Accessing Minimal-Impact Personal Audio Archives

Daniel P.W. Ellis and Keansub Lee
Columbia University

We've collected personal audio—essentially everything we hear—for two years and have experimented with methods to index and access the resulting data. Here, we describe our experiments in segmenting and labeling these recordings into episodes (relatively consistent acoustic situations lasting a few minutes or more) using the Bayesian Information Criterion (from speaker segmentation) and spectral clustering.

Preservation and recollection of facts and events are central to human experience and culture, yet our individual capacity to recall, while astonishing, is also famously fallible. As a result, technological memory aids date back to cave paintings and beyond. More recent trends include the shift from specific, active records (such as making notes) to transparent, comprehensive archives (such as the sent box of an email application)—which become increasingly valuable as the tools for retrieving the contents improve.

We have been investigating what we see as a natural extension of this trend to the large-scale collection of daily personal experiences in the form of audio recordings, striving to capture everything heard by the individual user during the time he or she collects archives (for example, during the working day). Our interest in this problem stems from work in content-based retrieval, which aims to make multimedia documents, such as movies and videos, searchable in much the same way that current search engines allow retrieval from text documents and archives. However, automatic indexing of movies has to compete with human annotations (for example, subtitles)—if the ability to search is important enough to people, it will be worth the effort to perform manual annotation. But for data of speculative or sparse value, where manual annotation would be out of the question, automatic annotation is a more compelling option. Recordings of everyday experiences—which may contain interesting material in much less than 1 percent of their span—are a promising target for automatic analysis.

The second circumstance that spurred our interest in this project was the sudden availability of devices capable of making these types of personal audio recordings at a low cost, with high reliability, and with minimal impact to the individual. Figure 1 shows one such device that we have used, an MP3 player with 1 Gbyte of flash memory and a built-in microphone able to record continuously for about 16 hours, powered by a single rechargeable AA battery. This kind of technology, along with the plummeting cost of mass storage, makes the collection of large personal audio archives astonishingly inexpensive and easy. However, using current tools, a 16-hour recording (less than 500 Mbytes at 64 kbps, which gives reasonable quality for a mono MPEG-Audio file) is singularly useless. To review a particular event would require loading the whole file into an audio browser and making some type of linear search: guessing the approximate time of the event of interest, then listening to little snippets, and trying to figure out whether to scan forward or backward. The time required for this type of search begins to approach the duration of the original recording, and renders any but the most critical retrieval completely out of the question.

Our interest is to develop tools and techniques that could turn these easily collected personal audio archives into something useful and worthwhile. As part of this, we're interested in imagining and discovering what uses as well as what pitfalls and limitations this data type presents. Our initial work, described in this article, considers the broad-scale information contained in such recordings, such as the user's daily locations and activities—the kind of information that someone might record in an appointment calendar. In particular, we describe our approach for dividing long-duration recordings into segments on the scale of minutes that contain consistent properties, and in clustering and classifying these segments into a few, recurrent activities. While we haven't yet developed sufficiently powerful tools to truly reveal the potential of these recordings, we're convinced that archives of this kind will, before long, become a commonplace addition to each individual's personal effects, and will become a routine source of valuable personal recollections.

Audio archives potential

Although our current experiments are limited

1070-986X/06/$20.00 © 2006 IEEE

Published by the IEEE Computer Society

in scope, it's worthwhile to consider the potential value and utility of these kinds of recordings, once the suitable indexing techniques are developed. Audio archives contrast with image or video archives in a number of important dimensions. First, they capture information from all directions and are largely robust to sensor position and orientation (and lighting), allowing data collection without encumbering the user. Second, the nature of audio is distinct from video, making certain kinds of information (for example, what is said) more accessible, and other information (for example, the presence of nonspeaking individuals) unavailable. In general, processing the content of an audio archive could provide a wide range of useful information:

- *Location*. We can characterize a physical location by its acoustic ambience, which might reveal finer gradations (for example, the same restaurant empty versus busy), although ambience is also vulnerable to confusion (for example, mistaking one restaurant for another).

- *Activity*. Different activities are in many cases easily distinguished by their sounds, for example, typing on a computer versus having a conversation versus reading.

- *People*. Speaker identification based on the acoustic properties of voice is a mature and successful technology. However, it requires some adaptation to work with the variable quality and noise encountered in personal audio.

- *Words*. Ideally, we would like to handle queries like "This topic came up in a discussion recently. What was that discussion about?" This would require not only recognizing all the words used in the earlier discussion, but summarizing and matching them. This is ambitious, although similar applications are being pursued for recordings of meetings.[1]

The "Background" sidebar (next page) reviews previous work in personal archive recording and in audio segmentation and classification.

