Video-aided model-based source separation in reverberant rooms

Mohsen Naqvi and Jonathon A. Chambers

Advanced Signal Processing Group
School of Electronic, Electrical and Systems Engineering
Loughborough University UK

Keynote Lecture
Wednesday 31st October 2012
The team of PhD and Postdoctoral researchers within the Advanced Signal Processing Group (ASPG) in the School of Electronic, Electrical and Systems Engineering, Loughborough University and our collaborators at Grenoble and Surrey Universities who have contributed.
Outline

1. Machine cocktail party problem

2. Background work on audio-visual source separation

3. Video-aided model-based source separation

4. Experimental results on real audio-video datasets (AV16.3) and conclusions
1. Machine Cocktail Party Problem
Motivation

- Using “Siri” the virtual assistant on the iPhone 4S/5: iHear, iSpeak
  “…Background noise can defeat the system, whether that’s noisy traffic or a room full of people talking…” , Ingear, Sunday Times, 23rd October 2011

- Hearing aids

- Teleconference and meeting_smart rooms
“One of our most important faculties is our ability to listen to, and follow, one speaker in the presence of others. This is such a common experience that we may call it ‘the cocktail party problem.’ No machine has been constructed to do just this, to filter out one conversation from a number jumbled together…”

Colin Cherry 1914-1979
Machine Cocktail Party Problem

Human Auditory Perception

Pioneers: Cherry, Sayers and Taylor, Imperial College, London, UK

Their key works appeared between 1953 and 1961

Highlighted benefits of **binaural hearing**

Potential of **lip reading** to improve hearing

Emphasized the critical role of **correlation in binaural fusion**
Our Target: “Machine Audition”

An Intelligent Machine Based on Audio and Video Signal Processing to Extract and Track a Sound Signal of Interest in a Room Environment

Localization: Where is the source?

Humans employ Acoustic and Visual Cues/Features: Interaural time difference (ITD); Interaural level difference (ILD) [Spatial Domain]

Sound Recognition: What is the source?

Humans utilize Acoustic Cues/Features: Intensity, Loudness; Periodicity; Onsets; AM/FM; Pitch; Timbre, Tone; Harmonicity; [Temporal/Spectral Domains]
Room is reverberant: microphone measurements affected by multipath

Reverberation time (RT60): function of dimensions of room and reflectivity of the wall surfaces together with the objects in the room
Machine Cocktail Party Problem

Our Motivation

Failure of Automatic Speech Recognition in Reverberant and Multispeaker Environments

“Voice control is absolutely the future for us” Nick Hedderman, Microsoft, 2011.
Machine Cocktail Party Problem

Application domain: audio-video conferencing environment

Information sensors: microphones and video cameras
2. Background work on audio-visual source separation
Loughborough University is leading the project with partners at University of Surrey, UK and the academic collaborators at the GIPSA Lab, Grenoble University, France.
Audio-Visual Source Separation

1. Camera Array
2. Video Localization
3. 3-D Visual-Tracking
4. Direction of Arrival and Velocity Information
5. Video Tracking
6. Mixed Audio Sources
7. Separated Sources
8. T-F Masking (Post Processing)
9. Robust Least Squares Frequency Invariant Data Independent Beamforming
10. Convulsive Blind Source Separation Method
11. Decision Making
**Convolutive Blind Source Separation (CBSS)**

**Compact form:**

\[ x(t) = H(t) * s(t) \]

**Expansion form:**

\[ x(t) = \sum_{j=1}^{N} \sum_{p=0}^{P-1} h_{ij}(p) s_j(t - p) \]

- Transform the CBSS Problem into the Frequency Domain (FD-CBSS) by applying STFT:

\[
\begin{bmatrix}
  x_1^{(k)}(n) \\
  \vdots \\
  x_M^{(k)}(n)
\end{bmatrix}
= \begin{bmatrix}
  h_{11}^{(k)}(n) & \cdots & h_{1N}^{(k)}(n) \\
  \vdots & \ddots & \vdots \\
  h_{M1}^{(k)}(n) & \cdots & h_{MN}^{(k)}(n)
\end{bmatrix}
\begin{bmatrix}
  s_1^{(k)}(n) \\
  \vdots \\
  s_N^{(k)}(n)
\end{bmatrix}
\]

