HARMONICITY AND DYNAMICS BASED AUDIO SEPARATION

S H Srinivasan[*] *Mohan Kankanhalli*

Department of Computer Science, National University of Singapore, Singapore
dcsshs@nus.edu.sg mohan@comp.nus.edu.sg

ABSTRACT

Audio signal source separation is an interesting task performed by humans. In this paper, we present a frequency grouping algorithm based on principles of harmonicity and dynamics: frequency components with a harmonic relation and similar dynamics belong to the same source. The grouping is demonstrated for a variety of sound mixtures.

1. INTRODUCTION

Human beings can separate audio inputs consisting of multiple sources. This is informally called "cocktail party effect": a listener can selectively track a signal originating from a given speaker in the presence of multiple interfering signals. It is still not clear how the human ear/brain achieves this. Sound source separation has several applications (other than cocktail parties). For example, speaker separation is an essential for robust speech recognition. Source separation also has several applications in multimedia indexing and retrieval. And finally, source separation is an indispensable part of model-based audio coding.

There have been several approaches to sound source separation. Psychoacoustic-based approach is called *auditory scene analysis* while the signal processing community uses the term *polyphonic separation*. The term "auditory scene" takes its inspiration from the term "visual scene". Just as a visual scene can be segmented into objects, auditory scene analysis attempts to segment the auditory signal into coherent *streams* [1]. The guiding principles of ASA [2] are[1]

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. Regularities 1 and 2 pertain to attack & decay and gradualness of change.

While these are guiding principles of auditory analysis, "computational auditory scene analysis" attempts to elucidate computational mechanisms for the same [3]. While it is impossible to summarize the research on CASA here, few major trends are readily identifiable. (See [4] for a recent review.) CASA is usually performed in two stages. In the first stage the signal is decomposed into several fragments or parts. This processing usually takes place in a time-frequency domain (like correlogram, Wigner–Ville transform, etc.). The second stage assembles the fragments belonging to common source signals. Different types of organizing principles can be used in the second stage. The popular choices for the second stage are:

Blackboard architectures: Blackboard is a framework for integrating diverse knowledge sources and data. The knowledge is typically provided in the form of hand-crafted rules. A recent example of use of blackboard architecture is in *prediction driven CASA* [5].

Model-based approaches: Model-based approaches contain models of acoustic data – usually in the form of Bayesian belief networks. While prior knowledge is encoded as rules in blackboard systems, parametric (e.g, probabilistic) knowledge representation is used in model-based systems. These parameters are usually learned. An example this approach can be found in [6].

While the above approaches are inspired by psychoacoustic research, signal separation is of general interest to the signal processing community. In the "blind signal separation" approaches, there are as many observed signals as sources. Independent Components Analysis [7] attempts to make the separated components as statistically independent as possible. Audio signals obey several constraints as listed above. In independent subspace analysis [8], a subspace constraint is imposed on statistically independent source signals.

In this paper, we revisit the original grouping principles and propose a simple model. While most of the other proposals are based on "sequential" – finding the dominant

[*] On leave from Applied Research Group, Satyam Computer Services, India.

[1] The terms in italics are our insertion.

component first and then finding related components – our formulation is more emergent. The resulting system exhibits good performance. Section 2 describes the model. This is followed by a presentation of experimental results (section 3). The last section (section 4) provides a critical evaluation of the results.

2. THE MODEL

We perform signal separation in the frequency domain using short-time Fourier transform (STFT). We view the separation problem as allocation of frequencies to different sources. The allocation is performed using *spectral magnitudes*. The frequency allocation is projected back to STFT which is then inverted to get the source signals (figure 1). In this paper, we use $s(t,i)$ to denote the short-time Fourier transform of the signal: t is the time index and i is the index of the spectral line. $|s(t,\cdot)|$ is the magnitude spectrum at time t.