Segmentation and clustering of personal audio

To ease the problem of locating and reviewing a particular event in a lengthy recording, we seek automatic means to generate a coarse index into the recording. At the broadest level, this index can divide a multihour recording into episodes consisting of, say, 5 minutes to an hour, during which statistical measures of the audio indicate a consistent location or activity. By segmenting the recording at changes in an appropriate statistic, then clustering the resulting segments to identify similar or repeated circumstances, a user could identify and label all episodes of a single category (for instance, attending lectures by Professor X) with minimal effort.

Figure 1. Data capture equipment. In the middle of the picture is the iRiver flash memory recorder. The larger unit to the right is a data logger recording ambient temperature, which we have considered as a proxy for more specific ground truth on location changes.

Features

Unlike audio analysis applications, such as speech recognition, that aim to distinguish audio events at a fine time scale, we're interested in segmenting and classifying much longer segments, and not becoming distracted by momentary deviations. We opted for a two-level feature scheme, with conventional short-duration features (calculated over 25-millisecond windows) summarized by statistics over a longer basic time frame of up to 2 minutes. Long time frames provide a more compact representation of long-duration recordings and might better represent background ambience properties when transient foreground events are averaged out over a longer window. We've experimented with several short-time features and several different statistics, comparing them empirically for their ability to support segmentation and clustering of our recorded episodes. We present the main results in this article; more details are available elsewhere.[2]

Our data consists of single-channel recordings resampled to 16 kHz. All features start with a conventional Fourier magnitude spectrum, calculated over 25-ms windows every 10 ms, but differ in how the 201 short-time Fourier transform (STFT)

Background

The concept of continuous, passive mechanical storage of experiences was initially articulated by Bush,[1] but it was not until almost five decades later that the technology to realize his vision became practical. Early experiments in live transmission from body-worn cameras developed into independent wearable computers,[2] but it was still several years before researchers could seriously propose comprehensive capture and storage portions of personal experience.[3]

In some early experiments, works by Clarkson et al.[4] and Clarkson[5] proposed analyzing continuous audio streams, including environmental sounds, which focused on identifying specific, distinctive acoustic events. This work eventually led to a project in which a continuous waking-hours record was collected for 100 days, and then segmented and clustered, but using features only from forward- and backward-facing fisheye video.[5]

Our work in segmenting and clustering based on recorded sound draws on work in audio segmentation. Early work on discriminating between speech and music in radio broadcasts[6] became important for excluding nonspeech segments from speech recognizers intended to work with news broadcasts.[7] Since speech recognizers can adapt their models to specific speakers, it was also important to segment speech into different speakers' turns and cluster the disjoint segments originating from the same speaker, by agglomerative clustering across likelihood ratios or measures such as the Bayesian Information Criterion, which compares likelihoods between models with differing numbers of parameters.[8] Other work in multimedia content analysis spans a number of projects to segment sound tracks into predefined classes such as speech, music, environmental sounds, and various possible mixtures.[9] Predefined classes allow model-based segmentation—for example, with hidden Markov models—but local measures of segment dissimilarity permit segmentation even when no prior classes are assumed.[10]

References

1. V. Bush, "As We May Think," *The Atlantic Monthly*, July 1945; http://www. theatlantic.com/unbound/flashbks/computer/bushf.htm.
2. S. Mann, "Wearable Computing: A First Step Toward Personal Imaging," *Computer*, Feb, 1997, pp. 25-32; http://doi.ieeecomputersociety.org/10.1109/2.566147.
3. J. Gemmel et al., "MyLifeBits: Fulfilling the Memex Vision," *Proc. ACM Multimedia*, ACM Press, 2002, pp. 235-238; http://research.microsoft.com/~jgemmell/pubs/MyLifeBitsMM02.pdf.
4. B. Clarkson, N. Sawhney, and A. Pentland, "Auditory Context Awareness via Wearable Computing," *Proc. Perceptual User Interfaces Workshop*, 1998; http://web.media.mit.edu/~nitin/NomadicRadio/PUI98/pui98.pdf.
5. B.P. Clarkson, "Life Patterns: Structure from Wearable Sensors," doctoral dissertation, MIT Media Lab, 2002.
6. E. Scheirer and M. Slaney, "Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing* (ICASSP), IEEE Press, 1997, pp. 1331-1334.
7. M.A. Siegler et al., "Automatic Segmentation, Classification and Clustering of Broadcast News Audio," *Proc. DARPA Broadcast News Workshop*, NIST, 1997, pp. 97-99; http://www.nist.gov/speech/publications/darpa97/pdf/siegler1.pdf.
8. S. Chen and P. Gopalakrishnan, "Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion," *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, NIST, 1998, pp. 127-132; http://www.nist.gov/speech/publications/darpa98/pdf/bn20.pdf.
9. T. Zhang and C.-C. J. Kuo, "Audio Content Analysis for Online Audiovisual Data Segmentation and Classification," *IEEE Trans. Speech and Audio Processing*, vol. 9, no. 4, 2001, pp. 441-457.
10. T. Kemp et al., "Strategies for Automatic Segmentation of Audio Data," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing* (ICASSP), IEEE Press, 2000, pp. III–1423-1426.

frequency bins resulting from each 400-point STFT are combined into a short-time feature vector. We compared the linear-frequency spectrum, auditory spectrum, mel-frequency cepstral coefficients (MFCCs), and spectral entropy.

