ROBUST MULTI-FEATURE SEGMENTATION AND INDEXING FOR NATURAL SOUND ENVIRONMENTS

Gordon Wichern, Harvey Thornburg, Brandon Mechtley, Alex Fink, Kai Tu, and Andreas Spanias

Arts, Media, and Engineering
Arizona State University
{Gordon.Wichern, Harvey.Thornburg}@asu.edu

ABSTRACT

Creating an audio database from continuous long-term recordings, allows for sounds to not only be linked by the time and place in which they were recorded, but also to sounds with similar acoustic characteristics. Of paramount importance in this application is the accurate segmentation of sound events, enabling realistic navigation of these recordings. We first propose a novel feature set of specific relevance to environmental sounds, and then develop a Bayesian framework for sound segmentation, which fuses dynamics across multiple features. This probabilistic model possesses the ability to account for non-instantaneous sound onsets and absent or delayed responses among individual features, providing flexibility in defining exactly what constitutes a sound event. Example recordings demonstrate the diversity of our feature set, and the utility of our probabilistic segmentation model in extracting sound events from both indoor and outdoor environments.

1. INTRODUCTION

The increasing affordability of high-fidelity audio recording and storage technology enables us to record natural sound environments, for instance offices, outdoor public spaces, and even wildlife habitats, over long periods of time. Such recordings can reveal much about patterns of activity occurring within these spaces. For example, monitoring daily patterns in the level of voice activity in a graduate student lab space can reveal when students collaborate, thus indicating to security guards when to patrol, to pizza companies when they should stay open, and so forth. Monitoring traffic noise vs. outdoor conversation patterns near a busy intersection may help gauge the relative density of cars vs. bicyclists or pedestrians, thus yielding important information for improving the safety and usability of the intersection. Audio information alone also finds use in ecological studies; for example Slabbekoorn and den Boer-Visser's recent work on the adaptation of city birds' songs in the presence of traffic noise has helped explain why certain populations are found in cities, while those with less adaptable sound-production mechanisms either migrate or become extinct [1].

Why make continuous recordings? One reason is that everything gets recorded, rather than just those events which one targets based on preconceptions about what *should* occur in the space. Preconceptions may help in testing specific hypotheses (such as the response of birds' songs to traffic patterns), but they become restrictive when doing exploratory analyses or when the causes of certain phenomena are unknown. A second reason for continuous recording is that once specific events are identified, one may recover the context in which they have occurred simply by listening to as much of the surrounding material as desired. The latter was in fact the impetus behind Vannevar Bush's MEMEX device proposed in 1945 [2, 3]: continuous recording as an aid to thought and memory. Context recovery becomes particularly useful when contextual events can be *linked* in some sense to related events, allowing the user to discover new connections.

A key drawback to continuous recordings remains size in the context of navigability. With recordings possibly in excess of several days, it is difficult for a human to find events of interest by the standard methods of listening or scanning visually via waveform display. Rather, an efficient indexing mechanism is needed which allows individual sound events to be segmented, categorized, and linked both to similar events and to the context of the original recording. While many such mechanisms have been discussed in the context of musical sound [4, 5, 6], our proposed method is specifically adapted to environmental sounds. It consists of a database in which individual sound events are automatically extracted and tagged with the following features: *loudness*, *spectral centroid*, *spectral sparsity*, *harmonicity*, *temporal sparsity*, and *transient index*. These features, of specific relevance for environmental sounds, are defined and discussed in Section 2. Three dimensions of relevance are as follows:

- **Diversity**: the particular feature or group should exhibit a large spread (entropy) in the context of real-

world data sets. In particular, we should avoid functionally redundant features (bandwidth and spectral sparsity, for instance).

- **Categoric relevance**: Different categories of sound (e.g. voice, construction sounds, traffic, wind noise) should have different average feature vectors.

- **Perceptual adaptation**: Sounds that sound different have different feature vectors; i.e., feature distance varies with perceptual distance. Reasonable efforts have been made to map out feature scales according to the concept of just-noticeable difference (JND).

Besides feature-based tags, start time, duration, calendar date, and GPS location (latitude-longitude pair) are indexed with each sound. Time-based tags allow sounds to be linked to their context in the original recording. Feature-based tags allow sounds, once retrieved, to be linked to sounds of similar type without requiring explicit human labeling. The critical step towards database construction is thus primarily one of segmentation, meaning the extraction of individual sound events.

