Automatically segmenting and clustering minimal-impact personal audio archives

Daniel P.W. Ellis and Keansub Lee
LabROSA, Dept. of Electrical Engineering
Columbia University
New York NY 10027 USA
{dpwe,kslee}@ee.columbia.edu

May 20, 2006

Abstract

To capture essentially everything that you hear takes little more than a $100 MP3 player with a built-in microphone; a year's worth of recordings is maybe 60 GB, or a small stack of writable DVDs. We have been collecting this kind of 'personal audio' on and off for a couple of years, and experimenting with methods to index and access the resulting data. Audio archives have several distinctive features when compared to other kinds of artificial memory, not least the serious privacy issues that arise when recording conversations – which may be one reason they have previously received almost no attention from the research community. At the same time, continuous audio archives are minimally intrusive to collect, contain a great deal of valuable information, and present some very interesting challenges in providing convenient and useful access. We describe our experiments in segmenting and labeling these recordings into 'episodes' (relatively consistent acoustic situations lasting a few minutes or more) using the BIC criterion (from speaker segmentation) and spectral clustering. We also briefly discuss our experiences in browsing and scanning the data, and our plans and ideas concerning future directions both for our research and for the technology of personal audio recording as a whole.

Keywords: Audio Segmentation, Audio Clustering, Personal Archives, Environment Recognition, Multimedia Content Analysis

1 Introduction

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-mail" box of an email appli-

cation) – 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 archives are collected (e.g. 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 fast and powerful retrieval from text documents and archives. However, automatic indexing of movies has to compete with human annotations (e.g. 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 much more compelling option. Recordings of everyday experiences, which may contain interesting material in much less than 1% of their span, are a promising target for automatic analysis.

The second spur to our interest in this project was the sudden availability of devices capable of making these kind of personal audio recordings at low cost, with high reliability, and with minimal impact to the individual. Figure 1 shows one device we have used, an MP3 player with 1GB 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 cheap and easy. However, using current tools a 16 hour recording (under 500MB at 64 kbps, which gives very 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 kind 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 forwards or backwards. The time required for this kind 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 are interested in imagining and discovering what kind of uses (and what pitfalls and limitations) this kind of data presents. Our initial work, described in this paper, considers the broad-scale information contained in such recordings, such as the daily locations and activities of the user – the kind of information that might be recorded 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 do not, as yet, feel that we have developed sufficiently powerful tools to truly reveal the potential of these recordings, we are 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.

1.1 The potential of audio archives

Although our current experiments are quite limited in scope, it is worthwhile taking a little time 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. Firstly, they capture information from all directions and are largely robust to sensor position and orientation (and lighting), allowing data collection without encumbering the user. Secondly, the nature of audio is distinct from video, making certain kinds of information (e.g. what is said) more accessible, and other information (e.g. the presence of nonspeaking individuals) unavailable. In general, processing the content of an audio archive could provide a wide range of useful information, including:

- **Location:** A physical location can be characterized by its acoustic ambience, which may even reveal finer gradations (e.g. the same restaurant empty vs. busy), although ambience is also vulnerable to confusions (e.g. mistaking one restaurant for another).
- **Activity:** Different activities are in many cases easily distinguished by their sounds e.g. typing on a computer vs. having a conversation vs. reading.
- **People:** Speaker identification based on the acoustic properties of voice is a mature and successful technology [Reynolds, 2002]. 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 of the earlier discussion, but summarizing and matching them. This is ambitious, although similar applications are being pursued for recordings of meetings [Renals and Ellis, 2003].

In the next section we review background, both in personal archive recording and in audio segmentation and classification. We then describe our processing of personal audio recordings, considering the features appropriate for long-duration recordings, identifying segmentation points, and clustering and classifying the resulting segments. Next, we discuss our initial efforts at displaying and interacting with this data, and in integrating it with other 'scavenged' data such as online calendars. Finally, we describe our current and future work that focusses more on specific events – particularly speech – and some of the important privacy issues raised by this kind of technology.

2 Background

The concept of continuous, passive mechanical storage of experiences was initially articulated by Bush [1945], 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 [Mann, 1997],

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.

but it was still several years before researchers could seriously propose comprehensive capture and storage portions of personal experience [Gemmel et al., 2002].

