UDRC Summer School, Edinburgh, 23-27 June, 2014

Audio-Visual and Sparsity based Source Separation

Wenwu Wang
Senior Lecturer in Signal Processing
Centre for Vision, Speech and Signal Processing
Department of Electronic Engineering
University of Surrey, Guildford

w.wang@surrey.ac.uk
http://personal.ee.surrey.ac.uk/Personal/W.Wang/

25/06/2014
Outline

- Introduction
  - Cocktail party problem, source separation, time-frequency masking
  - Why audio-visual BSS (AV-BSS)
- AV-ICA
- Dictionary learning (AVDL) based AV-BSS
  - Audio-visual dictionary learning
  - Time-frequency mask fusion
- Results and demonstrations
- Conclusions and future work
Introduction----Cocktail party problem

- Independent component analysis (ICA)
- Time-frequency (TF) masking

```
Unknown mixing
s_1(n) -> x_1(n) -> y_1(n)
```
```
Unknown mixing
s_2(n) -> x_2(n) -> y_2(n)
```

Sources Observations Source estimates

www.surrey.ac.uk
BSS using TF masking

Sparsity assumption ------ each TF point is dominated by one source signal.
\[
\frac{X_1(m, \omega)}{X_2(m, \omega)} \rightarrow \alpha(m, \omega), \beta(m, w)
\]
Adverse effects

- Acoustic noise
- Reverberations

Why AV-BSS?----AV coherence

Audio-visual (AV) coherence

Perception
Why AV-BSS?

- The audio-domain BSS algorithms degrade in adverse conditions.
- The visual stream contains complementary information to the coherent audio stream.

Objective

How can the visual modality be used to assist audio-domain BSS algorithms in noisy and reverberant conditions?

Potential applications

AV-BSS

Surveillance

Robot audition

Hello world

AV speech recognition

HCI

Key Challenges

- Reliable AV coherence modelling
- Bimodal differences in size, dimensionality and sampling rates
- Fusion of AV coherence with audio-domain BSS methods

www.surrey.ac.uk
Visual Information to Resolve the Permutation Problem

Observations $x(n)$

STFT

$X(f, t)$

ICA

$Y(f, t), W(f)$

Permutation alignment

$W(n)$

IDFT

$y(n) = \hat{S}(n)$

Audio feature extraction

$s(n)$

Features fusion

$p(a(t), v(t))$

Audio-Visual coherence

Off-line training process

Visual feature extraction

$v(t)$

Training visual stimulus

Training audio stimulus

Recorded visual stimulus

www.surrey.ac.uk
Feature Extraction

• **Visual feature extraction**
  – Internal lip Width and Height
  – 2-Dimensional
    \[ v_T(m) = [LW(m), LH(m)]^T \]

• **Audio feature extraction**
  – Mel-scale Frequency Cepstrum Coefficients (MFCCs)
  – Block processing (synchronize with each video frame)
  – L-dimensional
    \[ a_T(m) = [a_{T1}(m), ..., a_{TL}(m)]^T \]

• **Audio-visual space-----Feature Selection**
Robust AV Feature Selection
$p(a(m), v(m)) = p(u(m)) = \sum_{d=1}^{D} w_d N(u(m) | \mu_d, \Sigma_d)$
Resolution of the permutation problem

**Objective**

\[
\hat{P}(\omega) = \arg \max_P \sum_m \sum_{k=1}^{K} p(u_k(m))
\]

**Solution: An iterative sorting scheme**
FD-BSS using ICA

$$x_1(n)$$

$$x_2(n)$$

Short Time Fourier Transform (STFT)

ICA

$$\hat{S}_1(m, \omega) = S_1(m, \omega_1)$$

$$\hat{S}_2(m, \omega) = S_2(m, \omega_2)$$

$$\hat{S}_1(m, \omega) = S_2(m, \omega_2)$$

$$\hat{S}_2(m, \omega) = S_1(m, \omega_2)$$

$$\hat{S}_1(m, \omega) = S_1(m, \omega_1)$$

$$\hat{S}_2(m, \omega) = S_2(m, \omega_2)$$
Resolution of the permutation problem

This is the 1-st section of STEP 1.

(STEP N gets $2^{N-1}$ sections.)

