Monaural Source Separation using Spectral Cues

Barak A. Pearlmutter[1] and Anthony M. Zador[2]

[1] Hamilton Institute, National University of Ireland Maynooth, Co. Kildare, Ireland
[2] Cold Spring Harbor Laboratory, One Bungtown Rd, Cold Spring Harbor, NY 11724, USA

Abstract. The acoustic environment poses at least two important challenges. First, animals must localise sound sources using a variety of binaural and monaural cues; and second they must separate sources into distinct auditory streams (the "cocktail party problem"). Binaural cues include intra-aural intensity and phase disparity. The primary monaural cue is the spectral filtering introduced by the head and pinnae via the head-related transfer function (HRTF), which imposes different linear filters upon sources arising at different spatial locations.
Here we address the second challenge, source separation. We propose an algorithm for exploiting the monaural HRTF to separate spatially localised acoustic sources in a noisy environment. We assume that each source has a unique position in space, and is therefore subject to preprocessing by a different linear filter. We also assume prior knowledge of weak statistical regularities present in the sources. This framework can incorporate various aspects of acoustic transfer functions (echos, delays, multiple sensors, frequency-dependent attenuation) in a uniform fashion, treating them as cues for, rather than obstacles to, separation. To accomplish this, sources are represented sparsely in an overcomplete basis. This framework can be extended to make predictions about the neural representations required to separate acoustic sources.

1 Introduction

Organisms exploit a variety of binaural and monoaural cues to separate acoustic sources, a process sometimes referred to as "stream segregation" [1]. One set of cues that can be used to separate sources is the differential filtering imposed by the head and pinnae (the head-related transfer function, or HRTF) on sources at different positions in space [2]. It is often reasonable to assume that sound arriving from different locations should be treated as arising from distinct sources. While the importance of the HRTF in sound localisation has been studied extensively, its role in source separation *per se* has not received as much scrutiny.

Let us consider a formulation of source separation that includes the HRTF. Suppose there are N acoustic sources $x_i(t)$ located at distinct positions in space. Associated with each position is a distinct spectral filter, given by the corresponding head-related transfer functions $h_i(t)$. The received signal $y(t)$ is then the sum of the filtered signals

$$y(t) = \sum_{i=1}^{N} h_i(t) * x_i(t) \tag{1}$$

where $*$ indicates convolution. Our goal is to recover the underlying sources $x_i(t)$ from the observed signal $y(t)$, using knowledge of the directional filters[3] $h_i(t)$. Although the HRTF can also be exploited in multi-sensor situations, in the present work we focus only on the more difficult single-sensor case.

2 Monoaural separation using a weak prior

We solve this underdetermined system in a sparse separation framework, with L_1-norm optimisation as a sparseness measure [3–7]. The two-sensor underdetermined case has been addressed in this context [8, 9] but separating multiple sources from a single sensor is harder and requires stronger assumptions [10–13]. In this framework, we model the i-th source $x_i(t)$ as a weighted sum of elements $d_j(t)$ from an overcomplete dictionary,

$$x_i(t) = \sum_j c_{ij}\, d_j(t), \tag{2}$$

where the weighting associated with dictionary element $d_j(t)$'s contribution to source i is c_{ij}, and the c_{ij} are assumed to be sparse.

In particular, the signals in the dictionary, $d_j(t)$, are chosen with two criteria in mind. First, sources should be sparse when represented in this dictionary, meaning that the coefficients c_{ij} required to represent $x_i(t)$ will have a distribution with more zeros (and more large values) than might be naively expected. A common formalisation of this assumption is that the distribution of coefficients is governed by a Laplacian distribution ($p(c_i) \propto e^{-|c_i|}$); a Laplacian distribution has more elements close to zero (and far from zero) than does a Gaussian with the same variance. Second, dictionary elements should be chosen such that, following transformation by the HRTF, elements differ as much as possible; this is equivalent to minimising the condition number of the matrix $\mathbf{D}$ introduced below.

