Audio Hot Spotting and Retrieval Using Multiple Features

Qian Hu
MITRE Corporation
qian@mitre.org

Fred Goodman, Stanley Boykin,
Randy Fish, Warren Greiff
MITRE Corporation
fgoodman@mitre.org,
sboykin@mitre.org,
fishr@mitre.org
greiff@mitre.org

Abstract

This paper reports our on-going efforts to exploit multiple features derived from an audio stream using source material such as broadcast news, teleconferences, and meetings. These features are derived from algorithms including automatic speech recognition, automatic speech indexing, speaker identification, prosodic and audio feature extraction. We describe our research prototype – the Audio Hot Spotting System – that allows users to query and retrieve data from multimedia sources utilizing these multiple features. The system aims to accurately find segments of user interest, i.e., audio hot spots within seconds of the actual event. In addition to spoken keywords, the system also retrieves audio hot spots by speaker identity, word spoken by a specific speaker, a change of speech rate, and other non-lexical features, including applause and laughter. Finally, we discuss our approach to semantic, morphological, phonetic query expansion to improve audio retrieval performance and to access cross-lingual data.

1. Introduction

Audio contains more information than is conveyed by the text transcript produced by an automatic speech recognizer [Johnson et al., 2000; Hakkani-Tur et al., 1999]. Information such as: a) who is speaking, b) the vocal effort used by each speaker, and c) prosodic features and certain non-speech background sounds, are lost in a simple speech transcript. In addition, due to the variability of acoustic channels and noise conditions, speaker variance, language models the recognizer is based on, and the limitations of automatic speech recognition (ASR), speech transcripts can be full of errors. 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 confusing. Our approach is to automatically detect, index, and retrieve multiple features from the audio stream to compensate for the weakness of using speech transcribed text alone. The multiple time-stamped features from the audio include an automatically generated index derived from ASR speech transcripts, automatic speaker identification, and automatically identified prosodic and audio cues. In this paper, we describe our indexing algorithm that automatically identifies potential search keywords that are information rich and provide a quick clue to the document content. We also describe how our Audio Hot Spotting prototype system uses multiple features to automatically locate regions of interest in an audio or video file that meet a user's specified query criteria. In the query, users may search for keywords or phrases, speakers, keywords and speakers together, non-verbal speech characteristics, or non-speech signals of interest. The system

also uses multiple features to refine query results. Finally, we discuss our query expansion mechanism by using natural language processing techniques to improve retrieval performance and to access cross-lingual data.

2. Data

We use a variety of multimedia data for the experiments in order to test Audio Hot Spotting algorithms and performance under different acoustic and noise environments. These include broadcast news (e.g. HUB4), teleconferences, meetings, and MITRE corporate multimedia events. In some cases, synthetic noise was added to clean source material to test algorithm robustness.

3. Automatic Spoken Keyword Indexing

As automatic speech recognition is imperfect, automatic speech transcripts contain errors. Our indexing algorithm focuses on finding words that are information rich (i.e. content words) and machine recognizable. 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. To eliminate words that are information poor and prone to mis-recognition, our algorithm examines the speech recognizer output and creates an index list of content words. The index-generation algorithm takes the following factors into consideration: a) absolute word length by its utterance duration, b) the number of syllables, c) the recognizer's own confidence score, and d) the part of speech (i.e. verb, noun) using a POS tagger with some heuristic rules. Experiments we have conducted using broadcast news data, with Gaussian white noise added to achieve a 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 Word Error Rate (WER [Fiscus, et al.]) 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 (%)
Orig.	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 (IWER) has risen to 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 [Kubala et al., 2000] or an index list such as this could be quite valuable.

4. Detecting and Using Multiple Features from the Audio

Automatic speech recognition has been used extensively in spoken document retrieval [Garofolo et al., 2000; Rendals et al., 2000]. However, high speech WER 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. Furthermore, the retrieval of a whole document or a story still requires the user to read the whole document or listen to the entire audio file in order to locate the segments where relevant information resides. In our approach, we recognize that there is more information in the audio file than just the words and that other attributes such as speaker

identification, prosodic features, and the type of background noise may also be helpful for the retrieval of information. In addition, we aim to retrieve the exact segments of interest rather than the whole audio or document so that the user can zero in on these specific segments rapidly. One of the challenges facing researchers is the need to identify "which" non-lexical features have information value. Since these features 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.