We form the linear-frequency spectrum by summing the STFT bins across frequency in equal-sized blocks. The linear-frequency spectrum for time step n and frequency index j is

$$A[n,j] = \sum_{k=0}^{N/2+1} w_{jk} X[n,k] \quad (1)$$

where $X[n, k]$ are the squared magnitudes from the N point STFT, and the w_{jk} define a matrix of weights for combining the STFT bins into the more compact spectrum. We used 21 output bins to match the size of the other features.

The auditory spectrum is similarly formed as weighted sums of the STFT bins, but using windows that approximate the bandwidth of the ear—narrow at low frequencies, and broad at high frequencies—to obtain a spectrum whose detail approximates the information perceived by listeners. A spacing of 1 Bark per band gave us 21 bins, corresponding to a different matrix of w_{jk} in Equation 1.

MFCCs use a different (but similar) frequency warping, then apply a decorrelating cosine transform on the log magnitudes. MFCCs are the features most commonly used in speech recognition and other acoustic classification tasks.

To preserve some of the information lost when summing multiple STFT bins into a single value, we devised a feature to distinguish between energy distributed across the whole band, or concentrated in just a few of the component bins. By considering the distribution of energy within the subband as a probability density function (pdf), we define a short-time spectral entropy at each time step n and each spectral channel j as

$$H[n,j] = -\sum_{k=0}^{N/2+1} \frac{w_{jk}X[n,k]}{A[n,j]} \cdot log\left(\frac{w_{jk}X[n,k]}{A[n,j]}\right)$$

where the band magnitudes $A[n, j]$ from Equation 1 normalize the energy distribution within each weighted band to be pdf-like. This entropy feature can be calculated for either of the subband schemes described previously, that is, for any weight matrix w_{jk}. Spectral entropy has intent and properties similar to the well-known spectral flatness measure.[3]

To represent longer time frames of up to 2 minutes, we tried a number of statistics to combine the set of short-time feature vectors (calculated at 10-ms increments) into a single vector. We calculated the mean and standard deviation for each dimension before or after conversion to logarithmic units (dB), giving four summary vectors, μ_{lin}, σ_{in}, μ_{dB}, and σ_{dB}, respectively, all finally expressed in dB units. We also calculate the average of the entropy measure μ_H, and the entropy deviation normalized by its mean value, σ_H/μ_H. Figure 2 illustrates each of these statistics, based on the Bark-scaled auditory spectrum, for 8 hours of audio recorded on one day.

