Features for segmenting and classifying long-duration recordings of "personal" audio

Daniel P.W. Ellis and Keansub Lee

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

Abstract

A digital recorder weighing ounces and able to record for more than ten hours can be bought for a few hundred dollars. Such devices make possible continuous recordings of "personal audio" – storing essentially everything heard by the owner. Without automatic indexing, however, such recordings are almost useless. In this paper, we describe some experiments with recordings of this kind, focusing on the problem of segmenting the recordings into different 'episodes' corresponding to different acoustic environments experienced by the device. We describe several novel features to describe 1-minute-long frames of audio, and investigate their effectiveness at reproducing hand-labeled ground-truth segment boundaries.

1. Introduction

The idea of using technology to aid human memory extends back as far as the earliest precursors to writing; more recently, the idea of literal recordings of particular events has extended from photographs and recordings of 'special events' to the possibility of recording *everything* experienced by an individual, whether or not it is considered significant. The general idea of a device to record the multitude of daily experience was suggested in 1945 by Vannevar Bush [1], who noted:

> *Thus far we seem to be worse off than before–for we can enormously extend the record; yet even in its present bulk we can hardly consult it.*

The prospect of complete records of an individual's everyday environment raises many deep issues associated with our assumptions and beliefs about with what authority past events can be described, yet the ease with which such recordings can be made with current technology would almost certainly lead to their widespread collection if they were in fact useful; on the whole they are not, because the process of actually locating any particular item of information in, for instance, a complete audio record of the past year, or week, or even hour, is so excruciatingly burdensome as to make it unthinkable except in the most critical of circumstances. This is Bush's problem of consultation, and the problem we consider in this paper.

A number of recent projects have worked along these lines, such as "MyLifeBits" [2], an explicit attempt to meet Bush's vision. We have been influenced by the work of Clarkson [3, 4], who analyzed "ambulatory audio" to make environment classifications to support context sensitive applications. In later work, Clarkson recorded 100 days of audio and video with a special backpack [5]. This was segmented and clustered into recurring locations and situations; rank-reduced features from the fish-eye video capture were most useful for this task.

Our work is particularly stimulated by the recent acceleration in portable digital audio devices. Whereas prior work has involved relatively bulky and expensive custom setups, we are now in the startling situation of being able to buy, for a few hundred dollars, palm-sized devices that, out of the box, are ready to record a day's worth of audio, then upload it to a computer in a few minutes. This convenience, along with the natural omnidirectionality of audio, has led us to return to the issue of segmenting and classifying environments based on audio alone.

A major open question in this field is what the ultimate value and best applications for such personal audio recordings would be. In broad terms, the idea of a detailed audio record that could be consulted at will to clarify murky recollections is appealing, but devising specific beneficial scenarios is more difficult. We believe that the best way to illuminate and answer these questions is to begin collecting such recordings and developing the tools to be able to analyze and access them, and to see what uses emerge.

In the next section, we describe our formulation of the task as finding segmentation points within the audio at a resolution of one minute, and describe the data we are working with. Section 3 presents the novel features we have devised to describe audio at this very coarse timescale. We describe our Bayesian Information Criterion segmentation algorithm in section 4, and present our results in section 5. Section 6 contains our conclusions.

Label	minutes	segments	avg. duration
Library	981	27	36.3
Campus	750	56	13.4
Restaurant	560	5	112.0
Bowling	244	2	122.0
Lecturer 1	234	4	58.5
Car/Taxi	165	7	23.6
Street	162	16	10.1
Billiards	157	1	157.0
Lecturer 2	157	2	78.5
Home	138	9	15.3
Karaoke	65	1	65.0
Class break	56	4	14.0
Barber	31	1	31.0
Meeting	25	1	25.0
Subway	15	1	15.0
Supermarket	13	2	6.5
total	3753	139	27.0

Table 1: Segment classes, counts, and average duration (in minutes) from the manual annotation of the 62 hour test set.

2. Task and Data

The most immediate problem in handling continuously-recorded audio is navigation within the recording to find anything of interest. We therefore have started our investigation with the problem of segmentation, with the idea that we can use very coarse divisions into different locations and environments as a useful top-level indexing structure. As in [5], these segments can be clustered into recurrent situations and perhaps correlated with other data sources (such as occasional GPS location tracks) to provide appropriate automatic descriptions.