Analyzing continuous audio streams – including environmental sounds – was proposed in some early experiments of Clarkson et al. [1998] and Clarkson [2002] 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 fish-eye video.

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 [Scheirer and Slaney, 1997] became important for excluding non-speech segments from speech recognizers intended to work with news broadcasts [Siegler et al., 1997]. Since speech recognizers are able to '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 (BIC), which is better able to compare likelihoods between models with differing numbers of parameters [Chen and Gopalakrishnan, 1998]. 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 [Zhang and Kuo, 2001]. Predefined classes allow model-based segmentation e.g. with hidden Markov models (HMMs), but local measures of segment dissimilarity permit

segmentation even when no prior classes are assumed [Kemp et al., 2000].

3 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 multi-hour 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. Below, we describe our approaches for extracting features, locating segmentation points, and clustering the resulting episodes.

3.1 Features

Unlike audio analysis applications such as speech recognition that aim to distinguish audio events at a fine time scale, we are 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-time features (calculated over 25 ms windows) being summarized by statistics over a longer basic time-frame of up to 2 min. Long time-frames provide a more compact representation of long-duration recordings and also have the advantage that the properties of the background ambience may be better represented when transient foreground events are averaged out over a longer window. We have experimented with several different short-time features and several different statistics, and compared them empirically for their ability to support segmentation and clustering of our 'episodes'. The main results are presented below; for more details see Ellis and Lee [2004a].

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) frequency bins resulting from each 400-point STFT are combined together into a short-time feature vector. We compared:

- **Linear-Frequency Spectrum**, formed by summing the STFT bins across frequency in equal-sized blocks. The Linear-Frequency Spectrum for time step n and frequency index j is:

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

  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.

- **Auditory Spectrum**, 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, in some sense, 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 eqn. 1 above.

- **Mel-frequency Cepstral Coefficients (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.

- **Spectral Entropy**: 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 pdf, we define a short-time *spectral entropy* at each time step n and each spectral channel j as:

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

  where the the band magnitudes $A[n,j]$ from eqn. 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 above i.e. for any weight matrix w_{jk}. Spectral entropy has intent and properties similar to the well-known spectral flatness measure [Johnston, 1988].

To represent longer time frames of up to 2 min, 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_{lin}, 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 eight hours of audio recorded on one day.

3.2 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. The speaker segmentation algorithm of Chen and Gopalakrishnan [1998] uses BIC to compare every possible segmentation of a window that is 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.

