AUDIO FINGERPRINTING TO IDENTIFY MULTIPLE VIDEOS OF AN EVENT

Courtenay V. Cotton and Daniel P. W. Ellis[*]

LabROSA, Dept. of Electrical Engineering
Columbia University, New York NY 10027 USA
{cvcotton,dpwe}@ee.columbia.edu

ABSTRACT

The proliferation of consumer recording devices and video sharing websites makes the possibility of having access to multiple recordings of the same occurrence increasingly likely. These co-synchronous recordings can be identified via their audio tracks, despite local noise and channel variations. We explore a robust fingerprinting strategy to do this. Matching pursuit is used to obtain a sparse set of the most prominent elements in a video soundtrack. Pairs of these elements are hashed and stored, to be efficiently compared with one another. This fingerprinting is tested on a corpus of over 700 YouTube videos related to the 2009 U.S. presidential inauguration. Reliable matching of identical events in different recordings is demonstrated, even under difficult conditions.

Index Terms— Acoustic signal analysis, Multimedia databases, Database searching

1. INTRODUCTION

Any notable current public event is very likely to have been captured on the personal video recorders (cameras, cellphones, etc.) of some of the people present, and many of these recordings will subsequently be published on the internet through video sharing sites. We are interested in automatically discovering these multiple recordings of the same event. It would be extremely difficult to identify these conclusively using visual information, since different recordings may be taken from entirely different viewpoints. It is, however, possible to consider doing this with the audio, since the same basic acoustic event sequence should be captured consistently by any recording made in the same vicinity.

This soundtrack matching problem has similarities with that of identifying identical musical recordings in the presence of noise and channel variations. In both cases, we expect to see a lot of invariant underlying structure (e.g. spectral peaks) in the same relative time locations, but possibly corrupted with different channel effects and mixed with varying levels and types of noise. This problem is addressed by a number of prior works in audio fingerprinting [1]; our work is based on the approach of [2] which uses the locations of pairs of spectrogram peaks as robust features for matching. A similar approach was used to identify repeated events in environmental audio in [3], and a variant based on matching pursuit (MP) was presented in [4] to group similar but non-identical audio events. This work is closely related in spirit to use of audio fingerprints to synchronize multiple cameras in [5] and amateur rock concert videos in [6].

In section 2 we present a strategy for using MP to select salient elements of a signal, pairing these elements to create distinguishing landmarks, and efficiently searching for matching landmarks. In section 3 we describe the video data used to test this strategy, and in section 4 we examine the precision of our search results.

2. ALGORITHM

2.1. Matching Pursuit

MP [7] is an algorithm for sparse signal decomposition into an over-complete dictionary of basis functions. MP basis functions, called atoms, correspond to concentrated bursts of energy localized in time and frequency, but spanning a range of time-frequency tradeoffs. The MP algorithm iteratively selects atoms corresponding to the most energetic points in the signal, as long as they can be approximated by a basis function in the dictionary. In contrast to selecting peaks with a fixed-window spectrogram representation, MP can capture salient features in the signal at varying time-frequency scales.

In our fingerprinting, each video soundtrack is decomposed into an MP representation in order to identify a sparse set of the most salient elements it contains. It is most straightforward to decompose the entire length of a video at once, in order to avoid issues with windowing and boundary overlaps. A variable number of atoms are extracted from each video, roughly a few hundred atoms per second, although these are not uniformly distributed throughout the video. Selecting a larger number of elements than will actually be used from the signal as a whole will tend to sufficiently cover both louder and quieter portions of the signal; then a smaller number of atoms in each local area can be selected from these.

We use the efficient implementation of MP from [8]. Our

[*] This work was supported by the NSF (grant IIS-0716203), and the Eastman Kodak company.

dictionary contains Gabor atoms (Gaussian-windowed sinusoids) at nine length scales, incremented by powers of two. For data sampled at 22.05 kHz, this corresponds to window lengths ranging from 1.5 to 372 ms. These are each translated in time by increments of one eighth of the atom length over the duration of the signal.

2.2. Landmark Formation and Hashing

A landmark consists of a pair of atoms, and is defined only by their two center frequencies and the time difference between their temporal centers. These values are quantized to allow effecient matching between landmarks. The time resolution is 32 ms. The frequency resolution is 21.5 Hz, with only frequencies up to 5.5 kHz considered; this results in 256 discrete frequencies.