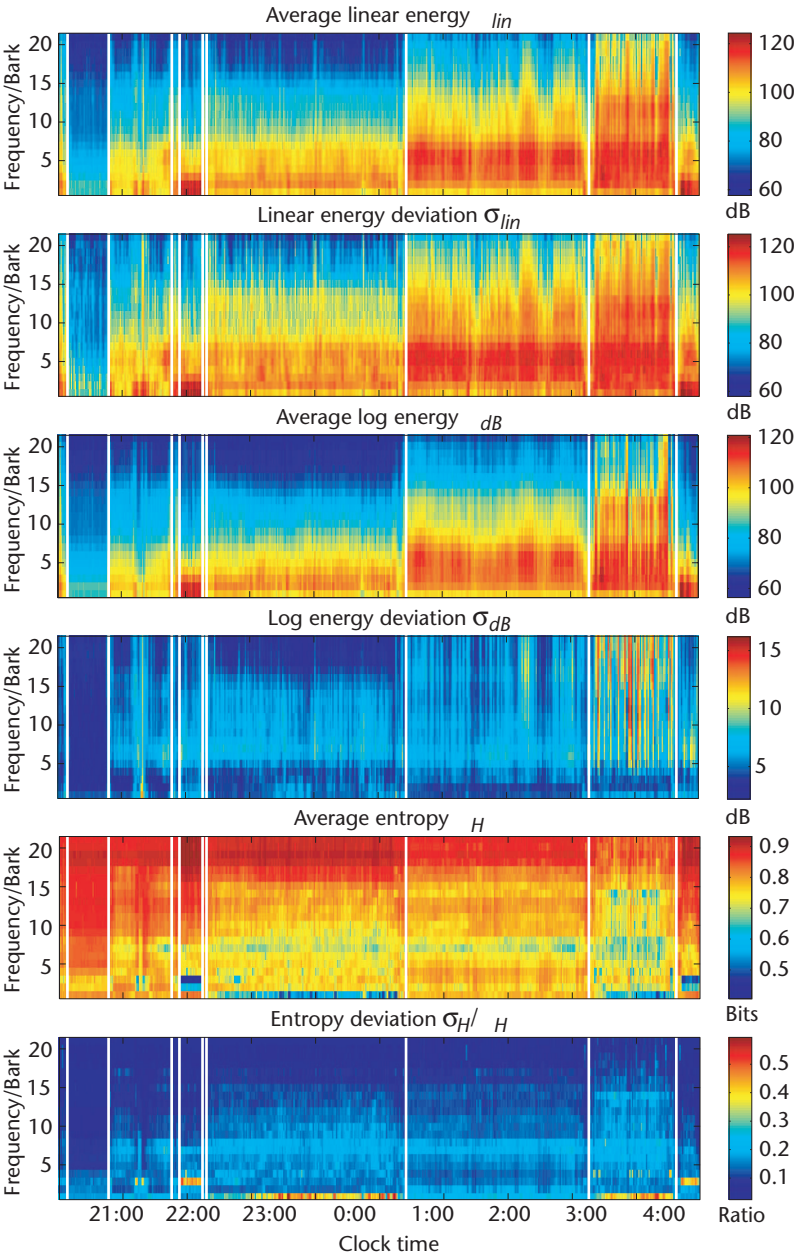

Figure 2. Examples of the six long-time-frame statistic features based on 21-band auditory (Bark-scaled) spectra. The underlying data is 8 hours of recordings including a range of locations. White vertical lines show our hand-marked episode boundaries.

Segmentation

To segment the recordings into episodes with internally consistent properties, we used the Bayesian Information Criterion (BIC). This provides a principled way to compare the likelihood of models with different numbers of parameters that describe different amounts of data. Chen and Gopalakrishnan's speaker segmentation algorithm uses the BIC to compare every possible segmentation of a window expanded until a valid boundary is found—meaning that the decisions are based on all time frames back to the previous boundary, and far enough forward until the decision is adequately confident.[4]

The BIC is a likelihood criterion penalized by model complexity as measured by the number of model parameters. If we are modeling data set $\mathcal{X} = \{x_i : i = 1, \ldots, N\}$ by some model M with $\#(M)$ parameters, and $\mathcal{L}(\mathcal{X}, M)$ is the likelihood of $\mathcal{X}$ under the best parameterization of M, then the BIC is defined as a property of the data set and model:

$$BIC(\mathcal{X}, M) = \log\mathcal{L}(\mathcal{X} \mid M) - \frac{\lambda}{2}\#(M) \cdot \log(N)$$

where λ determines the weight applied to model parameters, theoretically 1, but tunable in practice. Given several different candidate models to explain a single data set, the model with the largest BIC gives the best fit according to this criterion.

The BIC-based segmentation procedure is as follows: A sequence of d-dimensional audio fea-

ture vectors $\mathcal{X} = \{x_i \in \Re^d : i = 1, \ldots, N\}$ are modeled as independent draws from either one or two multivariate Gaussian distributions. The null hypothesis is that the entire sequence is drawn from a single distribution:

$$H_0 : \{x_1, \quad , x_N\} \sim \mathcal{N}(\ _0, \Sigma_0)$$

where $\mathcal{N}(\mu, \Sigma)$ denotes a multivariate Gaussian distribution with mean vector μ and full covariance matrix Σ, which is compared to the hypothesis that there is a segment boundary after sample t. That is, that the first t points are drawn from one distribution and that the remaining points come from a different distribution:

$$H_1 : \{x_1, \quad , x_t\} \sim \mathcal{N}(\ _1, \Sigma_1),$$
$$\{x_{t+1}, \ldots, x_N\} \sim \mathcal{N}(\ _2, \Sigma_2)$$

The difference in BIC scores between these two models is a function of the candidate boundary position t:

$$\Delta BIC(t) = \log\left(\frac{\mathcal{L}(\mathcal{X}|H_0)}{\mathcal{L}(\mathcal{X}|H_1)}\right) - \frac{\lambda}{2}\frac{d^2 + 3d}{2}\log(N)$$

where $\mathcal{L}(\mathcal{X}|H_0)$ is the likelihood of $\mathcal{X}$ under hypothesis H_0 and so on, and $(d^2 + 3d)/2$ is the number of extra parameters in the two-model hypothesis H_1. When $\Delta\text{BIC}(t) > 0$, we place a segment boundary at time t, and then begin searching again to the right of this boundary and reset the search window size N. If no candidate boundary t meets this criterion, we increase the search window size, and repeat the search across all possible boundaries t. This continues until we reach the end of the signal.

