# Minimal-Impact Audio-Based Personal Archives

Daniel P.W. Ellis
LabROSA, Dept. of Elec. Eng.
Columbia University
New York NY 10027 USA
dpwe@ee.columbia.edu

Keansub Lee
LabROSA, Dept. of Elec. Eng.
Columbia University
New York NY 10027 USA
kl2074@columbia.edu

## ABSTRACT

Collecting and storing continuous personal archives has become cheap and easy, but we are still far from creating a useful, ubiquitous memory aid. We view the inconvenience to the user of being 'instrumented' as one of the key barriers to the broader development and adoption of these technologies. Audio-only recordings, however, can have minimal impact, requiring only that a device the size and weight of a cellphone be carried somewhere on the person.

We have conducted some small-scale experiments on collecting continuous personal recordings of this kind, and investigating how they can be automatically analyzed and indexed, visualized, and correlated with other minimal-impact, opportunistic data feeds (such as online calendars and digital photo collections). We describe our unsupervised segmentation and clustering experiments in which we can achieve good agreement with hand-marked environment/situation labels.

We also discuss some of the broader issues raised by this kind of work including privacy concerns, and describe our future plans to address these and other questions.

## Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

## General Terms

Algorithms

## Keywords

Archives, Sound, Audio, Recording, Segmentation, Clustering, Diary

## 1. INTRODUCTION

We make huge demands of our brains' capacity to store and recall disparate facts and impressions, and it is not uncommon to be let down or frustrated by difficulty in recalling particular details; there are also aspects of past events that we rarely even hope to recall, for instance the number of hours spent in a particular location in a particular week, which might sometimes be of interest. Artificial memory aids, based on digital recordings, hold the promise of removing these limitations, but there are many technical obstacles to be overcome before this promise becomes reality.

In this paper, we consider some practical aspects of such technologies, focusing on the idea of continuous audio recordings, and other minimal- or zero-impact data collection techniques. In the next section we consider the pros and cons of using only audio-recordings (without images), as well as how this skeleton can be enhanced with other data streams. Section 2 describes our preliminary experiments in collecting and analyzing such audio data, with the goal of creating an automatic diary of locations and activities at a broad timescale. Then, in section 3, we discuss some of the issues that have arisen in this initial investigation, and give details of some specific plans for future developments.

### 1.1 Audio-based archives

We have conducted some small-scale experiments to help clarify and reveal the primary challenges in memory aid systems. We have chosen to use audio recordings, instead of video, as the foundation for our personal archive system. While the information captured by audio and video recordings is clearly complementary, we see several practical advantages to using audio only: Firstly, an omnidirectional microphone is far less sensitive to positioning or motion than a camera. Secondly, because audio data rates are at least an order of magnitude smaller than video, the recording devices can be much smaller and consume far less energy – as quantified in section 2.1. Thirdly, we note that in multimodal systems with both audio and video available, the audio can be equally or more useful than the video [13], if only because it presents a more tractable machine learning problem as a consequence of the preceding two observations.

Potentially, processing the content of an audio archive could provide a wide range of useful information:

- **Location:** Particular physical locations frequently have characteristic acoustic ambiences that can be learned and recognized, as proposed in [3]. The particular sound may even reveal finer gradations than pure physical location (e.g. the same restaurant empty vs. busy), although at the same time it is vulnerable to different confusions (e.g. mistaking one restaurant ambience for another).
- **Activity:** Different activities are in many cases easily


*CARPE'04*, October 15, 2004, New York, New York, USA.

distinguished by their sounds e.g. typing on a computer vs. having a conversation vs. reading or staring into space.

- **People:** Speaker identification based on the acoustic properties of voice is a relatively mature and successful technology [15]. An audio archive provides evidence to recognize the identity of individuals with whom the user has any significant verbal interaction.

- **Words:** The fantasy of a personal memory prosthesis is the machine that can fulfill queries along the lines of: "This topic came up in a discussion recently. What was that discussion about?", implying not only that the device has recognized all the words of the earlier discussion, but that is can also summarize the content and match it against related topics. This seems ambitious, although similar applications are being pursued for recordings of meetings [11, 14]. Capturing casual discussions also raises serious privacy concerns, to which we return in section 3.1.

  A more palatable and possibly more feasible approach is to mimic the pioneering Forget-me-not system [10] in capturing tightly-focused 'encounters' or events, such as the mention of specific facts like telephone numbers, web addresses etc. [7]. This could work as an automatic, ubiquitous version of the memo recorders used by many professionals to capture momentary ideas.

Taken together, this information can form a very rich 'diary' or personal history – all without cumbersome or custom hardware. Our vision, however is to use it as a skeleton framework into which other sources of information, discussed below, can be inserted. The guiding principle is that the data collection should have minimal impact – both in practical terms such as the weight, physical intrusion, and battery life of collection devices, and in terms of modifications required of the user's behavior to support being 'instrumented': Most users will be highly resistant to work patterns or devices that are imposed, leading to unsuccessful and short-lived products. Instead, we focus on exploiting data that is already available, or which the user can generate essentially transparently.