For every block of 32 time steps (around 1 second), the 15 highest energy atoms are selected. Each of these is paired with other atoms only in a local target area of the frequency-time plane. Here, each atom is paired with up to 3 others; if there are more than 3 atoms in the target area, the closest 3 in time are selected. This leads to approximately 45 landmarks per second. The target area is defined as the frequency of the initial atom, plus or minus 667 Hz, and up to 64 time steps after the initial atom.

The landmark values as quantized above can be described as a unique hash of 20 bits: 8 bits for the frequency of the first atom, 6 bits for the frequency difference between them, and 6 bits for the time difference. A hash table is constructed to store all the locations of each landmark hash value. Landmark locations are stored in the table with an identification number from the originating video and a time offset value, which is the time location of the earlier atom relative to the start of the video.

2.3. Query Matching

To find instances of the same events as in a query video, the query is decomposed with MP as described above. The video is then divided into five-second (non-overlapping) clips, and landmarks are formed from the atoms in each clip and hashed, as described in section 2.2. Each clip will contain an average of 225 landmarks. We break the query into these shorter pieces to improve the opportunity for matching subportions of videos, as well as to provide independent tests of matches between longer videos, as described below. The hash table is queried for each of the landmarks found in the five second clip. The start time of each query landmark is treated as a reference time; this is subtracted from the offset times for landmarks returned from the table. A likely match will therefore return multiple landmarks from the same video, all reporting the same relative offset time from their corresponding query landmarks.

3. VIDEO DATABASE

We wanted to test this algorithm on a set of videos likely to contain multiple versions of the same sequence of acoustic events. We chose to consider videos taken during the 2009 American presidential inauguration. We assumed there were likely to be many different professional and personal recordings of the ceremony available, given the massive public attendence and news coverage. We obtained a set of videos from YouTube using the query "inauguration obama". YouTube query results are limited to 1000 items; this and other complications (e.g. videos with no soundtrack) limited our actual database set to 733 videos. Other than this, no hand selection or filtering was done on the video set. The set comprises 56.2 hours of video. All audio is sampled at 22.05 kHz.

4. RESULTS

4.1. Match Evaluation

Each video was processed as above and stored in the hash table. Then each video was divided into five second (non-overlapping) segments, and used as a query to the database. Matches to the query video itself were discounted. Matches were returned based on the proportion of identical landmarks matched to the total number of landmarks in the query segment. A lower threshold proportion will result in more matches returned. In this experiment, all matches containing at least 5% of the query landmarks and at least 10 actual landmarks were considered.

Two videos with a long stretch of matching audio will result in a number of sequential query segments matching the same video, with the same time offset. For the purpose of simplifying evaluation, all matches occuring between the same two videos at the same offset are collapsed into a single match, spanning the time from the start of the first matching segment to the end of the last matching segment. This is a reasonable assumption, since it is unlikely for two videos to match at multiple points with the same offset unless they are truly part of the same long matching segment.

4.2. Estimating Precision

The procedure described above produced 34,247 individual matches. Fig. 1 shows a histogram of the number of matches found by proportion of matching landmarks to total query landmarks, with 5% being the minimum considered. There were 91 matches which occured above the level of 40%; manual examination of these results revealed them to be largely matches between videos in several different 'series', each with a signature introduction sound or music at the beginning of the video. There were also six pairs of identical or nearly identical videos in this set. All these matches are accurate, but not particularly interesting. The 'series' videos in general did not contain footage of the inauguration or related events. For simplicity, they and (one copy of) the six exact duplicate videos were all removed from the database. This left a set of 31,756 matches. Of these, 8186 (27%) matched over a longer time period than a single five-second clip. Given the small probability of two videos matching with the same time offset in multiple places by chance, it is reasonable to assume that most of these longer matches are accurate. This was confirmed by random checking of longer matches. Similarly, even short matches with relatively high proportion (over 15%) of matching landmarks seem generally accurate on the basis of casual spot-checks.

We therefore wanted to examine the precision of short matches (a single five-second clip) with low percentages of matching landmarks. In order to estimate the precision of these, we randomly sampled 1.5% of them to hand check; at least 20 samples were taken at each match percentage level. Fig. 2 shows the level of precision observed in this set of matches versus the proportion of landmarks matched. A large number of the incorrect matches were between clips which either both contained music or crowd noise. Further examination revealed that a large number of these spurious matches contain a long chain of landmarks in a single frequency bin. It seems likely that the large majority of these could be automatically identified and removed in future experiments, but this work has not been completed yet.

