Temporal Constraints for Sound Source Formation using the Normalized Cut

Mathieu Lagrange, Jennifer Murdoch, George Tzanetakis
Department of Computer Science
University of Victoria
3800 Finnerty Road
Victoria, B.C., Canada V8P 5C2
{lagrange,jmurdoch,gtzan}@uvic.ca

Abstract

In this paper, we explore the use of a graph algorithm called the normalized cut in order to organize prominent components of the auditory scene. We focus specifically on defining a time-constrained similarity metric. We show that such a metric can be successfully expressed in terms of the time and frequency masking phenomena and can be used to solve common problems in auditory scene analysis.

1 Introduction

Auditory scene analysis attempts to organize auditory sensory input in order to derive from it a useful representation of reality that may be used by higher-level cognitive processes such as planning. Bregman [1] tells us that organizing auditory stimuli according to the source of the stimuli is important to the development of an accurate representation. While human perception is able to accomplish this task quite well, computational algorithms designed to accurately model auditory perception are exceedingly difficult to develop because the cues that affect perception of an auditory stimuli are not yet clear to us and they are undoubtedly large in number; contextual parameters, and the level of "acoustic maturity" and motivation of a listener are some of the more, complex and difficult to quantify examples of such cues. Furthermore, the functions among these parameters that accurately model human perception are difficult to extract from psychological testing due to this complex cues.

Despite this complexity, it is clear that temporal cues play a critical role in this process of source identification and segmentation. This intuitively makes sense within the context of the physical world because acoustic events that are proximal in time are more likely to be related than those that occur far apart.

In this paper we explore the use of a graph algorithm called the *normalized cut* in order to organize prominent components of the auditory scene according to several constraints. The normalized cut criterion for graph partitioning was initially proposed for image segmentation [2]. 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 analysis, image segmentation and motion tracking. There are few applications of spectral clustering to audio processing that we are aware of: blind one-microphone speech separation [3] and unsupervised clustering of similar sounding segments of audio [4, 5]. Closer to our approach, harmonicity relationships and common fate cues underlie a devised similarity measure presented by Srinivasan [6]. However, the sound source formation process is done on a frame basis only. To integrate time constraints, it is proposed in [7] to cluster previously tracked partials to form 'blobs' according to onset cues. Normalized cut clustering is then carried out on the elements of this intermediate representation.

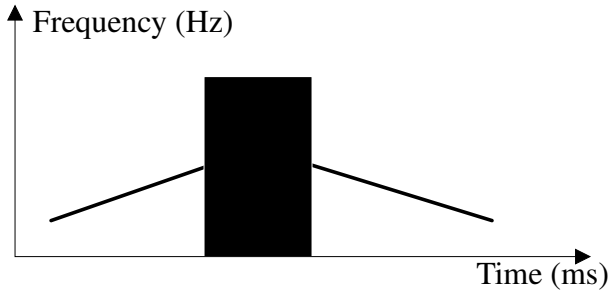

Figure 1: ASA scenario illustrating the temporal constraints in the simultaneous integration level. In this case, the noise burst leads to the perceptual continuity of the tone.

Alternatively, we propose an analysis framework for simultaneous sound source formation and tracking. We particularly focus on incorporating temporal constraints that model human perception of continuity and streaming into the affinity matrix W. We propose to address this issue by introducing a similarity metric expressed in terms of the masking phenomenon.

This paper is organized as follows: after presenting temporal aspects of auditory scene analysis in section 2, we describe our computational framework in section 3. We introduce in section 4, a dissimilarity metric that incorporates our temporal constraints, and we show that the use of this similarity metric can handle some known problems in auditory scene analysis.

2 Temporal Constraints in Auditory Scene Analysis

The choice to model aspects of spectral clustering algorithms after human perception can be justified on two levels. Firstly, human perception has evolved to organize acoustic events in a manner that is likely to mirror the organization of sound sources within the physical world. Secondly, the way the human auditory system is constructed gives us insight into how we should design our analysis framework.