## 1.2 Scavenging other data sources

Given the minimal impact of collecting audio archives, we have looked for other data sources to exploit. Since users are resistant to changing their work patterns, including the software they use, our goal was to find existing information streams that could be 'scavenged' to provide additional data for a personal history/diary. The basic framework of a timeline provided by the audio recordings can be augmented by annotations derived from any time-stamped event record. This is the idea of "chronology as a storage model" proposed in Lifestreams [8] as a method of organizing documents that exploits human cognitive strengths. While our interest here is more in recalling the moment rather than retrieving documents, the activities are closely related.

Some of the time-stamped data we have identified includes:

- **Online calendars:** Many users keep maintain their calendars on their computers, and this data can usually be extracted. The calendar is of course the most familiar interface for accessing and browsing time-structured data extending over long periods, and forms the basis of our preliminary user interface.

- **E-mail logs:** E-mail interaction typically involves a large amount of time-stamped information. We have extracted all the dates from a user's sent messages store to build a profile of when (and to whom) email messages were being composed.

- **Other computer interactions:** There are many other activities at the computer keyboard than can lead to useful time logs. For instance, web browser histories constitute a rich, easily reconstituted, record of the information seen by the user. As a more specific example, the popular outliner NoteTaker [1] is frequently used for real-time note taking, and records a datestamp (down to one-second resolution) for each line entered into the outline. Dense note-taking activity can thus be extracted and presented on the calendar interface, along with the titles of pages being modified, effortlessly providing a topic description. Moreover, instead of using the time-line to organize outline entries, the outline itself – a hierarchic structuring of information created by the user – can also be used, when available, as an alternative interface to the recorded archive. Replaying the recording from immediately prior to a particular entry being made in the outline would be a very useful enhancement to written notes of talks and lectures, along the lines of the Audio Notebook [18] – but without requiring users to acquire special hardware or change their current practice, and involving only a small amount of additional software to link the existing records.

- **GPS TrackLogs:** Inexpensive personal Global Positioning System (GPS) receivers can keep a log of their positions, synchronized to a highly-accurate clock, and with minimal impact to the user as long as the device is carried (and periodically uploaded). We initially investigated this as a way to collect ground-truth tags for the segmentation experiments described in section 2, but since GPS does not work indoors (and only intermittently on the streets of built-up cities), it was not so useful. None the less, when available, GPS information giving exact location as well as motion data can provide a rich input to an automatic diary.

- **Phone records:** Phone companies typically provide detailed logs of every phone call placed (as well as calls received on mobile phones), and this information is usually available in electronic form via the web; such data can be parsed and included in the timeline view.

- **Digital photos:** Cheap digital cameras have by now almost completely eliminated analog formats, at least for casual photographers. Since these pictures are usually datestamped and uploaded to the user's personal computer, the information about when pictures were taken – and thumbnails of the images themselves – can be added to the timeline.

The common theme, in addition to the temporally-based indexing, is that each of these data streams already exists

**Figure 1: Data capture equipment. On the left is the Neuros hard-disk recorder, shown with the Soundman in-ear microphones and interface module. Middle is the mobiBLU flash-memory recorder. On the right is a logging personal GPS unit.**

and requires only minimal additional processing to be incorporated. By the same token, since the data is being opportunistically scavenged rather than carefully collected expressly for the diary, it may offer unreliable, partial coverage; users will take photographs or make phone calls only sporadically. Even in our focused efforts to collect baseline audio archives, the recorders will be used only for a few hours each day, and certain files may become corrupted or lost. These are realities of personal information, and practical applications and user interfaces should be built to accommodate them, for instance by offering multiple, partially-redundant data streams, rather than being useful only when everything 'works as planned'.

## 2. UNSUPERVISED SEGMENTATION OF PERSONAL AUDIO

For the past year we have been periodically collecting personal audio recordings, experimenting with different equipment and techniques. We have collected hundreds of hours of recordings, most of which has never been replayed – underlining the inaccessibility of this raw data. In this section we describe our equipment setup and data sets. We also give details of our first set of experiments, which investigate unsupervised segmentation and clustering of the recordings, at a coarse, one-minute timescale, in order to recover the infrequently-changing location/activity states that can form the first-level description in an automatic diary.

### 2.1 Recording Equipment

The explosion in personal audio playback devices fueled by MP3 soundfiles and typified by Apple's iPod has resulted in a very wide range of portable digital audio devices. While most of these are designed for listening only, a number of them also include the ability to record, generally presented as a way to record lectures and other specific events for later review. We have experimented with two such devices, shown in figure 1. Both devices feature built-in microphones and the memory and battery capacity to record for many hours continuously onto files that can be directly uploaded via USB connections, and both are small enough to be conveniently carried e.g. clipped to a belt.