We used a "Neuros" personal audio computer, which has a built-in microphone, a 20G hard disk, and battery power sufficient to record for over 8 hours without recharging. By carrying this device on a belt hook for a week, we collected a database of more than 62 hours. This single channel data was originally recorded as a 64 Mbps MPEG-Audio Layer 3 file, then downsampled to 16 kHz.

We manually annotated the data with the kinds of gross environment changes we were hoping to detect e.g. entering and leaving buildings and vehicles etc. This resulted in 125 ground-truth boundaries (excluding end-of-file boundaries) and an average segment length of 27 minutes. Table 1 lists the different segment classes identified in the data along with the number of such segments and their average durations.

Using a temporal resolution of 1 minute, we defined our evaluation task as the detection of these environment change boundaries to within +/- 3 frames (i.e. a 7 minute window).

3. Features

The vast majority of work in audio analysis, for example speech recognition, has been concerned with much finer discriminations, and has been based on timescales of tens of milliseconds or less. Here we are considering the one minute timescale, more than a thousand times longer, and significantly different features may be appropriate. This section describes the variants we considered.

3.1. Short-time features

All our features started with a conventional Fourier magnitude spectrum, calculated over 25 ms windows every 10 ms, but differed in how the 201 short-time Fourier transform (STFT) frequency values were combined together into a smaller per-time-frame feature vector, and in how the 6000 vectors per minute were combined into a single feature describing each 1 minute frame.

We used several basic short-time feature vectors, each at two levels of detail.

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

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

  where $X[n, k]$ are the squared-magnitudes from the N point STFT, $N_F = N/2 + 1$ is the number of nonredundant points in the STFT of a real signal, and the w_{jk} define a matrix of weights for combining the STFT samples into the more compact spectrum. To match the dimensionality of the auditory features below, we created two versions of the Energy Spectrum; the first combined the 201 STFT values into 21 Energy Spectrum bins (each covering about 380 Hz or about 10 STFT bins); the second Energy Spectrum had 42 bins (of about 190 Hz).

- **Auditory Spectrum**, similarly formed as weighted sums of the STFT points, but using windows that approximate the bandwidth of the ear – narrow at low frequencies, and broad at high frequencies – to obtain spectrum whose detail approximates, in some sense, the information perceived by listeners. We used the Bark axis, so a spacing of 1 Bark per band gave us 21 bins, and 0.5 Bark/band gave 42 bins. Each of these variants simply corresponds to a different matrix of w_{jk} in eqn. 1 above.

- **Entropy Spectrum**: The low-dimensional spectral features collapse multiple frequency bands into one value; it might be useful to have a little more information about the spectral structure within those

bands, such as whether the energy was distributed across the whole band, or concentrated in just a few of the component STFT samples. We use entropy (treating the distribution of energy within the subband as a PDF) as a measure of the concentration (low entropy) or diffusion (high entropy) within each band, i.e. we define the short-time *entropy* spectrum at each time step n and each spectral channel j as:

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

where the the band magnitudes $A[n,j]$ from eqn. 1 serve to normalize the energy distribution within each weighted band to be PDF-like.

The entropy can be calculated for the bins of both the Energy Spectrum and the Auditory Spectrum; for the Auditory Spectrum, since the w_{jk} define unequal width bands, it is convenient to normalize each channel by the theoretical maximum entropy (of a uniform initial spectrum X) to aid in visualizing the variation in entropy between bands.

- **Mel-frequency Cepstral Coefficients (MFCCs)** use a different (but similar) frequency warping, then apply a decorrelating cosine transform on the log magnitudes. We tried the first 21 bins, or all 40 bins from the implementation we used. MFCCs are the features most commonly used in speech recognition and other acoustic classification tasks. (For the 'linear' averaging described below, we first exponentiated the cepstral values to obtain nonnegative features comparable to the spectral energies above).

3.2. Long-time features

We wished to represent each minute of our recordings (corresponding to 6000 of the 10 ms frames described above) with a single feature vector. We experimented with several different mechanisms for combining the multiple feature vectors into a single summary feature:

- **Average Linear Energy** μ_{lin}: The mean of the vector of energies for a minute of data from each individual channel. This value is then converted to logarithmic units (dB).