In what follows, we assume that each source appears at a unique position in space, and that there is only a single source at each position. The components $d_j(t)$ of each source might thus be subject to filtering by any of the HRTFs $h_i(t)$. We therefore construct a new dictionary by applying each possible filter to each original element. We denote the resulting dictionary elements

$$d'_{ij}(t) = h_i(t) * d_j(t). \tag{3}$$

Note that the number of elements in the new d' dictionary is equal to the number of original dictionary elements times the number of sources N; the original overcomplete basis has now become "more overcomplete" by the factor N.

The source separation problem can now be cast as decomposing $y(t)$ into this overcomplete dictionary by finding appropriate c_{ij} for

$$y(t) = \sum_{ij} c_{ij}\, d'_{ij}(t). \tag{4}$$

[3] The filter terms $h_i(t)$ may be interpreted to include not just the filtering of the head and pinnae, but also the filter function of the acoustic environment, and the audiogram of the ear itself.

Once the coefficients c_{ij} are known, the individual sources can be reconstructed directly from the unfiltered elements $d_j(t)$ using Eq. 2.

Source separation thus requires estimating the coefficients c_{ij}. Let us define $\mathbf{c}$ as a single column vector containing all the coefficients c_{ij}, with the elements indexed by i, j, and $\mathbf{D}$ as a matrix whose k-th row holds the elements $d'_{ij}(t_k)$. The columns of $\mathbf{D}$ are indexed by i and j, and the rows are indexed by k. Finally, let $\mathbf{y}$ be a column vector whose elements correspond to the discrete-time sampled elements $y(t)$. Thus $\mathbf{y} = \mathbf{Dc}$.

If the dictionary $d'_{ij}(t)$ formed a complete basis, $\mathbf{c}$ would be given by $\mathbf{c} = \mathbf{D}^{-1}\mathbf{y}$. However, by assumption the system is now underdetermined—many possible combinations of sources yield the observed sensor data $y(t)$—so in order to specify a unique solution we must have a way of choosing among them. We therefore introduce a regulariser that incorporates some weak prior information about the problem and renders it well-posed [14]. Here we express the regulariser in terms of an easily stated condition on the norm of the solution vector $\mathbf{c}$: Find the $\mathbf{c}$ that minimises the L_p norm $\|\mathbf{c}\|_p$ subject to $\mathbf{Dc} = \mathbf{y}$, where $\|\mathbf{c}\|_p = (\sum_{ij} |c_{ij}|^p)^{\frac{1}{p}}$.

Different choices for p correspond to different priors and so yield different solutions $\mathbf{c}$. A natural choice would seem to be $p = 2$, which corresponds to assuming that the source coefficients c_{ij} were drawn from a Gaussian distribution; this is the solution found by the the pseudo-inverse $\mathbf{c} = \mathbf{D}^*\mathbf{y}$. However, this choice does not exploit the sparseness assumption about the sources; rather, it seeks a solution in which the power is spread across the sources (Figure 1). With $p = 0$ ($\|\mathbf{c}\|_0$ is the number of nonzero elements of $\mathbf{c}$) we would exploit sparseness, but this can be a computationally intractable combinatorial problem, and moreover the solution would not be continuous in $\mathbf{y}$ and therefore not be robust to noise [15].

Instead, as shown in Figure 1, we use $p = 1$ (the L_1-norm), which is equivalent to a Laplacian prior on the coefficients $\mathbf{c}$. That is, we solve

$$\text{minimise} \sum_{ij} |c_{ij}| \text{ subject to } \mathbf{y} = \mathbf{Dc} \tag{5}$$

This has a single global optimum which can be found efficiently using linear programming [3], and is continuous in $\mathbf{y}$.