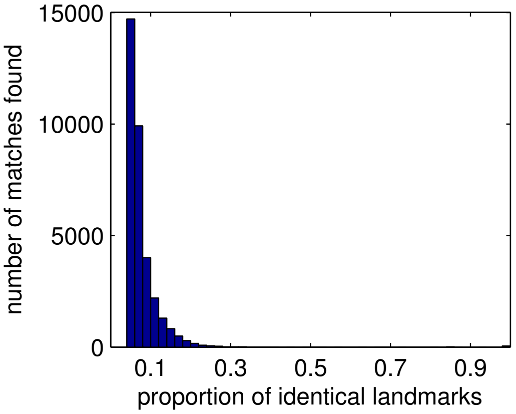

Fig. 1. Histogram of matches found, by proportion of landmarks matched.

4.3. Identifying Unique Recordings

In the process of examining matches above, a number of different types of accurate matches were observed. The most common at high landmark proportion levels were between videos of the same events taken by different news organizations. Another set were between videos which were obviously derived from the same original news recording, but with various levels of additional processing. Some of these were rebroadcasts by a news organization in another country, with additional narration or translation over the original footage. Others had been remixed with music. A surprisingly large number seemed to be videos taken of television screens. A very small number were discovered which had been taken by amateurs in attendance at the actual event.

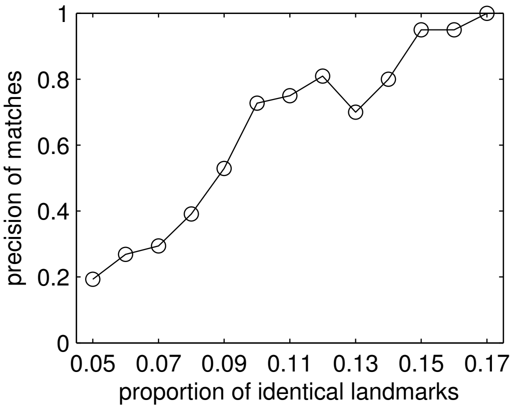

Fig. 2. Precision of five-second matches, based on manual examination of random samples.

An interesting question is how many of these independent recordings exist in the database. We observed that each of the various professional news recordings represented in the database tend to match each other well, since they are all very clean long-duration recordings of exactly the same chain of events. We attempted to estimate this subset of professional recordings by selecting any videos that match each other in at least 15% of the total landmarks, contain at least 25 actual matching landmarks, and are at least 20 seconds long. This described 691 matches, between 118 separate videos.

We expect amateur recordings to also match one or more of these professional videos, but likely for a shorter duration and/or at a smaller landmark percentage level. We therefore looked at the set of videos which match any of the presumed professional set described above, in at least 10% of the landmarks, with at least 20 actual landmarks, and with no minimum duration. This yielded a set of 2130 matches, between 189 videos (in additional to the 118 above). For each of these videos, the top (highest proportion of landmarks) match was returned for examination. Many of these videos turned out to be heavily processed or remixed versions of a professional recording. A few were actually incorrect, but commonly mistaken, videos containing either music or crowd noise. A number of them (14) were actually discovered to be independently recorded videos of the inauguration ceremony or related events, that were reliably matched with professional footage of the same events. Fig. 3 demonstrates the matching landmarks in one of these amateur video results; fig. 4 shows frames from each video. These and other examples of the matches described here can be viewed at `http://www.ee.columbia.edu/~cvcotton/vidMatch.htm`.