Now we get two source estimates $Y_1(f,t)$ and $Y_2(f,t)$. Suppose the blue curves represent frequency components coming from source 1, while the red curves are from source 2. We find that $Y_1(f,t)$ contains components from both source 1 and source 2.
AVDL based BSS

TF masking, Mandel et al. 2010.
Figures taken from ICASSP 2013 Tutorial 11, by Dai, Maihe and Wang. Likewise for next four pages. Acknowledgement to Wei Dai for making these figures.
A two-stage procedure

Sparse coding (approximation)

\[
\min \| X \|_0 \quad \text{s.t.} \quad \| Y - DX \|_F^2 \leq \epsilon.
\]

Greedy algorithms:
- Subspace pursuit (SP) W. Dai and O. Milenkovic 2009
- CoSaMP  D. Needell and J. Tropp 2009
- IHT  T. Blumensath and M. Davies 2009
Dictionary update: the formulation

- **Constraints:**
  - Fixed sparsity pattern
    \[
    \Omega = \{(i, j) : X_{i,j} \neq 0\}, \\
    \mathcal{X}_\Omega = \{X : X_{i,j} = 0, \forall (i, j) \in \Omega^c\}.
    \]
  - Unit norm codewords
    \[
    \mathcal{D} = \{D : \|D_{:,j}\|_2 = 1, \forall j \in [d]\}.
    \]
- **Dictionary Update:**
  \[
  \min_{D \in \mathcal{D}, X \in \mathcal{X}_\Omega} \|Y - DX\|_F^2.
  \]
Dictionary update: K-SVD algorithm

\[
\begin{align*}
\|Y - DX\|^2 &= \|Y - D_{:,j \neq i} X_{j \neq i} - D_{:,i} X_{i,:}\|^2 \\
&= \|Y_r - D_{:,i} X_{i,:}\|^2 \\
&= \|(Y_r)_{:,J} - D_{:,i} X_{i,J}\|^2 + c
\end{align*}
\]
Sparse assumption of AVDL
Flow of the AVDL

The coding process relies on the matching criterion, how well an atom fits the signal in the MP algorithm.

A scanning index is proposed to reduce the computational complexity.

The learning process uses two different update methods, to accommodate different bimodality sparsity constraints.

By mapping the AV sequence to the learned dictionary, a visual mask can be achieved.
The overall algorithm

Algorithm 1: Framework of the Proposed AVDL

Input: A training AV sequence $\psi = (\psi^a; \psi^v)$, an initial $D$ with $K$ atoms, and the number of non-zero coefficients $N$

Output: An AV dictionary $D = \{\phi_k\}_{k=1}^K$

1. **Initialization**: $\text{iter} = 1$, $\text{MaxIter}$
2. while $\text{iter} \leq \text{MaxIter}$ do
3. %Coding stage
4. Given $D$, decompose $\psi$ using (1) to obtain $\Omega$.
5. %Learning stage
6. Given $\Omega$ and the residual $v$, update $D = \{\phi_k\}$ for $k = 1, 2, \ldots, K$ to fit model (1).
7. $\text{iter} = \text{iter} + 1$
The coding process

\[ J^{av}(\vec{v}_{\bar{y}\bar{x}\bar{l}\bar{m}}, \phi_k) = J^a(\vec{v}_{\bar{m}}, \phi_k^a) J^v(\vec{v}_{\bar{y}\bar{x}\bar{l}}, \phi_k^v), \]

\[ J^a_{\text{Mon}} = |\langle \vec{v}_{\bar{m}}^a, \phi_k^a \rangle| \]

\[ J^v(\vec{v}_{\bar{y}\bar{x}\bar{l}}, \phi_k^v) = \exp \left\{ \frac{-1}{YXL} \left\| \vec{v}_{\bar{y}\bar{x}\bar{l}}^v - \phi_k^v \right\|_1 \right\} . \]

\[ [k_n, y_n, x_n, l_n, m_n] = \arg \max \ J^{av}(\vec{v}_{\bar{y}\bar{x}\bar{l}\bar{m}}, \phi_k), \]

\[ B(k_n, y_n, x_n, l_n) = 1 \]

\[ C(k_n, m_n) = J^a(\vec{v}_{m_n}^a, \phi_{k_n}^a). \]

\[ \overline{v}_{l_n}^a \leftarrow \overline{v}_{l_n}^a - C(k_n, l_n) \phi_{k_n}^a. \]
The coding process (algorithm)

