MONAURAL SEPARATION OF INDEPENDENT ACOUSTICAL COMPONENTS

Gert Cauwenberghs

Department of Electrical and Computer Engineering
and Center for Language and Speech Processing
Johns Hopkins University, Baltimore MD 21218
E-mail: gert@jhu.edu

ABSTRACT

The problem of blindly separating signal mixtures with fewer mixture components than independent signal sources is mathematically ill-defined, and requires suitable prior information on the nature of the sources. Recently, it has been shown that sparse methods for function approximation using a Laplacian prior can be effective, but the method fails to separate a single mixture without further prior information. Other techniques track harmonics, but assume separability in the time-frequency domain. We show that a measure of temporal and spectral coherence provides an effective cue for separating independent acoustical or sonar sources, in the absence of spatial cues in the monaural case. The technique is shown to successfully separate single mixtures of sources with significant spectral overlap.

1. INTRODUCTION

The human brain excels at segregating complex mixtures of unknown signals from sensory modalities that convey a convoluted and incomplete version of the co-existing signals from the environment. The fact that our auditory system and periphery is able to distinguish, even from a monaural signal, two simultaneous conversations or music instruments playing the same note, suggests that effective signal processing solutions exist in biology [1, 2]. This motivates a study of architectures for adaptive blind signal processing which emulate function and structure of information processing in biological neural systems [3].

Auditory scene analysis [2, 1] synthesizes extensive psychophysical experimentation into a descriptive theory of how the brain processes auditory information from a cochlear time-frequency representation to extract and track relevant signals amidst noise and interference. Separation occurs through binding and grouping of events in the time-frequency domain. This approach appears to be the most neuromorphic of techniques proposed for acoustic signal separation, and is the basis for several algorithms for separating acoustic sources.

This paper addresses the problem of separating a *single* mixture of signals, such as extraction of independent voice streams in a monaural acoustic channel, illustrated in Figure 1. This task is mathematically ill-defined since the problem carries more unknowns than supported by the data. The approach to solve the ambiguity problem, is to specify a suitable prior on the source distributions, and an answer to the question what is a suitable prior depends in on the type of signals considered. In principle, one could construct general models and train the parameters in the model from data. A generic Laplacian prior, expressing sparsity in the source distributions, was proposed in [4], but it requires at least two mixtures to separate three signals [5]. Real-time tracking of a group of co-modulated harmonics provides a means to separate one source from time-frequency clutter [6], but the success of the technique depends strongly on the degree of interference from other sources. This papers considers a source prior that expresses *spectral and temporal coherence*, which is capable of the monaural case (separating single mixtures) and applies to a wide range of acoustic and sonar signals. A related approach based on amplitude and frequency modulation is given in [7].

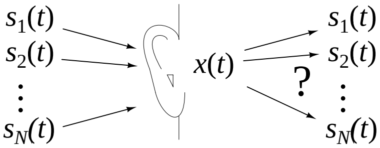

Figure 1: *Monaural separation of independent acoustic components.*

In Section 2 we outline a general mathematical framework for source separation, which we extend in Section 3 to the monaural case, accounting for a measure of coherence. Results on separating monaural mixtures of acoustic signals are presented in Section 4.

2. INDEPENDENT COMPONENT ANALYSIS AND SPARSE APPROXIMATION

2.1. Independent Component Analysis

The mathematical tool behind the problem of separating unknown mixtures of unknown sources is that of *independent component analysis* (ICA) [8]-[14], which applies information-theoretic concepts to extract mixture coefficients from the data, mostly using a linear or convoluting mixing model together with the assumption that the sources are *statistically independent.* Adaptive algorithms of this type offer abstract models of how the brain may separate multiple streams of sensory information [8], and scalable architectures can be formulated which are amenable to analog VLSI implementation [17, 18].

The task of separating linear mixtures of signals into independent components to reconstruct the sources of the signals is mathematically ill-defined in all but a few cases. Clearly, any observed signal can be arbitrarily decomposed as a sum of signals in

This work was sponsored by ONR (MURI N00014-95-1-040 and YIP, Harold Hawkins, program manager) and NSF (Career, MIP-9702346).

an infinite number of ways, and suitable prior knowledge has to be embedded in the model of the sources to yield meaningful values of the mixing parameters. Typically, the number of independent channels of observed signals is assumed to be at least the number of independent sources. This makes the task relatively straightforward, and in the case with more independent (sensor) observations than sources the mixing matrix can be uniquely determined, even from just second-order statistics of the signal data.