Convolutive BSS problem \(\rightarrow\) Multiple complex-valued instantaneous BSS problems
Independent Component Analysis (ICA)

- Independent components are estimated as:

\[
\hat{s}^{(k)}(n) = w^{(k)} \cdot x^{(k)}(n)
\]

- Define a Cost Function by Indepency:

\[
C = KL\left(p(\hat{s}_1^{(k)}, \hat{s}_2^{(k)}) \prod_{i=1}^{2} p(\hat{s}_i^{(k)})\right)
\]

As \(\hat{s}_i^{(k)}\) approaches to the independent component, The cost goes to zero.

- Using a Gradient Descent Algorithm:

\[
w_{ij}^{(k)\text{new}} = w_{ij}^{(k)\text{old}} + \eta \Delta w_{ij}^{(k)}
\]

\[
\Delta w_{ij}^{(k)} = -\frac{\partial C}{\partial w_{ij}^{(k)}} = w_{ij}^{(k)} - E\{\varphi(\hat{s}_i^{(k)})\} x_j^{(k)}
\]

- Using Natural Gradient Learning:

\[
\Delta w_{ij}^{(k)} = \sum_{l=1}^{2} (I_{il} - E\{\varphi(\hat{s}_i^{(k)}) \cdot \hat{s}_l^{(k)*}\}) w_{ij}^{(k)}
\]

Nonlinear Score Function

\[
\varphi(\hat{s}_i^{(k)}) = -\frac{\partial \log p(\hat{s}_i^{(k)})}{\partial \hat{s}_i^{(k)}}
\]
Independent Component Analysis (ICA)

Permutation and Scaling Problems

- ICA is conducted in each subband independently.
- Ordering and Scaling of outputs are arbitrary in ICA.
- 2x2 Frequency-Domain ICA.

- Gain indeterminacy can be managed by normalization.
- What is the best available method to solve the permutation ambiguity?
Independent Vector Analysis (IVA)

• Modified Multivariate Cost Function

\[ C = KL \left( p(\hat{s}_1, \hat{s}_2) \right) \prod_{i=1}^{2} p(\hat{s}_i) \]

Components → Vectors!

\[ \hat{s}_i = (\hat{s}_i^{(1)}, \ldots, \hat{s}_i^{(K)})^T \]

Conventional ICA Score Function

\[ \varphi(\hat{s}_i^{(k)}) = -\frac{\partial \log p(\hat{s}_i^{(k)})}{\partial \hat{s}_i^{(k)}} = \frac{\hat{s}_i^{(k)}}{|\hat{s}_i^{(k)}|} = \exp(j \cdot \text{arg}(\hat{s}_i^{(k)})) \]

• Using a Gradient Descent Algorithm:

\[ w_{ij}^{(k)\text{new}} = w_{ij}^{(k)\text{old}} + \eta \Delta w_{ij}^{(k)} \]

\[ \Delta w_{ij}^{(k)} = -\frac{\partial C}{\partial w_{ij}^{(k)}} = w_{ij}^{-\text{T}(k)} - E \{ \varphi^{(k)}(\hat{s}_i^{(1)} \ldots \hat{s}_i^{(K)}) \} x_j^{* (k)} \]

• Using Natural Gradient Learning:

\[ \Delta w_{ij}^{(k)} = \sum_{i=1}^{2} (I_{ij} - E \{ \varphi^{(k)}(\hat{s}_i^{(1)} \ldots \hat{s}_i^{(K)}) \hat{s}_j^{* (k)} \}) w_{ij}^{(k)} \]

IVA Score Function (Multivariate)

\[ \varphi^{(k)}(\hat{s}_i^{(1)} \ldots \hat{s}_i^{(K)}) = \frac{\partial \sqrt{\sum_{i=1}^{K} |\hat{s}_i^{(k)}|^2}}{\partial \hat{s}_i^{(k)}} = \frac{\hat{s}_i^{(k)}}{\sqrt{\sum_{i=1}^{K} |\hat{s}_i^{(k)}|^2}} \]

Multivariate function exploiting frequency dependency!