Clustering

Since recordings of daily activities are likely to contain many routine, repeated circumstances, we apply unsupervised clustering to group the automatically segmented episodes into recurrences of the same location or activity. Then, with a small amount of human input, appropriate labels can be automatically propagated within the browsing software to all members of a cluster.

We used spectral clustering, which starts from a matrix of affinities (similarities) between every segment to be clustered.[5] We begin with the symmetrized Kullback-Leibler (KL) divergence between single, diagonal-covariance Gaussian models fit to the feature frames within each segment. For Gaussians, the symmetrized KL divergence is given by

$$D_{KLS}(i,j) = \frac{1}{2}((\ _i - \ _j)'(\Sigma_i^{-1} + \Sigma_j^{-1})(\ _i - \ _j) + tr(\Sigma_i^{-1}\Sigma_j + \Sigma_j^{-1}\Sigma_i - 2\boldsymbol{I}))$$

where Σ_i is the unbiased estimate of the feature covariance within segment i, μ_i is the vector of per dimension means for that segment, $\boldsymbol{I}$ is the identity matrix, and $tr(\cdot)$ is the trace of a matrix. (Since some segments can be just a few frames long, we regularized our covariance estimates with a small empirically optimized constant added to the leading diagonal.) D_{KLS} is zero when two segments have identical means and covariances, and progressively larger as the distributions become more distinct. To convert these distances to affinities, we use a quadratic exponential mapping, so the affinity between segments i and j is given by

$$a_{ij} = exp\left(-\frac{1}{2}\frac{D_{KLS}(i,j)^2}{\sigma^2}\right)$$

where σ is a free parameter controlling the radius in distance space over which points are considered similar; increasing leads to fewer, larger clusters. We tuned it by hand to give reasonable results.

Clustering then consists in finding the eigenvectors of the affinity matrix. When the affinities indicate a clear clustering (most values close to zero or one), the eigenvectors will tend to have bimodal values, with each vector contributing a block on the diagonal of a reconstructed affinity matrix whose rows and columns have been reordered to make similar segments adjacent. In the simplest case, the nonzero elements in each of the top eigenvectors indicate the dimensions belonging to each of the top clusters in the original data. To deal with more general cases, we find K clusters in a set of K-dimensional points formed by the rows of the first K eigenvectors (taken as columns)—that is, each of the N original segments lies on a point defined by the values of the corresponding elements from the top K eigenvectors of the affinity matrix, and points with similar values across all these vectors will be clustered together. Choosing K, the desired number of clusters, is always problematic: we chose it automatically by considering every possible value up to some limit, using the size for which the Gaussian mixture model we employed for the final clustering had the best BIC score. (These details of our clustering scheme are drawn from other work.[2])

Long-duration recording experiments

Evaluating and developing our techniques required test data including ground truth for segmentation points and episode categories. We manually annotated some 62 hours of audio recorded over eight successive days, marking boundaries wherever there was a clear shift in environment and/or activity. This resulted in 139 segments (average duration 26 minutes) that we assigned to 16 broad classes such as street, restaurant, class, library, and so on. We note the risk of experimenter bias here, since the labeling was performed by the researchers who were already aware of the kinds of distinctions that would be possible or impossible for the system. Thus, although our results might be optimistic for this reason, we believe they are still indicative of these approaches' viability.

Features and segmentation results

We evaluated the BIC segmentation scheme for each of our base features and statistics by adjusting the parameter described previously to achieve a false alarm rate of one false boundary every 50 minutes (that is, 2 percent with 1-minute time frames, or a specificity of 98 percent), then looking at the resulting correct-accept rate (probability of marking a frame as a boundary given that it's a true boundary, also called sensitivity). We judged as correct a boundary placed within 3 minutes of the ground-truth position; otherwise, it was a false alarm, as were boundaries beyond the first near-to-a-ground-truth event. Table 1 compares the results from the three different short-time features (linear spectrum, auditory spectrum, and MFCC) represented by the six different summary statistics—except that spectral entropy was not calculated for the MFCCs, since the coefficients don't correspond to contiguous frequency bands.

While all features perform similarly when we use linear averaging, log domain averaging reveals a wide variation with the auditory spectrum clearly superior. The entropy measure statistics, describing the structure within each frequency band and its variation, prove the most successful basis for segmentation. We also tried combinations of the three best features—μ_{dB}, μ_H, and σ_H/μ_H—for the auditory spectrum, and used principal component analysis to compress the resulting high-dimensional feature vectors. Our best result came from combining μ_{dB} and μ_H reduced to three and four dimensions, respectively, giving a sensitivity of 0.874.