- **Linear Energy Deviation** σ_{lin}: The standard deviation of one minute's worth of each feature dimension, converted to dB.

- **Normalized Energy Deviation** σ_{lin}/μ_{lin}: Energy Deviation divide by Average Energy, in linear units. If two temporal profiles differ only by a gain constant, this parameter is unchanged.

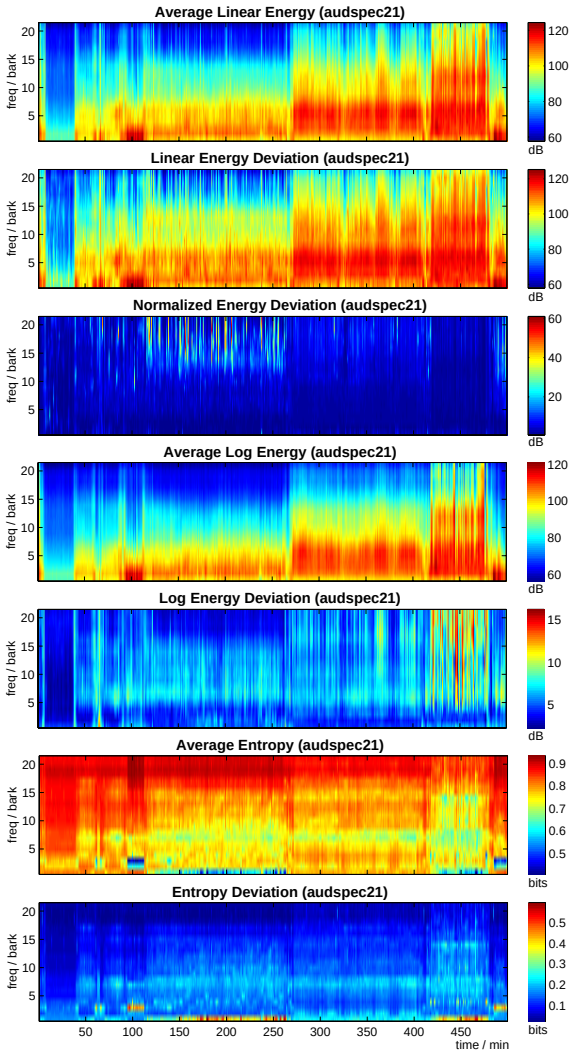

Figure 1: Examples of the seven long-time feature types based on 21-band auditory (Bark-scaled) spectra. The underlying data is over eight hours of recordings covering a range of locations.

- **Average Log Energy** μ_{dB}: We take the logarithm of the energy features first, then calculate the mean within each dimension over the full time frame.

- **Log Energy Deviation** σ_{dB}: The standard deviation of the log-domain version of the feature values. In general, log-domain mean and deviation place excessive emphasis on very quiet sections (which can become arbitrarily negative in log units) but the relatively high noise background in this data avoided this problem. Note that this feature is already invariant to overall gain changes.

- **Average Entropy** μ_H: We calculate the Average Entropy by taking the mean of each dimension of the Entropy Spectrum feature $H[n,j]$ of eqn. 2 over the

one minute frame window.

- **Entropy Deviation** σ_H/μ_H: The standard deviation of $H[n,j]$ within each window, normalized by its average.

4. Segmentation

To segment an audio stream we must detect the time indices corresponding changes in the nature of the signal, in order to isolate segments that are acoustically homogeneous. One simple approach is to measure dissimilarity (e.g. as log likelihood ratio or KL divergence) between models derived from fixed-size time windows on each side of a candidate boundary. However, the fixed window size imposes both a lower limit on detected segment duration, and an upper bound on the accuracy with which the statistical properties of each segment can be measured, limiting robustness. In contrast, the Bayesian Information Criterion (BIC) provides a principled way to compare the likelihood performance of models with different numbers of parameters and explaining different amounts of data e.g. from unequal time windows. The speaker segmentation algorithm presented in [6] uses BIC to compare every possible segmentation of a window that is expanded until a valid boundary is found, so that the statistics are always based on complete segments.

