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

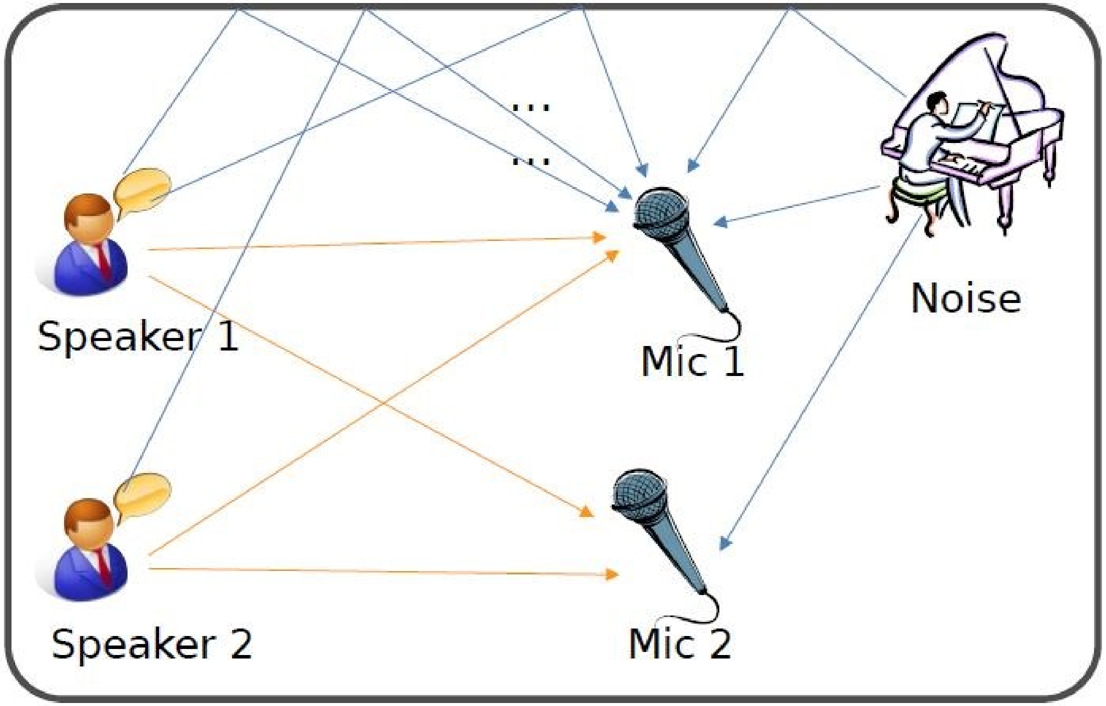
w.wang@surrey.ac.uk
http://personal.ee.surrey.ac.uk/Personal/W.Wang/

25/06/2014

Outline

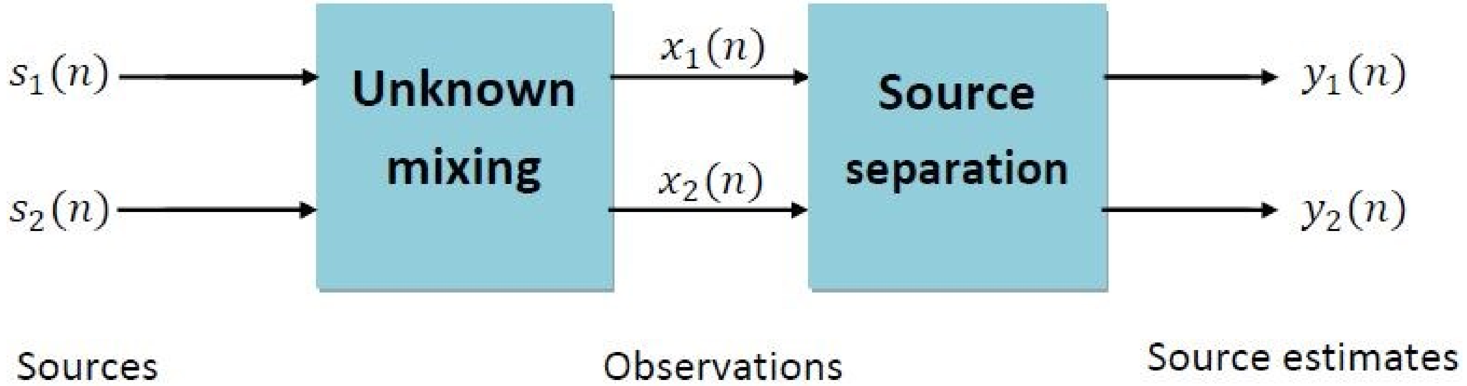
- Introduction
  - Cocktail party problem, source separation, time-frequency masking
  - Why audio-visual BSS (AV-BSS)
- AV-ICA
- Dictionary learning (AVDL) based AV-BSS
  - Audio-visual dictionary learning
  - Time-frequency mask fusion
- Results and demonstrations
- Conclusions and future work

