MULTI-CHANNEL AUDIO SEGMENTATION FOR CONTINUOUS OBSERVATION AND ARCHIVAL OF LARGE SPACES

Gordon Wichern, Harvey Thornburg, and Andreas Spanias

Arts, Media, and Engineering
Arizona State University
Tempe, AZ 85281 USA
{Gordon.Wichern, Harvey.Thornburg, spanias}@asu.edu

ABSTRACT

In most real-world situations, a single microphone is insufficient for the characterization of an entire auditory scene. This often occurs in places such as office environments which consist of several interconnected spaces that are at least partially acoustically isolated from one another. To this end, we extend our previous work on segmentation of natural sounds to perform scene characterization using a sparse array of microphones, strategically placed to ensure that all parts of the environment are within range of at least one microphone. By accounting for which microphones are *active* for a given sound event, we perform a *multi-channel* segmentation that captures sound events occurring in any part of the space. The segmentation is inferred from a custom dynamic Bayesian network (DBN) that models how event boundaries influence changes in audio features. Example recordings illustrate the utility of our approach in a noisy office environment.

Index Terms— Acoustic signal analysis, Acoustic signal detection, Bayes procedures, Acoustic arrays.

1. INTRODUCTION

Understanding the acoustic scene in fixed spaces has assisted researchers in diverse fields such as ecology [1], surveillance [2], and personal media archives [3]. These recordings are often continuously captured over days or weeks by microphones left in the space or carried on a person. However, the amount of information contained in most continuous recordings calls for automated solutions that allow users to determine what type of sound events are present and where they occur. Thus, segmentation and indexing remain important technical challenges, especially for the difficult and often ignored classes of natural and environmental sounds.

Regarding segmentation our goal is to identify individual sound events (source separation is not currently considered) along with their onset and end times. In our definition of a sound event we follow the *auditory stream* concept [4] where an event is perceptually equivalent to the sound emanating from a single physical source, thus, a footstep event would consist of a cluster of footstep sounds rather than individual footsteps. Although the majority of past research in audio segmentation has focused on speech [5] and music [6], several approaches for temporally isolating sound events from continuously recorded environments have recently appeared [3, 2, 7, 8]. All of these event-based segmentation approaches are *single-channel*, using a single microphone placed somewhere in the space. Many spaces such as office environments, however, are large enough that sounds originating in one part of the space may not be perceptible in another, or may be perceptible only at low SNR. In this paper, we extend our own previous work on environmental sound segmentation [8] to the multi-channel case where sound is continuously recorded using a microphone array distributed throughout the space.

We begin the explanation of our multi-channel segmentation algorithm by reviewing the acoustic features that form the basis of the algorithm in Section 2. By monitoring changes in these features our method jointly infers onsets and end times of the most prominent sound events in the space along with the *active subset* of microphones responsible for capturing each event. This method utilizes a custom dynamic Bayesian network (DBN) that models how event boundaries influence changes in audio features, which will be detailed in Section 3. An illustrative example recorded with five microphones in an office environment is shown in Section 4, along with segmentation results in additive noise. Finally, conclusions and future work are provided in Section 5.

2. AUDIO FEATURE EXTRACTION

The audio features used as input to the DBN segmentation algorithm were chosen to represent a large variety of sounds without specifically assuming particular categories (e.g., speech, music). In this work we use combinations of the following features: *RMS level*, Bark-weighted *spectral centroid*, *spectral sparsity* (the ratio of ℓ^∞ and ℓ^1 norms calculated over the short-time Fourier Transform (STFT) magnitude spectrum), *transient index* (the ℓ^2 norm of the MFCC difference between consecutive frames), *temporal sparsity* (the ratio of ℓ^∞ and ℓ^1 norms calculated over all short-term RMS levels computed in a one second interval),

and *harmonicity* (a probabilistic measure of whether or not the STFT spectrum exhibits a harmonic frequency structure). These features are computed either directly from windowed time series data, via STFT using overlapping 40ms Hamming windows hopped every 20ms, or using a sliding "super frame" to combine data from multiple 40ms frames. A description of how all features are computed can be found in [8].