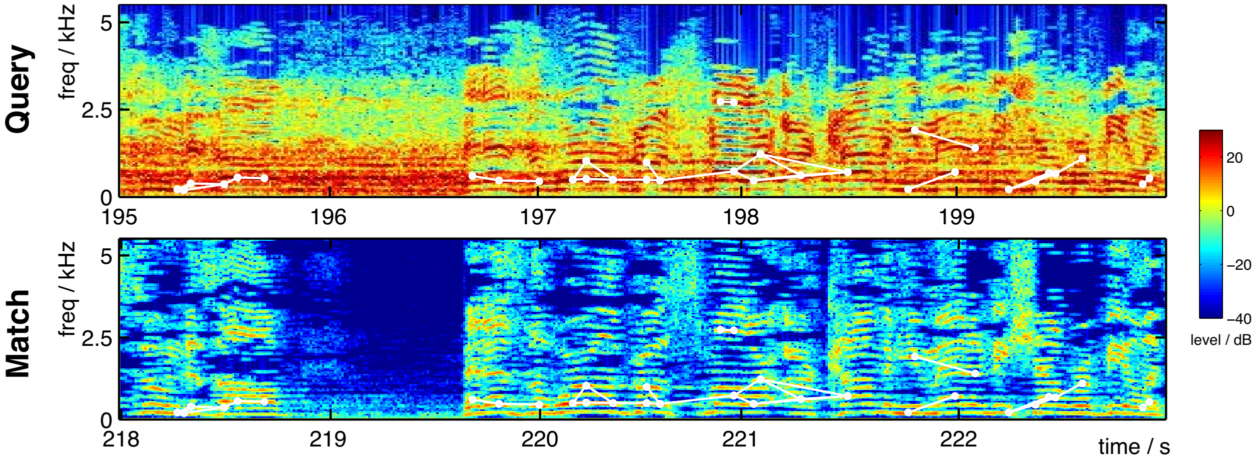

Fig. 3. Top: a clip of an amateur video of the inauguration speech; Bottom: a CNN broadcast. The two share 59 common landmarks over a 10 second period (only five seconds shown). The matching landmarks are drawn in white.

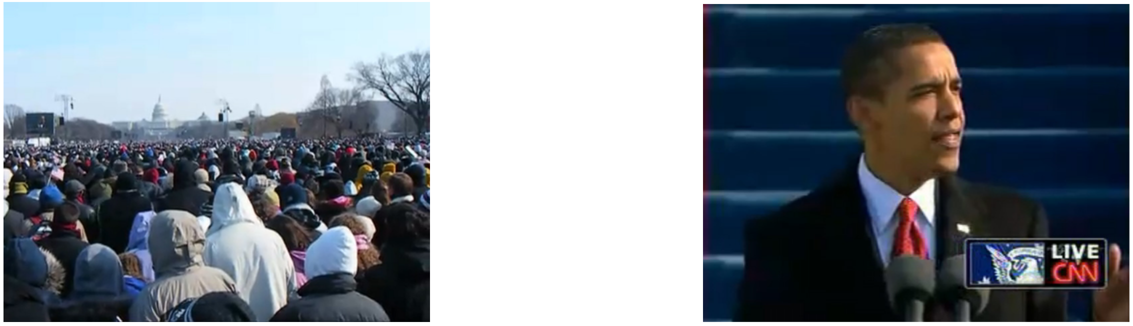

Fig. 4. Frames from each of the two matching videos.

5. CONCLUSIONS

The fingerprinting procedure outlined here was demonstrated to be robust to high levels of noise and channel differences. The system as demonstrated reliably returns accurate matches with very few false positives at a match threshold of around 15% of landmarks. The main shortcoming is the number of false positives that occur at lower match levels. We are, however, confident that many of these can be reliably removed in future experiments by filtering matches with landmarks occuring all or mostly in a single frequency bin.

6. REFERENCES

[1] P. Cano, E. Batlle, T. Kalker, and J. Haitsma, “A review of audio fingerprinting,” *J. VLSI Sig. Proc.*, vol. 41, no. 3, pp. 271–284, 2005.

[2] A. Wang, “The Shazam music recognition service,” *Comm. ACM*, vol. 49, no. 8, pp. 44–48, Aug. 2006.

[3] J. Ogle and D.P.W. Ellis, “Fingerprinting to identify repeated sound events in long-duration personal audio recordings,” in *Proc. ICASSP*, 2007, vol. I, pp. 233–236.

[4] C. Cotton and D.P.W. Ellis, “Finding similar acoustic events using matching pursuit and locality-sensitive hashing,” in *Proc. WASPAA*, 2009, pp. 125–128.

[5] P. Shrstha, M. Barbieri, and H. Weda, “Synchronization of multi-camera video recordings based on audio,” in *Proc. 15th Int. Conf. on Multimedia*. ACM, 2007, pp. 545–548.

[6] L. Kennedy and M. Naaman, “Less talk, more rock: automated organization of community-contributed collections of concert videos,” in *Proc. 18th Int. Conf. on World Wide Web*. ACM, 2009, pp. 311–320.

[7] S. Mallat and Z. Zhang, “Matching pursuits with time-frequency dictionaries,” *IEEE Tr. Sig. Proc.*, vol. 41, no. 12, Dec. 1993.

[8] Sacha Krstulovic and Rémi Gribonval, “MPTK: Matching Pursuit made tractable,” in *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP’06)*, Toulouse, France, May 2006, vol. 3, pp. III–496 – III–499.