Table 1. Sensitivity at specificity = 0.98 for each feature set.

Short-Time Features	μ_{lin}	σ_{lin}	μ_{dB}	σ_{dB}	μ_H	σ_H/μ_H
Linear spectrum	0.723	0.676	0.355	0.522	0.734	0.744
Auditory spectrum	0.766	0.738	**0.808***	0.591	**0.811**	**0.816**
MFCC	0.734	0.736	0.145	0.731	N/A	N/A

*Values greater than 0.8 are shown in bold. All features had 21 dimensions.

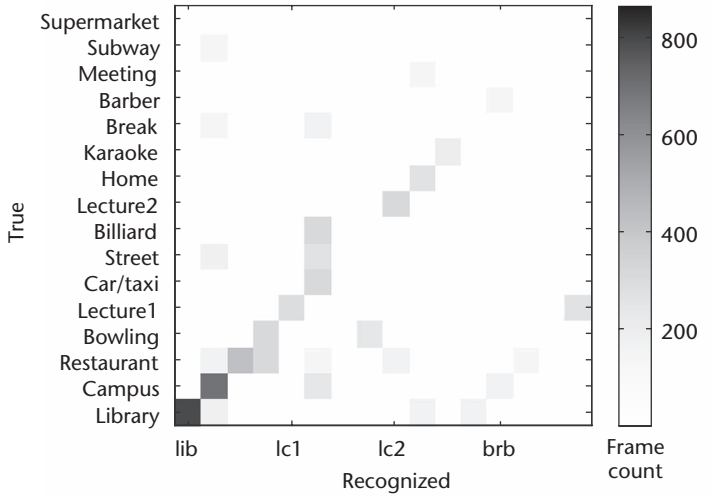

Figure 3. Confusion matrix for the 16 segment class labels, calculated over the 3,753 1-minute frames in the test data.

Clustering results

Our best segmentation scheme produced 127 automatically generated segments for our 62-hour data set. Spectral clustering (using the same average spectrum features as used for segmentation) then arranged these into 15 clusters. We evaluated these clusters by comparing them against the 16 labels used to describe the 139 ground-truth segments. Since there is no a priori association between the automatically generated segments and the hand-labeled ones, we chose this association to equate the most similar clusters in each set, subject to the constraint of a one-to-one mapping. This resulted in one ground-truth class (street) with no associated automatic cluster, and five more (billiards, class break, meeting, subway, and supermarket) for which no frames were correctly labeled, meaning the correspondences are arbitrary.

Since the automatic and ground-truth boundaries will not correspond, we evaluated the clustering at the frame level—that is, for each 1-minute time frame, the ground-truth and automatic labels were combined. Overall, the labeling accuracy at the frame level was 67.3 percent (which is also equal to the weighted average precision and recall, since the total number of frames is constant). Figure 3 shows an overall confusion matrix for the labels.

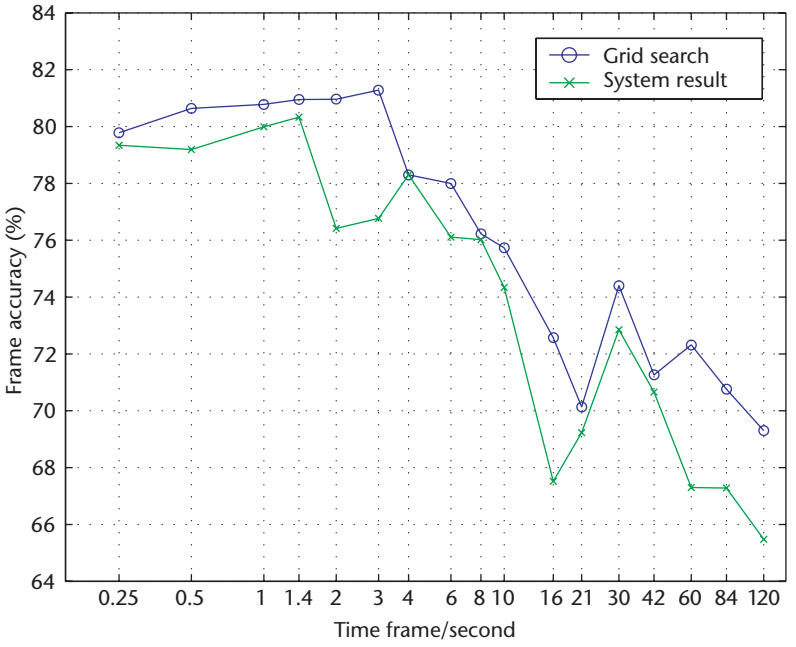

Figure 4. Effect on labeling frame accuracy of varying the basic time-frame duration.

