. First, if the user cannot find what he/she is looking for using keyword search but knows who spoke, the user can retrieve content defined by the beginning and ending labels of the chosen speaker; assuming enough speech exists to build a model for that speaker. Secondly, we provide summary speaker statistics indicating how many turns each speaker spoke and the total duration of each speaker's audio. Finally, we use speaker identification to refine the query result by allowing the user to query keywords and speaker together. For example, the user can find when President Bush spoke the word "anthrax".

2.3 Detection of Group Laughter and Applause

In addition to language and speaker identification, we wanted to illustrate the power of other non-lexical sounds in the audio track. As a proof-of-concept, we created detectors for crowd applause and laughter. The algorithms used both spectral information as well as the estimated probability density function (pdf) of the raw audio samples to determine when one of these situations was present. Laughter has a spectral envelope which is similar to a vowel, but since many people are voicing at the same time, the audio has no coherence. Applause, on the other hand, is spectrally speaking, much like noisy speech phones such as "sh" or "th." However, we determined that the pdf of applause differed from those individual sounds in the number of high ampli-

tude outlier samples present. Applying this algorithm to the 2003 State of the Union address, we identified all instances of applause with only a 2.6% false alarm rate. One can imagine a situation where a user would choose this non-lexical cue to identify statements that generated a positive response.

2.4 Speech Rate Estimation

The rate at which a person speaks is often indicative of the person's mental state. Fast speech is often associated with elevated emotions including stress, excitement, anger, or urgency. We have begun to look at speech rate as a separate attribute. Speech rate estimation is important, both as an indicator of emotion and stress, as well as an aid to the speech recognizer itself (see for example [14, 15, 23]). Currently, recognizer error rates are highly correlated to speech rate. For the user, marking that a returned passage is from an abnormal speech rate segment and therefore more likely to contain errors allows him/her to save time and ignore these passages if desired. However, if passages of high stress are desired, these are just the passages to be reviewed. For the recognizer, awareness of speech rate allows modification of HMM state probabilities, and even permits different sequences of phones.

One approach to determine the speech rate accurately is to examine the phone-level output of the speech recognizer. Even though the phone-level error rate is quite high, the timing information is still valuable for rate estimation. By comparing the phone lengths of the recognizer output to phone lengths tabulated over many speakers, we have found that a rough estimate of speech rate is possible [27]. Initial experiments have shown a rough correspondence between human perception of speed and the algorithm output. One outstanding issue is how to treat audio that includes both fast rate speech and significant silences between utterances. Is this truly fast speech?

2.5 Vocal Effort Estimation

Vocal effort is the continuum between whispering and shouting. Detecting this attribute is valuable in a variety of situations (e.g. speech recognition), but here we give the system user the opportunity to listen to (e.g.) all shouted segments, because shouting may indicate extra importance in some situations. To determine the vocal effort of a speaker, LPC (Linear Predictive Coding) analysis is performed using several different orders (13-19). The best fit to the incoming spectrum is then determined, and used for inverse-filtering. Inverse filtering removes (most of) the effects of the vocal tract shape, leaving a train of glottal pulses during voiced speech. We examine the shape of these pulses to determine the vocal effort. Higher vocal effort pulses have different time and frequency characteristics than lower effort (softer) speech. Softer glottal pulses are less peaky in the time domain and have a lower spectral bandwidth. Using the SUSAS [28] and Emotional Prosody [29] Corpora, we were able to get good separation (>95%) between the soft and loud cases. More subtle effort distinctions could not be reliably made, because of inter-speaker variations.

The issue of truly whispered speech (no voicing) hasn't been studied extensively, and is heavily dependent on signal quality. If the noise level is low, whispering can be detected by spectral shape, because the formants of the vowels still appear (though with low energy).

3. Keyword Retrieval Using Word-based and Phoneme-based Recognition Engines

Even with the option to use non-textual audio search cues, the primary criterion in a hot spotting search is still likely to be a key word or phrase. There are two methods for keyword retrieval from multimedia. The most common one is using automatic speech recognizer to turn speech into words. Another approach is to turn speech into phonemes without making a hard decision on the word. Thus a sequence of phonemes like [k-eh-r-i-a] can represent words like "career" and "Korea". There are advantages and disadvantages of both approaches. The word-based approach is dependent on the engine's lexicon and language model. Therefore, given a reasonable audio quality that matches the acoustic and language model of the recognition engine, the word-based approach will transcribe the speech accurately enough for keyword retrieval when the spoken words fall within the lexicon of the speech recognition engine. If those conditions are unmet, the speech transcripts will degrade significantly, leading to lower keyword retrieval accuracy. In other words, the word-based approach yields high precision but lacks recall due to its dependency on lexicon and language model. On the other hand, the phoneme-based approach is lexicon and language model independent. It can retrieve words that are not in the lexicon as long as the phonemes match. This is helpful especially for proper names. For example, if the proper name *Nesbit* is not in the speech recognizer vocabulary, the word will not be correctly transcribed. In fact, it was transcribed as *Nesbitt* (with two 't's). With the phoneme-based engine, *Nesbit* is retrieved.