The larger device is a Neuros recorder, including a 20G hard disk and a rechargeable battery, which retails for a few hundred dollars [5]. This is a carefully designed and engineered audio processor, able to directly encode MPEG-audio MP3 files, or alternatively write uncompressed WAV files directly to its disk. Since 20G represents 32 hours of full-rate uncompressed CD audio (and at least ten times that when compressed to MP3), the recording capacity is limited by the battery life, which was a little over 8 hours in our tests. Although the built-in microphone is mono (and picks up some noise from the internal hard disk, which spins up and down periodically), it has a high-quality stereo line input. While portable mic preamps are rare and/or expensive, we found the Soundman binaural mics (shown in the figure), with their supplied line-level adapter, to work well [17]. These mics are designed to be worn in the ears like mini headphones for high-realism binaural recordings; while multichannel recordings can capture more information about the acoustic environment, and offer the possibility of some signal separation and enhancement, using such mics obviates many of the minimal-impact advantages of audio archiving, and we have made only a few of these recordings.

The smaller device in the figure is a mobiBLU flash-memory recorder [9] with 256MB of internal memory and a slot for an "SD" memory card (although it is only able to record to internal memory). This very inexpensive device does not, in fact, encode to MP3 but instead to the lower-quality IMA ADPCM 4 bits/sample standard. Although it also has a line-in input, it cannot record in stereo. When run at 16 kHz sampling rate, it can record for over 9 hours and gives acceptable results; its long battery life (about 20 hours from two AAA rechargeable batteries) and light weight (a couple of ounces) make it an attractive alternative to the larger Neuros.

### 2.2 Data set and task

Of the many recordings made, our experiments have been focused on a particular dataset recorded over one week by author KL. This constitutes some 62 hours of data, originally recorded as 64 Mbps MP3, then downsampled to 16 kHz. The attraction of this data is that it has been manually divided into 139 segments, each corresponding to a distinct location or environment with an average segment duration of around 26 minutes. Each of these segments has also been labeled, and assigned to one of 16 broad classes, chosen to span the kinds of distinctions that would be useful in an automatic diary (such as 'street', 'restaurant', 'class', 'library' etc.).

Similar to the audio-video analysis in [4], the goal of our initial experiments was to see how well we could automatically recover these boundaries and class labels from the original data. In the absence of a larger quantity of annotated data, we chose an unsupervised approach, seeking a feature representation in which the boundaries between different episodes will be self-evident as significant shifts in statistical properties. We can then evaluate this system quantitatively by comparing it to our hand annotations, permitting fine distinctions between alternate approaches.

### 2.3 Features

For the automatic diary application, temporal resolution

on the order of one minute is plenty: most of the events we wish to identify are at least a quarter-hour long. We therefore constructed a system where the temporal frame rate was one per minute, rather than the 10 or 25 ms common in most audio recognition approaches. 25 ms is popular because even a dynamic signal like speech will have some stationary characteristics (e.g. pitch, formant frequencies) at that time scale. For characterizing acoustic environments, however, it is the stationary properties at a much longer timescale that concern us – the average level and degree of variation of energy at different frequency bands, measured over a window long enough to smooth out short-term fluctuations. Thus, we proposed a range of features applicable to one-minute frame rates, and evaluated their usefulness as a basis for unsupervised segmentation [6].

Of the features we considered, the most useful were log-domain mean energy measured on a Bark-scaled frequency axis (designed to match physiological and psychological measurements of the human ear), and the mean and variance over the frame of a 'spectral entropy' measure that provided a little more detail on the structure within each of the 21 broad auditory frequency channels. Specifically, starting from an initial short-time Fourier transform (STFT) over 25 ms windows every 10 ms to give a time-frequency energy magnitude $X[n,k]$ where $n$ indexes time and $k$ indexes the frequency bins, our auditory spectrum is:

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

where $j$ indexes the 21 one-Bark auditory frequency bands, and $w_{jk}$ specifies the weight matrix that maps STFT bins to auditory bins; $N_{FT}$ is the size of the discrete Fourier transform (DFT). Then, spectral entropy is defined as:

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

i.e. it is the 'entropy' (disorder) within each subband at a single timeframe if the original weighted-DFT magnitudes are considered as a probability distribution. The intuition here is that although the auditory bands are wide, the entropy value will distinguish between energy that is spread broadly across a subband (high entropy), versus one or two narrow energy peaks or sinusoids providing the bulk of the energy in the band. Humans are of course very sensitive to this distinction.

## 2.4 Segmentation

We used these features to identify segment boundaries in the data using the Bayesian Information Criteria (BIC) procedure originally proposed for speaker segmentation in broadcast news speech recognition [2]. In this scheme, every possible boundary position in a given window within an observation sequence is considered. New boundaries are placed when the BIC score indicates a modeling advantage to representing the features on each side of the boundary with separate statistical models instead of describing the entire window with a single model. If no acceptable boundary is found, the window is widened by advancing its upper limit until the end of the data is reached; when a new boundary is created, the lower limit of the window is moved up to the new boundary, and the search continues.

