SOUND SOURCE TRACKING AND FORMATION USING NORMALIZED CUTS

Mathieu Lagrange, George Tzanetakis

Department of Computer Science
University of Victoria
3800 Finnerty Road
Victoria, BC, Canada V8P5C2
{*lagrange,gtzan*}*@uvic.ca*

ABSTRACT

The goal of computational auditory scene analysis (CASA) is to create computer systems that can take as input a mixture of sounds and form packages of acoustic evidence such that each package most likely has arisen from a single sound source. We formulate sound source tracking and formation as a graph partitioning problem and solve it using the normalized cut which is a global criterion for segmenting graphs that has been used in Computer Vision. It measures both the total dissimilarity between the different groups as well as the total similarity within groups. We describe how this formulation can be used with sinusoidal modeling, a common technique for sound analysis, manipulation and synthesis. Several examples showing the potential of this approach are provided.

Index Terms— auditory scene analysis, sinusoidal modeling, sound source tracking, normalized cut

1. INTRODUCTION

A fundamental characteristic of the human hearing system is the ability to selectively attend to different sound sources in complex mixtures of sounds such as speech with multiple overlapping speakers in "natural" environments or music. This process has been termed *Auditory Scene Analysis* (ASA) by McGill psychologist Albert Bregman [1]. *Computational Auditory Scene Analysis* (CASA) refers to the process of modeling ASA computationally [2]. Effective CASA systems would result in improved speech recognition in noisy environments and facilitate the analysis of complex audio signals such as music or bioacoustic signals. Humans use a variety of cues for perceptual grouping in hearing such as similarity, proximity, harmonicity and common fate. However many of the computational issues of perceptual grouping for hearing are still unsolved. The *normalized cut* is a global criterion for graph partitioning that has been proposed for solving similar grouping problems in computer vision [3].

Thanks to NSERC and SSHRC for their support of this work.

Sinusoidal modeling is a technique for analysis and synthesis where sound is modeled as the summation of sine waves parameterized by time-varying amplitudes, frequencies and phases. In the classic McAulay and Quatieri method [4] these time varying quantities are estimated by performing a short-time Fourier transform (STFT) and locating the peaks of the associated magnitude function. Partial tracking algorithms can then be used to track the sinusoidal parameters from frame to frame, and to determine when new partials begin and existing ones terminate. If the goal is to identify potential sound sources then a separate stage of partial grouping is needed. Typically perceptual grouping cues such as common onset time and spectral proximity are used.

In this paper we use the term sound source formation and tracking to refer to these two processes of connecting peaks over time to form partials (tracking) and grouping them to form potential sound sources (formation). They roughly correspond to the simultaneous and sequential aspects of organization described by Bregman [1]. Although typically implemented as separate stages these two organizational principles directly influence each other. For example if we have knowledge that a set of peaks belong to the same source then their correspondence with the next frame as easier to find. Similarly the formation of sound sources is easier if peaks can be tracked perfectly over time. Methods such as [5, 6] that apply these two stages in fixed order tend to be brittle as they are sensitive to errors and ambiguity. To cope with this chicken-and-egg problem we show how both sound source tracking and formation can be jointly optimized within a unified framework using the normalized cut criterion.

We model the problem as a weighted undirected graph **G = (V,E)**, where the nodes of the graph are the peaks of the magnitude spectrum and an edge is formed between each pair of nodes. The edge weight $w(i, j)$, is a function of the similarity between nodes i and j and utilizes various grouping cues such as frequency and amplitude proximity. We use the term formation rather than separation as our goal is not to recover the original sound sources that comprise the mixture but rather to provide an intermediate representation for sound.

2. RELATED WORK

The normalized cut criterion for graph partitioning was initially proposed for image segmentation [3]. It is a representative example of spectral clustering techniques which use an *affinity matrix* W to encode topological knowledge about a problem. Spectral clustering approaches have been used in a variety of applications including high performance computing, web mining, biological data, image segmentation and motion tracking. There are few applications of spectral clustering to audio processing that we are aware of. In this section they are briefly described and we contrast them with our approach.

Spectral clustering has been used for blind one-microphone speech separation [7]. Rather than building specific speech models, the authors show how the system can separate mixtures of two speech signals by learning the parameters of affinity matrices based on various harmonic and non-harmonic cues. Rather than sound source separation the focus of our work is the formation of an intermediate audio representation [8] that combines ideas from sinusoidal partial tracking and grouping. Another important difference is the use of sinusoidal modeling as a front-end rather than the entire STFT magnitude spectrum. This results in more accurate and robust similarity relations as well as significantly smaller affinity matrices that are computationally more tractable.