The number of sources can not always be determined *a priori* for a given application, and a suitable number of linearly independent observations of the sources can not always be guaranteed. Research in ICA on the case of fewer independent observations than sources has been non-existing until recently, making use of *prior* information on the source distribution. We start by formulating source models as a prior in a general Bayesian framework.

2.2. A General Bayesian Framework

In the Bayesian approach to independent component analysis [16, 15, 4], prior model information on the source $\mathbf{s}$ is used to construct the optimal maximum *a posteriori* (MAP) estimator through application of Bayes's rule:

$$\max_{A,\mathbf{s}} : \quad P(\mathbf{s}|\mathbf{x}) \propto P(\mathbf{x}|\mathbf{s})\; P(\mathbf{s}) \tag{1}$$

The key is a suitable formulation of a source and mixture models. They could be specified, for instance, as an auto-regressive source model and linear mixture model

$$\begin{aligned} s_i(k) &= g_i(s_i(k-1), s_i(k-2), \ldots) + e_i(k) \quad \forall i \\ \mathbf{x}(k) &= \mathbf{A}\mathbf{s}(k) + \mathbf{n}(k) \end{aligned} \tag{2}$$

respectively, where the g_i are (possibly nonlinear) scalar functions characterizing the sources i, e_i represent generative errors in these source models, and $\mathbf{n}(k)$ represents the observation noise in the mixture model.

A full implementation of the general model (2) can be obtained, in principle, by expanding g_i in parametric form, and adapting the autoregressive parameters of g_i on-line. The problem with this approach is that training data in the regressive source model (2) is not available, since both g_i and $\mathbf{s}_i$ are unknown. A possible solution which estimates the parameters on-line is demonstrated in [15]. Parametric techniques of this type work well when there is sufficient data in the observations to estimate all model parameters.

2.3. Sparse Approximation in Overcomplete Bases

The case of source separation with fewer observation channels than sources has been addressed by techniques of sparse approximation in an overcomplete representation, assuming a Laplacian (L_1-norm) prior on the source distributions [4]. This is a special case of (2) where $g_i \equiv 0$, e_i is Laplacian distributed, and $\mathbf{n}$ is normally distributed. Then (1) yields

$$\max_{A,\mathbf{s}} : \quad \log P(\mathbf{s}|\mathbf{x}) \propto -\frac{\lambda}{2}|\mathbf{x} - \mathbf{A}\mathbf{s}|^2 - \mu \sum_i |s_i| \tag{3}$$

which can be solved through resursion of an algorithm closely related to ICA [4]. The technique has been applied in an experimental setting to separate three speech sources using only two microphones [5].

The success in separating three sources from only two observation channels is quite remarkable, especially because the methods makes no assumptions on the source distributions themselves ($g_i \equiv 0$). For acoustic signals, we can expect improvement from a source model that incorporates some general knowledge about acoustics. A more informative source model is necessary in the monaural case, because a Laplacian prior on itself fails to separate more than one speech signal from a single microphone recording.

In this paper, we consider a measure of *spectral and temporal coherence* as a suitable prior to separate monaural mixtures of acoustical signals in the time-frequency domain.

3. MONAURAL SEPARATION IN THE TIME-FREQUENCY DOMAIN

3.1. Time-Frequency Wavelet Decomposition and Reconstruction

Wavelets [20] provide a powerful mathematical tool for analyzing temporal data that contain pertinent features both in the time domain and the frequency domain, such as with acoustical [3] waveforms.

Wavelet decomposition and reconstruction can be performed on analog continuous-time temporal data using a complex gaussian kernel [21]:

$$\begin{aligned} x^k(t) &= \int_{-\infty}^{t} x(\tau)\; \exp\left(j\omega_k\tau - \alpha(\omega_k(t-\tau))^2\right) d\tau \\ x'(t) &= C \sum_k x^k(t)\; \exp\left(-j\omega_k t\right) \end{aligned} \tag{4}$$

where the center frequencies ω_k are equally spaced on a logarithmic scale. The constant α sets the relative width of the frequency bins in the decomposition, and can be adjusted together with C to accommodate compact support of the wavelet kernel.

The computation (4) is efficiently implemented in custom analog VLSI technology, using a parallel architecture. In previous research we have developed several versions of wavelet transform processors in analog and hybrid analog-digital technology [21].

Wavelet decomposition provides a convenient signal representation for blind separation in the time-frequency domain.

3.2. Time-Frequency Domain Coherent Signal Segregation

A distribution-invariant prior can be defined on the sources in the form of an error functional which expresses a direct measure of *short-term coherence* between sources in the time-frequency domain. This of course requires that short-term coherence is a meaningful metric, which it is for most acoustical and sonar signals.