This algorithm can be sensitive to sensor and background noise, as it insists on precisely accounting for the measured signal using some combination of dictionary elements, which can generate large artefacts. However, we can generalise the optimisation problem to include a noise process (simulations not shown) by changing the goal to

$$\text{minimise } \|\mathbf{c}\|_1 \text{ subject to } \|\mathbf{Dc} - \mathbf{y}\|_p \leq \beta \tag{6}$$

where β is proportional to the noise level and $p = 1$, 2, or ∞. The Gaussian noise case, $p = 2$, which can also be formulated as unconstrained minimisation, can be solved by Semidefinite Programming [16], or mixed L_1+L_2 optimisation methods used in control theory. Unfortunately these are too computationally burdensome for our purposes. Both $p = 1$ and $p = \infty$ can be solved using linear programming. All these are qualitatively similar, and in them all as $\beta \to 0$ the noise is assumed to be very small, and the solutions converge to that of the zero-noise solution, Eq. 5.

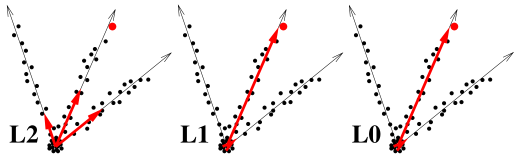

Fig. 1. Minimising the L_1-norm can provide a good and computationally tractable solution to data generated by a sparse prior. In this example, three non-orthogonal basis vectors (*black arrows*) are assumed, and each data point (*black points*) is generated by assuming only one non-zero coefficient c_i (the sparseness assumption), along with a small amount of noise. Since there are three basis vectors in two dimensions, there are many possible solutions, and additional constraints are required to specify the solution. The *red vectors* illustrate solutions found for the *red point* under three different constraints. (*Right*) Minimising the L_0-norm of $\mathbf{c}$ finds the sparse solution, but is computationally intractable (NP-complete). (*Left*) The L_2-norm can be efficiently minimised by the pseudoinverse, but yields a poor solution because it spreads the power across multiple basis vectors, in violation of the sparseness assumption. (*Centre*) The minimum L_1-norm solution can be found efficiently using linear programming, and under suitable assumptions finds a good approximation of the sparse solution.

Example: Harmonic comb prior We illustrate the algorithm with a simple example. Suppose that the sources can be modelled as simple "musical instruments" playing notes drawn from a 12-tone (Western) scale. Sources are defined by position—there is by definition only a single source at a given position—but each source may play more than one note simultaneously. Each note consists of a "harmonic comb"—a fundamental frequency F and its harmonics nF, $n = 2, 3, \ldots$, with amplitudes $1/n$. Each dictionary element, then, is given by

$$d_i = \sum_{n=1} \frac{1}{n} \sin(2\pi\, nF_i\, t). \tag{7}$$

where $F_i = 2^{i/12} F_0$ is the fundamental frequency of the i-th note in the equal-tempered scale, and F_0 is the frequency of the lowest note.

Figure 2 shows that such harmonic comb sources can be readily separated using knowledge of the spectral filtering, provided that one searches for a sparse solution vector $\mathbf{c}$ by minimising its L_1-norm. In this example three sources were assumed, each playing two "notes" selected from 72. Thus each source is fully described by the values of the two non-zero coefficients.

The top graph of Figure 2 shows the difference between L_1- and L_2-norm minimisation, in the absence of spectral filtering. The L_2-norm solution fits the received signal $y(t)$ using coefficients c_{ij} distributed in a roughly Gaussian fashion, whereas the L_1-norm solution found by linear programming finds a sparse solution in which the only non-zero dictionary coefficients correspond to notes actually present in at least one of the sources. However in the absence of the HRTF, even the L_1-norm solution has no way to assign the notes to the appropriate sources, so it assumes that an equal fraction of each note arises from each source. L_1-norm optimisation thus finds a more interpretable solution than L_2-norm optimisation even without an HRTF, but due to lack of any suitable cues it is equally unable to correctly separate the sources (see Table 1).