The key idea in BIC is to calculate the 'modeling advantage' as a likelihood gain (which should always be achieved when using the extra parameters of two models versus one) penalized by a term proportional to the number of added parameters. Specifically, the BIC score for a boundary at time $t$ (within an $N$ point window) is:

$$BIC(t) = log\left(\frac{\mathcal{L}(X_1^N|M_0)}{\mathcal{L}(X_1^t|M_1)\mathcal{L}(X_{t+1}^N|M_2)}\right) - \frac{\lambda}{2}\Delta\#(M)\cdot log(N) \quad (3)$$

where $X_1^N$ represents the set of feature vectors over time steps $1..N$ etc., $\mathcal{L}(X|M)$ is the likelihood of data set $X$ under model $M$, and $\Delta\#(M)$ is the difference in number of parameters between the single model ($M_0$) for the whole segment and the pair of models, $M_1$ and $M_2$, describing the two segments resulting from division. $\lambda$ is a tuning constant, theoretically one, that can be viewed as compensating for 'inefficient' use of the extra parameters in the larger model-set.

Using this procedure, any feature sequence can be segmented into regions that have relatively stable characteristics, and where the statistics of adjacent segments are significantly different. We used single, full-covariance Gaussian models to describe the multidimensional feature vectors formed from one or more of our basic spectral feature sets. Varying $\lambda$ controls the 'readiness' of the system to generate new segments i.e. trading oversegmentation (insertion of spurious segment boundaries) for undersegmentation (failure to place boundaries at 'true' changes in the underlying properties). To compare different systems, we tuned $\lambda$ to achieve a fixed false-accept rate of 2% (one spurious boundary per 50 one-minute non-boundary frames, on average), and evaluated the correct-accept rate (proportion of true boundaries marked, also called "sensitivity") of the different systems, where we accepted a boundary placed within 3 frames of the hand-marked position. Combining the two best features achieved a sensitivity of 84% for this 42-dimensional feature vector; using PCA to reduce the feature space dimensionality to 3 coefficients for the average spectral energies and 4 for the spectral entropies increased the sensitivity to 87.4%.

## 2.5 Clustering and Classification

Given a data stream divided into self-consistent segments, an automatic diary application needs to make some kind of labeling or classification for each segment. These labels will not be meaningful without some kind of supervision (human input), but even without that information, the different sequential segments can be clustered together to find recurrences of particular environments – something which is very common in a continuous, daily archive.

The automatic segmentation scheme from the previous section generated 186 segments in our 62 hour test set. We performed unsupervised clustering on these segments to identify the sets of segments that corresponded to similar situations, and which could therefore all be assigned a common label (which can be obtained from the user in a single interaction).

We used the spectral clustering algorithm [12]. First, a matrix is created consisting of the distance between each pair of segments. We use the symmetrized Kullback-Leibler (KL) divergence between single, diagonal-covariance Gaussian models fit to the feature frames within each segment.

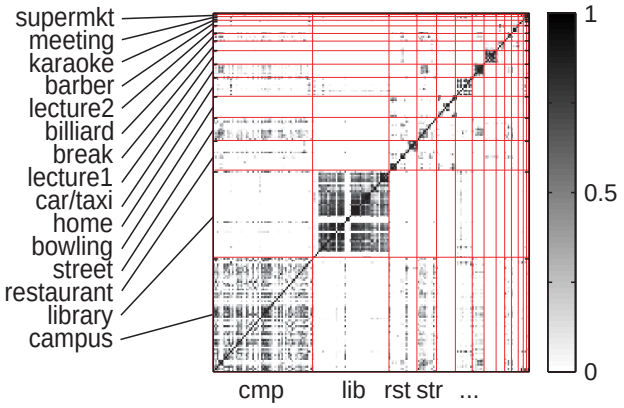


**Figure 2: Affinity matrix for the 186 automatic segments. Segments are ordered according to the dominant ground-truth label in each segment.**

For Gaussians, the symmetrized KL divergence is given by:

$$2D_{KLS}(i,j) = (\mu_i - \mu_j)'(\Sigma_i^{-1} + \Sigma_j^{-1})(\mu_i - \mu_j) + tr(\Sigma_i^{-1}\Sigma_j + \Sigma_j^{-1}\Sigma_i - 2\mathbf{I}) \quad (4)$$

where $\Sigma_i$ is the unbiased estimate of the feature covariance within segment $i$, $\mu_i$ 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. $D_{KLS}$ is zero when two segments have identical means and covariances, and progressively larger as the distributions become more distinct.

These distances are then converted to an 'affinity matrix' consisting of elements $a_{ij}$ which are close to 1 for similar segments (that should be clustered together), and close to zero for segments with distinct characteristics. $a_{ij}$ is formed as a Gaussian-weighted distance i.e.

$$a_{ij} = exp\left(-\frac{1}{2}\frac{D_{KLS}(i,j)^2}{\sigma^2}\right) \quad (5)$$

where $\sigma$ is a free parameter controlling the 'radius' in the distance space over which points are considered similar; increasing $\sigma$ leads to fewer, larger clusters. In the results below, $\sigma$ was tuned to give the best results. Figure 2 shows the affinity matrix between the 186 automatic segments.