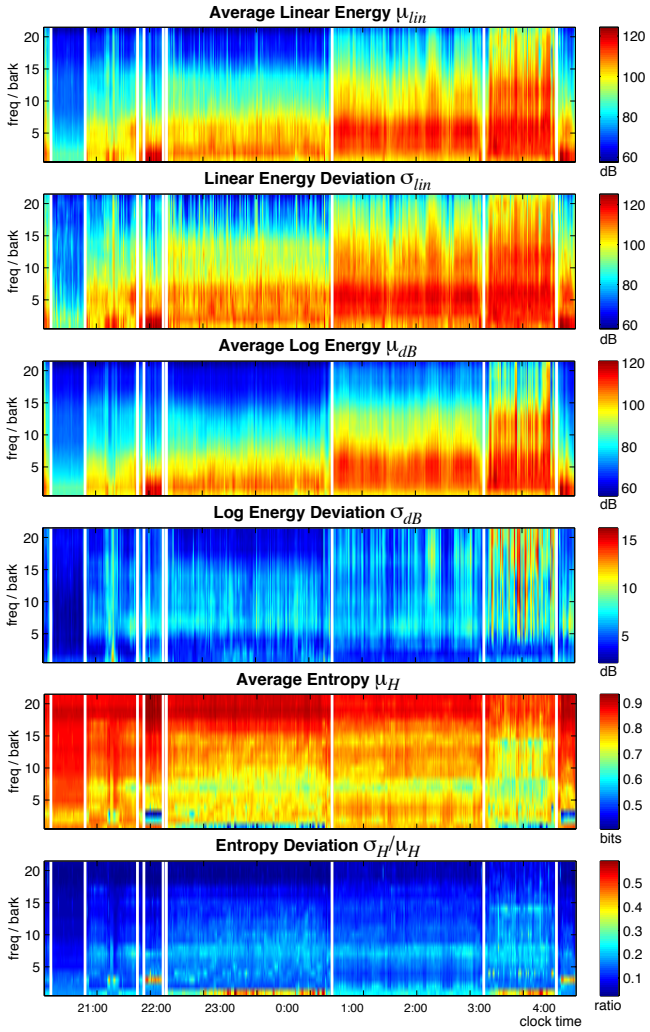

Figure 2: Examples of the six long-time-frame statistic features based on 21-band auditory (Bark-scaled) spectra. The underlying data is eight hours of recordings including a range of locations. White vertical lines show our hand-marked episode boundaries (see text).

BIC is a likelihood criterion penalized by model complexity as measured by the number of model parameters. If we are modeling a dataset $\mathcal{X} = \{x_i : i = 1, ..., N\}$ 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 BIC is defined as a property of the dataset and model:

$$BIC(\mathcal{X}, M) = log\mathcal{L}(\mathcal{X}, M) - \frac{\lambda}{2}\#(M) \cdot log(N) \tag{3}$$

λ determines the 'weight' applied to model parameters, theoretically 1, but tunable in practice. Given several different candidate models to explain a single dataset, 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 feature vectors $\mathcal{X} = \{x_i \in \Re^d : i = 1, ..., N\}$ 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:

$$H_0 : \{x_1, ..., x_N\} \sim \mathcal{N}(\mu_0, \Sigma_0) \tag{4}$$

(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 i.e. that the first t points are drawn from one distribution and that the remaining points come from a different distribution:

$$\begin{aligned} H_1 :&\{x_1, ..., x_t\} \sim \mathcal{N}(\mu_1, \Sigma_1), \\ &\{x_{t+1}, ..., x_N\} \sim \mathcal{N}(\mu_2, \Sigma_2) \end{aligned} \tag{5}$$

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

$$\Delta BIC(t) = log\left(\frac{\mathcal{L}(\mathcal{X}|H_0)}{\mathcal{L}(\mathcal{X}|H_1)}\right) - \frac{\lambda}{2}\frac{d^2 + 3d}{2}log(N) \tag{6}$$

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

3.3 Clustering

Since recordings of daily activities are likely to contain a great 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 propagated automatically to all members of a cluster.

We used spectral clustering [Ng et al., 2001] which starts from a matrix of 'affinities' (similarities) between every segment to be clustered. 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:

$$D_{KLS}(i,j) = \frac{1}{2}\Big((\mu_i - \mu_j)'(\Sigma_i^{-1} + \Sigma_j^{-1})(\mu_i - \mu_j) + tr(\Sigma_i^{-1}\Sigma_j + \Sigma_j^{-1}\Sigma_i - 2\mathbf{I})\Big) \quad (7)$$

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, $\mathbf{I}$ 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:

$$a_{ij} = exp\left(-\frac{1}{2}\frac{D_{KLS}(i,j)^2}{\sigma^2}\right) \quad (8)$$

σ is a free parameter controlling the radius in distance space over which points are considered similar; increasing σ 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) – i.e. 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 used for the final clustering had the best BIC score. (These details of our clustering scheme are drawn from Ellis and Lee [2004b].)

4 Experiments with Long-Duration Recordings

Evaluating and developing our techniques required test data including ground truth for segmentation points and episode categories. We manually annotated some 62 h of

audio recorded over 8 successive days (by author KL), marking boundaries wherever there was a clear shift in environment and/or activity. This resulted in 139 segments (average duration 26 min) which we assigned to 16 broad classes such as 'street', 'restaurant', 'class', 'library' etc. 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 may be optimistic for this reason, we believe they are still indicative of the viability of these approaches.

4.1 Features and Segmentation Results

We evaluated the BIC segmentation scheme for each of our base feature/statistics by adjusting the λ parameter described above to achieve a false alarm rate of one false boundary every 50 min (i.e. 2% with 1 min time-frames, or a specificity of 98%), then looking at the resulting correct-accept rate (probability of marking a frame as a boundary given that it is a true boundary, also called sensitivity). A boundary placed within 3 min of the ground-truth position was judged correct, 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.