Introduction----Cocktail party problem

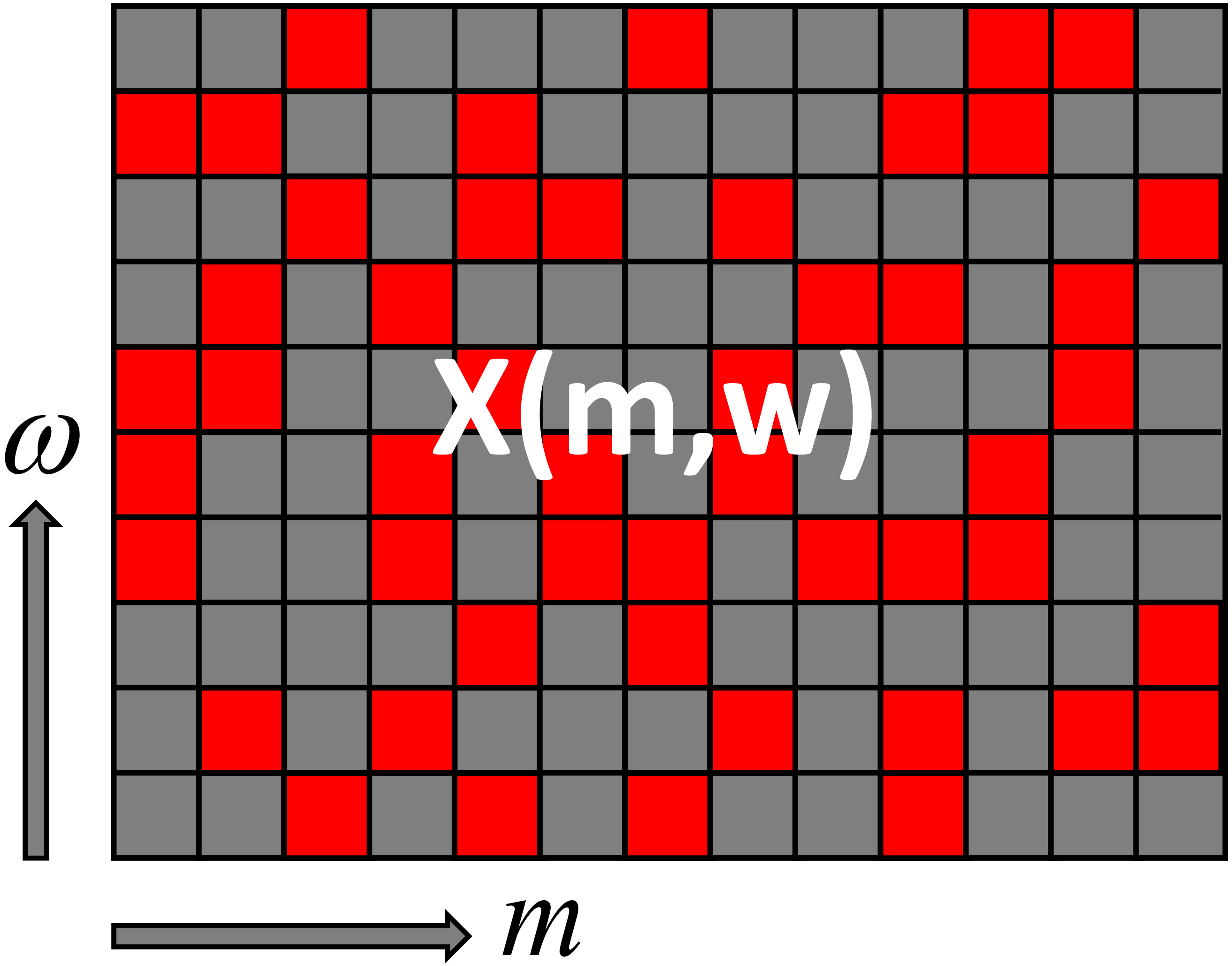


- ➢Independent component analysis (ICA)
- ➢Time-frequency (TF) masking

"Blind" source separation
BSS