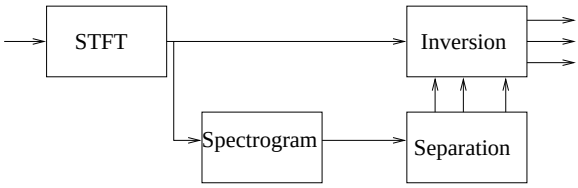

Fig. 1. Separation architecture

The harmonicity principle is modeled using a hmap (short for *harmonic map*). The hmap, $h(i,j)$, provides *harmonic similarity* between are spectral lines i and j.

$$h(i,j) = 1, \text{ if } i|j \text{ or } j|i$$

($a|b$ denotes "a divides b".) Since real-world spectrograms are noisy, we need to smooth the above map. Smoothing is performed by using a Gaussian instead of an impulse in the map. Smoothing is also frequency dependent: the variance of the Gaussian is made proportional to the center frequency. The smoothed harmonic similarity between spectral lines $i+\delta i$ and $j+\delta j$ due to the harmonic similarity between lines i and j is given by

$$h(i,j) \times \mathcal{N}(i, pi; \delta i) \times \mathcal{N}(j, pj; \delta j)$$

where $\mathcal{N}(\mu, \sigma; x)$ is the Gaussian function with parameters μ and σ and p is a constant. (In the experiments discussed below, the standard deviation of the Gaussian is 10% of the center frequency.) The smoothed harmonic map, $sh(i,j)$, is obtained by summing all the contributions of the neighboring spectral lines at (i,j). The map is normalized so that

$$\sum_j sh(i,j) = 1$$

Spectral components belonging to a given source obey the same temporal dynamics (temporal envelope, for example). The dynamic similarity is measured by, dmap (dynamics map). Let w be the size of temporal window. The vector $D(t,i)$, as given by,

$$|s(t-w,i)|, |s(t-w+1,i)|, \cdots |s(t,i)|, \cdots |s(t+w,i)| \tag{1}$$

can be considered as a measure of local evolution of spectral line i. The dynamic similarity between spectral lines i and j is given by the cosine of the angle between the vectors $D(t,i)$ and $D(t,j)$. Thus, $d_t(i,j)$, the similarity due to dynamics between lines i and j at time t is given by

$$d_t(i,j) = \frac{D(t,i).D(t,j)}{|D(t,i)||D(t,j)|}$$

The combined similarity between lines i and j at time t is defined as

$$S_t(i,j) = |s(t,i)|sh(i,j) + \alpha d_t(i,j) \tag{2}$$

where α is a parameter.

We can now subject the similarity matrix to clustering. In our experiments, we have used the clustering algorithm based on normalized cuts [9].

Frequencies are clustered at each frame of STFT. We need a mechanism for relating frequency clusters of adjacent frames. Let $C_{t,i}$ denote the set of frequencies in cluster i at time t. Given the clusters $\{C_{t,i}\}$ and $\{C_{t+1,j}\}$, we need to calculate the correspondence between clusters since cluster labels are arbitrary. Since the frequency content of sources does not change rapidly, we can use maximal intersection to get the correspondence. For each cluster i at time t, we calculate $C_{t,i} \cap C_{t+1,j}, \forall j$. The cluster at $t+1$ which has maximal intersection continues the cluster $C_{t,i}$. We order the intersections in decreasing order of energy.

3. EXPERIMENTS

We used the sound mixtures described in [10][2]. The database consists of a mix of 10 voiced sources and 10 intrusive sources for a total of 100 audio samples. The sampling frequency is 16 kHz. The voiced utterances consist of five sentences spoken by two male speakers. The table 1 (from [10]) describes the intrusion signals.

We use the following parameters in our experiments. The STFT is performed for every 256 samples (16 ms) with a overlap of 128 samples (8 ms). We use Hanning window as it results in better STFT inversion. For calculating dynamics-based similarity, we use $w = 3$ (equation 1). We use $\alpha = 0.1$ in equation 2.

[2]The samples are available from `http://www.dcs.shef.ac.uk/~martin/`

id	description	characterization
n0	1 kHz tone	NB, C, S
n1	white noise	WB, C, US
n2	series of brief noise bursts	WB, I, US
n3	teaching laboratory noise	WB, C, partly S
n4	new wavw music	WB, C, S
n5	FM signal (“siren”)	locally NB, C, S
n6	telephone	WB, I, S
n7	female TIMIT utterance	WB, C, S
n8	male TIMIT utterance	WB, C, S
n9	female utterance	WB, C, S

Table 1. Description of intrusion signals. NB: narrowband, WB: wideband, C: continuous, I: interrupted, S: structured, US: unstructured

To evaluate the performance of the separation algorithm, we use the cosine similarity between the extracted signals and the known original sources. For each noise type, we calculate the minimum and maximum cosine similarity. Figure 2 shows the results.