Table 1: Sensitivity @ Specificity = 0.98 for each feature set. Values greater than 0.8 are shown in bold. All features had 21 dimensions.

Short-time ftrs	μ_{lin}	σ_{lin}	μ_{dB}	σ_{dB}	μ_H	σ_H/μ_H
Linear Spec	0.723	0.676	0.355	0.522	0.734	0.744
Auditory Spec	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

It is interesting to note that while all features perform similarly when linear averaging is used, log-domain averaging reveals a wide variation with the auditory spectrum clearly superior. The statistics of the entropy measure, describing the structure within each frequency band and its variation, prove the most successful basis for segmentation. We also tried combinations of the 3 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 3 and 4 dimensions respectively, giving a sensitivity of 0.874.

4.2 Clustering Results

Our best segmentation scheme produced 127 automatically-generated segments for our 62 h 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

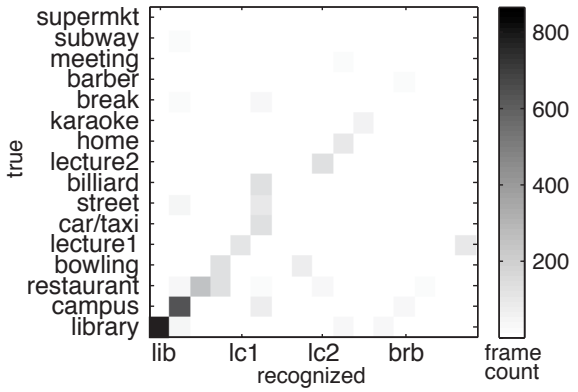

Figure 3: Confusion matrix for the sixteen segment class labels, calculated over the 3753 one-minute frames in the test data.

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 "supermkt") for which no frames were correctly labeled, meaning the correspondences are arbitrary.

Since the automatic and ground-truth boundaries will not correspond, we evaluate the clustering at the frame level i.e. for each 1 min time-frame, the ground-truth and automatic labels were combined. Overall, the labeling accuracy at the frame level was 67.3% (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.

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

4.3 Varying the time-frame

The results above are based on 60 s 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 s to 120 s 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% frame accuracy, with the best results achieved at the shortest time frames, and significant degradation for time-frames above 10 s. The upper trace shows the best result from

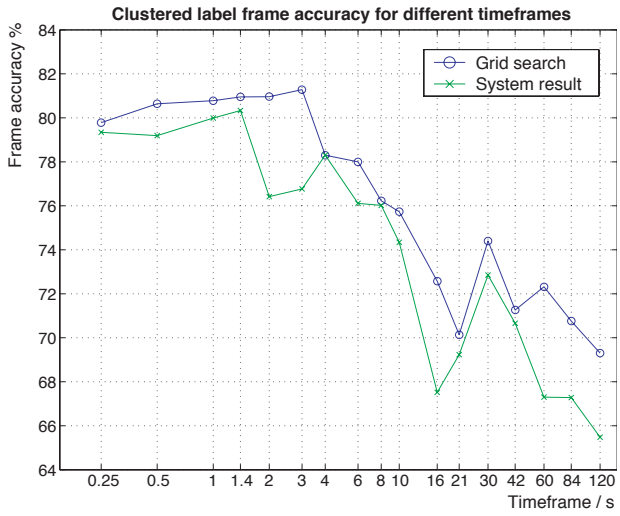

Figure 4: Effect on labeling frame accuracy of varying the basic time-frame duration.

an exhaustive grid search over the clustering parameters K and σ, giving an upper bound in performance. We see that 3 s is the time-frame with 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 s 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 above are based on the 21-dimensional log-domain average auditory spectrum μ_{dB}, which achieved a 76.8% frame-level labeling accuracy with the 3 s window. Using the normalized entropy deviation, σ_H/μ_H increased this to 82.5%, and combining both features with the mean entropy achieved the best result of 82.8%.

Note, however, that we have not reported the segmentation performance – shorter time frames gave many more inserted segmentation points, which did not, however, impact labeling accuracy since 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 eigen-solution of an $N \times N$ affinity matrix.