BSS using TF masking



CASA

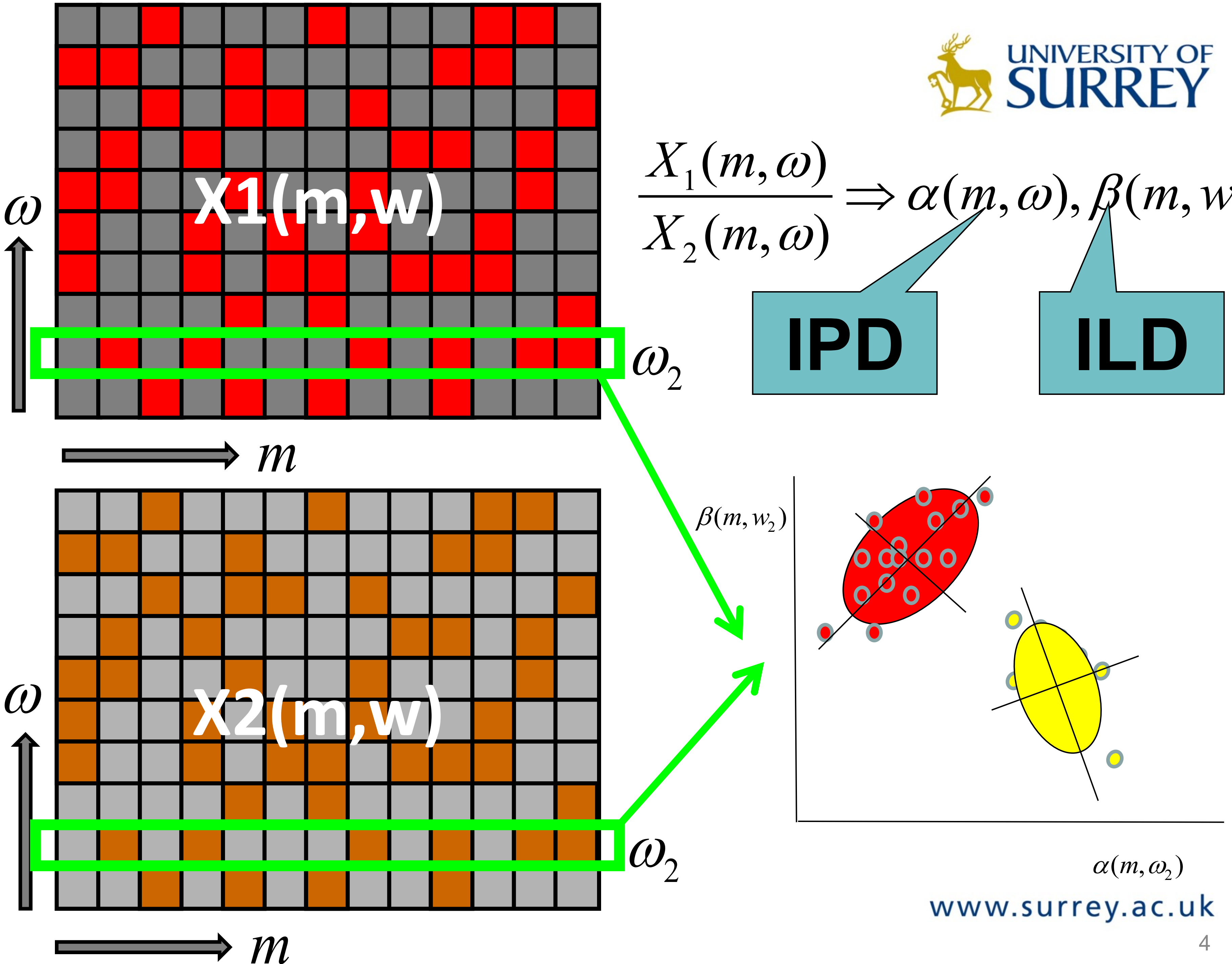
Onset
Periodicity
Harmonicity
Locations
Binarual cues

Sparsity assumption ------ each TF point is dominated by one source signal.