What a *sound event* means is itself highly dependent on context; for instance, events in speech may be phonemes, words, phrases, or periods where the identity of the speaker remains constant. Similar variations also occur in environmental sound [7]. In [3] a portable audio recording device is used as a memory aid when carried by a human subject. The average reported segment length for these preliminary experiments was approximately 26 minutes, with a segment typically corresponding to time spent in a given location, e.g., street or restaurant. However, in [8] an acoustic surveillance system continuously monitors an office environment, and detects patterns in very short-term audio events such as coughs and typing sounds. To systematically address the effects of context, we propose a dynamic Bayesian framework for the segmentation. This framework proves robust to real-world conditions regarding the flexibility of what constitutes an "event", the diversity of sources, and the presence of overlapping sounds. Specifically, our framework addresses the following: (1) event onsets might not be instantaneous, but distributed over several frames of data, (2) different features will be responsive to different types of sounds, depending on the environment where the sound was recorded, and (3) segment boundaries between responsive features for a given sound event can be delayed. How our multi-feature segmentation model encodes this information is detailed in Section 3.

2. AUDIO FEATURE EXTRACTION

A fundamental goal is that features indicate the presence of a large variety of sounds while not specifically adapting to a particular type of sound, e.g., speech or music. Due to the diversity of sound sources, we have found it necessary to calculate features at different time scales, from 20ms (short-term) to one second (long-term). Short-term features are computed either directly from the windowed time series data or via short-time Fourier Transform (STFT) using overlapping 20ms Hamming windows hopped every 10ms. Long-term features are computed using a one-second sliding window to combine the data from 99 of the 20ms windows. Using 99% overlap for the sliding window, (i.e., slide in 10ms steps), both long and short-term features remain synchronous.

2.1. Short-term features

The first of our short term features, which are calculated every 10ms using 20ms of data, is *loudness*. We define loudness as the RMS level of a windowed frame of data in decibels. A second feature, *spectral sparsity* is calculated from the zero-padded STFT data of each frame, via the ratio of L^∞ and L^1 norms calculated over the magnitude STFT spectrum (inspired by the GINI index and related sparsity metrics [9]). Defining $X_t(j)$ as the M-point STFT coefficient from frequency bin j for frame t, the spectral sparsity is defined as follows

$$\epsilon_t \triangleq \frac{\max(|X_t(1)|, ..., |X_t(M)|)}{\sum_{j=1}^{M} |X_t(j)|}. \quad (1)$$

Spectral sparsity should be large for pure sine tones, voice, or bells, and smaller for sounds with significant "noise" characteristics which imply a wide frequency spectrum.

Finally, we compute from STFT data the Bark-weighted *spectral centroid*:

$$\mu_t \triangleq \frac{\sum_{j=1}^{M} b_j(b_j - b_{j-1})|X_t(j)|^2}{\sum_{j=1}^{M}(b_j - b_{j-1})|X_t(j)|^2} \quad (2)$$

where b_j is the frequency value of the center of STFT bin j in units of Barks, a psychoustically-adapted frequency measure [10].

2.2. Long-term features

As mentioned previously, the long-term features are calculated every 10ms using one second worth of data by combining data from N=99 of the 20ms frames. First, *temporal sparsity* is defined as the ratio of L^∞ and L^1 norms calculated over the N small window RMS values in a given one second window, i.e.,

$$\tau_t \triangleq \frac{\max(RMS_{t-(N+1)}, ..., RMS_t)}{\sum_{k=t-(N+1)}^{t} RMS_k}. \quad (3)$$

Temporal sparsity should be large for sounds such as footsteps in relative silence and useful for detecting and segmenting these types of sounds.

Second, the *transient index* is computed by combining Mel frequency cepstral coefficients (MFCC's) [11] from several frames of data. We define the transient index for frame t as follows.

$$\kappa_t \triangleq \sum_{k=t-(N+2)}^{t} \| MFCC_k - MFCC_{k-1} \|_2 \quad (4)$$

where $MFCC_k$ is the MFCC vector for frame k, and N signifies the number of short frames over which the transient index is computed. The transient index should be useful in detecting and segmenting sounds whose spectral characteristics exhibit consistent fluctuations between adjacent frames, e.g., crumpling newspaper.

Finally, *harmonicity* is used to measure probabilistically whether or not the STFT spectrum for a given frame exhibits a harmonic frequency structure. The algorithm begins by choosing the L most prominent peaks from the STFT magnitude spectrum of frame t and storing the frequency of each peak in Hz, in the set $\rho_t = \{f_1, .., f_L\}$. We then estimate the fundamental frequency $\hat{f}_o$ and likelihood $P(\rho_t|\hat{f}_o)$ using an approximation of Goldstein's algorithm [12, 13]. A detailed discussion of this likelihood approximation appears in Appendix A. As a measure of harmonicity we use the marginal distribution $P(\rho_t)$:

$$P(\rho_t) = \int P(\rho_t|f_o)P(f_o)df_o. \quad (5)$$

Approximating $P(\rho_t|f_o)$ using only Goldstein's algorithm is computationally expensive, so we use the following approximation. First, we compute $P(\rho_t|f_o)$ with Goldstein's algorithm for a small number of f_o values. The f_o values chosen for this first estimate of the likelihood surface should contain several rational multiples of $\hat{f}_o$ and a small number of uniformly spaced points on the bark scale. Choosing these points should allow for a thorough coverage of common fundamental frequencies perceptible to the human ear and also include the likely peaks of $P(\rho_t|f_o)$. This initial estimate of the likelihood surface is smoothed using a variable bandwidth K-nearest neighbor method with an Epanechnikov kernel, to obtain a large number of points uniformly distributed on the Bark scale. Assuming a uniform prior for $P(f_o)$ (in Barks) we can easily approximate the integral in (5). As a short-term feature we define the harmonicity β_t for frame t as $\beta_t = P(\rho_t)$, but we find it more useful to look at harmonicity on a larger scale by averaging over N frames:

$$\beta_t \triangleq \frac{1}{N} \sum_{k=t-(N+1)}^{t} P(\rho_t). \quad (6)$$

Harmonicity should be large for speech, music, and certain machine noises and smaller for most other types of environmental audio, as well as some musical sounds (bells).

3. FEATURE FUSION SEGMENTATION MODEL

Once the features from the previous section have been extracted from the audio signal, we can use them to segment the recording into appropriate environmental sound clips. We employ a generative probabilistic model where true event onsets and end points influence observed audio feature trajectories.

3.1. Probabilistic Model

We first define $t \in 1:T$ to be the time index of the audio frame, for a recording of length T. Letting K represent the number of features extracted from each frame of the signal, $Y_t^{(i)}$, for $i \in 1:K$ are the observed audio features at time t. In our underlying model we are not simply concerned with separating regions of silence from regions where the sound is on, but also accounting for new sound clips beginning before the previous clip has turned off. This information is modeled by assigning a global *mode*, M_t to each audio frame of the recording. We represent M_t as a discrete random variable, which is responsible for controlling the underlying dynamics of all K features. The mode, M_t can take three possible values:

- $\emptyset$-There are no perceptible events in this frame.
- $O1$-The onset, or beginning of a new sound event.
- $C1$-The continuation of a sound event between two successive frames.

Even if there is a sound present at time t, it is very likely that some of the audio features will fail to respond. We can represent the responsiveness of a given feature for a given sound by $H_t^{(i)} \in \{0, 1\}$ for $i \in 1:K$, which is unique to each feature. If feature i is responsive at frame t, then $H_t^{(i)} = 1$, while $H_t^{(i)} = 0$ represents a feature that behaves no differently from silence or ambient noise, even when there is a sound event present.

Because of the variation in time scales and meaning of the different features, it is possible that certain features lag behind the overall mode M_t when turning on or off. The discrete random variables $\mu_t^{(i)}$, for $i \in 1:K$ are essentially gates modeling delays between the onset and end times of the different features. Similarly to M_t, $\mu_t^{(i)} \in \{\emptyset, O1, C1\}$, where the definitions of $\emptyset, O1$, and $C1$ are the same as for M_t.

We also define the hidden *states* $S_t^{(i)}$ for $i \in 1:K$, which are continuous random variables mediating the effect of individual features' onsets/end times on the actual feature

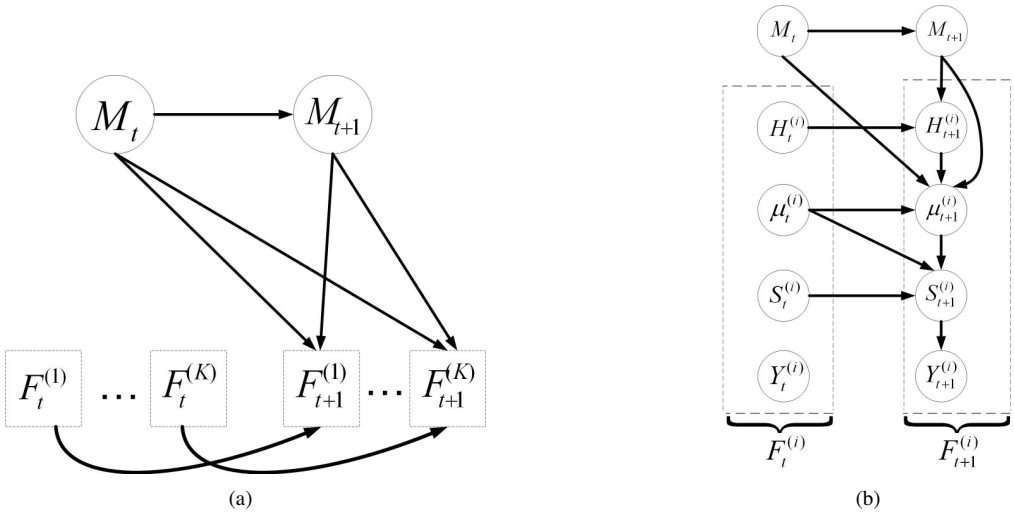


Fig. 1. Directed acyclic graph for feature fusion audio segmentation model. (a) Overall segmentation model for K features (b) Expansion of $F_t^{(i)}$ into individual components.

observations $Y_t^{(i)}$. The latter are modeled as *inherent* features corrupted by noise. The state $S_t^{(i)}$ is then composed of this inherent feature plus an auxiliary variable enabling $S_t^{(i)}$ to be encoded as a first order Gauss-Markov process in time.