5 Discussion

5.1 Visualization and browsing

We have developed a prototype browsing interface, shown in figure 5. A day-by-day pseudo-spectrogram visualization of the audio (using a coloring that reflects both in-

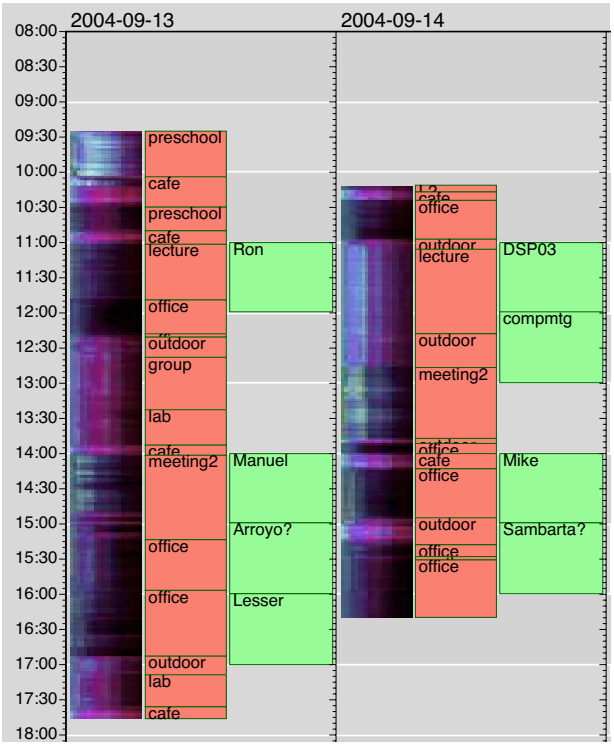

Figure 5: Screenshot from our experimental browser. Recorded audio is shown by a pseudocolor spectrogram with a vertical time axis. Next to this are the automatically-derived segments along with their per-cluster manual labels. The display also shows appointments read from the user's online calendar – a useful prompt in navigating the recordings and interpreting the automatic segments.

tensity 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/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 notetaking (as in Stifelman et al. [2001]), or other timeline-oriented data such as documents and emails.

5.2 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 information and 'reminiscence' purposes. To this end, we are developing a robust speech detector that we intend to be able to identify 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 – e.g. by making use of the techniques being developed for meeting transcription [Renals and Ellis, 2003].

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 an application of this kind can become widely accepted and useful. While segmentation requires only the long-time-frame statistics (which do not contain sufficient information for resynthesis to audio), much of the usefulness of the data 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 [1994]. If recorders become more pervasive, they could be made to respect an "opt-out" (or opt-in) beacon along the lines of Brassil [2005].

We are also looking at ways of securing the recordings against unauthorized access. An intriguing technique for co-operative computing breaks the data into two individually-useless parts (e.g. by adding and subtracting the same random sequence to the original waveform) which are distributed to two agents or locations, then permits computation of derived features (such as our time-frame statistics) without either party having access to the full data [Du and Atallah, 2001].

6 Conclusions

We have described a vision of personal audio archives and presented our initial work on providing automatic indexing based on the statistics of frequency-warped short-time energy spectra calculated over windows of seconds or minutes. Our automatically-clustered segments can be grouped into similar or recurring classes which, once the unknown correspondence between automatic and ground-truth labels is resolved, gives frame-level accuracies of over 80% on our 62 h hand-labeled test set.

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 are intrigued and encouraged by our investigations so far.

Acknowledgment

Our thanks go to 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 paper appeared as Ellis and Lee [2004b].

References

J. Brassil. Using mobile communications to assert privacy from video surveillance. In *Proc. 1st Intl. Workshop on Security in Systems and Networks*, April 2005. URL `http://www.hpl.hp.com/personal/Jack_Brassil/cloak.pdf`.

V. Bush. As we may think. *The Atlantic Monthly*, July 1945. URL `http://www.theatlantic.com/unbound/flashbks/computer/bushf.htm`.

S. Chen and P. Gopalakrishnan. Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion. In *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, 1998. URL `http://www.nist.gov/speech/publications/darpa98/pdf/bn20.pdf`.

