AUDITORY BLOBS

S H Srinivasan

Applied Research Group
Satyam Computer Services Ltd, Bangalore
SH_Srinivasan@satyam.com

ABSTRACT

Auditory scene analysis tries to segment an auditory signal (scene) into objects. Most of the intermediate representations currently proposed based on ASA are difficult to compute. In this paper, we propose auditory strands and blobs as intermediate representations. Auditory blobs are parts of an audio signal which have the same onset. By the principles of computational auditory scene analysis, they belong to the same object. In this paper we show how auditory blobs can be extracted and define harmonicity, dynamics, and onset features for auditory blobs. We also demonstrate their application to audio separation.

1. INTRODUCTION

Audio signals are notoriously difficult to analyze. The signal processing approaches to audio analysis are dominated by linear prediction and cepstrum analysis. For example, linear prediction is at the heart of speech coding and cepstral features are at the heart of speech recognition. Other tasks like genre classification also use these features to a great extent. A fundamental limitation of this approach is that these techniques are valid only when the audio signal consists of a single source. If the audio signal consists of multiple sources, the features do not have the intended interpretation.

Blind separation techniques of signal processing can be used for audio separation. Independent Components Analysis (ICA) tries to make the extracted sources as statistically independent as possible. ICA requires as many mixtures as the number of sources. ICA is a general purpose signal processing technique and auditory-specific constraints need to be incorporated for use in auditory separation. See, for example, [1].

Auditory scene analysis (ASA) [2] provides a refreshing alternative to dominant audio processing approaches. One major limitation of ASA models is that most of them use complex auditory filter banks [3] [4] [5]. Most auditory filter banks are nonlinear, complex, and difficult to invert [6]. Models which work in the familiar short-time Fourier transform (STFT) domain have wider applicability. [1] uses a subspace constraint along with independent components analysis in the STFT domain. We have recently proposed a model for audio separation based on ASA in STFT domain [7]. This model does not use onset information. In this paper we propose a model which is primarily based on onset information. The onset-based objects are called *blobs*.

This paper is organized as follows. Section 2 discusses the principles and models for ASA. Section 3 describes our technique for measuring onsets and defines intermediate auditory representations: *strands* and *blobs*. Blob features and blob similarity measures are described in section 4. Section 5 provides the experimental results for audio separation. Section 6 closes the paper with a discussion.

2. AUDITORY SCENE ANALYSIS

Just as a visual scene consists of objects, auditory scene also consists of "auditory objects". The principles of ASA [8] are[1]:

Regularity 1 *(Onsets & Offsets)*: Unrelated sounds seldom start or stop at exactly the same time.

Regularity 2 *(Dynamics)*: Gradualness of change: (a) A single sound tends to change its properties smoothly and slowly. (b) A sequence of sounds from the same source tends to change its properties slowly.

Regularity 3 *(Harmonicity)*: When a body vibrates with a repetitive period, its vibration give rise to an acoustic pattern in which the frequency components are multiples of a common fundamental.

Regularity 4 *(Dynamics)*: Many changes that take place in an acoustic event will affect all the components of the resulting sound in the same way and at the same time.

Just as a visual scene can be segmented based on color and texture, auditory scenes can be clustered based on onsets & offsets, dynamics, and harmonicity. (In this paper we address *monaural separation* - separation of a single audio mixture. Binaural separation uses two mixtures and has more cues that can be used: interaural time difference, in-

[1] Terms in parenthesis are ours.

teraural intensity difference. The availability of several mixtures also makes the problem amenable to the use of techniques like independent component analysis [9].)

The above four principles are the most important principles of ASA. More principles can be found in [10].

2.1. ASA models

There are several models based on these principles.

Cooke [3] uses a gammatone filter, sigmoid nonlinearity, and a hair cell model. The adaptation and recovery of nerve cell firing provide a convenient way to detect onsets and offsets.[2] Spectro-temporal grouping is based on fundamental frequency, amplitude modulation rate, etc. No onset information is used. The intermediate representation is called *synchrony strands*.

Brown and Cooke [4] use the gammatone filterbank and Meddis hair cell model. The intermediate representations are constructed for firing rate, frequency transition, onset/offset, etc in the form of maps. The onsets are detected using neural cells that receive initial excitation and subsequent inhibition. They construct symbolic representations called *auditory elements*. Auditory elements are formed using onset/offset information and periodicity.

Ellis [5] uses the a linear model of cochlea. The derived features are onset maps, correlogram, and periodogram representation. Grouping again is based on fundamental frequency. The intermediate representation is called *wefts* for periodic sounds. (The other representations are *noise clouds* and *noise clicks* for textures and transients.)