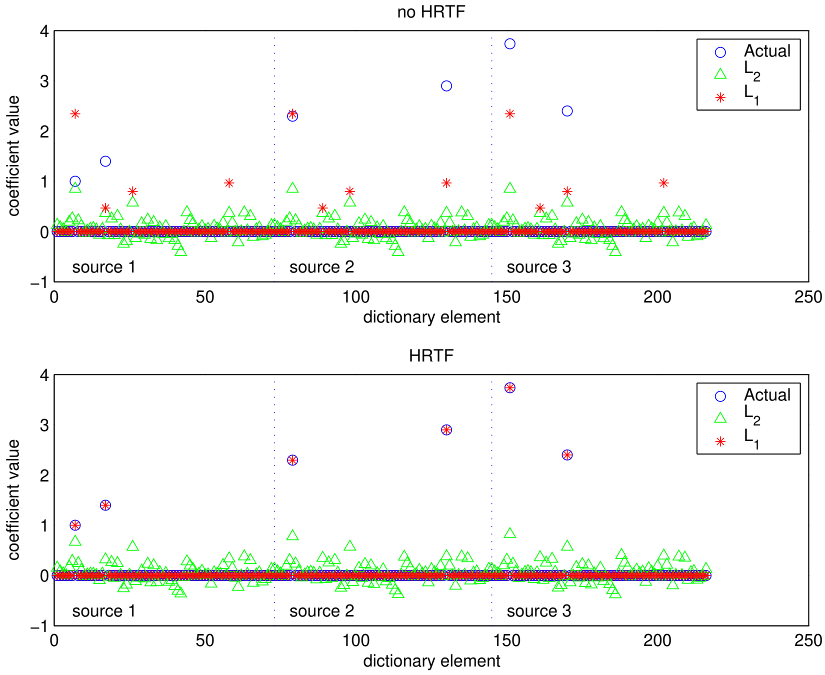

Fig. 2. Spectral cues can be exploited by assuming a sparse prior. The input to the microphone consisted of the sum of three sources (x-axis), each playing two notes but with different amplitudes (y-axis). **(Top)** If no spectral filtering is applied, the algorithm minimising the L_1-norm of the solution **c** accounts for the signal using a small number of coefficients, but cannot assign the correct amplitude to each source. It therefore assumes equal weight among the sources. By contrast, minimising the L_2-norm spreads the energy across many dictionary elements, leading to an uninterpretable solution. **(Bottom)** When a different spectral filter is applied to each source L_1-norm minimisation finds the exact solution, while minimising the L_2-norm yields a solution that remains both uninterpretable and unseparated.

The lower graph of Figure 2 shows how the spectral filtering due to the HRTF can enhance separation. In this case, the L_1-norm constraint is able to separate the sources almost perfectly, while the L_2-norm solution remains poor (see Table 1.) This example, although highly idealised, is intended to capture key features of many real-world problems in which sources have characteristic spectrotemporal signatures. In this framework, more sophisticated models of spectrotemporal structure can be readily accommodated by adding dictionary elements.

3 Discussion

We have described an algorithm for using the head-related transfer function to improve the separation of acoustic sources at different spatial locations. We show how, in certain special cases, the added cues provided by the HRTF permit otherwise unseparable

Table 1. SNR in dB of sources recovered using the proposed algorithm, in a synthetic acoustic environment with versus without an HRTF. Large positive numbers indicate better performance; the best performance is achieved by the algorithm that exploits the HRTF and minimises the L_1-norm of the solution.

norm	SNR without HRTF	SNR with HRTF
L_1	1.78	106.69
L_2	-4.86	-5.19

sources to be separated. We also show how, in the more general case, the cues can be used to improve separation.

The novel contribution of this work is a specific proposal for how the HRTF can be used for source separation, a process related to but distinct from localisation. It has long been known that the HRTF provides important cues for localisation [17–20]. Acoustic sources that bypass the HRTF (*e.g.* those presented with headphones) are typically perceived inside the head, unlike real sounds which are perceived outside the head [20, 21]. The HRTF is not, however, strictly required for localisation; under some conditions, binaural cues are sufficient to localise sounds even in the absence of the HRTF. Conversely, source separation can occur even without spatial cues, for example when selecting out the individual instruments of a concerto presented over a single speaker. Nevertheless, it is clear that the HRTF cues, when present, help in source separation [2].