We consider three common levels of sound integration. The first integration level is the spectral level (spectrogram, correlelogram, etc.) which mimics the output of the inner ear. Bregman proposed two additional higher-level integration schemes: simultaneous integration and sequential integration [1]. Simultaneous integration facilitates the identification of sound sources, while sequential integration organizes these sources through time, forming what Bregman describes as streams.

The temporal properties of acoustic events play an important role in the way in which we may perceptually form (simultaneous integration) and group (sequential integration) these events. Here we describe one particular auditory scene analysis scenario that was originally proposed by Bregman which clearly illustrates integration at the simultaneous level.

The ASA scenario, illustrated in figure 1, is comprised of an oscillating pitch glide with a discrete break positioned within the glide. The break may contain silence or be filled with noise. While the glide is audibly disjoint in the case where the break contains silence, the glide may be perceived as a single acoustic entity that is continuous through the noise if the noise is of sufficient power. However, the duration of the break within the pitch oscillation also contributes to this effect; in the case where the break contains silence, the glide may still be perceived as continuous at the spectral level if the duration of the silence is reduced sufficiently. Similarly, the glide may be perceived as disconnected in the case where noise is introduced in the break if the duration of the break is made sufficiently long.

3 Computational Framework

We now present our computational framework based on the use of the normalized cut to cluster sinusoidal components in order to form sound sources.

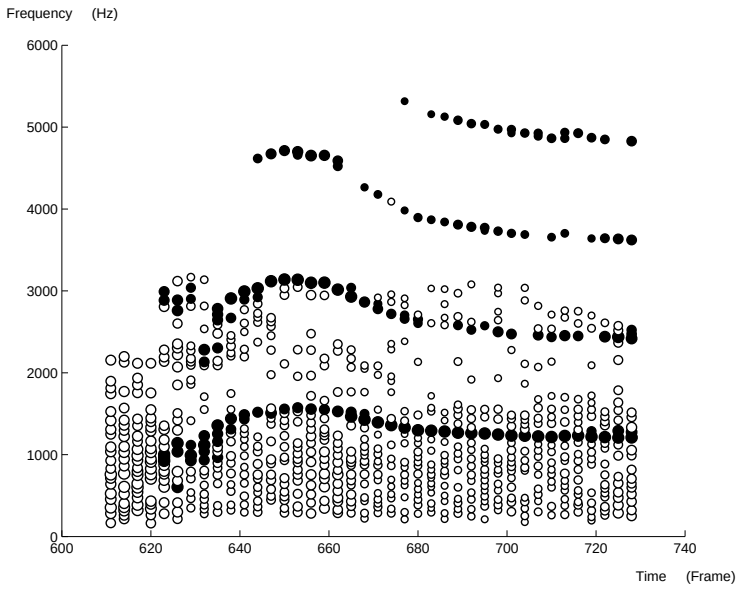

Figure 2: Orca Vocalization tracked using the normalized cut.

Sinusoidal modeling aims to represent 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:

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

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. The fn and an are respectively the frequency and the amplitude of the l-th sine wave, both of which 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*.

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 [8], 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 the proximity of both frequency and amplitude:

$$W(p_l^k, p_{l'}^{k'}) = W_f(p_l^k, p_{l'}^{k'}) + W_a(p_l^k, p_{l'}^{k'}) \tag{2}$$

Notice that edges are formed both for peaks within a frame and peaks across frames, and the number of peaks for each frame may vary. Figure 2 shows the clustering of several harmonics within the same source (an orca vocalization) in the presence of significant noise by also incorporating into the similarity calculation harmonicity informations.

4 Temporal Constraints

In addition to the aforementioned similarities, we now introduce a similarity measure that considers time and continuity constraints. Our primary motivation stems from the fact that most illusions of continuity, or "perceptual" continuity, exhibited by the ASA scenario (see Section 2) are due to the

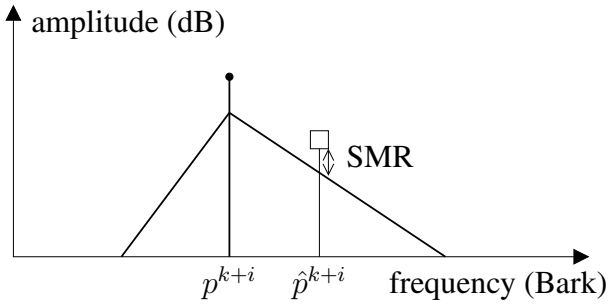

Figure 3: Frequency masking curve.

perceptual phenomenon called masking which is widely used in the audio coding area. Finally, we propose an interpretation of the perceptual mechanism underlying the ASA scenario illustrated in figure 1 that is based on the masking phenomenon, and we illustrate how our algorithm models it.