For comparison, direct clustering of 1-minute frames without any prior segmentation, and using an affinity based on the similarity of feature statistic distributions among 1-second subwindows, gave a labeling accuracy of 42.7 percent—better than the a priori baseline of guessing all frames as a single class (26.1 percent), but far worse than our segmentation-based approach.

Varying the time frame

The previous results are based on 60-second windows, our arbitrary initial choice motivated by the granularity of the task. Returning to this parameter, we ran the entire system (both segmentation and clustering) for time frames varying from 0.25 to 120 seconds to see how this affected performance, holding other system parameters constant. Figure 4 shows the overall frame accuracy of the clustering as a function of time-frame length. The lower trace gives the system results, showing variation from 65 to over 80 percent frame accuracy, with the best results achieved at the shortest time frames, and significant degradation for time frames above 10 seconds. The upper trace shows the best result from an exhaustive grid search over the clustering parameters K and σ, giving an upper bound in performance. We see that the 3-second time frame has the best performance—arguably still long enough to capture background ambience statistics by averaging over foreground transients, but much shorter than (and distinctly superior to) the 60-second window we had used thus far.

We also experimented with basing the clustering on different features, which of course need not be the same as those used in segmentation. The results presented previously are based on the 21-dimensional log-domain average auditory spectrum μ_{dB}, which achieved a 76.8 percent frame-level labeling accuracy with the 3-second window. Using the normalized entropy deviation, σ_H/μ_H, increased this to 82.5 percent, and combining both features with the mean entropy achieved the best result of 82.8 percent.

However, we have not reported the segmentation performance—shorter time frames gave many more inserted segmentation points, which did not impact labeling accuracy because the resulting short segments were still correctly clustered on the whole. For the indexing application, however, excess segment boundaries are a problem, so labeling frame accuracy is not the only metric to consider. Larger numbers of segments also severely impact the running time of spectral clustering, which is based on the eigensolution of an $N \times N$ affinity matrix.

Visualization and browsing

We developed a prototype browsing interface, shown in Figure 5. A day-by-day pseudospectrogram visualization of the audio (using a coloring that reflects both intensity and spectral entropy distribution) lies alongside the automatically derived segments and cluster labels, as well as the user's calendar items. Audio can be reviewed by clicking on the spectrogram, along with the usual fast-forward and rewind transport controls. Our informal experiences with this interface have been mixed. It greatly facilitates finding particular events in a recording compared to the timeline slider provided by a basic media player. However, the interface has an effective resolution no better than a minute or two, and having to listen through even this much audio to reach the desired moment is still painful and boring, and would benefit from the addition of time-scaling techniques for faster review. Future directions for the interface include the addition of further data streams, such as synchronization with explicit note taking (as in Stifelman et al.[6]), or other timeline-oriented data such as documents and emails.

Speech and privacy

Initially, our interest was in the nonspeech background ambience in the audio signals as we

consider this a neglected topic in audio analysis. However, it has become clear that the speech content is the richest and most engaging information in our recordings, both for practical and reminiscence purposes. To this end, we're developing a robust speech detector that we intend to use for identifying fragments of speech amid noisy and reverberant backgrounds as encountered in our data. Dividing into speech and nonspeech segments allows both purer modeling of background ambience (for location recognition) as well as more focused processing of speech. Identifying interactions with particular speakers would be useful for access, as, of course, would recognizing the spoken content—for example, by using the techniques being developed for meeting transcription.[1]

This, however, brings us squarely into the domain of privacy concerns. This project readily arouses resistance and suspicion from acquaintances who find the idea of recording conversations threatening and creepy. We must address such concerns before people can widely accept and use this type of application. While segmentation requires only the long-time-frame statistics (which do not contain sufficient information for resynthesis to audio), much of the data's usefulness is lost unless users have the ability to listen to the original audio. Sufficiently accurate speaker identification could enable the retention of intelligible utterances only if the speaker has given explicit permission, along the lines of the "revelation rules" in the location-tracking system of Lamming and Flynn.[7]

We're also looking at ways of securing the recordings against unauthorized access. An intriguing technique for cooperative computing breaks the data into two individually useless parts—for example, by adding and subtracting the same random sequence to the original waveform—which are distributed to two agents or locations, who then permit computation of derived features (such as our time-frame statistics) without either party having access to the full data.[8]

Conclusions

Ubiquitous, continuous recordings seem bound to become a part of our arsenal of personal records as soon as the retrieval and privacy issues are tackled, since, for audio-only recordings, the collection technology is already quite mature. While the most compelling applications for this data remain to be clarified, we're intrigued and encouraged by our investigations so far. **MM**

