# A Method for Direct Audio Search with Applications to Indexing and Retrieval


**Sue Johnson & Phil Woodland**

Cambridge University Engineering Department

*{sej28 , pcw}@eng.cam.ac.uk*


## SUMMARY


A technique for searching audio data to find an exact match for a given piece of *cue-audio* is described and shown to be fast, accurate and robust. The method is used on a database of broadcast news audio to successfully

1) locate the end points of the news shows
2) eliminate commercials and increase subsequent retrieval performance


## INTRODUCTION

We would like to be able to retrieve specific portions of *cue-audio* from within a large audio database both accurately and quickly. Applications when the audio is from broadcast TV and radio news include:

- Locating a given event known to have happened in the audio
  *e.g. to find where you are in the overall recording*
- Locating events which are characterised by a certain piece of audio
  *e.g. to find the start or end of a news show using a station's signature tune*
- Finding repeated events within an audio database
  *e.g. to aid navigation, identify broadcast structure, or find commercials*

The method must only find exact matches for the cue-audio, which correspond to the cue-audio itself, or an identical copy (e.g. re-broadcast).

## DIRECT AUDIO SEARCH

The audio is sampled at 16kHz. Searching for a known sequence of these samples is computationally intractable so a higher-order representation of the audio, such as is used in content-based audio retrieval is sought.

The samples are grouped into frames (25ms long, with a shift of 10ms) and represented by their mel-frequency PLP cepstral coefficients for subsequent speech recognition.

The covariance matrix of these coefficients has been shown to capture general information in a segment of audio for use in e.g. speaker clustering. The *arithmetic-harmonic sphericity* (AHS) distance:

$$d(X,Y) = \log[tr(\Sigma_y \Sigma_x^{-1}) \cdot tr(\Sigma_x \Sigma_y^{-1})] - 2\log(D)$$

is used to define the closeness between segments, with 0 distance implying an exact match.

A repetition is defined when the distance between a segment of *cue-audio* and a segment of *target-audio* of the same length is less than a very small threshold.

## EXPT. 1 - END POINT DETECTION

Since the overall audio properties, rather than a specific sequence of values are being used, a match can still be found when a slight offset between the cue-audio and the repeat occurs if the threshold is increased. This means fewer potential target-audio segments must be compared against the cue-audio to locate the repeats.

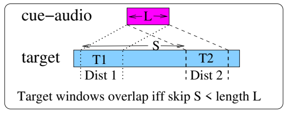


**Figure 1 :** *Comparing a segment of cue-audio with a target audio stream*

A 6.6 second segment of cue-audio which is usually played at the end of the shows was used to automatically locate the end of the 19 ABC-news broadcasts from the 1998 TREC-7 SDR evaluation data. A sliding window system (see Figure 1) was used on the target audio, with a skip of 0.5s, to find the minimum distance score over the last 5 minutes of each show.

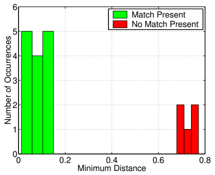


**Figure 2 :** *Distribution of minimum AHS distance over last 5 minutes of ABC broadcasts*

The method correctly identified every occurrence of the cue-audio with a *separability ratio* of 4.53 where:

$$\text{Separability Ratio} = \frac{\text{min. distance when cue-audio not present}}{\text{max. distance when cue-audio found}}$$

| Features Used | Distance Metric | Covariance Representation | Speed (xRT) | Separability Ratio |
|---|---|---|---|---|
| static+deltas | AHS | Full | 1/39 | 4.53 |
| static+deltas | Euclidean | Diagonal | 1/120 | 1.50 |
| static only | AHS | Full | 1/190 | 3.01 |
| static only | Euclidean | Diagonal | 1/330 | 1.50 |

**Table 1 :** *The effect on speed and separability of using a simple Euclidean distance on the diagonal covariance matrix and removing delta parameters from the feature vector*

### A) SPEEDING UP THE SEARCH

Reasonable separability can still be achieved with the AHS distance even when the search is speeded up by a factor of around 10 by increasing the window skip. This would not be possible with a sequence-based method.

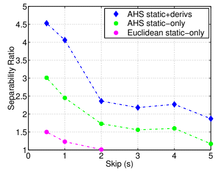


**Figure 3 :** *Effect on separability of increasing the window skip to speed up the search*

### B) REDUCING THE SIZE OF CUE-AUDIO

We tested how reliable the method was when the size of cue-audio was reduced. Results using a shift of 0.5s, given in Figure 4, show the method still works well when the cue-audio is only 1.0s long, whilst the speed is almost doubled.

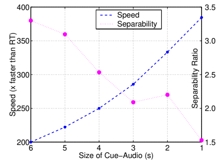


**Figure 4 :** *Effect on separability and speed of reducing the size of cue-audio*

Reducing the cue-audio further requires a reduction in skip size and becomes unreliable for audio of <0.2s long (see Table 2) as the covariance matrix becomes more ill-conditioned.

| Size (s) | Shift (s) | Misses | Speed (xRT) | Separability Ratio |
|---|---|---|---|---|
| 0.5 | 0.1 | 0 | 1/240 | 2.18 |
| 0.3 | 0.1 | 0 | 1/270 | 1.51 |
| 0.2 | 0.1 | 1 | 1/294 | 1.07 |
| 0.2 | 0.05 | 1 | 1/200 | 1.06 |

**Table 2 :** *Effect on accuracy, speed and separability of reducing the size of the cue-audio*

## EXPT. 2 - FINDING NON-NEWS EVENTS

The method was applied recursively over the ABC data to find repeated segments, which were assumed to be commercials, since they had been re-broadcast. Table 3 shows the results when compared to the manually generated "news" and "non-news" labels from the TREC-7 evaluation.

| Windowing Params | Time (CPUs) | Non-news | News |
|---|---|---|---|
| L=5s, S=1s | 3514 | 5459s = 60.8% | 110s = 0.5% |
| L=10s, S=2s | 967 | 5480s = 61.0% | 114s = 0.5% |
| L=20s, S=4s | 284 | 4598s = 51.2% | 122s = 0.5% |

**Table 3:** *Amount of non-news and news rejected on the ABC TREC-7 SDR data by recursively searching the audio*

The results when applying this technique recursively with some smoothing between the detected repeats, over all the ABC and CNN data for the 2000 TREC-9 SDR evaluation are given in Table 4.

| Broadcaster | Non-news | News | Total |
|---|---|---|---|
| ABC | 15.6hrs = 79.8% | 15.7sec = 0.01% | 25% |
| CNN | 45.8hrs = 62.4% | 44.1min = 0.43% | 19% |

**Table 4:** *Amount of non-news and news rejected on TREC-9 SDR data by recursively searching the audio*

### EFFECT ON RETRIEVAL

A system was built for the 1999 TREC-8 SDR evaluation, to automatically remove commercials by detecting previously seen audio segments using this technique with some smoothing between the repeats.

- 42.3 hours of test data was removed (8.4% of the total)
- 97.8% of this data was non-news (compared to manual labels)
- This information helped determine the structure of broadcasts
- Retrieval was significantly improved by removing commercials

## CONCLUSIONS

- We have described a technique for finding exact matches in audio
- The method is very accurate and can run over 200x faster than real time
- Less than 0.5s of cue-audio is needed for a simple search
- Applications include automatically detecting the end of news shows
- Removing commercials for Spoken Document Retrieval by this method reduces subsequent processing time, helps indicate broadcast structure and significantly improves retrieval performance