[Kim 2007]
Independent Vector Analysis (IVA)

- IVA [Kim 2007] mitigates the permutation problem if sufficient audio samples are available.
- If speakers are moving?
Adaptive Beamforming Concept

Desired signal

Interference signal

Array element 1
Array element 2
Array element N

∑
y(t)

- d(t)

+ e(t)

Weight approximation
Adaptive algorithm

λ = 40
λ = 20
λ = 0
Robust Beamformer

RLSFIDI Beamformer with Circular Microphone Array

• The least squares problem is:

\[ \min_{w(\omega)} \| H^T(\omega)w(\omega) - r_d(\omega) \|^2_2 \]

where \( r_d(\omega) \) is a desired response vector and can be designed from a 1D window e.g. Dolph-Chebyshev or Kaiser windows.

and

\[ H(\omega) = [d(\omega, \theta_1, \phi_1), \ldots, d(\omega, \theta_M, \phi_M)] \]

• The 3-D positions of the microphones are provided in the matrix form:

\[
U = \begin{bmatrix}
    u_{x_1} & u_{y_1} & u_{z_1} \\
    \vdots & \vdots & \vdots \\
    u_{x_N} & u_{y_N} & u_{z_N}
\end{bmatrix}
\]
Robust Beamformer

RLSFIDI Beamformer with Circular Microphone Array

• The beamformer response vector is:

\[
d(\omega, \theta_i, \phi_i) = \begin{bmatrix}
    \exp(-j\kappa(\sin(\theta_i)\cos(\phi_i).u_{x1} + \sin(\theta_i)).
    \sin(\phi_i).u_{y1} + \cos(\theta_i).u_{z1}) \\
    \vdots \\
    \exp(-j\kappa(\sin(\theta_i)\cos(\phi_i).u_{xN} + \sin(\theta_i)).
    \sin(\phi_i).u_{yN} + \cos(\theta_i).u_{zN})
\end{bmatrix}
\]

where \( \kappa = \omega/c \) and \( c \) is the speed of sound in air at room temperature.

• The least squares problem is optimized subject to the constraints for wider main lobe for the source of interest and wider attenuation pattern

• The white noise gain (WNG) constraint measures the sensitivity of the beamformer
Advantages of RLSFIDI beamformer

1. Data independent and frequency invariant design.
2. Superior spatial selectivity at low frequencies and some variability in microphone positioning.
3. White noise gain (WNG) constraint control the beamformer sensitivity.
4. Control the uncertainties in source localization and DOA information by using convex optimization approach.

Beampatterns.
Robust Beamformer

Visual Information

DOA

\[ \theta_1, \phi_1 \]

\[ \theta_N, \phi_N \]

Convolutive Environment

Audio mixtures

Linear filtering

Robust Beamformer

\[ s_1[n], s_2[n], ..., s_N[n] \]

\[ x_1[n], x_2[n], ..., x_M[n] \]

\[ U_1[n], U_2[n], ..., U_N[n] \]

\[ y_1[n], y_2[n], ..., y_N[n] \]

- The RLSFIDI beamformer performs well at lower reverberation time.
- If reverberation time is high?
Linear to Nonlinear Separation

- Linear Separation: Multichannel ICA/IVA/Beamforming
- Nonlinear Separation: Using a time frequency mask
In 1946, Gabor proposed, “a new method of analysing signals is presented in which time and frequency play symmetrical parts”.

![Time Frequency Signal Representation Diagram]
“Imagine two narrow channels dug up from the edge of a lake, with handkerchiefs stretched across each one. Looking only at the motion of the handkerchiefs, you are to answer questions such as: How many boats are there on the lake and where are they?” [Bregman, 1990]

• It's a challenge!
Computational Auditory Scene Analysis (CASA)

- CASA systems approach sound separation based on ASA principles

Cues:

1. Spatial (binaural): ITD, ILD

2. Brain uses signal-intrinsic cues (onset, harmonicity) to form sources

[Ellis & Mandel, 2009]
Audio-Visual Source Separation

Our Proposed Solution

Visual Information

Audio mixtures → Linear filtering → Non-linear filtering
3. Video-aided model-based source separation
**Objective:** Obtain improved soft time-frequency (TF) masks for each source

- The ILD, IPD and mixing vectors are probabilistically modeled
- The model parameters are refined using an expectation-maximization (EM)
- The algorithm utilizes the source location information estimated by video
- Masks output are used to reconstruct the sources
- **Assumptions:** W-disjoint orthogonality, sources are physically stationary
Competing audio-visual source separation approaches