However, the same sequences of phonemes, particularly short sequence of phonemes, may have multiple or partial matches to the query words. So it may yield false positives while recall is good. While the word-based approach provides a transcript beneficial to down-stream processes such as keyword indexing and translation, the phoneme-based approach produces no transcript. The word-based speech recognition costs significantly more processing time than phoneme-based indexing. While the speech-to-text speech recognition takes real time – one hour takes one hour of processing time, the phoneme engine only takes about a few minutes to process an hour of audio.

The Audio Hot Spotting research prototype takes advantage of both word-based and phoneme-based indexing engine by merging outputs from both engines and rank ordering the results to improve both precision and recall of keyword retrieval.

In section 3.1 we discuss the keyword indexing algorithm to take advantage of word-based transcripts. In section 3.2 we report the preliminary findings of keyword retrieval and compare the retrieval results from a word-based approach with those using a phoneme-based engine. In Section 3.3, we discuss our fusing approach using both word-based and phoneme-based recognition engines..

3.1. Keyword Indexing

Knowing that not every machine-transcribed word is correct and that not every one has equal information value, we developed a keyword indexing algorithm to assist the user in the selection of keywords for search. The algorithm finds words that are information rich (i.e. content words) with high likelihood of being correctly recognized by the recognition engine. The Audio Hot Spotting prototype examines speech recognizer output and creates an index list of content words. Our approach is based on the principle that short duration and weakly stressed words are much more likely to be mis-recognized, and are less likely to be important. [10, 11] To eliminate words that are information poor and prone to mis-recognition, our index-generation algorithm takes the following factors into consideration: a) absolute word length, b) the number of syllables, c) the recognizer's own confidence score, d) the part of speech (i.e. verb, noun) using a POS tagger. Experiments we have conducted using Broadcast data, with Gaussian white noise added to achieve the desired Signal-to-Noise Ratio (SNR), indicate that the index list produced typically covers about 10% of the total words in the ASR output, while more than 90% of the indexed words are actually spoken and correctly recognized given a WER of 30%. The following table illustrates the performance of the automatic indexer as a function of Signal-to-Noise Ratio during a short pilot study.

SNR (dB)	ASR WER (%)	Index Coverage (%)	IWER (%)
Inf	26.8	13.6	4.3
24	32.0	12.3	3.3
18	39.4	10.8	5.9
12	54.7	8.0	12.2
6	75.9	3.4	20.6
3	87.9	1.4	41.7

Table 1 Indexer SNR Performance

where Index Coverage is the fraction of the words in the transcript chosen as index words and IWER is the index word error rate.

As expected, increases in WER result in fewer words meeting the criteria for the index list. However, the indexer algorithm manages to find reliable words even in the presence of very noisy data. At 12dB SNR, while the recognizer WER has jumped up to 54.7%, the index word error rate has risen modestly, to only 12.2%. Note that an index word error indicates that an index word chosen from the ASR output transcript did not in fact occur in the original reference transcription.

Whether this index list is valuable will depend on the application. If a user wants to get a feel for a 1-hour conversation in just a few seconds, automatically generated topic terms such as those described in [24] or an index list such as this could be quite valuable.

3.2. Keyword Retrieval from Meeting Room Data Using Word-Based and Phoneme-based Recognition Engines

This section describes preliminary results of our attempts to retrieve meaningful information from meeting room data [7]. This search for information included the use of both a word and a phoneme based automatic speech recognition system. For a limited test set, our results suggest that the word based recognizer is better than the phoneme based system at retrieving information based on single keyword queries for word error rate (WER) up to about 75%. As the WER degrades, the phoneme based system performance surpasses the word based system. When the information search is based upon multi-word phrases, the phoneme based recognizer is superior at all three examined word error rates.