Results from psychophysical experiments of hearing discrimination [1, 2] show that coherence and harmonic relationships between acoustic sources plays a crucial role in identifying independent sounds. Developments of this idea have led to several algorithms for music separation, *e.g.,* [6, 7]. Most techniques, however, ignore relative phase information present in interfering signal components, and thus discard one of the most discriminatory features to distinguish independent signals. The argument to what extent and in what form the auditory periphery employs phase information is subject to an intense debate among neurophysiologists; and it is clear that phase information plays a crucial role, *e.g.,* for localization in the auditory periphery of bats [22].

3.2.1. *Incoherence and Independence*

The key idea to solve the underconstrained source separation problem is to exploit *temporal* rather than *spatial* correlations in distinguishing among independent signal components. The principle is illustrated in Figure 2. We consider scalar mixtures of signals in the monaural case; this assuption is not essential and simplifies the analysis. Information on temporal structure is conveniently obtained from the time-frequency decomposition (4):

$$s(t) = \sum_k s^k(t)\ \exp\left(-j\omega_k t\right) \tag{5}$$

where the constant C is dropped for notational convenience. Temporal incoherence between two signals expresses randomness in the relative phase (or, more specifically, time difference) between the two signals. We generalize this notion to include both random time θ and amplitude A fluctuations:

$$\begin{aligned} \tilde{s}(t) &= A(t)\ \bar{s}(t - \theta(t)) \qquad (6)\\ &= A(t)\ \sum_k \bar{s}^k(t)\ \exp\left(j\omega_k \theta(t)\right)\ \exp\left(-j\omega_k t\right) \end{aligned}$$

where $\bar{s}$ represents a periodic waveform[1]. While amplitude variations scale the spectrum uniformly, we see that time fluctuations (jitter) modulate the spectrum nonuniformly. Incoherence in mixtures of independent sources thus provides a key to distinguish components based on a measure of fluctuations in relative amplitude and phase.

3.2.2. *Time-Frequency Separation of Quasi-Coherent Sources*

Separation of mixtures of sources based on incoherence (6) only works provided that the sources $\bar{s}$ themselves are sufficiently coherent over an extended period of time, for fluctuations in relative $A(t)$ and $\theta(t)$ to be observable. In particular, let the observed signal $x(t)$ be composed of an incoherent mixture of sources $\tilde{s}_i(t)$, $i = 1 \cdots N$:

$$\begin{aligned} x(t) &\approx \sum_i \tilde{s}_i(t)\ =\ \sum_i A_i(t)\ \bar{s}_i(t - \theta_i(t)) \qquad (7)\\ &= \sum_i \sum_k A_i(t)\ \bar{s}_i^k(t)\ \exp\left(j\omega_k \theta_i(t)\right)\ \exp\left(-j\omega_k t\right) \end{aligned}$$

which can be reformulated in the time-frequency domain as

$$x^k(t) \approx \sum_i A_i(t)\ \bar{s}_i^k(t)\ \exp\left(j\omega_k \theta_i(t)\right)\ . \tag{8}$$

The only knowledge available about the coefficients $\bar{s}_i^k(t)$ is that they represent periodic sources $\bar{s}_i$. From (8), the assumption of incoherence across sources reduces the complex autocorrelation of x^k to

$$\begin{aligned} &E(x^k(t)\, x^{k*}(t - \delta t)) \\ &\approx \sum_i E\Big(A_i(t)\, A_i(t - \delta t)\ \bar{s}_i^k(t)\ \bar{s}_i^{k*}(t - \delta t)\ \times \\ &\qquad \exp\left(j\omega_k(\theta_i(t) - \theta_i(t - \delta t))\right)\Big) \qquad (9)\\ &\approx \sum_i E(|A_i(t)|^2)\ E(\bar{s}_i^k(t)\ \bar{s}_i^{k*}(t - \delta t)) \end{aligned}$$