Another use of spectral clustering methods for audio processing has been the unsupervised clustering of similar sounding segments of audio [9, 10]. Each audio frame is characterized by a feature vector and a self-similarity matrix is constructed and used as the basis for clustering. This approach has also been linked to the singular value decomposition of feature matrices to form audio basis vectors [11]. In all these approaches the audio mixture is characterized statistically without any attempt to use spectral clustering for forming and tracking individual sound sources.

3. THE NORMALIZED CUT

The normalized cut algorithm, presented in [3], aims to partition an arbitrary set of data points into n clusters. The data set is modeled as a complete weighted undirected graph **G = (V, E)**, the nodes representing the data points and each edge weight $w(i, j)$ representing the relative similarity between the two end nodes i and j. The graph is represented internally by an affinity matrix, W, that specifies all edge weights. The partitioning is achieved by recursively dividing one of the connected components of the graph into two until n complete components exist. The criterion that is minimized in order to establish the optimal partitioning at any given level is the normalized cut disassociation measure ($Ncut$):

$$Ncut(A, B) = \frac{cut(A, B)}{asso(A, V)} + \frac{cut(A, B)}{asso(B, V)} \tag{1}$$

where $asso(X, V) = \sum_{u \in X, t \in V} w(u, t)$ is the total of the weights from nodes in cluster X to all nodes in the graph. An analogous measure of the association within clusters is the following ($Nasso$):

$$Nasso(A, B) = \frac{asso(A, A)}{asso(A, V)} + \frac{asso(B, B)}{asso(B, V)} \tag{2}$$

where $asso(X, X)$ is the total weight of edges connecting nodes within cluster X. We note the following relationship between $Ncut$ and $Nasso$:

$$Ncut(A, B) = 2 - Nasso(A, B) \tag{3}$$

Hence, the attempt to minimize the disassociation between clusters is equivalent to maximizing the association within the clusters. The formulation of the $Ncut$ measure addresses the bias toward partitioning out small sets of isolated nodes in a graph which is inherent in the simpler minimal cut disassociation measure (cut):

$$cut(A, B) = \sum_{u \in A, v \in B} w(u, v) \tag{4}$$

The hierarchical clustering of the data set via the minimization of the $Ncut$ measure, or the equivalent maximization of the $Nasso$ measure, may be formulated as the solution to an eigensystem. One of the advantage of the normalized cut over clustering algorithms such as K-means or mixtures of Gaussians estimated by the EM algorithm is that there is no assumption of convex shapes in the feature representation.

4. SOUND SOURCE FORMATION AND TRACKING

Sinusoidal modeling aims at representing a sound signal as a sum of sinusoids characterized by amplitudes, frequencies, and phases. A common approach is to segment the signal into successive frames of small duration so that the stationarity assumption is met. The discrete signal $x^k(n)$ at frame index k is then modeled as follows:

$$x^k(n) \quad = \quad \sum_{l=1}^{L^k} a_l^k \cos\left(\frac{2\pi}{F_s} f_l^k \cdot n + \phi_l^k\right) \tag{5}$$

where F_s is the sampling frequency and ϕ_l^k is the phase at the beginning of the frame of the l-th component of L_k sine waves, f_l^n and a_l^n are respectively the frequency and the amplitude. Both are considered as constant within the frame.

For each frame k, a set of sinusoidal parameters $\mathcal{S}^k = \{p_1^k, \cdots, p_{L^k}^k\}$ is estimated. The system parameters of this Short-Term Sinusoidal (STS) model $\mathcal{S}^k$ are the L^k triplets $p_l^k = \{f_l^k, a_l^k, \phi_l^k\}$, often called *peaks*. These parameters can be efficiently estimated by picking some local maxima from a Short-Term Fourier Transform (STFT).

The precision of these estimates is further improved using phase-based frequency estimators which utilize the relationship between phases of successive frames [12]. Using this enhanced frequency, the rough amplitude estimate provided by the magnitude of the local maximum is also corrected.