Clustering then consists in finding the eigenvectors of the affinity matrix, which are the vectors whose outer products with themselves, scaled by the corresponding eigenvalues, sum up to give the affinity matrix. When the affinity matrix indicates a clear clustering (most values close to zero or one), the eigenvectors will tend to have binary values, with each vector contributing a block on the diagonal of the reconstructed affinity matrix whose rows and columns have been reordered to make the 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.

Assigning the data to $K$ clusters can then be achieved by fitting $K$ Gaussian components to the data using the standard EM estimation for Gaussian mixture models. This fit is performed on a set of $K$-dimensional points formed by the rows of the first $K$ eigenvectors (taken as columns). Similar segments will have similar projections in this space – along each of the axes in the simplest case – and will cluster together. The choice of $K$, the desired number of clusters, is always problematic: we considered each possible value of $K$ up to some limit, then evaluated the quality of each resulting clustering using the BIC criterion introduced above, penalizing the overall likelihood achieved by describing the data with $K$ Gaussians against the number of parameters involved.

This approach clustered the 186 automatically-generated segments into 15 clusters. We evaluate these clusters by comparing them against the 16 labels used to describe the 139 ground-truth segments. As discussed above, there is no *a priori* association between the automatically-generated segments and the hand-labeled ones; we choose this association to equate the most similar clusters in each set, subject to the constraint of a one-to-one mapping.

The results are presented in table 1. The fifteen automatic clusters were mapped onto the sixteen manual clusters, leaving "subway" with no automatic equivalent, although there were four other clusters ("car", "lecture 1", "karaoke", and "meeting") for which no frames were correctly labeled i.e. these correspondences are arbitrary.

**Table 1: Automatic clustering results, compared to manual ground truth. The first column gives the names assigned during manual annotation; the next two columns give the total number of segments, and the total number of frames (minutes) receiving this label in the ground truth. The next two columns describe the automatic segments assigned to each label based on maximum overlap. The final three columns give the raw count of correct frames for each label, and this as a proportion of the true frames (recall) and automatic frames (precision). Overall frame accuracy is 61.4% (rounded to 61% at the bottom right of the table).**

| Label | Manual | | Auto | | corr | rec | prc |
|---|---|---|---|---|---|---|---|
| | min | seg | min | seg | min | | |
| Library | 981 | 27 | 882 | 38 | 864 | 88% | 98% |
| Campus | 750 | 56 | 906 | 62 | 537 | 72% | 59% |
| Rest'r'nt | 560 | 5 | 303 | 6 | 302 | 54% | 99% |
| Bowling | 244 | 2 | 473 | 13 | 152 | 62% | 32% |
| Lect'r 1 | 234 | 4 | 104 | 5 | 0 | 0% | 0% |
| Car/taxi | 165 | 7 | 38 | 4 | 0 | 0% | 0% |
| Street | 162 | 16 | 146 | 9 | 35 | 22% | 24% |
| Billiards | 157 | 1 | 158 | 5 | 114 | 73% | 72% |
| Lect'r 2 | 157 | 2 | 176 | 9 | 152 | 97% | 86% |
| Home | 138 | 9 | 292 | 15 | 103 | 75% | 35% |
| Karaoke | 65 | 1 | 21 | 2 | 0 | 0% | 0% |
| Class brk | 56 | 4 | 51 | 3 | 26 | 46% | 51% |
| Barber | 31 | 1 | 68 | 6 | 8 | 26% | 12% |
| Meeting | 25 | 1 | 38 | 4 | 0 | 0% | 0% |
| Subway | 15 | 1 | 0 | 0 | 0 | 0% | - |
| Suprmkt | 13 | 2 | 97 | 5 | 10 | 77% | 10% |
| total | 3753 | 139 | 3753 | 186 | 2303 | 61% | 61% |

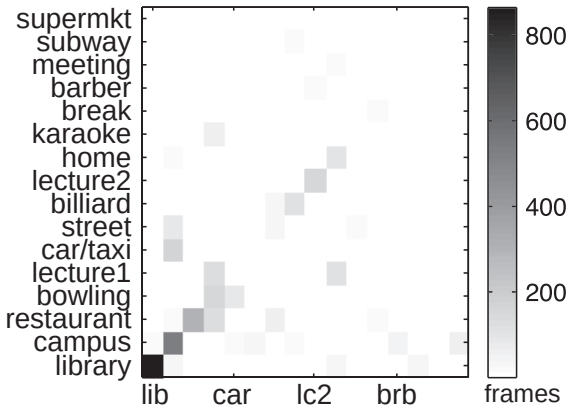


**Figure 3: Confusion matrix for the sixteen segment class labels, calculated over the 3753 one-minute frames in the test data.**

Since the automatic and ground-truth boundaries do not, in general, align, there is no perfect assignment of ground-truth segment labels to automatic segments. Instead, scoring was done at the frame level i.e. each 1 minute frame in the data was assigned a ground-truth label Overall, the labeling accuracy at the frame level was 61.4% (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.