Assuming T available data frames and K features, the sequence of observations can be written as $Y_{1:T}^{(1:K)}$, meaning a sequence of T observations for each of the K features. Similar notation is used to represent the sequences of hidden variables, $S_{1:T}^{(1:K)}$, $\mu_{1:T}^{(1:K)}$, and $H_{1:T}^{(1:K)}$, while all features share a common global mode sequence, $M_{1:T}$. The directed acyclic graph (DAG) for our proposed feature fusion segmentation model with K features is shown in Figure 1(a), and leads to the following factorization of the joint distribution:

$$P(M_{1:T}, H_{1:T}^{(1:K)}, \mu_{1:T}^{(1:K)}, S_{1:T}^{(1:K)}, Y_{1:T}^{(1:K)}) = P(M_1)\prod_{i=1}^{K} P(F_1^{(i)}|M_1) \times \prod_{t=2}^{T} P(M_{t+1}|M_t)\prod_{i=1}^{K}\Big[P(F_t^{(i)}|M_t, M_{t-1})\Big]. \quad (7)$$

where $F_{1:T}^{(i)}$ denotes all of the random variables for feature i, whose DAG is shown in Figure 1(b), and $P(F_t^{(i)}|M_t, M_{t-1})$ at frame t factors as

$$P(F_t^{(i)}|M_t, M_{t-1}) = \prod_{i=1}^{K}[P(H_t^{(i)}|H_{t-1}^{(i)}, M_t)P(\mu_t^{(i)}|\mu_{t-1}^{(i)}, H_t^{(i)}, M_t, M_{t-1}) \times P(S_t^{(i)}|S_{t-1}^{(i)}, \mu_t^{(i)}, \mu_{t-1}^{(i)})P(Y_t^{(i)}|S_t^{(i)})]. \quad (8)$$

The prior $P(F_1^{(i)}|M_1)$ from (7) is defined as

$$P(F_1^{(i)}|M_1) \triangleq P(H_1^{(i)}|M_1)P(\mu_1^{(i)}|H_1^{(i)}, M_1)P(S_1^{(i)}|\mu_1^{(i)})P(Y_1^{(i)}|S_1^{(i)}). \quad (9)$$

3.2. Distributional specifications

To define the segmentation model of Figures 1(a) and (b) we must specify all conditional probability distributions from (7-9). We begin with the frame likelihood

$$P(Y_t^{(i)}|S_t^{(i)}) \sim \mathcal{N}(v_t^{(i)}, R^{(i)}) \quad (10)$$

where $v_t^{(i)}$ is one component of the continuous random state vector $S_t^{(i)}$, i.e., $S_t^{(i)} = [u_t^{(i)}, v_t^{(i)}]^T$. The continuous state variables $u_t^{(i)}$ and $v_t^{(i)}$ satisfy the following recursive relations

$$u_t^{(i)} = u_{t-1}^{(i)} + q_t^{(i)}(\mu_t^{(i)}, \mu_{t-1}^{(i)})$$
$$v_t^{(i)} = (1-\alpha^{(i)})u_t + \alpha^{(i)}v_{t-1}^{(i)} \quad (11)$$

where $\alpha^{(i)}$ is a low-pass filter coefficient, which allows for rapid but non-instantaneous change of the inherent feature across segments. Process noise $q_t^{(i)}(\mu_t^{(i)}, \mu_{t-1}^{(i)})$ is distributed according to

$$q_t^{(i)}(\mu_t^{(i)}, \mu_{t-1}^{(i)}) \sim \mathcal{N}(0, Q^{(i)}(\mu_t^{(i)}, \mu_{t-1}^{(i)})). \quad (12)$$

It is also possible to describe the dynamics of the state vector $S_t^{(i)}$ defined in (11), using a state-space equation

$$S_t^{(i)} = A^{(i)}S_{t-1}^{(i)} + B^{(i)}q_t^{(i)}(\mu_t^{(i)}, \mu_{t-1}^{(i)}) \quad (13)$$

where

$$A^{(i)} = \begin{bmatrix} 1 & 0 \\ 1-\alpha^{(i)} & \alpha^{(i)} \end{bmatrix}, \quad B^{(i)} = \begin{bmatrix} 1 \\ 1-\alpha^{(i)} \end{bmatrix}. \tag{14}$$