B. Clarkson, N. Sawhney, and A. Pentland. Auditory context awareness via wearable computing. In *Proc. Perceptual User Interfaces Workshop*, 1998. URL `http://web.media.mit.edu/~nitin/NomadicRadio/PUI98/pui98.pdf`.

B. P. Clarkson. *Life patterns: structure from wearable sensors*. PhD thesis, MIT Media Lab, 2002.

W. Du and M. J. Atallah. Privacy-preserving co-operative statistical analysis. In *Proc. 17th Annual Computer Security Applications Conf.*, pages 102–110, New Orleans,

Louisiana, USA, December 10-14 2001. URL http://citeseer.ist.psu.edu/article/du01privacypreserving.html.

D. P. W. Ellis and K.-S. Lee. Features for segmenting and classifying long-duration recordings of "personal" audio. In *Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing SAPA-04*, Jeju, Korea, October 2004a. URL http://www.ee.columbia.edu/~dpwe/pubs/sapa04-persaud.pdf.

D. P. W. Ellis and K.-S. Lee. Minimal-impact audio-based personal archives. In *Proceedings of the 1st ACM Workshop on Continuous Archival and Retrieval of Personal Experiences (CARPE)*, New York, NY, October 2004b. URL http://www.ee.columbia.edu/~dpwe/pubs/carpe04-minimpact.pdf.

J. Gemmel, G. Bell, R. Lueder, S. Drucker, and C. Wong. MyLifeBits: Fulfilling the Memex vision. In *Proc. ACM Multimedia*, pages 235–238, Juan-les-Pins, France, Dec 2002. URL http://research.microsoft.com/~jgemmell/pubs/MyLifeBitsMM02.pdf.

J. D. Johnston. Transform coding of audio signals using perceptual noise criteria. *IEEE J. Selected Areas in Comm.*, 6(2):314–323, Feb 1988.

T. Kemp, M. Schmidt, M. Westphal, and A. Waibel. Strategies for automatic segmentation of audio data. In *Proc. IEEE Int. Conf. on Acous., Speech, and Sig. Proc. (ICASSP)*, pages III–1423–1426, Istanbul, 2000.

M. Lamming and M. Flynn. Forget-me-not: Intimate computing in support of human memory. In *Proc. FRIEND21, 1994 Int. Symp. on Next Generation Human Interface*, Meguro Gajoen, Japan, 1994. URL http://www.lamming.com/mik/Papers/fmn.pdf.

S. Mann. Wearable computing: A first step toward personal imaging. *IEEE Computer Magazine*, pages 25–32, Feb 1997. URL http://ieeexplore.ieee.org/iel4/2/12264/00566147.pdf.

A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in NIPS*. MIT Press, Cambridge MA, 2001. URL http://citeseer.ist.psu.edu/ng01spectral.html.

S. Renals and D. P. W. Ellis. Audio information access from meeting rooms. In *Proc. IEEE Int. Conf. on Acous., Speech, and Sig. Proc. (ICASSP)*, Hong Kong, 2003. URL http://www.dcs.shef.ac.uk/~sjr/pubs/2003/icassp03-mtg.html.

D. Reynolds. An overview of automatic speaker recognition technology. In *Proc. IEEE Int. Conf. on Acous., Speech, and Sig. Proc. (ICASSP)*, Orlando, FL, 2002.

E. Scheirer and M. Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. In *Proc. IEEE Int. Conf. on Acous., Speech, and Sig. Proc. (ICASSP)*, 1997.

M. A. Siegler, U. Jain, B. Raj, and R. M. Stern. Automatic segmentation, classification and clustering of broadcast news audio. In *Proc. DARPA Broadcast News Workshop*, 1997. URL `http://www.nist.gov/speech/publications/darpa97/pdf/siegler1.pdf`.

L. Stifelman, B. Arons, and C. Schmandt. The audio notebook: Paper and pen interaction with structured speech. In *Proc. ACM SIGCHI Conf. on Human Factors in Comp. Sys.*, pages 182–189, Seattle, WA, 2001. URL `http://portal.acm.org/citation.cfm?id=365096`.

T. Zhang and C.-C. J. Kuo. Audio content analysis for online audiovisual data segmentation and classification. *IEEE Tr. Speech and Audio Proc.*, 9(4):441–457, 2001.