Our previous work [7] does not try to estimate fundamental frequency. Harmonicity and dynamic similarity measures are defined between spectral lines for each frame of STFT. The spectral lines are clustered using the normalized cut algorithm [11]. The clusters of adjacent frames are continued using maximal overlap. The formulation does not use the notion of onsets. In this paper we propose a formulation which uses onset as the primary cue. The representation using only onsets is called *blob*. We then define harmonicity and dynamics features of blobs.

3. ONSETS

Onsets are a dominant cue [10]. But detecting onsets is difficult. Though there are several models for onset detection, they can be classified into two: based on energy differences between successive frames [5] [12] [13] or neural-network based approaches [14]. Finding onsets based on energy differences is not reliable. Learning onsets using neural networks generalizes poorly in the presence of multiple sources. Also, as an aid for spectral grouping for separation, onsets need to be defined for *each spectral line* - a considerably more difficult task.

[2] A disadvantage of such models is the presence of a large number of parameters.

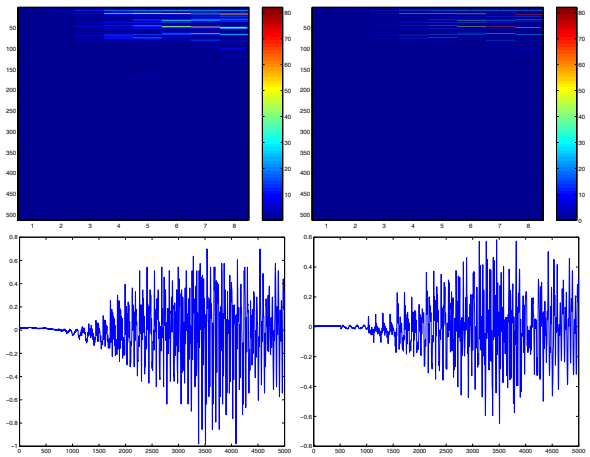

Fig. 1. Strands. The top row shows the spectrogram of the original signal and the spectrogram containing only the spectral peaks. Each "connected component" in this peak map is called a strand. The original signal and a reconstruction based only on strands is shown in the bottom row. The signal used is the male speech + female speech mixture ("v0n9.au" in the database) from the auditory separation database of Cooke (`http://www.dcs.shef.ac.uk/~martin/`). Only 5000 samples are shown.

The following discussion is motivated by the following observation. Spectral peaks are perceptually important. We identify and track spectral peaks. In fact, in the initial stages of processing only spectral peaks are used. Peaks are separated by valleys. The energy in the valleys is used in the final "assembly".

3.1. Strands & Blobs

We define a continuous evolution of a spectral peak as a *strand*. (The term *synchrony strands* has been used in [3] for the result of spectral grouping based on fundamental frequency; such strands may contain multiple spectral lines. Strands, as defined here, contain one spectral peak per frame. They may last several frames.) Thus strands have a well-defined onset and offset (start and end frames). Figure 1 shows the original speech signal and strand-based reconstruction.

We define an *auditory blob* as the collection of strands which have the same onset. Since all the strands in a blob have the same onset, they are grouped as belonging to the same source. Auditory blobs of the basic "events" which take place in the auditory stream. Higher level events can be obtained by grouping blobs. Grouping blobs instead of grouping spectral lines or strands is also more efficient and

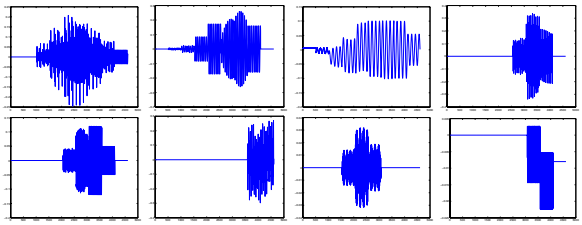

Fig. 2. Blobs. The blobs are constructed from the strands shown in figure 1. Sometimes, blobs correspond to perceptible auditory objects. The first blob, for example, accounts for major part of one source (male voice). Most of the remaining blobs constitute the second source (female voice).

reliable. It is more efficient since there are fewer blobs than spectral lines or strands. It is more reliable since blobs carry more source-specific cues than individual spectral lines or strands. Blobs, as defined here, are extremely easy to extract. Figure 2 shows the blobs corresponding to the mixture signal of figure 1.