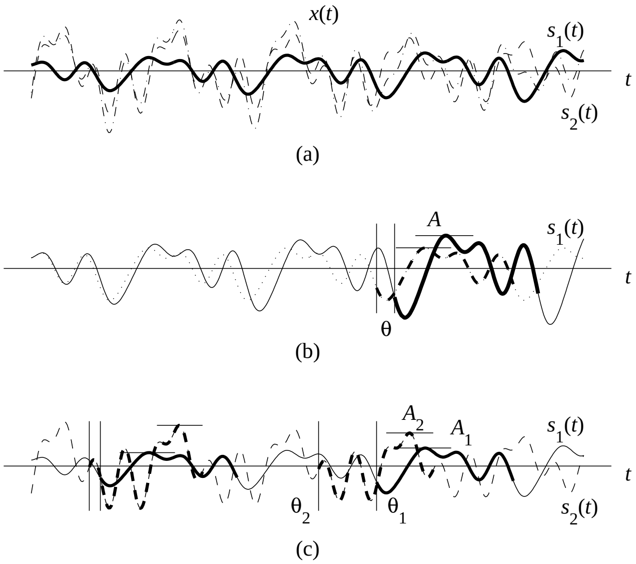

Figure 2: *Coherence-based monaural separation. (a) Source separation from a single mixture $x(t)$ is ambiguous, unless the sources $x_i(t)$ have a known structure. (b) Short-term coherence of a source can be expressed in terms of time θ and amplitude A fluctuations of a periodic waveform. (c) These fluctuations allow to distinguish and separate the sources even if they overlap in the spectral domain.*

where the last equality assumes that δt is shorter than the coherence time, *i.e.,* characteristic time scale of A_i and θ_i. Finally, if $\bar{s}_i^k(t) \approx \bar{S}_i^k(t) \exp(-j\Omega_i^k t)$ represents the harmonic precession of source i in band k, then

$$E(\bar{s}_i^k(t)\ \bar{s}_i^{k*}(t - \delta t)) \approx |\bar{S}_i^k(t)|^2\ \exp(-j\Omega_i^k\ \delta t)\ . \tag{10}$$

and the coefficients $\bar{s}_i^k(t)$ can be completely determined, up to a constant phase, from a number of autocorrelation observations (9) larger than the number of sources. The remaining unknowns are then derived from (7).

3.2.3. *Time-Domain Separation*

A simpler procedure is obtained by directly implementing (7) in the time domain. For any fragment of $s_i(t)$, a greedy correlation-based search over neighboring time intervals produces a matching segment $s_i(t - \theta_i)$, for a value θ_i that best satisfies the constraint (6) for s_i. Both segments are updated towards each other in order to refine the constraint (6), along with additional soft constraints that enforce smoothness in the sources[2]. The procedure is repeated until all pairs of matching segments converge to within the noise level of condition (6).

4. EXPERIMENTS

Best results were obtained using the time-domain separation method. For the experiments, we considered single mixtures of 2

[1] We cannot assume that the coefficients $\bar{s}^k(t)$ are constant. Periodicity implies, to first order, that the $\bar{s}^k(t)$ oscillate with constant frequency.

[2] In particular, an update $\delta s_i(t)$ towards $s_i(t - \theta_i)$ is one of two fixed amplitudes: the larger one if the update is in the direction of the spatial average, and the smaller one otherwise. The same applies for the update $\delta s_i(t - \theta_i)$ towards $s_i(t)$.

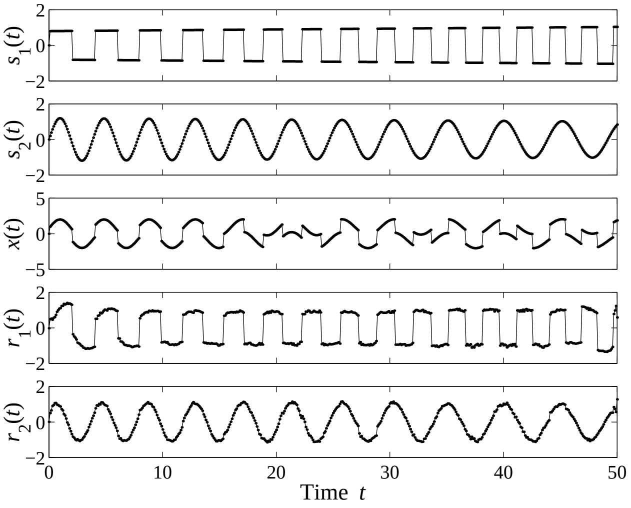

Figure 3: *Monaural separation in the time domain.* Top: *Two sources with overlapping spectra.* Center: *Single mixture.* Bottom: *Reconstructed sources from the mixture.*

sources with overlapping spectra. Figure 3 demonstrates successful separation of two frequency-chirped and amplitude-modulated, quasi-periodic waveforms. The separation succeeds even though harmonics of the instantaneous frequencies of the sources coincide at two instances in time. We are continuing experiments to characterize the method on real-world signals, such as acoustical and sonar recordings with overlapping spectra, in the presence of background noise.

5. CONCLUSIONS

We proposed and demonstrated a technique for monaural separation of acoustical sources, based on a metric of coherence that expresses each source as a periodic waveform with random short-term time and amplitude fluctuations. We plan to implement a coherence-based separation network in parallel VLSI, and further research is directed towards simple on-line variants of the presented algorithms both in the frequency and time domain.