Using (13) we now specify the conditional distribution of $P(S_t^{(i)}|S_{t-1}^{(i)}, \mu_t^{(i)}, \mu_{t-1}^{(i)})$ as

$$P(S_t^{(i)}|S_{t-1}^{(i)}, \mu_t^{(i)}, \mu_{t-1}^{(i)}) \sim \mathcal{N}(A^{(i)}S_t^{(i)}, Q^{(i)}(\mu_t^{(i)}, \mu_{t-1}^{(i)})). \tag{15}$$

Via $P(\mu_t^{(i)}|\mu_{t-1}^{(i)}, H_t^{(i)}, M_t, M_{t-1})$, assuming feature i is responsive ($H_t^{(i)} = 1$), we model possible lags between when a particular gate, $\mu_t^{(i)}$, turns on after M_t has turned on as Poisson. Letting $p_{\text{lag}+}^{(i)}$ be the probability that the lag will continue for an additional frame, the expected lag becomes $1/p_{\text{lag}+}^{(i)}$. Similarly, we model possible lags between when a particular gate, $\mu_t^{(i)}$, turns off after M_t has turned off as Poisson, with $p_{\text{lag}-}^{(i)}$ as the probability that the lag will continue for an additional frame.

A summary of $P(\mu_t^{(i)}|\mu_{t-1}^{(i)}, H_t^{(i)}, M_t, M_{t-1})$ for all possible combinations of $\mu_{t-1}^{(i)}$, M_t, and M_{t-1}, given that feature i is responsive (i.e., $H_t^{(i)} = 1$) is shown in Table 1. If feature i is unresponsive, as symbolized by $H_t^{(i)} = 0$, then $P(\mu_t^{(i)} = \emptyset|\mu_{t-1}^{(i)}, H_t^{(i)} = 0, M_t, M_{t-1}) = 1$, i.e., even if a sound is on, feature i behaves as if the recording is in a silent region.

In order to define $P(H_t^{(i)}|H_{t-1}^{(i)}, M_t)$, we consider two distinct cases. First, if the global mode indicates the onset of a new sound event, then feature i will be responsive with probability $p_H^{(i)}$, hence, $P(H_t^{(i)} = 1|H_{t-1}^{(i)}, M_t = O1) \sim \text{Bernoulli}(p_H^{(i)})$. If a sound is in a silent region, or continuing on, as indicated by $M_t \in \{\emptyset, C1\}$, then the responsiveness of feature i is not allowed to change, and $H_t^{(i)} = H_{t-1}^{(i)}$, or formally $P(H_t^{(i)} = H_{t-1}^{(i)}|H_{t-1}^{(i)}, M_t \neq O1) = 1$.

Finally, we specify $P(M_{t+1}|M_t)$ as in Table 2, which models Poisson onset times and event durations. In Table 2, p_{new} is the prior probability of a new onset, while p_{off} is the prior probability of a sound turning off, given that it is currently on.

3.3. Inference Methodology

Segmentation is achieved by estimating the global mode sequence $M_{1:T}$. Ideally, our estimation criterion should preserve the correct number of segments, and the detected segment boundaries should be near the true segment locations. In order to achieve these goals we choose the maximum-a-posteriori (MAP) or global segmentation criterion to estimate $M_{1:T}$. The MAP criterion can be defined as

$$\hat{M}_{1:T} = \arg\max_{M_{1:T}} P(M_{1:T}|Y_{1:T}^{(1:K)}) \tag{16}$$

and is typically more effective in estimating segmentation sequences that do not contain boundaries in adjacent frames, in contrast to a local frame error segmentation criterion. A detailed discussion of this topic can be found in [14].

Unfortunately, computing the exact MAP estimate requires exponential-time complexity. A linear-time approximation nevertheless exists, using the approximate Viterbi inference scheme of [15]. The latter is based on the following assumption:

$$\begin{aligned} &P(Y_{t+1}^{(1:K)}|\Psi_{1:t+1}, Y_t^{(1:K)}) \approx \\ &P(Y_{t+1}^{(1:K)}|\Psi_{1:t-1}^*(\Psi_t), \Psi_t, \Psi_{t+1}, Y_{1:t}^{(1:K)}) \end{aligned} \tag{17}$$

where

$$\Psi_{1:t-1}^*(\Psi_t) \approx \arg\max_{\Psi_{1:t-1}} P(\Psi_{1:t-1}|\Psi_t, Y_{1:t}^{(1:K)}) \tag{18}$$

In (17) and (18), Ψ_t represents the Cartesian product of the $2K+1$ discrete nodes, M_t, $H_t^{(1:K)}$, and $\mu_t^{(1:K)}$ from the DAG in Figure 1. General computational details can be found in [14, 15].

4. RESULTS AND DISCUSSION

We have applied the proposed multi-feature segmentation to both indoor and outdoor continuous recordings. Sound files were captured at 16 bits/44.1kHz, uncompressed. Prior to segmentation, all extracted audio features were normalized to $[0, 1]$.

