Audio Source Separation: Solutions & Problems
by
Mike Davies & Nikolaos Mitianoudis
Digital Signal Processing Lab
Queen Mary, University of London
Outline of Talk

**Introduction:**
- Audio Source Separation: ICA & CASA

**A Frequency Domain Framework**
- Unmixing in the frequency domain
- Source Modelling & permutation problem

**More Sources than Sensors:**
- Exploitation of sparsity

**A Reverberant BSS Example:**
- Physical Channels
- ICA as a beamformer

**Unresolved problems**
- Dealing with Reverberation
- Nonstationarity
- Noise etc.

**Discussion**
Cocktail Party Problem
Cocktail Party Problem
Computational tools for audio source separation

- Computational Auditory Scene Analysis (CASA)
  - Typically extracts one source from a single channel of audio using heuristic grouping rules based upon psychological observations

- Blind Source Separation (BSS aka ICA)
  - Uses spatial diversity (strong similarity to beamforming) based on source independence. Extensions include: convolutional mixing, overcomplete mixtures.

I will concentrate on BSS – though I will return to CASA in the discussion.
ICA is a generative model

Audio observations are linear convolution (plus additive noise)

\[ x_i(t) = \sum_{j=1}^{m} \sum_{k=1}^{q} a_{ij}(k)s_j(t-k) + e_i(t) \]

Unmixing filter uses an FIR approximation (complete case):

\[ \hat{s}_j(t) = \sum_{i=1}^{n} \sum_{k=1}^{p} w_{ji}(k)x_i(t-k) \]
Frequency (subband) filtering

The unmixing filtering can be efficiently performed within a subband framework.

This does not necessarily imply a frequency domain model for the sources.
Various authors have suggested the simple gradient-based algorithm for ICA:

$$\Delta W \propto (I - \varphi(s)s^T)W$$

This can be viewed as a Maximum Likelihood estimator with

$$\varphi(s) = \frac{\partial}{\partial y} \log P(s)$$

$$\varphi(s)$$ often takes a tanh-like shape → superGaussian prior.

For convolutive mixing this can be adapted to:

$$\Delta W(\omega) \propto \left( I - \text{fft}(\phi(s(t)), \omega)s(\omega)^H \right)W(\omega)$$

(time domain source model)
Frequency (subband) filtering

\[ \Delta W(\omega) \propto \left( I - \text{fft}(\phi(s(t)), \omega)s(\omega)^H \right) W(\omega) \]

Time domain modelling e.g Lee et al. 1997
Frequency Domain Source Model

An alternative strategy is to model the sources in the frequency domain (e.g. Smaragdis 1997).

\[ \Delta W(\omega) \propto \left( I - \phi(s(\omega))s(\omega)^H \right) W(\omega) \]

Advantages:

- Computational Efficiency
- Sparser Statistics (→ better estimates)
Frequency (subband) filtering

Frequency domain modelling (e.g. Smaragdis 1997).

\[ \Delta W(\omega) \propto \left( I - \phi(s(\omega))s(\omega)^H \right) W(\omega) \]

Disadvantage: The Permutation Problem.
Permutation Problem

Source Modelling Solutions

**Time Domain** → no permutation problem (Lee *et al.* 1997).

**Time-Frequency** → couples adaptive filters,
- using signal envelopes (Ikeda *et al.* 1999) or
- TF generative models (Mitianoudis & Davies 2001).

Permutation problem can persist with gradient learning (Davies 2002).

Channel Modelling Solutions

**Smooth filter Transfer Functions** → couples adaptive filters
- Heuristic (Smaragdis 1997)
- Constrained filter model (Parra & Spence 1998)

Solutions tend to get trapped in local minima (Ikram & Morgan 2000)
- Directivity patterns to resolve permutation (Kurita *et al.* 2000)

Problems at high frequencies and with high reverberation
Permutation Problem

Two speech signals mixed with a single echoes of about ~ 5ms

Smaragdis Alg.  Mitianoudis & Davies Alg.
Overcomplete case

How do we deal with more sources than sensors?

• Same ideas apply: estimate $A$ by nonlinear decorrelation
• Requires estimates for sources given the mixing model $A$
  - Sparse sources are very important!
  - Need to cope with noise
Choosing the right transform domain

3 Audio mixtures are not well isolated...

...But in an appropriate transform domain (MDCT)
Overcomplete Audio Example

2 (synthetic) mixtures of three audio sources.

Sparsity achieved by working with mDCT coefficients

Mixing matrix, A, learnt using sparse MoG model and GEM algorithm

Yellow lines indicate learnt mixture directions

Source 1    Source 2    Source 3
Source data allocation

Winner-takes all: independent & sparse \approx \text{mutually exclusive}

(note \text{>90\% coefficients treated as noise})
A Real Reverberant Example

Let's consider a real (complete) Source Separation Example
Reverberant BSS experiment I

~ 7.5 m

1m

2m

~ 6m
Reverberant BSS experiment II

~ 7.5 m

1.5 m

2 m

~ 6 m

1 m
Beamforming

We can examine the phase component of the unmixing system to deduce a directivity pattern. The learnt unmixing for a simple delay gives:

**Frequency Dependent Directivity Patterns (ideal)**
Beamforming

For a real reverberant room we get:

**Frequency Dependent Directivity Patterns (real)**

Separated sources: 🎧 🎧
Misaligned beamformer

When one of the sources is shifted by 50cm the previously learnt beamformer will be misaligned. For low frequencies this is not a problem. For frequencies greater than ~400Hz even a very small change will render the beamformer useless.

Comparison of Directivity Patterns

100Hz

800Hz
Unresolved Issues

• Reverberation:
  – Model it or treat as noise?

• Moving Sources:
  – NOT just an issue of faster convergence!
  – higher level priors (moving towards CASA)?

• And others:
  Estimating number of sources, de-noise, extracting a single (desired) source, permutation on-line, dereverberation, computational issues…