One alternative to this segment-then-cluster procedure would be to perform clustering directly on the one-minute frames to find common patterns. This approach has the disadvantage of making no effort to assign a single label to temporally-adjacent stretches of frames. For comparison, however, we attempted this process. Using spectral clustering, the affinity between each frame was calculated by fitting a single Gaussian to the features from the 60 one-second subframes within each of one-minute frame. Measuring similarity at this finer one-minute scale resulted in only two clusters; scoring these using the same procedure as above gave a frame-level accuracy of 42.7% – better than the *a priori* baseline of guessing all frames as a single class (which gives a frame accuracy of 981/3753 or 26.1%), but far worse than the segmentation-based approach.

## 2.6 Visualization

In our initial experiments, we have sought to adapt existing tools for visualization and browsing, in preference to embarking on the development of a complex, special-purpose user interface. Figure 4 shows one example, where the data from audio segmentation and clustering are converted into calendar entries, tagged with the event class name derived from earlier labeling, and displayed in a standard program. Also displayed are the user's hand-entered appointments, and outgoing email messages identified from the "saved mail" spool.

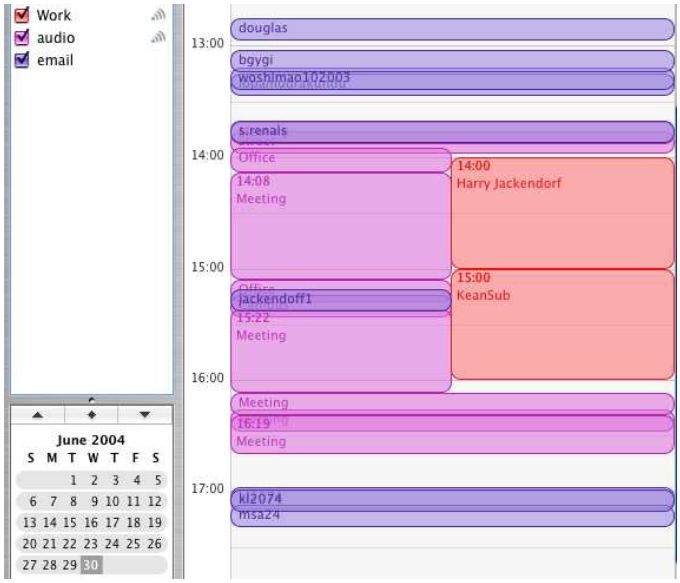


**Figure 4: Example of diary display. Red items are user-entered diary events; purple regions show segments automatically extracted from audio along with their assigned labels. Blue items are outgoing email events scavenged from the user's savebox. Audio was available only between 13:41 and 16:40.**

For access to the recorded audio, a more detailed timescale is required. Using the features described above, we have developed alternatives to the conventional spectrogram display that convey more of the information used in automatic segmentation [6]. An example timeline display is shown in figure 5; audio can be reviewed by clicking the spectrogram in this display.

# 3. DISCUSSION AND FUTURE WORK

A major goal of our initial investigations was to gain familiarity with this kind of data collection – what was easy and hard, what possible uses might occur, and what kind of influence it would have on us. This section discusses some resulting impressions, which lead naturally to our planned future developments.

## 3.1 Speech and Privacy

Initially, our interest tended toward the nonspeech background ambience in the audio signals as we consider this a neglected topic in audio analysis. However, after working with the data, it has become clear that the speech content is the richest and most engaging information in our recordings. Part of this is the manifest factual content of some speech, but another part is more nebulous – there can be a nostalgic pleasure in re-experiencing an earlier conversation, somewhat analogous to looking at snapshot photographs.

We are therefore focusing on speech in the immediate future. We have an existing system that distinguishes segments of speech from nonspeech by their characteristic properties at the output of a speech recognizer's acoustic model [20], which has the advantage of being already trained on separate data but the disadvantage that the speech it was trained on is mostly studio-quality broadcasts, acoustically

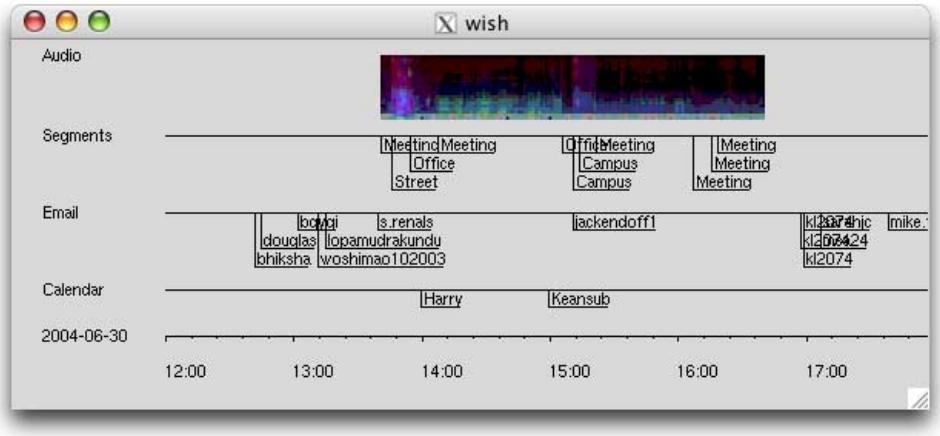


**Figure 5: Example of timeline display. The same time period as figure 4, but viewed in the custom timeline viewer. In addition to the automatic segments, audio is rendered as a psuedo-spectrogram, with intensity depending on average energy and color based on spectral entropy.**

unlike the kind of speech recorded by our body-worn devices. Better results will be obtained by training a classifier on labeled data from our own data collection, the only drawback being the obligation to perform more hand annotation. A support vector machine classifier is an excellent match to this problem and we anticipate high-accuracy results, particularly for the wearer's speech.