**[Wilson, 2001]:** 32-microphone array and multiple stereo cameras for audio-video tracking and separation

**[Sodoyer, 2004]:** separate acoustic source by utilizing coherence with speaker’s lip movements. Convolutive case not considered

**[Rivet, 2007]:** combined audio-video coherence and BSS for separation from convolutive mixtures. Separated two sources only

**[Casanovas, 2010]:** separation using camera video signal and synchronous one-microphone audio recording. Separated only two sources with full-frontal close-up video

**Limitations:** determined/overdetermined, anechoic mixing or limited reverberation, two sources only, close-up full-frontal images
Video-aided model-based source separation

Video:

- **Track speakers** in a room environment and **obtain their locations** when judged physically stationary. Markov chain Monte Carlo based particle filter (MCMC-PF) has been used for tracking.

Audio:

- The ratio of the audio measurements in the TF domain, gives the observed ILD and IPD. The **ILD, \( \alpha(\omega, t) \)**, and the **phase residual**, \( \hat{\phi}(\omega, t; \tau) \), at time \( t \) and frequency \( \omega \) are each modeled as normal distributions.
  - ILD: \( p(\alpha(\omega, t) | \mu(\omega), \eta^2(\omega)) = \mathcal{N}(\alpha(\omega, t) | \mu(\omega), \eta^2(\omega)) \)
  - Phase residual: \( p(\phi(\omega, t) | \tau(\omega), \sigma^2(\omega)) = \mathcal{N}(\hat{\phi}(\omega, t; \tau) | \xi(\omega), \sigma^2(\omega)) \)

where \( \tau \) is the frequency-dependent delay.
Video-aided model-based source separation

- The mixture $x(\omega, t)$, formed by concatenating $L(\omega, t)$ and $R(\omega, t)$, can be written as [Sawada, 2007], $x(\omega, t) = \sum_{i=1}^{I} h_i(\omega)s_i(\omega, t)$ and approximated as

$$x(\omega, t) \approx h_d(\omega)s_d(\omega, t)$$

where $h_d(\omega) = [h_{ld}(\omega), h_{rd}(\omega)]^T$ is the mixing vector from the dominant source $s_d(\omega, t)$ to the left and right sensor.

- The mixing vectors are modeled as [O’Grady, 2004],

$$p(x(\omega, t)|d_i(\omega), \zeta_i^2(\omega)) = \frac{1}{\pi \zeta_i^2(\omega)} \exp \left( - \frac{\|x(\omega, t) - (d_i^H(\omega)x(\omega, t))d_i(\omega)\|^2}{\zeta_i^2(\omega)} \right)$$

where $d_i(\omega)$ is the direction vector calculated using the source location estimates from video,

$$d_i(\omega) = \begin{bmatrix} \exp(-j\kappa(\sin(\theta_i) \cdot \cos(\phi_i) \cdot p_{x1}^j + \sin(\theta_i)) \cdot \sin(\phi_i) \cdot p_{x1}^j \cdot \cos(\theta_i) \cdot p_{x1}^j) \\ \exp(-j\kappa(\sin(\theta_i) \cdot \cos(\phi_i) \cdot p_{y2}^j + \sin(\theta_i)) \cdot \sin(\phi_i) \cdot p_{y2}^j \cdot \cos(\theta_i) \cdot p_{y2}^j) \end{bmatrix}$$

where $p_{xj}^j$, $p_{yj}^j$ and $p_{zj}^j$ for $j = 1, 2$ are the 3-D positions of the sensors and $\kappa = \omega/c_s$ and $c_s$ is the speed of sound in air at room temperature. The vector $d_i(\omega)$ is normalized to unity length before it is used in the model.
Video-aided model-based source separation

- **Model Parameters** can be collected as, 
  \[ \tilde{\Theta} = \{\mu_i(\omega), \eta_i(\omega), \xi_{i\tau}(\omega), \sigma_{i\tau}(\omega), d_i(\omega), \varsigma_i(\omega), \psi_{i\tau}\} \]
  where \( \mu_i, \xi_{i\tau}, \) and \( d_i \) and \( \eta_i^2, \sigma_{i\tau}^2 \), and \( \varsigma_i^2 \) are respectively the means and variances of the ILD, IPD, and mixing vector models.