4.1 Temporal and Frequency Masking

The perceptual behavior of the ASA scenario may be interpreted with consideration to the frequency and temporal masking effects described in [9]. The principal behind the frequency masking effect is that every harmonic component of a sound has the ability to "mask", or render inaudible, harmonics having a similar frequency and smaller amplitude; the amplitude threshold of a harmonic that can be masked is approximately inversely proportional to the frequency difference between the masked and masking harmonics when frequency is measured using the Bark scale, see Figure 3. An important parameter that measures this phenomenon is the signal-to-mask ratio (SMR) which increases as the audibility of a tone increases such as the one labeled with a square stem in Figure 3.

The notable result that may be drawn from the frequency masking effect is that if we perceive an harmonic at some frequency, then based on perception alone, we cannot say anything about the frequency content of the sound within the thresholded maskable region of the Bark-dB scale that surrounds the harmonic. This has interesting implications with respect to the ASA scenario. When the breaks in an oscillating pitch glide contain silence, we do not experience frequency masking during the break, and thus, we are able to infer reliably that no harmonics are actually present. In this case, we hear the silent break which results in the perception of the pitch glide as disconnected.

Now consider the case in which the break in the oscillating pitch glide are filled with white noise of sufficient energy to cause the perception of a pitch glide that is continuous through the noise. In this case, the white noise provides a continuum of frequency components, all of which are capable of inducing masking effects. Due to these masking effects, we are not able to infer anything about the presence or absence of harmonics having less energy than the noise itself, and yet we perceive a continuous pitch glide! We hypothesize that in this situation the human brain cannot determine the presence or absence of the pitch glide through the noise, and it therefore tries to reconcile this information using other cues. One such cue is likely temporal continuity; since the disjoint portions of the pitch glide are separated by a relatively small break in time, the brain infers that the glide is continuous.

In addition to the frequency masking discussed above, temporal masking effects also exist, of which there are two types. Post-masking occurs when the masking sound disappears. In fact, its effect persists in time during some number of milliseconds (See figure 4). As a consequence, even if the masking sound is not present, the masking effects persist, though they decrease with time. Perhaps more surprisingly, pre-masking also exists. More precisely, the masking effect is active a few milliseconds before the masking sound really appears. However, this phenomenon is less significant than post-masking.

4.2 Similarity Metric

Following these remarks, we introduce a similarity metric that deals with the problem of perceptual continuity. First, a given masking curve that considers both frequency and temporal masking phe-

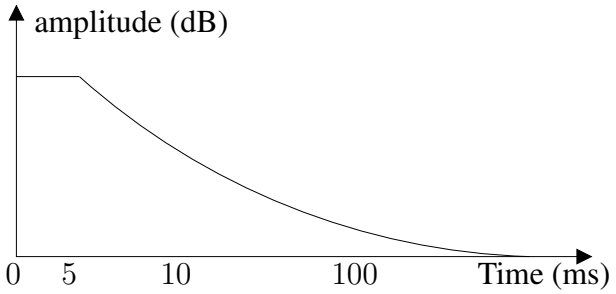

Figure 4: Approximated curve of temporal post-masking.

nomena is computed for each analyzed frame. The masking power of neighboring tones are additive, but in a non-linear way. In our implementation, the masking curve at a given frequency is set to the maximal masking contribution of the neighboring tones in frequency and time [10, 11].