Excerpts of the selected audio files were processed by a commercial word-based speech recognition engine with no tuning of the language or acoustic models for the meeting room domain. NIST's SCLITE alignment tool was used to align the output of the recognizer with the NIST-provided transcripts and to determine the word error rate for each file. The same files were also processed by a commercial phoneme-based recognition engine, also without any tuning [2]. The reference transcripts were used to select information-bearing keywords as query terms. The keywords were manually selected by two researchers working independently; the agreed upon list of single keywords is shown in Table 1 and the list of selected query phrases is shown in Table 2.

Our system response to a keyword query is a temporal pointer into the multimedia file. We consider this audio hot spot to be *correct* if the queried keyword actually exists in the audio at a time within half of the audio segment duration from the returned pointer. For our evaluation the desired duration of a returned audio segment is six seconds; therefore, the keyword must exist in the audio within +/-3 seconds of the returned temporal pointer [26]. A response is flagged as a *missed detection* if the keyword exists in the audio but no temporal pointer within half the audio segment duration is returned. Finally a response is flagged as a *false alarm* if the temporal pointer is too far away from an actual occurrence of the queried keyword in the audio or the queried keyword doesn't exist in the audio.

File: 20020627	File: 20020214	File: 20020304
agonizing	computer	castle
backslashes	dsk	detail
computer	door	evil
debugging	help	healed
decorations	knobs	idea
Emily	move	king
function	office	kingdom
graffiti	officemate	love
InstallShield	paper	mother
Joe	problem	people
keyboard	room	prince
Linux	shelf	princess
meeting	space	queen
messages	survived	road
module	table	stepmother
onscreen	temperature	story
operate	vent	village
package	window	
Palm		
PWS		
remote		
unpack		
VIP		
web		
Wednesday		
Windows		

Table 1: Key Words Selected from the Reference Transcripts

File: 20020214	File: 20020627
air conditioner	having trouble
real problem	high priority
leg room	July seventeenth
plant division	onscreen keyboard
prime location	scrolling frame
control the temperature	successfully packaged
storage problem	
	File:20020304
	evil spell
	wonderful deeds

Table 2: Key Phrased Selected from the Reference transcripts

For the keyword query performance we report precision and recall. Precision indicates the percentage of returned audio pointers which actually identify audio segments containing the query term. Recall indicates the percentage of the total number of audio segments containing the query term which are actually identified by the returned audio pointers. We also report F-measure which is the harmonic mean of precision and recall. F-measure performance as a function of WER for our three excerpts is shown in Figure 1.

Since the word based recognizer makes a hard decision about each word in the transcript, searching for a particular keyword is very straightforward; the word is either there or it is not. When using the phoneme-based recognizer, the return for each query is a list of possible matches sorted by a confidence measure. Without some stopping criterion, the recognizer will return an endless list of possible matches resulting in excessive false alarms. We have investigated using both fixed and relative confidence thresholds as well as a fixed number of false alarms for each keyword. We have found that the fixed false alarm threshold gives the best performance. In the work reported here, phoneme-based returns were ignored after a single false alarm for each keyword.

Table 3 shows a precision/recall comparison between word-based and phoneme-based systems on the single word retrieval task.

WER	Prec. Word	Prec. Phone	Rec. Word	Rec. Phone	F-M Word	F-M Phone
90.3	1.00	0.438	0.091	0.318	0.167	0.368
89.4	0.83	0.381	0.102	0.327	0.182	0.352
75.9	0.85	0.500	0.370	0.369	0.516	0.425

Table 3: Single Word Retrieval Performance: Word-Based vs Phoneme Based Recognition. (WER = Word Error Rate, Prec. =Precision, Rec.=Recall, F-M = F-Measure, Word = Word-Based, Phone = Phoneme-based)

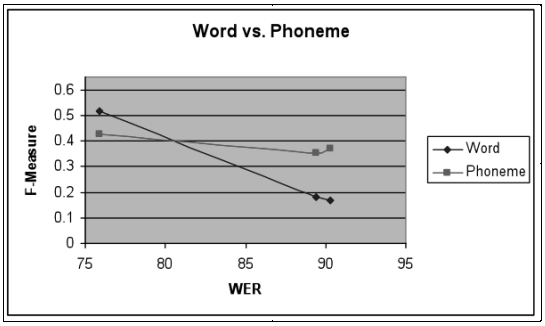

Figure 1: Single Word Retrieval Performance (F-Measure) for Word Based vs. Phoneme Based Recognizers.