BIC is a likelihood criterion penalized by model complexity as measured by the number of model parameters. Let $\chi = \{x_i : i = 1, ..., N\}$ be the data set we are modeling and $\mathcal{M} = \{m_i : i = 1, ..., K\}$ be the candidate models we wish to choose between. Let $\#(M_i)$ be the number of parameters in model M_i, and $\mathcal{L}(\chi, M_i)$ be the total likelihood of χ under the optimal parameterization of M_i. BIC is defined as:

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

where λ is a weighting term for the model complexity penalty which should be 1 according to theory. By balancing the expected improvement in likelihood for more complex models by the penalty term, choosing the model with the highest BIC score is, by this measure, the most appropriate fit to the data.

The BIC-based segmentation procedure described in [6] proceeds as follows. We consider a sequence of d-dimensional audio feature vectors $\chi = \{x_i \in R^d : i = 1, ..., N\}$ covering a portion of the whole signal as independent draws from one or two multivariate Gaussian processes. Specifically, the null hypothesis is that the entire sequence is drawn from a single distribution:

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

which is compared to the hypothesis that the first i points are drawn from one distribution and that the remaining points come from a different distribution i.e. there is a segment boundary after sample t:

$$H_1 : \{x_1, ..., x_t\} \sim N(\mu_1, \Sigma_1), \{x_{t+1}, ..., x_N\} \sim N(\mu_2, \Sigma_2) \quad (5)$$

where $N(\mu, \Sigma)$ denotes a multivariate Gaussian distribution with mean vector μ and full covariance matrix Σ.

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

$$BIC(t) = log\left(\frac{\mathcal{L}(\chi|H_0)}{\mathcal{L}(\chi|H_1)}\right) - \frac{\lambda}{2}\left(d + \frac{d(d+1)}{2}\right) log(N) \quad (6)$$

where $\mathcal{L}(\chi|H_0)$ is the likelihood of χ under hypothesis H_0 etc., and $d + d(d+1)/2$ is the number of extra parameters in the two-model hypothesis H_1. When $BIC(t) > 0$, 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 N is increased, and the search across all possible boundaries t is repeated. This continues until the end of the signal is reached.

The weighting parameter λ provides a 'sensitivity' control which can be adjusted to make the overall procedure generate a larger or smaller number of boundaries for a given signal.

5. Results

Our six short-time spectral representations – linear frequency, auditory spectrum, and Mel cepstra, each at either 21 or 42 (40 for MFCC) elements per frame – summarized by our seven "long-time" feature summarization functions gave us 38 different compact representations of our 60 hour dataset. (The entropy-based measures were not calculated for the cepstral features, since the concept of 'subband' does not apply in that case.) BIC segmentation was applied to each version, and the λ parameter was varied to control the trade-off between finding too many boundaries (false alarms in the boundary detection task) and too few boundaries (false rejection of genuine boundaries). Table 2 shows the Sensitivities (Correct Accept rate, the probability of marking a frame as a boundary given that it is a true boundary) of each system when λ is adjusted and the results interpolated to achieve a Specificity of 98% (the probability of marking a frame as a non-boundary given that it is not a boundary, or equivalently a False Alarm rate of 2%).

There is a wide range of performance among the different features; mean and deviation of the linear energy perform quite well across all underlying representations, and their ratio does not. Log-domain averaging performs very well for the auditory spectrum but not for the other representations, and log-domain deviation is most useful for MFCCs. However, the spectral entropy features,

Short-time features	μ_{lin}	σ_{lin}	σ_{lin}/μ_{lin}	μ_{dB}	σ_{dB}	μ_H	σ_H/μ_H
21-bin Energy Spectrum	0.723	0.676	0.386	0.355	0.522	0.734	0.744
42-bin Energy Spectrum	0.711	0.654	0.342	0.368	0.505	0.775	0.752
21-bin Auditory Spectrum	0.766	0.738	0.487	**0.808**	0.591	**0.811**	**0.816**
42-bin Auditory Spectrum	0.761	0.731	0.423	0.792	0.583	0.800	**0.816**
21-bin MFCC	0.734	0.736	0.549	0.145	0.731	N/A	N/A
40-bin MFCC	0.714	0.640	0.498	0.166	0.699	N/A	N/A

Table 2: Sensitivity @ Specificity = 0.98 for each feature set. Values greater than 0.8 are shown in bold.

describing the sparsity of the spectrum within each subband give the best overall performance, particularly when based on the auditory spectra.