Let us consider the case of two peaks (p_l^k, p_m^{k+n}) several frames apart. If these peaks belong to the same partial, then the energy of this partial within the frame interval will generate peaks or be masked by another source. Let us consider some virtual peaks that model the continuous partial between p_l^k and p_m^{k+n} with parameters linearly interpolated: $\hat{p}^{k+i}, i \in [1, n-1]$. The temporal similarity between two peaks is then:

$$W_t(p_l^k, p_m^{k+n}) = F\left(\sum_{i=1}^{n-1} \text{SMR}(\hat{p}^{k+i})\right) \tag{3}$$

where F is in our case a radial basis function and SMR is the Signal-to-Mask-Ratio of the virtual peaks, see Figure 3.

5 Experiments

To exhibit the advantageous properties of the proposed metric, we now consider three test cases taken from the ASA scenario. In the first case, a sinusoidal tone is interrupted for 5 ms. According to our listening tests, the gap is perceived, but the tones before and after the gap are perceived as only one tone. The second test case considers a longer gap of approximately 50 ms where the tones are perceived separately. In the last case, we consider even longer gaps (120 ms) filled with high amplitude low-pass-filtered noise. In this case, the two separated tones are perceived as continuous.

The results of our algorithm for these three test cases are shown in Figure 5 from top to bottom. We used the combination of frequency, amplitude and time similarity metrics and requested 5 clusters. For the sake of clarity, the peak clusters are illustrated with filled or unfilled dots for each test case. For example, all the peaks of the second experiment belong to the same cluster, whereas the peaks of the first and last experiments are separated into two different clusters.

These experiments show that a time-constrained similarity metric can be successfully expressed in terms of the masking phenomenon. The advantage of this is the ability to simultaneously deal with time and continuity constraints in a perceptive way.

6 Future Work

The use of intermediate time-frequency representations calculated using our method is an interesting approach to reduce the complexity of polyphonic sounds and provide a more high level description. By using perceptively motived similarity metrics, we aim at providing a representation that is easier to interpret and to interact with.

We have shown that the use of the masking phenomena can be succesfully considered to express the ”continuity illusion” phenomenon. Other aspects of temporal constraints should be investigated in the near future. For example, we would like to reduce the complexity of the algorithm by considering

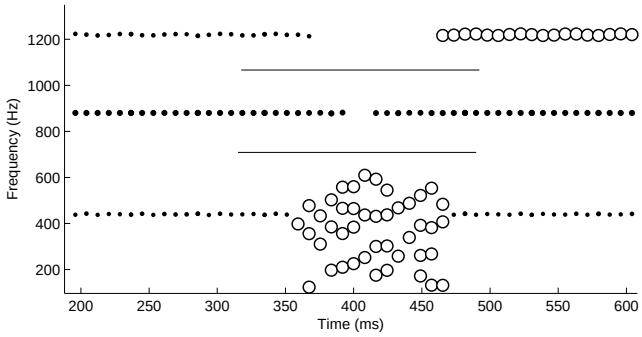

Figure 5: Results of the proposed algorithm for the three test cases using temporal and continuity constraints.

the computation of the affinity matrix for several fixed time intervals instead of considering the whole signal duration.

We believe that the use of such a front-end for more complex machine learning or inference algorithms may lead to methods of practical complexity for applications such as speech enhancement, bioacoustics (see Figure 2), and music information retrieval.

References

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

[2] 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.

[3] 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.

[4] 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.

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

[6] S.H. Srinivasan and M. Kankanhalli, "Harmonicity and dynamics based audio separation," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, 2003, vol. 5, pp. v–640 – v–643.

[7] S.H. Srinivasan, "Auditory blobs," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). 2004 IEEE International Conference on*, 2004, vol. 4, pp. iv–313 – iv–316.

[8] 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.

[9] E. Zwicker and H. Fastl, *Psychoacoustics Facts and Models*, Springer Verlag, 1990.

[10] G. Garcia and J. Pampin, "Data compression of sinusoidal modeling parameters based on psychoacoustic masking," in *International Computer Music Conference, 1999. Proceedings. (ICMC '99).*, 1999, pp. 40–43.

[11] M. Lagrange and S. Marchand, "Real-time additive synthesis of sound by taking advantage of psychoacoustics," in *Digital Audio Effects, 2001. Proceedings. (DAFx '01). 2001 Conference on*, 2001, pp. 5–9.