Figure 2 displays results for a recording made while brewing coffee. The time-domain waveform for this sound file is shown in the upper plot of Figure 2(a), while the extracted global mode sequence $M_{1:T}$ is displayed directly beneath the time domain waveform. Mode values of $M_t = \emptyset$ are plotted as zero, $M_t = C1$ are plotted as one, and $M_t = O1$ are represented by dashed lines. The top plots of Figures 2(b)-(d) display the the audio feature sequences corresponding to spectral centroid, spectral sparsity, and harmonicity, respectively. The middle and bottom plots of Figures 2(b)-(d) contain the inferred feature fusion parameters $H_{1:T}^{(i)}$ and $\mu_{1:T}^{(i)}$, respectively, for i=1,2,3.

An interesting observation from Figure 2 is the ability of the different spectral features to detect different events, partially illustrating the diversity of our chosen features. The spectral centroid (Figure 2(b)) detects the sound of coffee beans pouring into the metal grinder between approximately 11 and 17 seconds, while failing to detect the grinding of the beans between approximately 26 and 38 seconds. Spectral sparsity (Figure 2(c)) appears to detect both sounds, along with the momentary decrease of pressure applied to the grinder around 33 seconds. The plots for harmonicity in Figure 2(d) show that it only contributes to the overall segmentation during the grinding sound. We note that

Tab. 1. Mode transition probabilities for $P(\mu_t^{(i)}|\mu_{t-1}^{(i)}, H_t^{(i)} = 1, M_t, M_{t-1})$.

M_t	M_{t+1}	$\mu_t^{(i)}$	$P(\mu_{t+1}^{(i)} = \emptyset)$	$P(\mu_{t+1}^{(i)} = O1)$	$P(\mu_{t+1}^{(i)} = C1)$
$\emptyset$	$\emptyset$	$\emptyset$	1	0	0
$\emptyset/O1/C1$	$\emptyset$	$O1/C1$	$1 - p_{\text{lag}-}^{(i)}$	0	$p_{\text{lag}-}^{(i)}$
$\emptyset/O1/C1$	$O1/C1$	$\emptyset$	$1 - p_{\text{lag}+}^{(i)}$	$p_{\text{lag}+}^{(i)}$	0
$\emptyset/C1$	$O1$	$O1/C1$	$p_{\text{lag}+}^{(i)} - (p_{\text{lag}-}^{(i)} \cdot p_{\text{lag}+}^{(i)})$	$1 - p_{\text{lag}+}^{(i)}$	$p_{\text{lag}-}^{(i)} \cdot p_{\text{lag}+}^{(i)}$
$C1$	$C1$	$O1/C1$	0	0	1
$O1$	$C1$	$O1$	0	0	1
$O1$	$C1$	$C1$	$p_{\text{lag}+}^{(i)} - (p_{\text{lag}-}^{(i)} \cdot p_{\text{lag}+}^{(i)})$	$1 - p_{\text{lag}+}^{(i)}$	$p_{\text{lag}-}^{(i)} \cdot p_{\text{lag}+}^{(i)}$

Tab. 2. Mode transition probabilities for $P(M_{t+1}|M_t)$.

M_{t+1}	$P(M_{t+1} = \emptyset)$	$P(M_{t+1} = O1)$	$P(M_{t+1} = C1)$
$\emptyset$	$1 - p_{\text{new}}$	p_{new}	0
$O1$	0	0	1
$C1$	$p_{\text{off}}(1 - p_{\text{new}})$	p_{new}	$1 - p_{\text{off}} - p_{\text{new}} + p_{\text{off}}p_{\text{new}}$

this group of spectral features does miss the low-amplitude, brief shuffling sounds at the beginning of the clip (see Figure 2(a)).

Results for a second example of picking up and dropping gravel recorded during a walk in a park on a windy day are shown in Figure 3. This recording is particularly corrupted by wind noise. We hear three faint but distinct segments where gravel is picked up and dropped, approximately corresponding to the M_t-sequence shown in Figure 3(a). The other high energy regions of the time domain waveform correspond to wind gusts. Because of the wind noise present in this example, loudness alone cannot segment the important events when gravel is picked up and dropped (Figure 3 (b)). However, both the spectral centroid shown in Figure 3(c), and the transient index displayed in Figure 3(d) are able to accurately segment these events. When comparing the bottom plots of Figures 3(c) and (d), we notice the transient index tends to lag behind the spectral centroid, but this does not seem to adversely effect the inferred global mode sequence shown in Figure 3(a), demonstrating the utility of our feature fusion model.