The present formulation can be readily extended to include binaural information. Each HRTF function is made single-input two-output, and the lengths of the column vectors corresponding to the post-HRTF dictionary elements d'_{ij} and the data vector $\mathbf{y}$ are doubled. In this way, intra-aural time and level disparity can be used to separate sources. Information from two (or more) sensors can thus be naturally incorporated into the present framework. Similarly, although presented here as a batch algorithm, an online variant which gradually estimates coefficients as the signal becomes available would be straightforward to develop.

3.1 Assumptions about the HRTF

One of the main limitations of the present algorithm is that it requires that the precise HRTF $h_i(t)$ associated with each source be known. This requires knowing both the dependence of HRTF on spatial position, and the spatial position of each source.

The first assumption, that organisms learn their own HRTF, is reasonable and supported by extensive experimental evidence [22–24]. When $h_i(t)$ is interpreted to include not only the HRTF but also the properties of the acoustic environment (reverberations, *etc.*) then this assumption becomes considerably stronger. Animals have, however, been shown to estimate some properties of their acoustic environments quite quickly [25].

The second assumption, that the precise positions of each source are known, is more restrictive. There are, however, several ways in which the source positions might be determined. One possibility is that they might be established by prior or additional knowledge, perhaps using visual information. Indeed, the spatial cues provided by vision can override those inferred from audition, as demonstrated by the "ventriloquist effect." A second possibility is that the positions of the sources could be established through auditory preprocessing, using for example the binaural cues available to the auditory brainstem. Finally, the positions of the sources, as well as the properties of the acoustic environment, could be jointly estimated along with the content of each source; this joint estimation might be made easier by moving the head slightly so as to perturb the HRTFs by some known angle without changing the source positions.

3.2 The signal dictionary and neural representations

We have not considered the question of how an appropriate signal dictionary might be obtained. Fortunately there is a rich literature on finding a basis matched (in the sense of yielding sparse representations) to an ensemble of signals [5, 26, 27].

The algorithm was developed here in the signal processing framework, with little attention to possible neural implementation. However, overcomplete representations have been suggested for visual areas V1 [27] and IT [28]. Signal dictionaries have been interpreted in terms of models of receptive fields, and receptive field properties have been predicted from the principles of sparse representations [26, 29]. Similarly, the signal elements derived from optimising the matrix $\mathbf{D}$ for separating ensembles of natural sounds filtered through the HRTF offers predictions for auditory representations. The extension of such models to auditory cortex is intriguing [30].

Acknowledgements

We thank Didier Depireux, Tomas Hromadka and Mike Deweese for helpful comments. Supported by Higher Education Authority of Ireland and Science Foundation Ireland grant 00/PI.1/C067 (BAP), and grants from the Sloan Foundation, Mathers Foundation, NIH, Packard Foundation and the Redwood Neuroscience Institute (AMZ).

References

[1] Albert S. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press, Cambridge, Massachusetts, 1990. ISBN 0-262-02297-4.

[2] W. A. Yost, Jr. Dye, R. H., and S. Sheft. A simulated "cocktail party" with up to three sound sources. *Percept Psychophys*, 58(7):1026–1036, 1996.

[3] Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1999.

[4] T.-W. Lee, M. S. Lewicki, M. Girolami, and T. J. Sejnowski. Blind source separation of more sources than mixtures using overcomplete representations. *IEEE Signal Processing Letters*, 4(5):87–90, 1999.

[5] M. S. Lewicki and T. J. Sejnowski. Learning overcomplete representations. *Neural Computation*, 12(2):337–365, 2000.

[6] Michael Zibulevsky and Barak A. Pearlmutter. Blind source separation by sparse decomposition in a signal dictionary. *Neural Computation*, 13(4):863–882, April 2001.