As an example of this, by extending a research speaker identification algorithm [Reynolds, 1995], 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 timestamps associated with the specified speaker; assuming enough speech exists to build a model for that speaker. Secondly, the system automatically generates speaker participation statistics indicating how many turns each speaker spoke and the total duration of each speaker's audio. Finally, the system uses speaker identification to refine the query result by allowing the user to query keywords and speaker together. For example, using the Audio Hot Spotting prototype, the user can find the audio segment in which President Bush spoke the word "anthrax".

In addition to speaker identification, we wanted to illustrate the information value 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 amplitude 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 (results were compared with hand-labeled data). One can imagine a situation where a user would choose this non-lexical cue to identify statements that generated a positive response.

Last year, we began to look at speech rate as a separate feature. 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 [Mirghafori et al., 1996; Morgan, 1998; Zheng et al., 2000]). Currently, recognizer word 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 by ignoring these passages or reading them with discretion if desired. However, if passages of high stress are of interest, 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 [Mirgafori et al. 1996]. Initial experiments using MITRE Corporate event data 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?

We are currently conducting research to detect other prosodic features by estimating vocal effort. These features may indicate when a speaker is shouting suggesting elevated emotions or near a whisper. Queries based on such features can lead to the identification of very interesting audio hot spots for the end user. Initial experiments are examining the spectral properties of detected glottal pulses obtained during voiced speech.

5. Query Expansion and Retrieval Tradeoffs

5.1 Effect of Passage Length

TREC SDR found both a linear correlation between speech word error rate and retrieval rate [Garofolo et al., 2000] 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 full-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 passage was allowed to be as long as 120sec.

In our Audio Hot Spotting prototype, we experimented with semantic, morphological, and phonetic 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.

5.2 Keyword Query Expansion

The Audio Hot Spotting prototype integrated the Oracle 9i 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 these 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. We are aware of the limitations of Oracle's phonetic expansion algorithms, which are simply based on spelling. This doesn't work well when text is a mis-transcription of the actual speech. Hypothetically, a phoneme-based recognition engine may be a better candidate for phonetic query expansion. We are currently evaluating a phoneme-based audio retrieval system and comparing its performance with a word-based speech recognition system. The comparison will help us to determine the strengths and weaknesses of each system so that we can leverage the strength of each system to improve audio retrieval performance.

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.

5.3 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.

6. Future Directions

As our research and prototype evolve, we plan to develop algorithms to detect more meaningful prosodic and audio features to allow the users to search for and retrieve them. We are also developing algorithms that can generate speaker identify in the absence of speaker training

data. For example, given an audio script, we expect the algorithms to automatically identify the number of different speakers present and the time speaker X changes to Y. For semantic query expansion, we are considering using more comprehensive thesauri and local context analysis to locate relevant segments to compensate for high ASR word error rate. We are also considering combining a word-based speech recognition system with a phoneme-based system to improve the retrieval performance especially for out of vocabulary words and multi-word queries.

7. Conclusion

In this paper, we have shown that by automatically detecting multiple audio features and making use of these features in a relational database, our Audio Hot Spotting prototype allows a user to begin to apply the range of cues available in audio to the task of multi-media information retrieval. Areas of interest can be specified using keywords, phrases, speaker identity, prosodic features, and information-bearing background sounds, such as applause and laughter. 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". With the advance of component technologies such as automatic speech recognition, speaker identification, and prosodic and audio feature extraction, there will be a wider array of audio features for the multimedia information systems to query and retrieve, allowing the user to access the exact information desired rapidly.

References

1. John Garofolo, et al. Nov., 2000. *The TREC Spoken Document Retrieval Track: A Successful Story*. TREC 9.

2. Julia Hirschberg, Steve Whittaker, Don Hindle, Fernando Pereira and Amit Singhal. April 1999. *Finding Information In Audio: A New Paradigm For Audio Browsing/Retrieval*, ESCA ETRW workshop Accessing information in spoken audio, Cambridge.

3. Sue Johnson, Pierre Jourlin, Karen Sparck Jones, and Philip Woodland. Nov., 2000. *Spoken Document Retireval for TREC-9 at Cambridge University*. TREC-9.

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

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

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

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

8. 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.

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

10. K. N. Stevens and 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.

11. 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

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

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