The separated signals are available at the URL `http://www.comp.nus.edu.sg/~sengam/icassp2003/`. Here we give a visual indication of the quality of separation. Figure 3 shows the original and recovered speech signals. Figure 4 shows the original and recovered intrusion signals.

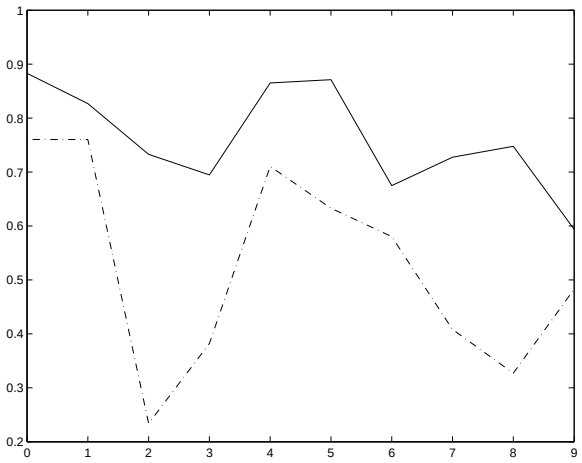

Fig. 2. Average maximum (solid line) and minimum (broken line) cosine similarity for each noise type. The X-axis shows the noise type. For an explanation of noise types, see table 1. The scores are obtained as follows. The cosine similarities of both extracted signals with both sources (speech and intrusion) are calculated. The maximum of these gives the maximum similarity. We now exclude the source and the recovered signal for which the similarity is maximum. The cosine similarity of the other extracted signal with the other source gives the minimum similarity. These scores are averaged for a given intrusion type and plotted.

4. DISCUSSION

It can be seen from figures 3 and 4 that the recovered signals capture the variations in the source signals well.[3] It will be best to compare our results with those of [10] which also used the same mixtures. Unfortunately [10] uses a time–frequency representation called *synchrony strands* and the results are reported in terms of the presence or absence of strands in the separated sources. So direct comparison is impossible. Let us consider one performance metric used: “$\cdots$ the model is able to group between 67% and 78% of the speech components, depending on the type of intrusive source” [10]. While we have a similar lower bound in our case (except for intrusive signal “n9”), our upper bound is higher. [10] also reports the metric when separating a single source signal with no intrusion at “$\cdots$ around four-fifths $\cdots$” which is an upper bound on the system’s performance. The figure in our case is 81%. This is not an upper bound on our models performance since we currently try find two sources in *any* input signal. This figure is of interest as an independent performance measure. Of course, we use cosine similarity.

[3]Caveat:The ears are more sensitive than eyes. Time-frequency domain plots of audio signals can hide distortions.

When we analyze figure 2, we find that the performance is poor for those intrusions which are interrupted. This is due to the fact that our clustering finds two sources for all input mixtures. The model needs to be modified by explicitly using start and stop times to handle such cases.

Our formulation is plausible in a neurobiological sense since sequential search is minimized. The clustering can be performed using competitive learning, for example. The prevalent approach in the field has been finding dominant frequency and then searching for related frequencies. It should be noted that the ear/brain may not have a need to explicitly *invert* the time-frequency representation for which we use the not-too-realistic STFT.

In our formulation, the transform and inversion are simple since they are based on the familiar Fourier transform. In all the separation schemes we are aware of: (1) The transforms used are computationally more complex. (2) Both separation and inversion are performed on the same representation leading to complex inversion schemes (see, for example, [11]).