Phrases presented a problem for the word-based recognizer. In this high WER domain, having all of the contiguous words in a phrase correctly recognized is rare. None of the phrases listed in Table 2 were identified by the word-based recognizer. The phoneme-based system recovered an average of 75% of the phrases across the three tested WERs.

3.3. Fusing of Word-based and Phoneme-based Recognition to Compensate on Precision and Recall

To take advantage of the high precision of a word-based recognition engine and high recall of phoneme-based recognition engine, we experimented with merging the results from both recognition engines and rank ordering the results. In our approach, words retrieved by both engines were given the highest rank. The next highest ranked words were those returned by the word-based recognizer. Finally, words returned by only the phoneme based system were given a rank proportional to their confidence score.

For the fused search, Audio Hot Spotting prototype aligns results from both recognition engines; thus we not only retrieve the exact segment of the audio but also the accompanying transcript.

4. Query Expansion

TREC SDR found both a linear correlation between Speech Word Error Rate WER [5] and retrieval rate [4] and that retrieval was fairly robust to WER. However, the robustness was attributed to the fact that misrecognized words are likely to also be properly recognized in the same document if the document is long enough. Since we limit our returned passages to roughly 10 seconds, we do not benefit from this long document phenomenon. The relationship between passage retrieval rate and passage length was studied by searching 500 hours of broadcast news from the TREC SDR corpus. Using 679 keywords, each with an error rate across the corpus of at least 30%, we found that passage retrieval rate was 71.7% when the passage was limited to only the query keyword. It increased to 76.2% when the passage length was increased to 10sec and rose to 83.8% if the returned document was allowed to be as long as 120sec.

In our Audio Hot Spotting prototype, we experimented with query expansion to achieve two purposes, 1) to improve the retrieval rate of related passages when exact word match fails, and 2) to allow cross lingual query and retrieval.

4.1 Keyword Query Expansion

The Audio Hot Spotting prototype made use of the Oracle 10g Text engine to expand the query semantically, morphologically and phonetically. For morphological expansion, we activated the stemming function. For semantic expansion, we utilized expansion to include hyponyms, hypernyms, synonyms, and semantically related terms. For example, when the user queried for "oppose", the exact match yielded no returns, but when semantic and morphological expansion options are selected, the query was expanded to include *anti, anti-government, against, opposed, opposition,* and returned several passages containing those expanded terms.

To address the noisy nature of speech transcripts, we used the phonetic expansion, i.e. "sound alike" feature from the Oracle database system. This is helpful especially for proper names. For example, if the proper name *Nesbit* is not in the speech recognizer vocabulary, the word will not be correctly transcribed. In fact, it was transcribed as *Nesbitt* (with two 't's). By phonetic expansion, *Nesbit* is retrieved.

Obviously more is not always better. Some of the expanded queries are not exactly what the users are looking for and the number of passages returned increases. In our Audio Hot Spotting implementation we made query expansion an option allowing the user to choose to expand semantically and/or, morphologically, or phonetically.

4.2 Cross-lingual Query Expansion

In some applications it is helpful for a user to be able to query in a single language and retrieve passages of interest from documents in several languages. We treated translingual search as another form of query expansion. We created a bilingual thesaurus by augmenting Oracle's default English thesaurus with Spanish dictionary terms. With this type of query expansion enabled, the system retrieves passages that contain the keyword in either English or Spanish. A straightforward extension of this approach will allow other languages to be supported.

5. Audio Hot Spotting Research Prototype

The Audio Hot Spotting research prototype consists of two primary components; the Audio Hot Spotting pre-processor and the Audio Hot Spotting query processor.

The Audio Hot Spotting Pre-processor allows users to prepare media files for search and retrieval. It automatically detects information from the media file for the downstream processes. This includes the media format, acoustic channel (telephone, microphone), primary language spoken in the media. With this information, the pre-processor will determine which pre-processing engines are available for the type of the media. For example, given a multimedia file in microphone speech for English; the pre-processors available will be speech to-text transcription, phoneme indexing, keyword indexing, background noise detection, and vocal effort detection. Once the media file is pre-processed, the file is available for the user to search and retrieve segments of interest by different types and combinations of searches using the Audio Hot Spotting Query engine.