3. MULTI-CHANNEL SEGMENTATION

Let $t \in 1:T$ be the frame index, K the number of features extracted from each frame of the signal, and N the number of microphones. The multi-channel DBN model begins with the *global mode*, M_t, which is shared by all features and microphones. The three values M_t can take are $\emptyset, O1$, and $C1$ meaning there is no sound event, the onset of a new sound event, and the continuation of a sound event between contiguous frames, respectively. We define the time-varying N-dimensional *active subset* vector A_t, whose elements $A_{t,n} \in \{0,1\}$ indicate whether microphone n is active at frame t. The active subset A_t can take 2^N possible values, one for each possible combination of active/inactive microphones. If there are acoustically isolated microphones, and several acoustically isolated events occur simultaneously, only the sound with the highest SNR will be considered active.

Due to the variation in time scales and meaning of the different features, it is possible that certain features lag behind the global mode M_t when turning on or off. Furthermore, even if a sound is present at time t, it is likely that some features will fail to respond at all. The discrete N-dimensional vectors $\mu_t^{(i)}$, for $i \in 1:K$, serve as *feature gating variables*, whose elements $\mu_{t,n}^{(i)} \in \{\emptyset, O1, C1\}$ are constrained by the active subset. For instance, suppose $N = 3$ and $A_t = [1,0,1]^T$; then $\mu_t^{(i)} \in \{[\emptyset,\emptyset,\emptyset]^T, [O1,\emptyset,O1]^T, [C1,\emptyset,C1]^T\}$. That is only active microphones can have features that behave differently from silence.

The *observed features* $Y_t^{(i)}$ are N-dimensional vectors, where the observation of feature i at frame t for the nth microphone is denoted by $Y_{t,n}^{(i)}$. We also define the hidden *states* $S_{t,n}^{(i)}$, which are continuous random variables mediating the effect of individual features' onsets/end times on the actual observations $Y_{t,n}^{(i)}$. The latter are modeled as *inherent* features corrupted by noise. The state $S_{t,n}^{(i)}$ is then composed of this inherent feature plus an auxiliary variable enabling $S_{t,n}^{(i)}$ to be encoded as a first order Gauss-Markov process in time. The global mode M_t and active subset vector A_t are shared by all features and microphones, while the linear dynamic systems described by $Y_{t,n}^{(i)}$ and $S_{t,n}^{(i)}$ are independent over features and microphones as summarized by the DAG of Figure 1.

3.1. Distributional specifications

We begin with the frame likelihood

$$P(Y_{t,n}^{(i)} | S_{t,n}^{(i)}) \sim \mathcal{N}(v_{t,n}^{(i)}, R^{(i)}) \quad (1)$$

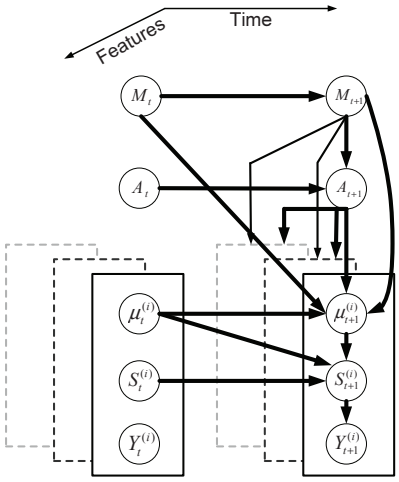

Fig. 1. **Directed acyclic graph for generative multi-channel audio segmentation model.**

where $v_{t,n}^{(i)}$ is the inherent feature and one component of the continuous random state vector $S_{t,n}^{(i)}$, i.e., $S_{t,n}^{(i)} = [u_{t,n}^{(i)}, v_{t,n}^{(i)}]^T$. The continuous state variables $u_{t,n}^{(i)}$ and $v_{t,n}^{(i)}$ satisfy the following stochastic recursive relations

$$\begin{aligned} u_{t,n}^{(i)} &= u_{t-1,n}^{(i)} + q_t^{(i)}(\mu_{t,n}^{(i)}, \mu_{t-1,n}^{(i)}) \\ v_{t,n}^{(i)} &= (1-\alpha^{(i)})u_{t,n} + \alpha^{(i)} v_{t-1,n}^{(i)} \end{aligned} \quad (2)$$