In order to simultaneously optimize partial tracking and source formation we construct a graph over the entire duration of the sound mixture of interest. Unlike approaches based on local information [4] we utilize the global normalized cut criterion to partition the graph. Each partition is a set of peaks that are grouped together such that the similarity within the partition is minimized and the dissimilarity between different partitions is maximized. The edge weight connecting two peaks p_l^k and $p_{l'}^{k'}$ (k is the frame index and l is the peak index) depends on both frequency and amplitude proximity:

$$W(p_l^k, p_{l'}^{k'}) = W_f(p_l^k, p_{l'}^{k'}) * W_a(p_l^k, p_{l'}^{k'}) \quad (6)$$

We use radial basis functions (RBFs) to model the frequency and amplitude similarities:

$$W(p_l^k, p_{l'}^{k'}) = e^{-\left(\frac{f_l^k - f_{l'}^{k'}}{\sigma_f}\right)^2} * e^{-\left(\frac{a_l^k - a_{l'}^{k'}}{\sigma_a}\right)^2} \quad (7)$$

Notice that edges are formed both for peaks within a frame and peaks across frames and the number of peaks for each frame can be variable. We also use a variation of the amplitude similarity weight we term mean scaled difference function (msdf). It considers pairs of high amplitude peaks more similar than pairs of low amplitude peaks that are equally different from each other. The rationale is that high amplitude peaks in audio tend to be more perceptually important.

$$W_{msdf}(p_l^k, p_{l'}^{k'}) = e^{-\left(\frac{a_l^k - a_{l'}^{k'}}{\sigma_a \cdot (a_l^k + a_{l'}^{k'})}\right)^2} \quad (8)$$

In the results presented below we also incorporate into the similarity calculation harmonicity information. There is not enough space to describe the method in detail but the basic idea is to try to increase the similarity of sounds that have shared harmonic peaks.

5. EXPERIMENTS

For the experiments described in this section the frame size is 2048 samples with a hop size of 360 samples at 44100Hz sampling rate. For each frame a set of spectrum peaks are selected up to a maximum of 20 selected by decreasing amplitude. We utilize an experimental setup inspired by the "old+new" heuristic described by Bregman [1]. Each sample is created by mixing two sound sources in the following way: for the first 20 frames only the "old" sound is played followed by the addition of the "new" sound (old+new). The idea is to use knowledge obtained from the "old" source to separate the "old" from the "old+new" mixture. In the normalized cut case we cluster the entire clip into 5 clusters with using both the "old" and "old+new" parts. Afterward, we identify the clusters that have peaks in the initial "old" part and only use the peaks belonging to them to resynthesize the "old" part from the "old+new". The quality of the resynthesized "old" part is an indication of the sensitivity of the tracking and formation algorithm to the introduction of a new interfering source and is measured as the Signal-to-Noise Ratio (SNR) between the resynthesized "old" source and the original. To put our results in context we also use the classic McAulay and Quatieri (MQ) [4] partial tracking technique and only keep partials from the initial "old" part in order to separate the "old" from the "old+new". It is important to note that the MQ approach only provides information about tracking whereas the normalized cut approach also attempts to do source formation. Figure 2 shows the clustering of several harmonics within the same source (an orca vocalization) in the presence of noise.

	SN	VS	VX	VN	SV	SN
MQ	7.16	-11.40	-9.89	-5.28	10.77	6.36
A	7.22	2.86	**2.59**	3.12	**2.77**	3.28
MA	6.05	**3.05**	1.01	3.12	1.49	5.97
F	8.36	1.00	1.01	3.51	1.49	6.25
FBrk	8.57	1.00	1.01	3.39	1.49	6.79
A,F	8.28	2.92	1.71	**3.71**	2.69	7.81
MA,F	**8.80**	1.00	1.01	3.32	1.49	**8.77**
H	-4.79	1.00	1.01	3.12	1.49	5.97
A,F,H	7.94	1.00	1.01	3.12	1.49	5.97
A,F,H(2)	6.05	1.00	1.01	3.12	1.49	5.97

Table 1. SNR result for old+new interference

Table 1 provides SNR measurements using different configurations of the similarity function. The following conventions are used for the measures and and their combinations: *A* (amplitude), *MA* (mean scaled amplitude), *F* (frequency), *FBrk* (frequency in Barks), and *H* (harmonicity). The columns correspond to different configurations of old+new mixtures. For example *SN* means an harmonic sweep (old) that is mixed after 20 frames with noise (old+new). They are: *S*(harmonic sweep), *V*(violin), *X*(sax), and *N*(noise). The number of clusters is set to 5 except the *A,F,H(2)* entry which is 2. For the resynthesis we select only the clusters that are present in the old part before the introduction of the new. These results indicate that our approach can perform tracking of partials that is equally good or better than local partial tracking (MQ). In addition it also performs grouping of the peaks for source formation. Figure 1 shows the results of our proposed approach (filled circles on top) and classic MQ tracking (connected circles at bottom) with only the 10 highest amplitude partials represented by solid lines. Because of the local nature of the MQ approach it can not group partials across frequency, time and through the noise bursts as our approach does.