- **Maximum likelihood** parameter estimation problem, EM is employed

- The log value of the likelihood function (L) given the observations is written as 
  \[
  \mathcal{L}(\tilde{\Theta}) = \sum_{\omega,t} \log p(\alpha(\omega,t), \phi(\omega,t), x(\omega,t)|\tilde{\Theta}) \\
  = \sum_{\omega,t} \log \sum_{i,\tau}[ \mathcal{N}(\alpha(\omega,t)|\mu_i(\omega), \eta_i^2(\omega)) \cdot \mathcal{N}(\phi(\omega,t;\tau)|\xi_{i\tau}(\omega), \sigma_{i\tau}^2(\omega)) \cdot \mathcal{N}(x(\omega,t)|d_i(\omega), \varsigma_i^2(\omega), \psi_{i\tau} ) ]
  \]

- **Expectation (E-step)**: calculate probability of the TF units belonging to sources and delays given the observations and the estimated parameters from the M-step

- **Maximization (M-step)**: re-estimate model parameters for each source and delay with the estimated occupation likelihood from the E-step

- **Soft masks** produced are used to reconstruct the sources
Results

Comparison of SDR (in decibels) performance as a function of RT60 using the proposed algorithm with and without dereverberation utilizing two microphones and the Naqvi, Maganti and RLSFIDI methods employing four and eight microphones for mixtures of three sources.

<table>
<thead>
<tr>
<th>RT60 (ms)</th>
<th>160 ms</th>
<th>210 ms</th>
<th>300 ms</th>
<th>485 ms</th>
<th>600 ms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed</td>
<td>7.02</td>
<td>6.89</td>
<td>6.25</td>
<td>4.33</td>
<td>3.40</td>
</tr>
<tr>
<td>Naqvi 4-mics</td>
<td>0.81</td>
<td>0.55</td>
<td>0.09</td>
<td>-0.76</td>
<td>-1.12</td>
</tr>
<tr>
<td>Naqvi 8-mics</td>
<td>2.93</td>
<td>2.56</td>
<td>1.92</td>
<td>0.82</td>
<td>0.37</td>
</tr>
<tr>
<td>Maganti 4-mics</td>
<td>1.07</td>
<td>0.92</td>
<td>0.38</td>
<td>-0.46</td>
<td>-0.79</td>
</tr>
<tr>
<td>Maganti 8-mics</td>
<td>3.17</td>
<td>2.97</td>
<td>2.29</td>
<td>1.05</td>
<td>0.47</td>
</tr>
<tr>
<td>RLSFIDI 4-mics</td>
<td>7.38</td>
<td>5.74</td>
<td>3.92</td>
<td>1.85</td>
<td>1.16</td>
</tr>
<tr>
<td>RLSFIDI 8-mics</td>
<td>10.15</td>
<td>8.28</td>
<td>6.30</td>
<td>4.08</td>
<td>3.34</td>
</tr>
</tbody>
</table>
4. Experimental results on real audio – video datasets AV16.3 and conclusions
Experimental Setup (Audio-Visual Configuration)

Audio-Visual Configuration

Cam1

Cam2

Cam3

Microphone Array
Results

Physically stationary speakers

Cam1

Cam2

Cam 3

Mixtures

Est. Sources

Two active speakers

Three active speakers
Conclusion

• We have introduced background work on audio – video source separation in which both linear and non-linear filtering is exploited with multiple microphone measurement.

• We have demonstrated that the video modality can improve the quality of the soft masks obtained in sparsity-based time-frequency source separation, which only requires binaural mixtures.

• More extensive evaluation is required for highly reverberant situations and further enhancements to the video processing in multi-source environments is required.
Journal Publications


• Ata-ur-Rehman, S. M. Naqvi, M. Yu, M. S. Khan, L. Mihaylova and J. A. Chambers, Multi-target tracking for variable number of targets with occlusion reasoning by using variational Bayesian clustering and a feature-based JPDAF, IEEE Journal of Selected Topics in Signal Processing 2012 (submitted).
Thank you for your attention!

Acknowledgement: Various sources for visual material