The Audio Hot Spotting Query Processor allows users to search for and retrieve information from the preprocessed media files. It supports single media and cross media search as long as the media contains speech, cross-lingual search, and multiple attributes search such as searching for a key word by a particular speaker. The supported cross-lingual

search in the current Audio Hot Spotting prototype supports English, Spanish, Modern Standard Arabic, Iraqi Arabic, and Gulf States Arabic. The query processor supports a variety of search criteria. It allows the user to search the pre-processed media by language, by speaker, by word-based search, by phoneme-based search, by fused search, by background noise such as applause and laughter, by speech rate, and by vocal effort such as shouting and whispering. The system already supports cross-attribute search such as keyword and speaker. With relational database as the back-end of the Audio Hot Spotting system, we plan to link other attributes for the user to search information such as shouting by a particular speaker and the other combinations of the multiple attributes.

Both processors work in a web-based environment to support current users for processing and query. The system provides the interface for both interactive query and pre-processing as well as batch query and pre-processing so that the user can submit multiple queries and pre-process multiple files without having to set the query and processing criteria every time.

Although the Audio Hot Spotting research prototype offers query and retrieval capability for multiple attributes, each module can function alone with the exceptions of keyword indexing and speech rate estimation, which depend on the output of automatic speech recognition.

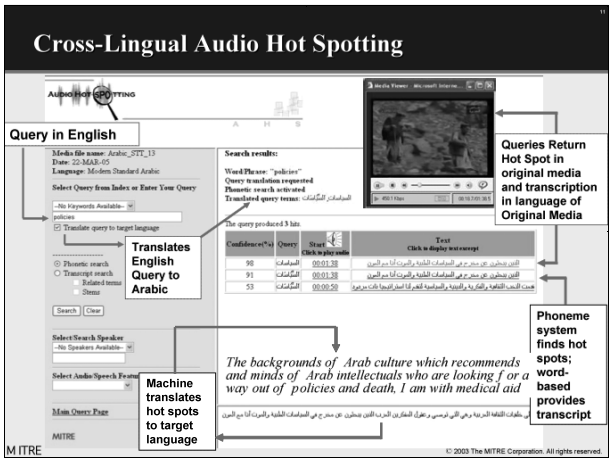

Figure 2. Audio Hot Spotting Prototype

6. Conclusion

By detecting multiple audio attributes, our Audio Hot Spotting prototype allows a user to begin to apply the range of cues available in audio to the task of information retrieval. Areas of interest can be specified using keywords, phrases, language and speaker ID, information-bearing background sounds, such as applause and laughter, and prosodic information such as shouting and whispering. When matches are found, the system displays the recognized text and allows the user to play the audio or video in the vicinity of the identified "hot spot". See figure 2 for a screenshot of the Audio Hot Spotting prototype.

With the advance of component technologies such as automatic speech recognition, language and speaker identification, and audio feature extraction, there will be a wider array of audio attributes for the multimedia information systems to query and retrieve, allowing the user to access the exact information desired both in post processing and live monitoring modes.

References

1. J. Allan. Knowledge Management and Speech Recognition.. *IEEE Computer*, 35(4):60-61, 2002.

[2] P. Cardillo, M. Clements, M.. Miller, "Phonetic Searching vs Large Vocabulary Continuous Speech Recognition," *International Journal of Speech Technology* , pp 9-22, January 2002.

3. T. Colthurst et. al. The 2000 BBN Byblos LVCSR System. In Hub-5, 2001.

4. J. Garofolo,et. al. The GREC Spoken Document Retrieval Track: A Successful Story. TREC 9, Nov, 2000.

5. Fiscus et. al. Speech Recognition Scoring Toolkit (http://www.nist.gov/speech/tools/)

6. D. Hakkani-Tur , G. Tur, A.Stolcke, E. Shriberg. Combining Words and Prosody for Information Extraction from Speech. *Proc. EUROSPEECH'99,*

7. Q. Hu, F. Goodman, S. Boykin, R. Fish, W. Greiff. Audio Hot Spotting and Retrieval Using Multiple Audio Features and Multiple ASR Engines. Processings of ICASSP 2004 NIST Meeting Recognition

Workshop. Montreal, Canada.

8. Q. Hu, F. Goodman, S. Boykin, R. Fish, W. Greiff. Audio Hot Spotting and Retrieval Using Multiple Features. Processings of the HLT-NAACL 2004 Workshop on Interdisciplinary Approaches to Speech Indexing and Retrieval. Boston, USA.

9. Q. Hu, F. Goodman, S. Boykin, R. Fish, W. Greiff. Information Discovery by Automatic Detection, Indexing, and Retrieval of Multiple Attributes from Multimedia Data, *3rd International Workshop on Multimedia Data and Document Engineering*, September 2003. Berlin, Germany.