Algorithm 2: The Coding State of the Proposed AVDL

**Input:** An AV sequence \( \psi \), the dictionary \( \mathcal{D} = \{ \phi_k \}_{k=1}^{K} \), the threshold \( \delta \), the number of non-zero coefficients \( N \)

**Output:** The coding parameter set \( \Omega = \{ B, C \} \) and residual \( \nu \)

1. **Initialization:** Set \( \Omega \) with zero tensors,
   \[ \nu = \psi, n = 1, J_{opt} = J_{max} = 0 \]
2. Calculate \( S^{av} \) using (10) to (13).
3. **while** \( n \leq N \) and \( J_{opt} \geq \delta J_{max} \) **do**
4. **% Projection**
5. \[ \mathcal{L} = \begin{cases} \{1 : L_s\}, & n=1 \\ l_{n-1} + \{1 - L : L - 1\}, & \text{otherwise} \end{cases} \]
6. **for** \( k \leftarrow 1 \) **to** \( K \) **do**
7. **foreach** \( \tilde{l} \in \mathcal{L} \) **do**
8. **Calculate** \( J^a(\tilde{v}_{\tilde{m}}, \phi^a_k) \), where \( \tilde{m} \) is tied with \( \tilde{l} \) via set (2).
9. **foreach** \( (\tilde{y}, \tilde{x}), \tilde{y} \in \{1 : Y_s\}, \tilde{x} \in \{1 : X_s\} \) **do**
10. **if** \( S^{av}(\tilde{y}, \tilde{x}, \tilde{l}) = 1 \) **then**
11. Obtain \( J^v(\tilde{v}_{\tilde{y}\tilde{x}\tilde{l}}, \phi^v_k) \) via (6) and \( J^{av}(\tilde{v}_{\tilde{y}\tilde{x}\tilde{l}m}, \phi_k) \) via (5).
12. **% Selection**
13. Obtain \( [y_n, x_n, l_n, k_n, m_n] \) via (7).
14. Update \( \Omega \) via (8).
15. Residual calculation via (9).
16. \( J_{opt} = J^{av}(\tilde{v}_{y_n x_n l_n m_n}, \phi_k) \)
17. **if** \( n = 1 \) **then**
18. \( J_{max} = J^{av}(\tilde{v}_{y_1 x_1 l_1 m_1}, \phi_k) \)
19. \( n = n + 1 \)
Algorithm 3: The Learning Stage of the Proposed AVDL.

**Input:** The parameter set $\Omega = \{B, C\}$, the residual $v$, the old dictionary $D = \{\phi_k\}_{k=1}^K$

**Output:** A new dictionary $D$

1 **Initialization:** $k = 1$
2 **while** $k \leq K$ **do**
3    Update $\phi_k^a$, $C$ and $v$ via K-SVD using (14) to (17).
4    Update $\phi_k^v$ via the K-means algorithm
5    $\phi_k^v = \text{Mean} \left( b_{k \tilde{y} \tilde{x} \tilde{l}} \hat{v}_{k \tilde{y} \tilde{x} \tilde{l}}^v \right)$, subject to $b_{k \tilde{y} \tilde{x} \tilde{l}} \neq 0$, $\forall (\tilde{y}, \tilde{x}, \tilde{l})$
6    $k = k + 1$

\[
\bar{v}_m^a \leftarrow \bar{v}_m^a + c_{km} \phi_k^a, \forall m. \quad \phi_k^a \leftarrow \text{ivec}(u_k | \phi_k^a).
\]

\[
\gamma_k \approx \lambda_k u_k v_k^T, \quad \bar{v}_m^a \leftarrow \bar{v}_m^a - c_{km} \phi_k^a, \forall m.
\]
AVDL evaluations

Synthetic data

(a) AV: /a/
(b) AV: /i/
(c) AV: /o/
(d) Visual only
(e) Audio only:

(f) The generated AV synthetic sequence (only one second data is shown)
AVDL evaluations

Additive noise added

(a) AVDL: /a/

(b) AVDL: /i/

(c) AVDL: /o/

(d) Monaci: /a/

(e) Monaci: /i/

(f) Monaci: /o/
AVDL evaluations

Convolutive noise added

(a) AVDL1
(b) AVDL2
(c) AVDL3
(d) Monaci1
(e) Monaci2
(f) Monaci3
(g) Monaci4
AVDL evaluations

The approximation error metrics comparison of AVDL and Monaci's method over 50 independent tests over the synthetic data.