[7] M. Lewicki and B. A. Olshausen. Inferring sparse, overcomplete image codes using an efficient coding framework. In *Advances in Neural Information Processing Systems 10*, pages 815–821. MIT Press, 1998.

[8] P. Bofill and M. Zibulevsky. Underdetermined blind source separation using sparse representations. *Signal Processing*, 81(11):2353–2362, 2001.

[9] S. T. Rickard and F. Dietrich. DOA estimation of many W-disjoint orthogonal sources from two mixtures using DUET. In *Proceedings of the 10th IEEE Workshop on Statistical Signal and Array Processing (SSAP2000)*, pages 311–314, Pocono Manor, PA, August 2000.

[10] G. Cauwenberghs. Monaural separation of independent acoustical components. In *Proc. IEEE Int. Symp. Circuits and Systems (ISCAS'99)*, volume 5, pages 62–65, Orlando FL, 1999.

[11] Sepp Hochreiter and Michael C. Mozer. Monaural separation and classification of mixed signals: A support-vector regression perspective. In Te-Won Lee, Tzyy-Ping Jung, Scott Makeig, and Terrence J. Sejnowski, editors, *3rd International Conference on Independent Component Analysis and Blind Signal Separation*, San Diego, CA, December 9-12 2001.

[12] Gil-Jin Jang and Te-Won Lee. A maximum likelihood approach to single-channel source separation. *Journal of Machine Learning Research*, 4:1365–1392, December 2003.

[13] Sam T. Roweis. One microphone source separation. In *Advances in Neural Information Processing Systems 13*, pages 793–799. MIT Press, 2001.

[14] T. Poggio, V. Torre, and C. Koch. Computational vision and regularization theory. *Nature*, 317(6035):314–319, 1985.

[15] D. L. Donoho and M. Elad. Maximal sparsity representation via l1 minimization. *Proceedings of the National Academy of Sciences*, 100:2197–2202, March 2003.

[16] R. Fletcher. Semidefinite matrix constraints in optimization. *SIAM J. Control and Opt.*, 23: 493–513, 1985.

[17] P. M. Hofman and A. J. Van Opstal. Bayesian reconstruction of sound localization cues from responses to random spectra. *Biol Cybern*, 86(4):305–316, 2002.

[18] E. I. Knudsen and M. Konishi. Mechanisms of sound localization in the barn owl. *Journal of Comparative Physiology*, 133:13–21, 1979.

[19] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman. Localization using nonindividualized head-related transfer functions. *J Acoust Soc Am*, 94(1):111–123, 1993.

[20] F. L. Wightman and D. J. Kistler. Headphone simulation of free-field listening. II: Psychophysical validation. *J Acoust Soc Am*, 85(2):868–878, 1989.

[21] A. Kulkarni and H. S. Colburn. Role of spectral detail in sound-source localization. *Nature*, 396(6713):747–749, 1998.

[22] A. J. King, C. H. Parsons, and D. R. Moore. Plasticity in the neural coding of auditory space in the mammalian brain. *Proc Natl Acad Sci USA*, 97(22):11821–11828, 2000.

[23] B. A. Linkenhoker and E. I. Knudsen. Incremental training increases the plasticity of the auditory space map in adult barn owls. *Nature*, 419(6904):293–296, 2002.

[24] P. M. Hofman, J. G. Van Riswick, and A. J. Van Opstal. Relearning sound localization with new ears. *Nat Neurosci*, 1(5):417–421, 1998.

[25] B. G. Shinn-Cunningham. Models of plasticity in spatial auditory processing. *Audiology and Neuro-Otology*, 6(4):187–191, 2001.

[26] Anthony J. Bell and Terrence J. Sejnowski. The 'independent components' of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338, 1997.

[27] B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37(23):3311–3325, 1997.

[28] M. Riesenhuber and T. Poggio. Models of object recognition. *Nature Neuroscience*, 3 Suppl:1199–1204, 2000.

[29] B. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

[30] B. A. Olshausen and K. N. O'Connor. A new window on sound. *Nature Neuroscience*, 5: 292–293, 2002.