5. CONCLUSIONS AND FUTURE WORK

In long-term continuous recording applications, construction of a general database framework should allow sounds to be linked in context to the time in which they were recorded and also to similar sounds in the database. The first step in generating this database is the development of both the novel audio feature set, and probabilistic multi-feature segmentation scheme, proposed in sections 2 and 3, respectively. As the provided sample recordings illustrate, our algorithm possesses the ability to account for absent or delayed individual features, without adversely impacting the overall segmentation of sound events. Furthermore, the varied response of our features to different sound events and environments demonstrates their ability as a useful tool in labeling segmented audio clips.

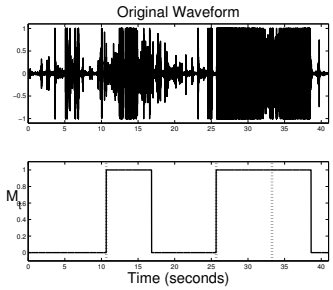
Future extensions include use of the EM algorithm for automatically learning from audio feature data, the distributional parameters required to specify the probabilistic segmentation model (cf. [14]). Additionally, because the Viterbi inference algorithm presently employed for obtaining segment boundaries is an inherently off-line approach, computationally efficient on-line inference algorithms, for instance multiple model approaches [16, 17], shall be developed. This will allow continuously recorded audio to be stored as individual clips in real-time with the audio feature values serving as the indexing method for an environmental sound database. Finally, psychoacoustic trials will be performed to evaluate the validity of our proposed feature set.

6. ACKNOWLEDGMENTS

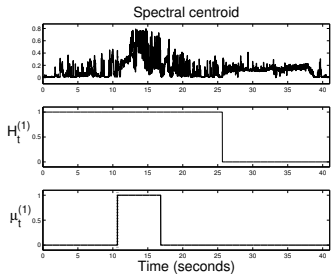
This material is based upon work supported by the National Science Foundation under Grant No. 0504647. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation (NSF).

7. REFERENCES

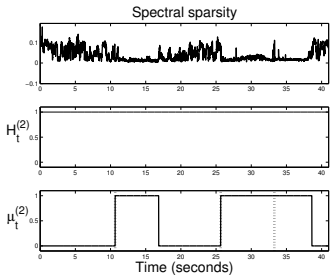
[1] H. Slabbekoorn and A. den Boer-Visser. Cities change the songs of birds. *Current Biology*, 16(23):2326–2331, 2006.



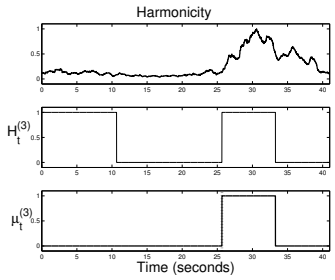
(a) Global segmentation



(b) Spectral centroid

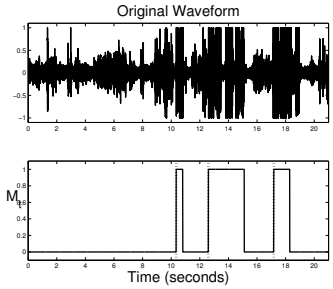


(c) Spectral sparsity

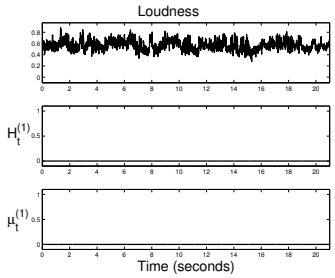


(d) Harmonicity

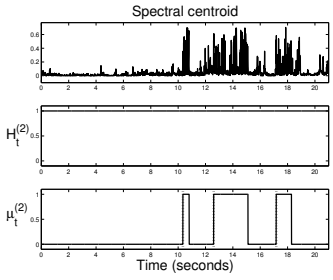
Fig. 2. Spectral Feature segmentation example for coffee sounds. (a) Signal waveform (upper panel) and global mode $M_{1:T}$ (lower panel). (b)-(d) Extracted audio features (upper panel) along with fusion gating parameters, $H_{1:T}^{(1:3)}$ (middle panel) and $\mu_{1:T}^{(1:3)}$ (lower panel).



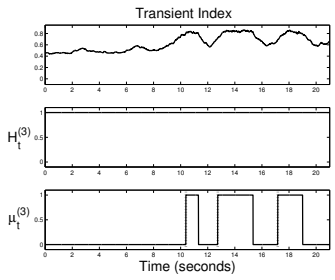
(a) Global segmentation



(b) Loudness



(c) Spectral Centroid



(d) Transient index

Fig. 3. Multi-feature segmentation example for gravel sounds. (a) Signal waveform (upper panel) and global mode $M_{1:T}$ (lower panel). (b)-(d) Extracted audio features (upper panel) along with fusion gating parameters, $H_{1:T}^{(1:3)}$ (middle panel) and $\mu_{1:T}^{(1:3)}$ (lower panel).