The proposed AVDL outperforms the baseline approach, giving an average of 33% improvement for the audio modality, together with a 26% improvement for the visual modality.
AVDL evaluations
AV mask fusion for AVDL-BSS

**Audio mask**
Statistically generated by evaluating the IPD and ILD of each TF point.

**Visual mask**
Mapping the observation to the learned AV dictionary via the coding stage in AVDL.
Visual mask generation

\[ M^v(m, \omega) = \begin{cases} 
1, & \text{if } \hat{\psi}^a(m, \omega) > \psi^a(m, \omega) \\
\hat{\psi}^a(m, \omega) / \psi^a(m, \omega), & \text{otherwise.}
\end{cases} \]

The first AV atom represents the utterance "marine" /ˈmɑːriːn/ while the second one denotes the utterance "port" /pɔːt/.

Sheerman-Chase et al. LILiR Twotalk database 2011

Lip tracking, Ong et al. 2008
Demonstration of TF mask fusion in AVDL-BSS

Why do we choose the power law combination, instead of, e.g., a linear combination?
AVDL-BSS evaluations----SDR

Noise-free

10 dB Gaussian noise
AVDL-BSS evaluations----OPS-PEASS

Noise-free

10 dB Gaussian noise

Overall perceptual score (OPS-PEASS) vs. Angle [deg]

- Mandel
- AVDL-BSS
- Visual-only
- Monaci-BSS
- AV-LIU
Some examples

<table>
<thead>
<tr>
<th></th>
<th>Mixture</th>
<th>Ideal</th>
<th>Mandel</th>
<th>AV-LIU</th>
<th>AVDL-BSS</th>
<th>Rivet</th>
<th>AVMP-BSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td><img src="image1.png" alt="Icon" /></td>
<td><img src="image2.png" alt="Icon" /></td>
<td><img src="image3.png" alt="Icon" /></td>
<td><img src="image4.png" alt="Icon" /></td>
<td><img src="image5.png" alt="Icon" /></td>
<td><img src="image6.png" alt="Icon" /></td>
<td><img src="image7.png" alt="Icon" /></td>
</tr>
<tr>
<td>B</td>
<td><img src="image1.png" alt="Icon" /></td>
<td><img src="image2.png" alt="Icon" /></td>
<td><img src="image3.png" alt="Icon" /></td>
<td><img src="image4.png" alt="Icon" /></td>
<td><img src="image5.png" alt="Icon" /></td>
<td><img src="image6.png" alt="Icon" /></td>
<td><img src="image7.png" alt="Icon" /></td>
</tr>
<tr>
<td>C</td>
<td><img src="image1.png" alt="Icon" /></td>
<td><img src="image2.png" alt="Icon" /></td>
<td><img src="image3.png" alt="Icon" /></td>
<td><img src="image4.png" alt="Icon" /></td>
<td><img src="image5.png" alt="Icon" /></td>
<td><img src="image6.png" alt="Icon" /></td>
<td><img src="image7.png" alt="Icon" /></td>
</tr>
<tr>
<td>D</td>
<td><img src="image1.png" alt="Icon" /></td>
<td><img src="image2.png" alt="Icon" /></td>
<td><img src="image3.png" alt="Icon" /></td>
<td><img src="image4.png" alt="Icon" /></td>
<td><img src="image5.png" alt="Icon" /></td>
<td><img src="image6.png" alt="Icon" /></td>
<td><img src="image7.png" alt="Icon" /></td>
</tr>
</tbody>
</table>
Summary

- AV provides alternative solutions to address permutation ambiguities in BSS
- AVDL offers an alternative and effective method for modelling the AV coherence within the audio-visual data.
- The mask derived from AVDL can be used to improve the BSS performance for separating reverberant and noisy speech mixtures

Future work

- To achieve dictionary adaptation and source separation simultaneously
Acknowledgement

• Collaborators: Dr Qingju Liu, Dr Philip Jackson, Dr Mark Barnard, Prof Josef Kittler, Prof Jonathon Chambers (Loughborough University), Dr Syed Mohsen Naqvi (Loughborough University), and Dr Wei Dai (Imperial College London)

• Financial support: EPSRC & DSTL, UDRC in Signal Processing
Thank you

Q & A

w.wang@surrey.ac.uk