4. BLOB FEATURES AND SIMILARITY

To use blobs for classification and separation tasks, we need to define features for blobs. We define the frequency, harmonicity, and dynamics features of blobs (F, H, D_e, and D_f) as follows. In the following discussion, we use the terms "frequency" and "frequency index" interchangeably. Frequency indexes are integers.

Frequency content: A blob is a collection of points in time-frequency plane defined by STFT (with certain continuity and onset constraints). We can denote a blob as $(t_1, f_{11}, f_{12}, \cdots, f_{1n_1}), (t_2, f_{21}, f_{22}, \cdots, f_{2n_2}), \cdots (t_m, f_{m1}, f_{m2}, \cdots, f_{mn_m})$. Here $t_1, t_2, \cdots, t_m$ are the frame indexes and f_{ij} is the jth frequency for frame t_i. The collection $f_{11}, f_{12}, \cdots, f_{1n_1}, \cdots, f_{21}, f_{22}, \cdots, f_{2n_2}, \cdots, f_{m1}, f_{m2}, \cdots, f_{mn_m}$ describes the frequency content of the blob. The histogram of these frequencies defines the feature F.

Harmonicity: A fundamental feature of a blob is its harmonicity which can be defined as the underlying fundamental frequency of the blob. To calculate the fundamental frequency, we cannot mix frequencies at different frame indexes. (Frequencies at different frame indexes were combined for the feature F above.) We define the difference histogram as a measure of the harmonicity of the blob. Consider the differences: $f_{12}-f_{11}, f_{13}-f_{12}, \cdots, f_{1n_1}-f_{1,n_1-1}$, $f_{22}-f_{21}, f_{23}-f_{22}, \cdots, f_{2n_2}-f_{2,n_2-1}, \cdots f_{m2}-f_{m1}, f_{m3}-f_{m2}, \cdots, f_{mn_m} - f_{m,n_m-1}$. We define the histogram of these differences to be a measure of harmonicity. In the example above, the differences are: 5, 5, 7, 16. This accuracy of this representation increases as the number of time-frequency points increases. This feature, H, is also able to capture time varying periodicities.

Energy Dynamics: A blob has multiple strands (of possibly different lengths). The sum of the energies of these strands defines the energy dynamics, D_e, of the blob. If we use the above notation, the vector of sums $[\sum_{j=1}^{n_1} E(t_1, f_{1j}), \sum_{j=1}^{n_2} E(t_2, f_{2j}), \cdots, \sum_{j=1}^{n_m} E(t_m, f_{mj})]$ defines the energy dynamics where $E(t, f)$ denotes the energy of frequency index f at time t.

Frequency dynamics The frequency dynamics, D_f, is defined as the vector of weighted sums $[\sum_{j=1}^{n_1} f_{1j}E(t_1, f_{1j}), \sum_{j=1}^{n_2} f_{2j}E(t_2, f_{2j}), \cdots, \sum_{j=1}^{n_m} f_{mj}E(t_m, f_{mj})]$

This vector is component-wise normalized by the energy vector, D_e, computed above.

4.1. Blob similarity

We can define similarities of two blobs, (F^1, H^1, D_e^1, D_f^1) and (F^2, H^2, D_e^2, D_f^2) as the similarities of their features. In particular, we define the following similarities.

Frequency similarity (S_F): This is the intersection of the histograms F^1 and F^2.

Harmonic similarity (S_H): This is the intersection of the histograms H^1 and H^2.

Dynamic similarity (S_D): The blobs may have different onsets and offsets. The common temporal interval for both the blobs is computed. The cosine similarity between the subvectors corresponding to this common interval is taken as the dynamic similarity. This is computed for both D_e and D_f. We require that the size of the common interval be at least two for this similarity to be computed. Otherwise, the similarity measure is set to zero.

The cumulative similarity is calculated as $e^{\frac{S_F}{\sigma_F}} \times e^{\frac{S_H}{\sigma_H}} \times e^{\frac{S_D}{\sigma_D}}$ where the σs are normalizing constants.

That the above measure takes onsets and offsets indirectly (through S_D). A direct way to measure onset similarity is to use the following: $e^{\frac{|t_1-t_2|}{\tau}}$ where t_1 and t_2 are the onset times of the two blobs and τ is a constant.

5. EXPERIMENTS

The term "blob" is inspired by "blobworld" representation of images [15]. Image segmentation is a difficult problem. "Visual Blobs" of [15] are regions which correspond to objects or parts of objects. The motivation for auditory blobs is the same: auditory blobs correspond to auditory objects or parts of objects. We expect auditory blobs to find applications in audio retrieval, coding, etc. As an example, we consider audio separation in this paper.