6. REFERENCES

[1] S. McAdams, "Segregation of Concurrent Sounds: 1. Effects of Frequency-Modulation Coherence," *Journal of the Acoustical Society of America,* vol. **86** (6), pp 2148-2159, 1989.

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

[3] K.S. Wang and S.A. Shamma, "Spectral Shape Analysis in the Central Auditory System," *IEEE T. Speech and Audio Processing,* vol. **3** (5), pp 382-395, 1995.

[4] M.S. Lewicki and T.J. Sejnowski, "Learning overcomplete representations." to appear in *Neural Computation,* 1999.

[5] T.-W. Lee, M.S. Lewicki, M. Girolami and T.J. Sejnowski, "Blind source separation of more sources than mixtures using overcomplete representations," submitted to *IEEE Sig. Proc. Lett.,* 1998.

[6] A.L.C. Wang, "Instantaneous and Frequency-Warped Signal Processing Techniques for Auditory Source Separation," Ph.D. Dissertation, Dept. of Electrical Engineering, Stanford University, 1994.

[7] M. Abe and S. Ando, "Auditory Scene Analysis Based on Time-Frequency Integration of Shared FM and AM," *Proc. 1998 IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP'98),* vol. **4**, pp 2421-2424, 1998.

[8] C. Jutten and J. Herault, "Blind Separation of Sources .1. an Adaptive Algorithm Based on Neuromimetic Architecture," *Signal Proc.*, vol. **24** (1), pp 1-10, 1991.

[9] P. Comon, "Independent Component Analysis, a New Concept," *Signal Proc.*, vol. **36** (3), pp 287-314, Apr 1994.

[10] A.J. Bell and T.J. Sejnowski, "An Information Maximization Approach to Blind Separation and Blind Deconvolution," *Neural Comp.*, vol. **7** (6), pp 1129-1159, Nov 1995.

[11] A. Cichocki and R. Unbehauen, "Robust Neural Networks with On-Line Learning for Blind Identification and Blind Separation of Sources," *IEEE Trans. Circuits and Systems I,* vol. **43** (11), pp 894-906, November 1996.

[12] J. Karhunen, E. Oja, L.Y. Wang, R. Vigario and J. Joutsensalo, "A Class of Neural Networks for Independent Component Analysis," *IEEE T. Neural Networks,* vol. **8** (3), pp 486-504, May 1997.

[13] J.F. Cardoso, "Infomax and Maximum-Likelihood for Blind Source Separation," *IEEE T. Sig. Pr.*, vol. **4** (4), pp 112-114, Apr 1997.

[14] T.W. Lee, *Independent Component Analysis,* Norwell, MA: Kluwer Academic, 1999.

[15] B.A. Pearlmutter and L.C. Parra, "A context-sensitive generalization of ICA," *1996 International Conference on Neural Information Processing,* Hong Kong, Sept. 1996.

[16] D.J.C. MacKay, "Maximum Likelihood and Covariant Algorithms for Independent Component Analysis,"University of Cambridge, Cavendish Laboratory Tech. Rep., 1996.

[17] M.H. Cohen and A.G. Andreou, "Current-Mode Subthreshold MOS Implementation of Herault-Jutten Autoadaptive Network," *IEEE J. Solid-State Circuits,* vol. **27**, pp 714-727, May 1992.

[18] M. Cohen and G. Cauwenberghs, "Blind Separation of Linear Convolutive Mixtures through Parallel Stochastic Optimization," *Proc. IEEE International Symposium on Circuits and Systems (ISCAS'98),* Monterey CA, 1998.

[19] T. Poggio and F. Girosi, "A Sparse Representation for Function Approximation," *Neural Comp.*, vol. **10** (6), pp 1445-1454, Aug 1998.

[20] I. Daubechies, "The Wavelet Transform, Time-Frequency Localization and Signal Analysis," *IEEE T. Info. Theory,* vol. **36** (5), pp 961-1005, 1990.

[21] R.T. Edwards and G. Cauwenberghs, "Analog VLSI Processor Implementing the Continuous Wavelet Transform," in *Adv. Neural Information Processing Systems,* Cambridge, MA: MIT Press, vol. **8**, May 1996.

[22] J.A. Simmons, M.J. Ferragamo and C.F. Moss, "Echo-Delay Resolution in Sonar Images of the Big Brown Bat, Eptesicus-Fuscus," *Proc. Nat. Academy Sciences USA,* vol. **95** (21), pp 12647-12652, Oct. 1998.