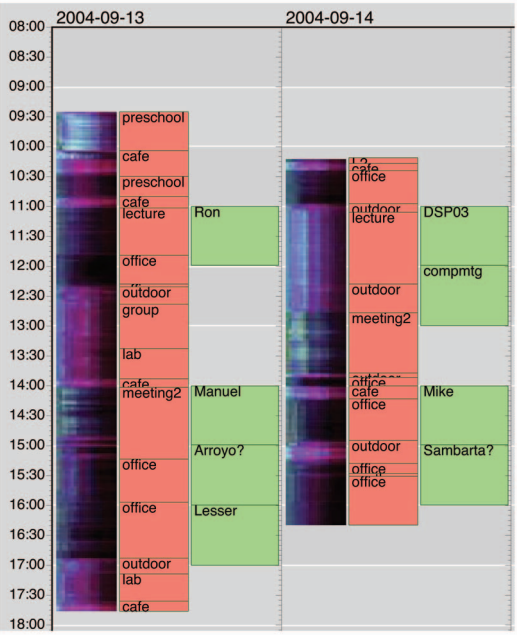

Figure 5. Screen shot from our experimental browser.

Acknowledgments

We thank the editors and reviewers for their helpful comments. This material is based in part upon work supported by the National Science Foundation (NSF) under grant no. IIS-0238301 "The Listening Machine." Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF. An earlier version of this article appeared elsewhere.[2]

References

1. S. Renals and D.P.W. Ellis, "Audio Information Access from Meeting Rooms," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing* (ICASSP), vol. IV, IEEE Press, 2003, pp. 744-747; http://www.dcs.shef.ac.uk/~sjr/pubs/2003/icassp03-mtg.html.
2. D.P.W. Ellis and K.-S. Lee, "Minimal-Impact Audio-Based Personal Archives," *Proc. 1st ACM Workshop Continuous Archival and Retrieval of Personal Experiences* (CARPE), ACM Press, 2004, pp. 39-47; http://www.ee.columbia.edu/~dpwe/pubs/carpe04-minimpact.pdf.
3. J.D. Johnston, "Transform Coding of Audio Signals Using Perceptual Noise Criteria," *IEEE J. Selected Areas in Comm.*, vol. 6, no. 2, 1988, pp. 314–323.
4. S. Chen and P. Gopalakrishnan, "Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion,"

Proc. DARPA Broadcast News Transcription and Understanding Workshop, NIST, 1998, pp.127-132; http://www.nist.gov/speech/publications/darpa98/pdf/bn20.pdf.
5. A. Ng, M. Jordan, and Y. Weiss, "On Spectral Clustering: Analysis and an Algorithm," *Advances in NIPS*, MIT Press, 2001; http://citeseer.ist.psu.edu/ng01spectral.html.
6. L. Stifelman, B. Arons, and C. Schmandt, "The Audio Notebook: Paper and Pen Interaction with Structured Speech," *Proc. ACM SIGCHI Conf. Human Factors in Computing Systems*, ACM Press, 2001, pp. 182-189; http://portal.acm.org/citation.cfm?id=365096.
7. M. Lamming and M. Flynn, "Forget-Me-Not: Intimate Computing in Support of Human Memory," *Proc. FRIEND21, Int'l Symp. Next Generation Human Interface*, Rank Xerox, 1994, pp. 125-128; http://www.lamming.com/mik/Papers/fmn.pdf.
8. W. Du and M.J. Atallah, "Privacy-Preserving Cooperative Statistical Analysis," *Proc. 17th Ann. Computer Security Applications Conf.*, IEEE CS Press, 2001, pp. 102-110; http://citeseer.ist.psu.edu/article/du01privacypreserving.html.

Daniel P.W. Ellis is an associate professor in the electrical engineering department at Columbia University in New York, where he runs LabROSA, and is an external fellow at the International Computer Science Institute, Berkeley. His research interests include extracting high-level information from audio, speech, and music. Ellis has a PhD from the Massachusetts Institute of Technology.

Keansub Lee is a PhD candidate at Columbia University. His research interests include auditory scene analysis of real-world personal audio. Lee has a BS in electronics engineering from Kyung-Hee University and an MS in electrical engineering from Korea University.

Readers may contact Daniel P.W. Ellis at the Dept. of Electrical Engineering, Columbia Univ., Mail Code 4712, Room 1312, 500 W. 120th St., New York, NY 10027; dpwe@ee.columbia.edu.

IEEE MultiMedia

FUTURE ISSUES

January—March 2007
Innovative Multimedia

April—June 2007
Visionary Media

July—September 2007
Multimedia Breakthroughs

October—December 2007
Multimedia Signal Processing in Life Sciences and Healthcare