Since the 21 bin Auditory Spectrum were the best underlying short-term features, our remaining results use only this basis. We experimented with using combinations of the best individual feature sets, to perform BIC segmentation on a higher-dimensional feature formed by simple concatenation. Table 3 shows the results of all possible combinations of the three best features, Average Log Energy μ_{dB}, Average Entropy μ_H, and Entropy Deviation σ_H/μ_H.

Feature Set	Sensitivity
μ_{dB}	0.808
μ_H	0.811
σ_H/μ_H	0.816
$\mu_{dB} + \mu_H$	0.816
$\mu_{dB} + \sigma_H/\mu_H$	**0.840**
$\mu_H + \sigma_H/\mu_H$	0.828
$\mu_{dB} + \mu_H + \sigma_H/\mu_H$	0.836

Table 3: Sensitivity @ Specificity = 0.98 for different combinations of the three best statistics based on the 21-bin Auditory Spectrum.

Although all the combinations yield broadly similar results, our best combination involves just two of the three features, namely the Average Log Energy plus the Entropy Deviation.

Figure 2 shows the Receiver Operating Characteristic (ROC) curve for our best performing detectors, illustrating the trade-off between false alarms and false rejects as the BIC penalty weight λ is varied. (A better performance lies closer to the top-left corner, and random guessing follows the leading diagonal). We see that the $\mu_{dB}+\sigma_H/\mu_H$ combination is the best overall, although the differences from the best individual feature sets are quite small.

6. Conclusions

These first investigations into handling very long (and relatively uneventful) recordings have raised a number of

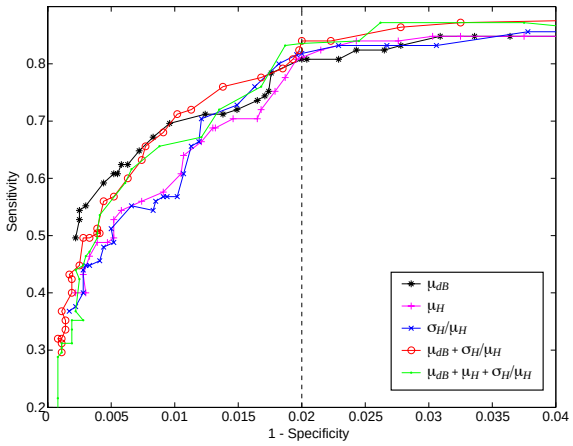

Figure 2: ROC curves for segment boundary detection based on different summaries and combinations of the 21 bin Auditory Spectral features.

further issues. One aspect we are considering is visualization of this data, perhaps by combining our two best features into a pseudo-spectrogram. Figure 3 shows a modified pseudocolor visualization of the data from figure 1, where each pixel's intensity reflects the average log energy, the saturation (vividness of the color) depends on the mean spectral entropy, and the hue (color) depends on the entropy deviation. We are experimenting with displays of this kind in order to find an effective, browsable representation for these sounds.

Our feature vectors are relatively large, particularly when feature combinations are used. We are currently pursuing rank-reduction of the features using PCA prior to the BIC segmentation. In an initial investigation, we obtained a Sensitivity of 0.874 (at Specificity = 0.98) for a combination of the first 3 principal components of μ_{dB} combined with the first 4 principal components of μ_H (which proved more useful than σ_H/μ_H in this case).