Once the audio record has been divided into speech and background segments, we can perform different analyses on each part. Excluding the speech segments from the ambience/location classification system described above should improve its accuracy since their will be no need to learn separate clusters for "location X" and "location X with speech in the foreground".

The speech portions constitute a conventional speaker segmentation task, which can be clustered to form separate classes for common interlocutors, the most common being the user herself. Automatic tags indicating who was involved in each conversation/meeting will certainly enhance the automatic diary function.

This, however, brings us squarely into the domain of privacy concerns. We have frequently encountered shock and resistance from acquaintances when we describe our project to create continuous audio archives. Traditionally, conversation is ephemeral, and it is disconcerting when this deeply-ingrained assumption is overturned by veridical records. These values have even been legally codified; although US Federal law permits recording of conversations in which the person making the recording is a participant, some states impose a higher standard – such as California, where it is always illegal to record a conversation without explicit permission of all parties.

Thus, we must also interpret our goal of 'minimal impact' as extending to having minimal negative impact on the privacy rights of all people with an involvement in the data collected. This issue was explicitly considered in the Forget-me-not system [10] which included the concept of "revelation rules" to govern how much information could be retained about different events at a fine level of detail. The goal of these rules was to mirror implicit assumptions about what is recalled, for instance, it may be acceptable for the automatic diary to record that a certain person was present in a particular meeting (based on voice identification), but their speech should only be retained when explicit permission has been obtained, both globally and on a per-event basis.

Privacy concerns can rightly act as a major impediment to development of these technologies, thus we also see it as a priority to provide proactive, reliable technical solutions for these issues. Given a reliable method for speech segment identification, we propose to have our system's default behavior be to scramble such segments to render the words unintelligible. One possible low-complexity technique is to break the audio into windows of, say, 50 ms, then randomly permute and reverse these segments over a 1 s radius. A large overlap between adjacent frames would make the process virtually impossible to undo. At the same time, the resulting signal, while unintelligible, would have a short-time spectrum with statistical properties similar to the original speech, meaning that speaker identification could still be applied.

Such scrambling could be suppressed for speakers for whom the system has 'permission' to make a full record, for instance the wearer. However, to err on the side of caution, we will want to tune our speech detector to make very few false rejects (so that almost all captured speech is correctly identified as such), and our identification of specific speakers should make very few false accepts (so that only speech that truly comes from 'authorized' speakers is retained unscrambled). As a failsafe starting point, we could scramble the entire audio signal and implement the scrambling within the signal capture routine so that no unscrambled signal is ever stored.

However, for speech that is captured in intelligible form, automatic speech recognition could provide valuable information for future summarization and retrieval applications. Recent progress in the recognition of natural meetings recorded with table-top mics has addressed very similar problems; we are involved in such a project and can investigate applying the same techniques [19].

## 3.2 Data capture

We are continuing our data collection activities – since,

according to the 'minimal impact' philosophy, it is easy to do so. We now have duplicate capture systems in use by several people.

A major lesson of the work so far is the expense and value of ground-truth data: anything that can be done to minimize or avoid time spent making annotations on recorded data is of great interest. As mentioned above, we have considered GPS position logs, recorded at the same time as the audio, as one way to automatically collect information on changes in location and activity. This data usually turns out to be pretty coarse, however, since it tells us only when the user was outside.

Other options we have considered include real-time annotation via date-stamped notes collected on a PDA. While preferable to offline annotation, this is intrusive and easily forgotten. A slightly less cumbersome approach is to embed in-band marker signals with some kind of 'bleeper'.[1] For instance, touch tones generated by a cell phone or self-contained 'blue box' can be played close to the recorder's microphone, then automatically located in the recorded signal, using DTMF detection routines modified to tolerate higher levels of background noise. At the very least, these can be used to quickly locate important regions during subsequent manual annotation.

## 3.3 Browsing and applications

We need a more sophisticated custom browser. The Snack audio toolkit [16] provides convenient high-level tools for incorporating sound and visual representations in graphical user interfaces, and a common XML file format will allow us to integrate timestamped event data scavenged from a wide range of other sources. Increasingly sophisticated secondary analyses, such as segment classifications and speech detection, will be incorporated as they become available.