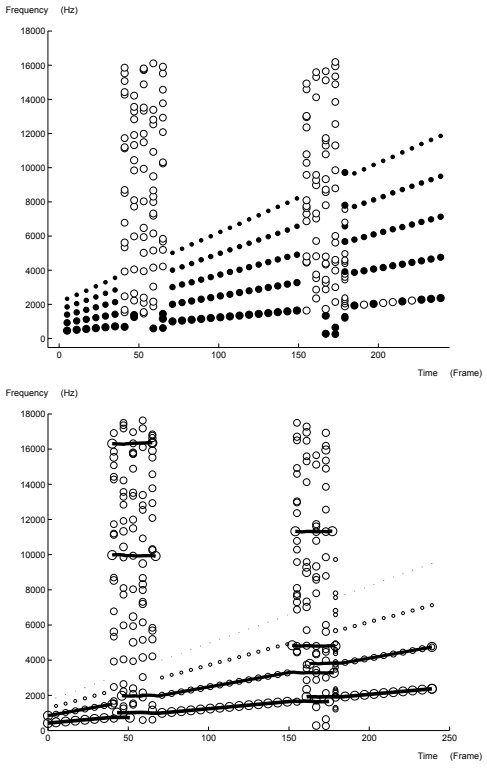

Fig. 1. Sweep through Noise using the Normalized Cut (top) and McAuleyQuatieri tracking (bottom)

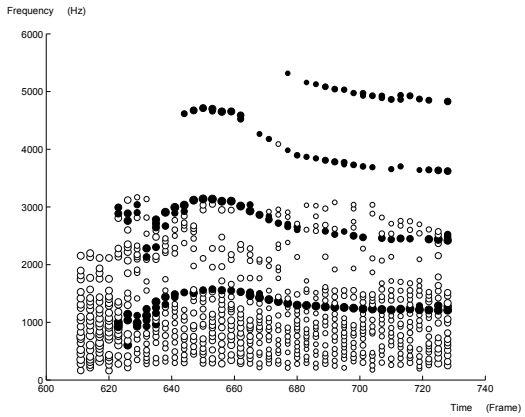

Fig. 2. Orca Vocalization tracked using the normalized cut

6. FUTURE WORK

We believe our work shows the potential of using the normalized cut criterion for simultaneous sound source formation and partial tracking. There are many directions for future work. We plan to explore the use of additional cues such as common onset and common fate for the similarity calculations as well as incorporate time decay and windowing to handle longer time scales. Another interesting direction is the use of prior models to inform the sound formation [13]. Finally we are exploring the use of intermediate time-frequency representations calculated using our method as a front-end for applications such as speech enhancement, bioacoustics (see Figure 2), and music information retrieval.

7. REFERENCES

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

[2] D.F. Rosenthal and H.G. Okuno, Eds., *Computational Auditory Scene Analysis*, 1998.

[3] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22(8), pp. 888–905, 2000.

[4] R.J. McAulay and T.F. Quatieri, "Speech analysis/synthesis based on sinusoidal representation," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 34(4), pp. 744–754, 1986.

[5] S.H. Srinivasan and M. Kankanhalli, "Harmonicity and dynamics based audio separation," in *IEEE ICASSP '03*, 2003, vol. 5, pp. v–640 – v–643.

[6] S.H. Srinivasan, "Auditory blobs," in *IEEE ICASSP '04*, 2004, vol. 4, pp. iv–313 – iv–316.

[7] F.R. Bach and M. I. Jordan, "Blind one-microphone speech separation: A spectral learning approach," in *Proc. Neural Information Processing Systems (NIPS)*, Vancouver, Canada, 2004.

[8] D.K. Mellinger, *Event formation and separation in musical sound*, Ph.D. thesis, Stanford University, 1991.

[9] D. Ellis and K. Lee, "Minimal-impact audio-based personal archives," in *Proc. ACM Workshop on Continuous Archival and Retrieval of Personal Experience (CARPE)*, New York, USA, 2004.

[10] R. Cai, L. Lu, and A. Hanjalic, "Unsupervised content discovery in composite audio," in *Proc. ACM Multimedia*, 2005.

[11] S. Dubnov and T. Appel, "Audio segmentation by singular value clustering," in *Proc. of Int.Conf. on Computer Music (ICMC)*, 2004.

[12] S. Marchand and M. Lagrange, "On the equivalence of phase-based methods for the estimation of instantaneous frequency," in *Proc. European Conference on Signal Processing (EUSIPCO'2006)*, 2006.

[13] E. Vincent, "Musical source separation using time-frequency priors," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14(1), pp. 91–98, 2006.