10. Q. Hu, F. Goodman, S. Boykin, Margot Peet. Multimedia Indexing and Retrieval Technologies Using the Audio Track. IEEE 2002 Conference on Technologies for Homeland Security. Boston.

11. Q. Hu, F. Goodman, S. Boykin, Margot Peet. The MITRE Audio Hot Spotting Prototype - Using Multiple Speech and Natural Language Processing Technologies. Interna-

tional Conference on Text, Speech and Dialog (TSD 2002). Burno.

12. S. .E, Johnson, P. Jourlin, S.S. Jones, and P.C. Woodland. Spoken Document Retireval for TREC-9 at Cambridge University. TREC-9, Nov., 2000.

13. K. Koumpis et. al. Extractive Summarization of Voicemail using Lexical and Prosodic Feature Subset Selection. *Proc. EUROSPEECH 2001.*

14. N. Mirghafori, E. Fosler, and N. H. Morgan. Towards Robustness to Fast Speech in ASR, *Proc. ICASSP,* Atlanta, GA, May 1996.

15. N. Morgan and E. Fosler-Lussier. Combining Multiple Estimators of peaking Rate, *Proc. ICASSP-98*, pp. 729-732, Seattle, 1998

16. M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds. Modeling of the Glottal Flow Derivative Waveform with Application to Speaker Identification, *IEEE Trans. On Speech and Audio Processing,* September 1999.

17. S. Rendals, and D. Abberley. The THISL SDR System at TREC-9. TREC-9, Nov., 2000.

18. E. Shriberg , A. Stolcke , D. Hakkani-Tur , G. Tur. Prosody-Based Automatic Segmentation of Speech into Sentences and Topics. *Speech Communication* Vol. 32, Nos. 1-2, September, 2000.

19. K. N. Stevens & H. .M. Hanson. Classification of Glottal Vibration from Acoustic Measurements. In *Vocal Fold Physiology: Voice Quality Control,* Fujimura O., Hirano H. (eds.), Singular Publishing Group, San Diego 1995.

20. A. Stolcke , E. Shriberg , D. Hakkani-Tur , G.Tur, Z. Rivlin , K. Sunmez. Combining Words and Speech Prosody for Automatic Topic Segmentation. *Proc. DARPA Broadcast News Workshop*, Herndon, VA, 1999.

21. A. Stolcke , E. Shriberg , R. Bates , M. Ostendorf , D.Hakkani , M. Plauchu , G. Tur , Y. Lu. Automatic Detection of Sentence Boundaries and Disfluencies based on Recognized Words. *Proc. ICSLP'98*, Sidney, Australia, 1998.

22. G. Tur , D. Hakkani-Tur , A. Stolcke , E. Shriberg. Integrating Prosodic and Lexical Cues for Automatic Topic Segmentation. *Journal of Computational Linguistics* , Vol. 27, No. 1, March, 2001.

23. J. Zheng, H. Franco, F. Weng, A. Sankar and H. Bratt.. Word-level Rate-of-Speech Modeling Using Rate-Specific Phones and Pronunciations, *Proc. ICASSP,* vol 3, pp 1775-1778, 2000

24. F. Kubala, S. Colbath, D. Liu, A. Srivastava, J. Makhoul. Integrated Technologies For Indexing Spoken Language, *Communications of the ACM*, February 2000.

25. D. Reynolds. Speaker Identification And Verification Using Gaussian Mixture Speaker Models, *Speech Communications*, vol.17, pp.91, 1995

26. J. Hirschberg, S. Whittaker, D. Hindle, F. Pereira and A. Singhal. Finding Information In Audio: A New Paradigm For Audio Browsing/Retrieval, *ESCA ETRW workshop Accessing information in spoken audio*, Cambridge, April 1999.

27. N Mirgafori, E. Fosler, and N. Morgan. Towards Robustness To Fast Speech. *Proc. ICASSP* pp335-338 Atlanta, Georgia. May, 1996

28. J.H.L. Hansen, S. Bou-Ghazale, "Getting Started with SUSAS: A Speech Under Simulated and Actual Stress Database," EUROSPEECH-97: Inter. Conf. On Speech Communication and Technology, vol. 4, pp. 1743-1746, Rhodes, Greece, Sept. 1997.

29. Mark Liberman et. al, Emotional Prosody Speech and Transcripts, Linguistic Data Consortium Catalog #LDC2002S28, 2002