[2] V. Bush. As we may think. *Atlantic Monthly*, July 1945.

[3] D. P. W. Ellis and K. Lee. Minimal-impact audio-based personal archives. First ACM workshop on Continuous Archiving and Recording of Personal Experiences CARPE-04, New York, October 2004.

[4] M. F. McKinney and J. Breebaart. Features for audio and music classification. Proceedings of the 4th International Conference on Music Information Retrieval, Baltimore, MD, October 2003.

[5] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002.

[6] D. Li, I. K. Sethi, N. Dimitrova, and T. McGee. Classification of general audio data for content-based retrieval. *Pattern Recognition Letters*, 22:533–544, 2001.

[7] P. Hanna, N. Louis, M. Desainte-Catherine, and J. Benois-Pineau. Audio features for noisy sound segmentation. Proceedings of the 5th International Conference on Music Information Retrieval, Barcelona, Spain, October 2004.

[8] A. Harma, M. F. McKinney, and J. Skowronek. Automatic surveillance of the acoustic activity in our living environment. IEEE International Conference on Multimedia and Expo, Amsterdam, The Netherlands, July 2005.

[9] S. Rickard and M. Fallon. The GINI index of speech. Conference on Information Sciences and Speech, Princeton, NJ, March 2004.

[10] J. O. Smith III and J. S. Abel. Bark and ERB bilinear transforms. *IEEE Transactions on Speech and Audio Processing*, 7:697–708, 1999.

[11] ETSI standard document. Speech processing, transmission and quality aspects (STQ); distributed speech recognition; front-end feature extraction algorithm; compression algorithms. *ETSI ES 201 108 v1.1.3 (2003-09)*, 2003.

[12] J. L. Goldstein. An optimum processor theory for the central formation of the pitch of complex tones. *Journal of the Acoustical Society of America*, 54:1496–1516, 1973.

[13] H. Thornburg and R. J. Leistikow. A new probabilistic spectral pitch estimator: Exact and MCMC-approximate strategies. In U. Kock Wiil, editor, *Lecture Notes in Computer Science 3310*. Springer-Verlag, 2005.

[14] H. Thornburg. *Detection and Modeling of Transient Audio Signals with Prior Information*. PhD thesis, Stanford University, 2005.

[15] V. Pavlovic, J. M. Rehg, and T. Cham. A dynamic Bayesian network approach to tracking learned switching dynamic models. Proceedings of the International Workshop on Hybrid Systems, Pittsburgh, PA, 2000.

[16] Y. Bar-Shalom and T. E. Fortmann. *Tracking and Data Association*. Academic Press, 1988.

[17] F. Gustaffsson. *Adaptive Filtering and Change Detection*. Wiley, New York, 2001.

A. GOLDSTEIN LIKELIHOOD APPROXIMATION

Denoting the set of L most prominent peaks from a signal spectrum as $\rho = \{f_1, .., f_L\}$, our goal is to estimate the fundamental frequency f_o, for which the nth harmonic has frequency nf_o. As discussed in [13], accurately computing the Goldstein pitch estimate requires searching over all possible combinations of the L frequency peaks, and $2L$ possible harmonics. Our approximation limits this search region to pairwise combinations of the L frequency peaks, and k_{max} possible harmonics. Assuming the pair of peak frequencies denoted by $f_{\ell,1}$ and $f_{\ell,2}$, are harmonics of f_o, with corresponding harmonic numbers of k_1 and k_2, the Goldstein likelihood [12, 13] can be computed from

$$P(\rho|f_o, k_1, k_2) = \prod_{j=1}^{2} \mathcal{N}(k_j f_o, \sigma_j^2) \tag{19}$$

where the variance terms σ_j^2 are considered to be proportional to frequency [12], i.e.,

$$\sigma_j^2 = C k_j f_o \tag{20}$$

with C a constant. Given this form of σ_j^2, it was shown in [12], that the fundamental frequency estimate $\hat{f}_o$ for any pair of frequency peaks is

$$\hat{f}_o = \frac{(f_{\ell,1}/k_1)^2 + (f_{\ell,2}/k_2)^2}{f_{\ell,1}/k_1 + f_{\ell,2}/k_2}. \tag{21}$$

In order to obtain an approximation of $P(\rho|f_o)$ we estimate $f_{\ell,1}$, $f_{\ell,2}$, k_1, and k_2, by maximization of

$$\begin{aligned} P(\rho|\hat{f}_o) = & \underset{f_{\ell,1}, f_{\ell,2}, k_1, k_2}{\arg\max} P(\rho|\hat{f}_o, k_1, k_2) \\ & \text{subject to :} \\ & \{f_{\ell,1}, f_{\ell,2}\} \in \rho \\ & 1 < k_1 < k_2 < k_{max} \\ & f_{\ell,1} < f_{\ell,2}. \end{aligned} \tag{22}$$