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 $q_t^{(i)}(\mu_{t,n}^{(i)}, \mu_{t-1,n}^{(i)})$ is distributed according to

$$q_t^{(i)}(\mu_{t,n}^{(i)}, \mu_{t-1,n}^{(i)}) \sim \mathcal{N}(0, Q^{(i)}(\mu_{t,n}^{(i)}, \mu_{t-1,n}^{(i)})). \quad (3)$$

where $Q^{(i)}(\mu_{t,n}^{(i)}, \mu_{t-1,n}^{(i)})$ is a variance, which will be large during event onset and end times.

Via $P(\mu_t^{(i)} | \mu_{t-1}^{(i)}, A_t, M_t, M_{t-1})$, we model possible time differences (lags) between when a particular feature gate, $\mu_{t,n}^{(i)}$, 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. The dependence on A_t allows only microphones that are included in the active subset to have their gate $\mu_{t,n}^{(i)}$ behave differently from silence, while all microphones that are included in the active subset must share the same value for $\mu_{t,n}^{(i)}$. A summary of $P(\mu_t^{(i)} | \mu_{t-1}^{(i)}, A_t, M_t, M_{t-1})$ is shown in Table 1.

The temporal dynamics of the active subset A_t follow $P(A_{t+1} | A_t, M_{t+1})$, which has three distinct forms depending on M_{t+1}. When there are no sound events observed by any of the microphones ($M_{t+1} = \emptyset$) the active subset will be

Table 1. **Transitions for $P(\mu_t^{(i)}|\mu_{t-1}^{(i)}, A_t, M_t, M_{t-1})$.**

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
$O1$	$C1$	$C1$	$p_{\text{lag}+}^{(i)}-(p_{\text{lag}-}^{(i)}\cdot p_{\text{lag}+}^{(i)})$	$1-p_{\text{lag}+}^{(i)}$	$p_{\text{lag}-}^{(i)}\cdot p_{\text{lag}+}^{(i)}$

Table 2. **Transition probabilities for $P(M_{t+1}|M_t)$.**

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

empty with probability one, i.e., $P(A_{t+1}=\emptyset|A_t, M_{t+1}=\emptyset)=1$. When a sound event continues between consecutive frames ($M_{t+1}=CI$) the active subset is constrained to be $P(A_{t+1}=A_t|A_t, M_{t+1}=CI)=1$, i.e., the active subset must be constant over an entire sound event. Thus, the active subset only changes during event onsets ($M_{t+1}=OI$).

If a sound event has an onset at time $t+1$, the active subset at time $t+1$ is independent of the active subset at time t, and $P(A_{t+1}|A_t, M_{t+1}=OI)=P(U)$, where $U\in\{1,2,...,2^{N-1}\}$ are all the non-empty active subset possibilities. To determine $P(U)$ we first define one active microphone as the anchor or reference microphone denoted by $\gamma\in\{1,..,N\}$ for each non-empty active subset. We can then write $P(U)=\sum_\gamma P(U|\gamma)P(\gamma)$ where the anchor microphone is uniformly distributed, i.e., $P(\gamma)=1/N$. The probability of each possible active subset given an anchor microphone is

$$P(U|\gamma)=\prod_{A_{t,n}=1} p_{on}^{(n)} \prod_{A_{t,n}=0}(1-p_{on}^{(n)}) \qquad (4)$$

where $A_{t,n}=1$ signifies that the nth microphone is active at time t, and similarly $A_{t,n}=0$ signifies inactivity. The probability that the nth microphone is active $p_{on}^{(n)}$ is given by an inverse sigmoidal function $p_{on}^{(n)}=\frac{1}{1+\exp(a\varepsilon^{(n)}-b)}$ where a and b are parameters controlling the slope and offset of the sigmoid function, and $\varepsilon^{(n)}$ is the Euclidean distance between microphone n and the anchor microphone. By choosing the inverse sigmoid model for $p_{on}^{(n)}$ we encode the prior knowledge that only those microphones located close to the anchor microphone will have a high probability of observing the same sound event.