5. REFERENCES

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

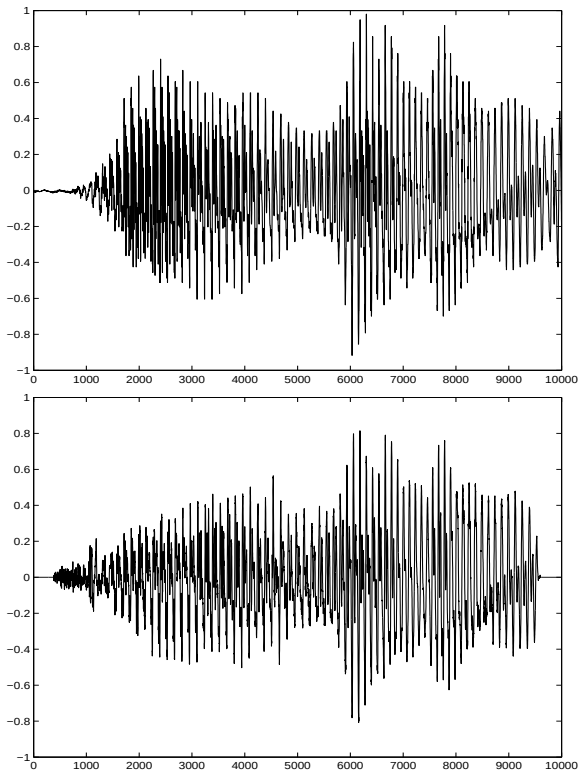

Fig. 3. Original and recovered speech signals. The audio sample used in this experiment is "v0n4". The recovered signal is zero for a few frames at the beginning and end due to temporal windowing for dmap calculation.

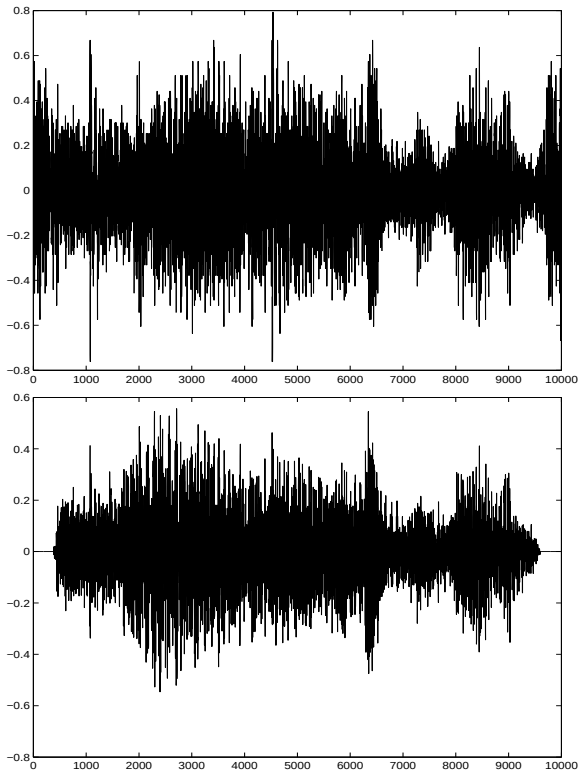

Fig. 4. Original and recovered intrusion signals for the mixture "v0n4". The recovered signal is zero for a few frames at the beginning and end due to temporal windowing for dmap calculation.

[2] A S Bergman, "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.

[3] D F Rosenthal and H G Okuno, Eds., *Computational auditory scene analysis*, Lawrence Erlbaum Associates, 1998.

[4] M Cooke and D P W Ellis, "The auditory organization of speech and other sources in listeners and computational models," *Speech Communication*, vol. 35, pp. 141–177, 2001.

[5] D P W Ellis, *Prediction-driven computational auditory scene analysis*, Ph.D. thesis, Massachusetts Institute of Technology, 1996, http://web.media.mit.edu/∼dpwe/pdcasa/.

[6] K Kashino, K Nakadai, T Kinoshita, and H Tanaka, "Applications of the Bayesian probability network to music scene analysis," in *Computational auditory scene analysis*, David F Rosenthal and Hiroshi G Okuno, Eds., pp. 115–137. Lawrence Erlbaum Associates, 1998.

[7] A Hyvarinen, J Karhunen, and E Oja, Eds., *Independent component analysis*, John Wiley, 2001.

[8] M A Casey and A Westner, "Separation of mixed audio sources by independent subspace analysis," in *International Computer Music Conference*, 2000.

[9] J Shi and J Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

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

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