One aspect we have yet to investigate is measures of temporal structure within the 1 minute segments; our current features are invariant to arbitrary permutations of the short-time frames within each longer frame. A natural choice is to look at modulation spectrum features

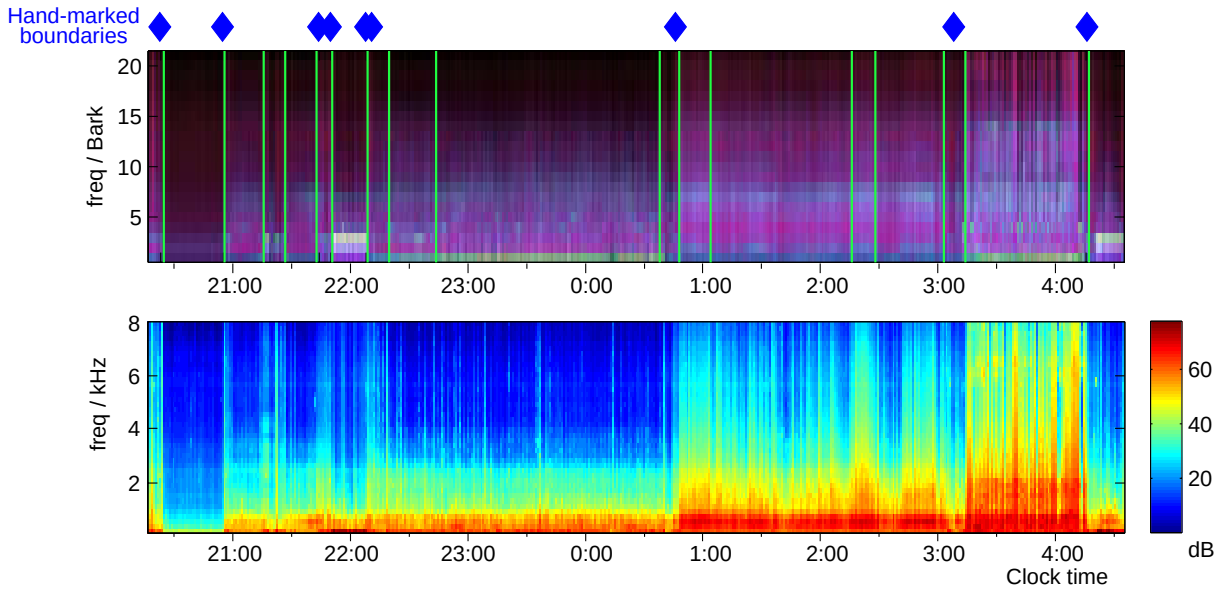

Figure 3: Upper panel: visualization of the analysis of an eight hour recording, combining average energy (pixel intensity), mean spectral entropy (saturation), and entropy deviation (hue). Also shown are the automatic boundaries found by the system (vertical lines) and the hand-marked ground truth boundaries for this segment (diamonds above the figure). The lower panel shows a conventional pseudocolor spectrogram based on fixed-bandwidth Fourier analysis, for comparison. The contrast between the two is particularly clear around 22:00.

i.e. a second Fourier transform of the intensity variations within the 1 minute segment along each subband.

In summary, we have proposed new feature extraction procedures to describe long-duration audio recordings at the scale of one minute. Based on frequency-warped short-time energy spectra, we calculate the mean, deviation and entropy statistics for the spectral parameters within each one minute window. Our segmentation system based on the Bayesian Information Criterion shows that the proposed features can successfully detect the acoustic changes associated with changes in location in everyday audio with good accuracy. Our best results combined Average Log Energy and Entropy Deviation features.

The motivation for this work is to develop automatic descriptions, summaries, and indexes for the kind of very long duration recordings we have been working with. Our current work is looking at clustering and labeling the segments found by this procedure, and we hope to incorporate these results into an information access system that can uncover true utility in such recordings.

7. 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. Thanks to the anonymous reviewers for their helpful comments.

8. References

[1] Vannevar Bush, "As we may think," *The Atlantic Monthly*, July 1945.

[2] Jim Gemmell, Gordon Bell, Roger Lueder, Steven Drucker, and Curtis Wong, "Mylifebits: Fulfilling the memex vision," in *ACM Multimedia*, Juan-les-Pins, France, December 2002, pp. 235–238.

[3] B. Clarkson, N. Sawhney, and A. Pentland, "Auditory context awareness via wearable computing," in *Proc. Perceptual User Interfaces Workshop*, 1998.

[4] B. Clarkson and A. Pentland, "Unsupervised clustering of ambulatory audio and video," in *Proc. ICASSP*, Seattle WA, 1999.

[5] Brian P. Clarkson, *Life patterns: structure from wearable sensors*, Ph.D. thesis, MIT Media Lab, 2002.

[6] 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.