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.

3.2. 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.

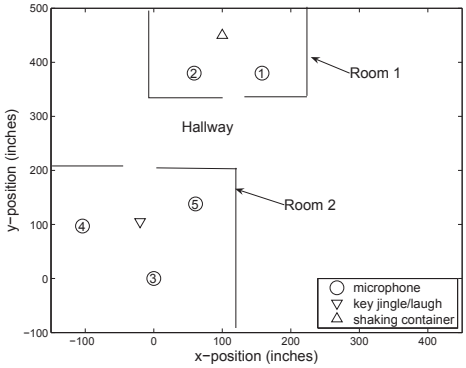

Fig. 2. **Office environment with microphone array positions and approximate locations of two sound sources.**

In order to achieve these goals we choose the maximum-a-posteriori (MAP) criterion [6],

$$\hat{M}_{1:T}=\arg\max_{M_{1:T}} P(M_{1:T}|Y_{1:T}^{(1:K)}). \qquad (5)$$

Unfortunately, computing the exact MAP estimate requires exponential-time complexity. A linear-time approximation nevertheless exists, using an approximate Viterbi inference scheme [9].

4. PRELIMINARY RESULTS

Our examples were recorded using a five microphone array in an indoor office environment consisting of two meeting rooms separated by a hallway (Figure 2). All microphones in the array had equal elevation, and (x,y) coordinates in inches of $(158,380),(59,380),(0,0),(-104,97),(61,138)$. To test noise robustness we added white Gaussian noise to the signal at each microphone at SNR levels between -10dB to 10dB.

Figure 3 displays the time-domain waveforms from each of the five microphones in the array with 0dB SNR. Additionally, the global mode sequence $M_{1:T}$ inferred from the Viterbi algorithm is shown in the bottom panel of Figure 3, while the inferred active subset variables ($A_{t,n}$) are plotted below the corresponding channel waveforms. Values when the global mode is off ($M_t=\emptyset$) are plotted as zero, values when the global mode is on ($M_t=C1$) are plotted as one, and onsets ($M_t=O1$) are plotted as dotted lines. From Figure 3 three distinct events from this example recording can be recognized. First, there is the sound of jingling keys (1-5.5 seconds) on channels three, four, and five. Second, there is a laugh (5.5-6.5 seconds) on channels three, four, and five. Third, there is a shaking food container (9.5-13 seconds) on channels one and two. We see that the global mode sequence in general does a good job of capturing these events in the presence of additive noise, while the active subset accurately detects the microphones that are observing the sounds.

In the example of Figure 3 the RMS level, spectral centroid, and spectral sparsity feature sequences extracted from each of the five channels are used as the DBN observations. The RMS level takes high values during a sound event, but

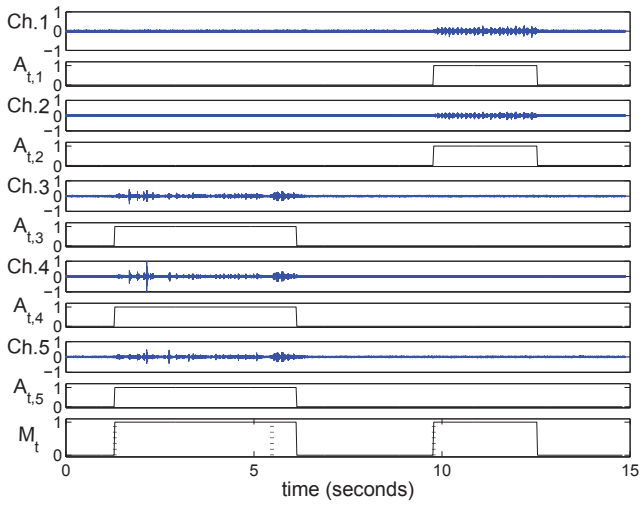

Fig. 3. **Signal waveforms for five microphone indoor recording along with the corresponding active subset variables $A_{t,n}$, and the global mode sequence $M_{1:T}$.**

has a hard time distinguishing between contiguous events. The spectral centroid detects only the shaking keys sound, and the spectral sparsity clearly distinguishes the harmonic laugh sound, but appears unresponsive to the more noise-like key shaking and plastic container events. This illustrates the necessity of using multiple features to detect environmental sound events, while the fact that the laugh event is captured as different from the key jingling event demonstrates the improvements of the proposed algorithm over a simple level detection segmentation approach.