One hope in beginning to grow archives of this kind is that previously unimaginable uses will become apparent; for us, however, the 'killer app' remains elusive. One of the main arguments for continuous data collection is that certain data will only become of interest in retrospect, as distinct from scheduled events such as lectures and important meetings that can be deliberately and explicitly recorded. A possible area for this retrospective data analysis could be health-related. For instance, if the user falls sick with a cold on a particular day, it may be possible to analyze and extract a pattern of infrequent sneezes that were actually early signs of the infection; once the pattern has been analyzed, automatic recognition of these subliminal signs could alert the user, allowing her to promptly medicate and reduce the impact of the disease.

Another use that does not merit an explicit data collection but could be of interest if the data were available 'for free' is a kind of weekly summary report, detailing how much time was spent in different activities such as meetings, handling email, travel etc. Tracking these indices over extended periods could help individuals refine and diagnose their time management issues.

# 4. CONCLUSIONS

We have described our experiments with using low-cost commercial portable audio technology to collect large-scale continuous personal recordings, and presented some initial results showing the feasibility of analyzing this data to create an 'automatic diary' of activities through the day. The main advantage of this approach in comparison with audio-video recordings is its near-zero impact on the behavior and practice of the user, while still capturing rich information. Extending this idea to include other readily available time-stamped data from sources such as email logs and phone bills brings the vision of a comprehensive automatic diary much closer.

By the same token, success and progress in this area will require sensitive consideration of the way in which these new technologies will influence existing social relations and practice. In particular, the idea of systematically recording hitherto evanescent casual conversations is deeply disturbing to many people. Finding an acceptable solution to this problem is an urgent priority, and we have proposed an approach of automatic scrambling of the short-term structure of the speech that can retain most of the essential ambient information in the audio track without violating the intimate privacy of conversations.

[1]Suggested by Harry Jackendoff.

## Acknowledgments

This material is based in part upon work supported by the National Science Foundation (NSF) under Grant No. IIS-0238301. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

# 5. REFERENCES

[1] AquaMinds Software. NoteTaker: An outlining program, 2003. http://www.aquaminds.com/.

[2] 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. http://www.nist.gov/speech/publications/darpa98/pdf/bn20.pdf.

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

[4] B. P. Clarkson. *Life patterns: structure from wearable sensors*. PhD thesis, MIT Media Lab, 2002. http://web.media.mit.edu/~clarkson/thesis.pdf.

[5] Digital Innovations. The Neuros digital audio computer, 2003. http://www.neurosaudio.com/.

[6] D. P. W. Ellis and K. 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 2004. http://www.ee.columbia.edu/~dpwe/pubs/sapa04-persaud.pdf.

[7] M. Flynn, 2004. Personal communication.

[8] E. Freeman and D. Gelernter. Lifestreams: A storage model for personal data. *SIGMOD Record (ACM Special Interest Group on Management of Data)*, 25(1):80–86, March 1996. http://citeseer.ist.psu.edu/article/freeman96lifestreams.html.

[9] Hyun Won Co. (mobiBLU). MusicMasterFM DAH-420 MP3 player, 2004. http://itave.com/mp3_players/musicmaster.html.

[10] M. Lamming and M. Flynn. Forget-me-not: Intimate computing in support of human memory. In *Proc. FRIEND21, 94 International Symposium on Next Generation Human Interface*, Meguro Gajoen, Japan, 1994. http://www.lamming.com/mik/Papers/fmn.pdf.

[11] N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg, and A. Stolcke. The meeting project at ICSI. In *Proc. HLT*, pages 246–252, 2001.

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

[13] N. Oliver and E. Horvitz. Selective perception policies for limiting computation in multimodal systems: A comparative analysis. In *Proceedings of Int. Conf. on Multimodal Interfaces (ICMI'03)*, Vancouver, CA, Nov 2003.

[14] S. Renals and D. P. Ellis. Audio information access from meeting rooms. In *Proc. ICASSP*, Hong Kong, 2003. http://www.dcs.shef.ac.uk/~sjr/pubs/2003/icassp03-mtg.html.

[15] D. Reynolds. An overview of automatic speaker recognition technology. In *Proc. IEEE ICASSP-02*, Orlando, FL, 2002.

[16] K. Sjölander and J. Beskow. Wavesurfer - an open source speech tool. In *Proc. ICSLP-2000*, Beijing, 2000. http://www.speech.kth.se/wavesurfer/wsurf_icslp00.pdf.

[17] Soundman. OKM I Mic set with A3 adapter, 2004. http://www.outwardsound.com/product/microphones/72/.

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

[19] A. Stolcke, C. Wooters, N. Mirghafori, T. Pirinen, I. Bulyko, D. Gelbart, M. Graciarena, S. Otterson, B. Peskin, and M. Ostendorf. Progress in meeting recognition: The ICSI-SRI-UW spring 2004 evaluation system. In *NIST 2004 Meeting Recognition Workshop*, Montreal, May 2004. http://www.speech.sri.com/papers/nist2004-meeting-system.ps.gz.

[20] G. Williams and D. Ellis. Speech/music discrimination based on posterior probability features. In *Proc. Eurospeech-99*, 1999. ftp://ftp.icsi.berkeley.edu/pub/speech/papers/euro99-mussp.pdf.