Auditory blobs need to be grouped into objects. As mentioned in section 2, auditory sources obey harmonicity, dynamics, and onset constraints. We use the similarity

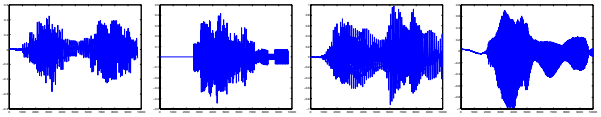

Fig. 3. Blob-based separation. The first two figures shows the separated and reconstructed source signals. The next two show the original signals. The blobs of figure 2 were clustered into two clusters using the normalized cut segmentation algorithm using the similarity measures defined in section 4. Only the spectral peaks were used in inversion. It is possible to improve the quality of this "skeletal reconstruction" by "filling in" using the spectral lines in the valleys between the peaks.

measure defined in section 4.1 to construct the blob similarity matrix and subject the similarity matrix to normalized cuts segmentation [11]. The results are shown in figure 3.

6. DISCUSSION

In this paper we have used ideas from auditory scene analysis (peaks and onset times) to propose a new representation scheme for auditory blobs. We have defined a set of features for auditory blobs - again inspired by ASA. We have demonstrated the use of auditory blobs and their features in the task of speech separation. Our formulation makes auditory scene analysis very close to visual segmentation not only in spirit but also in details. More experimental support is needed to show that the formulation is robust.

It is instructive to compare blobs as intermediate representations with those mentioned in section 2.1.

1 All the models mentioned in section 2.1 use the fundamental frequency as the major cue. Fundamental frequency is computed using autocorrelation techniques. Our own view is that finding fundamental frequency of blobs is more reliable than finding the fundamental frequency of ungrouped mixture. As mentioned in section 4, the collection of strands that constitute the blob carry fundamental frequency information in a more readily extractable form since they belong to the same object. We have used difference histograms to characterize harmonicity. Autocorrelation techniques can also be used.

2 All the other intermediate representations are obtained after substantial processing. Blob can be obtained more easily. There are very few parameters. This enables them to be usable in other applications.

3 The other representations are based on auditory filter banks. STFT-based blobs have wider applicability.

Blobs do not account for all audio signals. Audio textures do not show persistent evolution of spectral peaks. Hence is is difficult to use blob-based representations for auditory textures. Other grouping principles are needed to identify audio textures.

7. REFERENCES

[1] M A Casey and A Westner, "Separation of mixed audio sources by independent subspace analysis," in *ICMC*, 2000.

[2] A S Bregman, *Auditory scene analysis: The perceptual organization of sound*, MIT Press, 1990.

[3] M Cooke, *Modelling auditory processing and organisation*, Cambridge University Press, 1993.

[4] G J Brown and M Cooke, "Computational auditory scene analysis," *Computer Speech and Language*, 1994.

[5] D P W Ellis, *Prediction-driven computational auditory scene analysis*, Ph.D. thesis, MIT, 1996.

[6] M Slaney, D Naar, and R F Lyon, "Auditory model inversion for sound separation," in *ICASSP*, 1994.

[7] S H Srinivasan and M Kankanhalli, "Harmonicity and dynamics based audio separation," in *ICASSP*, 2003.

[8] A S Bregman, "Auditory scene analysis: hearing in complex environments," in *Thinking in sound: The cognitive psychology of human audition*, S McAdams and E Bigand, Eds., pp. 10–36. Clarendon Press, 1992.

[9] A J W van der Kouwe, D L Wang, and G J Brown, "A comparison of auditory and blind separation techniques for speech segregation," *IEEE SAP*, 2001.

[10] D K Mellinger and B M Mont-Reynaud, "Scene analysis," in *Auditory Computation*, H L Hawkins et al, Ed. Springer, 1996.

[11] J Shi and J Malik, "Normalized cuts and image segmentation," *IEEE PAMI*, 2000.

[12] A Klapuri, "Sound onset detection by applying psychoacoustic knowledge," in *ICASSP*, 1999.

[13] C Duxbury, M Sandler, and M Davies, "A hybrid approach to musical note onset detection," in *DAFX*, 2002.

[14] M Marolt, A Kavcic, and M Privosnik, "Neural networks for note onset detection in piano music," in *ICMC*, 2002.

[15] C Carson, S Belongie, H Greenspan, and J Malik, "Blobworld: Image segmentation using expectation-maximization and its application to image querying," *IEEE PAMI*, 2002.