Next we examine the performance of our algorithm in additive noise by comparing it to a ground truth (human annotated) segmentation. We use a histogram distance metric[1] to evaluate the Viterbi algorithm by comparing the ground truth global mode sequence $M_{1:T}$ to the estimated global mode sequence $\hat{M}_{1:T}$, as $\delta(M_{1:T}, \hat{M}_{1:T}) = \sqrt{\sum_{j=1}^{3}(hist(j) - \hat{hist}(j))^2}$ where $hist(j)$ is the number of frames spent in global mode j divided by T. Table 3 summarizes the histogram distance, the number of onsets, and the frames where a sound event is on for the example recording shown in Figure 3. From Table 3 we observe that the number of onsets, which can also be interpreted as the number of sound segments detected, remains relatively constant even at low SNR. As the SNR decreases, the number of frames where sound events are on becomes smaller as portions of the key jingle sound become completely overshadowed by the additive noise. As expected, we notice a large jump in histogram distance $\delta(M_{1:T}, \hat{M}_{1:T})$, as the SNR becomes negative.

5. CONCLUSIONS AND FUTURE WORK

In order to characterize the auditory scene in a fixed space, a single microphone is often insufficient. In this paper we demonstrate a method to temporally isolate sound events recorded with a microphone array that covers multiple rooms of an office environment. In the future we hope to integrate a fine-scale likelihood-based localization technique into the DBN model, so that sound events can be localized in both time and space. We are also currently investigating solutions based around a local MAP criterion to retain or even surpass the quality of the Viterbi segmentation results.

[1] This metric is adapted from the MPEG-7 audio standard[10]

Table 3. **Multi-channel segmentation performance in various levels of additive noise for the example environment.**

SNR (dB)	onsets	frames on	histogram distance
truth	3	408	0
10	3	410	0.0038
5	4	402	0.0106
0	3	377	0.0592
-5	5	302	0.2004
-10	4	203	0.3903

6. REFERENCES

[1] H. Slabbekoorn and A. den Boer-Visser, "Cities change the songs of birds," *Current Biology*, vol. 16, no. 23, pp. 2326–2331, 2006.

[2] A. Harma, M. F. McKinney, and J. Skowronek, "Automatic surveillance of the acoustic activity in our living environment," in *IEEE ICME*, Amsterdam, 2005.

[3] D. P. W. Ellis and K. Lee, "Minimal-impact audio-based personal archives," in *ACM CARPE*, New York, 2004.

[4] A.S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*, MIT Press, Cambridge, 1990.

[5] R. Andre-Obrecht, "A new statistical approach for automatic segmentation of continuous speech signals," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 36, pp. 29–40, 1988.

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

[7] L. Lu, R. Cai, and A. Hanjalic, "Audio elements based auditory scene segmentation," in *IEEE ICASSP*, Toulouse, France, 2006.

[8] G. Wichern, H. Thornburg, B. Mechtley, A. Fink, K. Tu, and A. Spanias, "Robust multi-feature segmentation and indexing for natural and environmental sounds," in *IEEE CBMI*, Bordeaux, France, 2007.

[9] V. Pavlovic, J. M. Rehg, and T. Cham, "A dynamic Bayesian network approach to tracking learned switching dynamic models," in *Intl. Wkshp. on Hybrid Systems*, Pittsburgh, PA, 2000.

[10] H. G. Kim, N. Moreau, and T. Sikora, *MPEG-7 Audio and beyond: audio content indexing and retrieval*, John Wiley & Sons Ltd., West Sussex, England, 2005.