$$\frac{X_1(m,\omega)}{X_2(m,\omega)} \Rightarrow \alpha(m,\omega), \beta(m,w)$$

IPD

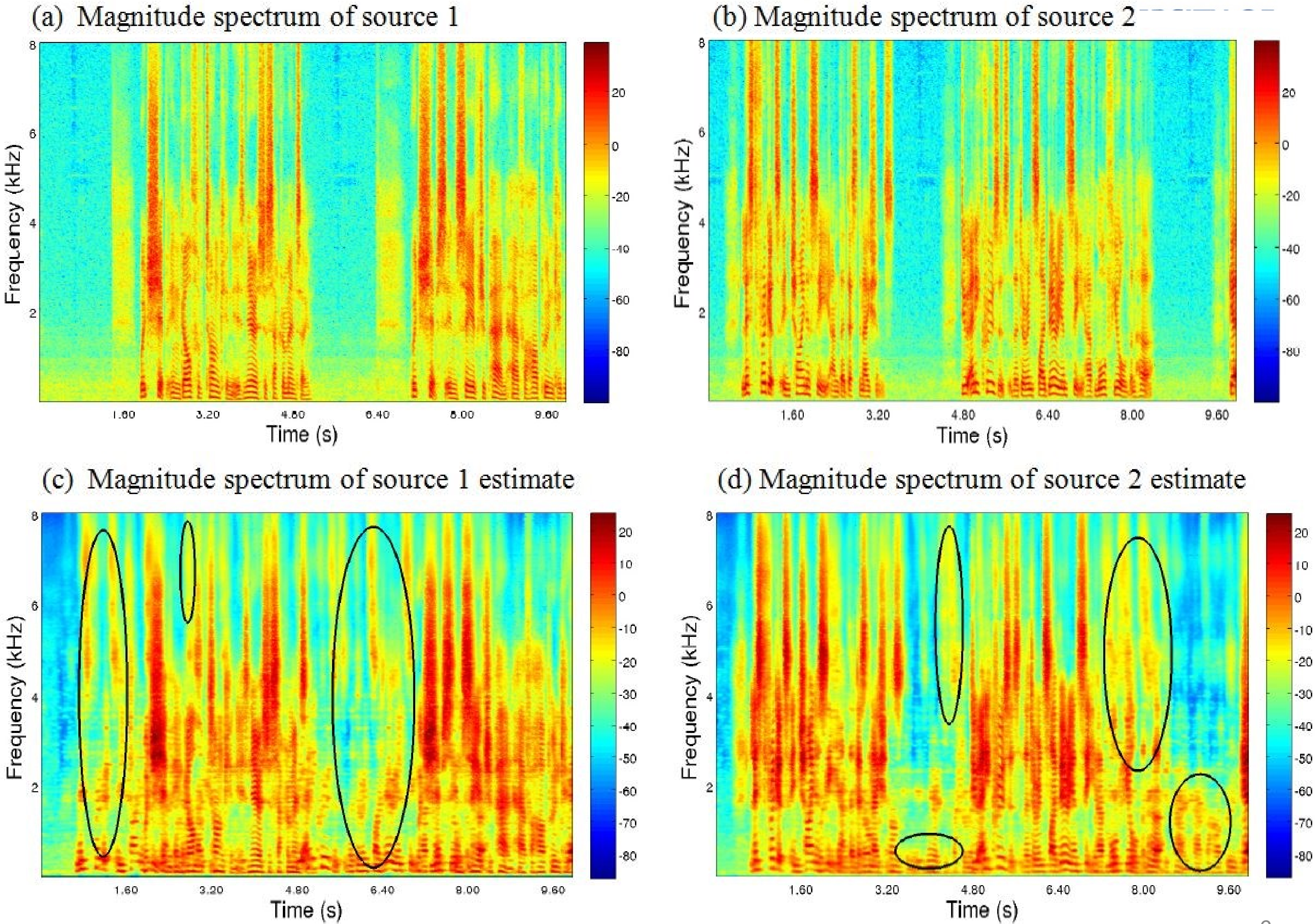
ILD



Adverse effects

- Acoustic noise
- Reverberations

- W. Wang, D. Cosker, Y. Hicks, S. Sanei, and J. A. Chambers, "Video Assisted Speech Source Separation," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP 2005), vol. V, pp.425-428, Philadelphia, USA, March 18-23, 2005.
- Q. Liu, W. Wang, and P. Jackson, "Use of Bimodal Coherence to Resolve Permutation Problem in Convolutive BSS," Signal Processing, vol. 92, vol. 8, pp. 1916-1927, 2012.
- Q. Liu, W. Wang, P. Jackson, M. Barnard, J. Kittler, and J.A. Chambers, “Source separation of convolutive and noisy mixtures using audio-visual dictionary learning and probabilistic time-frequency masking", *IEEE Transactions on Signal Processing*, vol. 61, no. 22, pp. 5520-5535, 2013.
- B. Rivet, W. Wang, S.M. Naqvi, and J.A. Chambers, "Audio-Visual Speech Source Separation", *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 125-134, 2014.
- Q. Liu, A. Aubery, and W. Wang, "Interference Reduction in Reverberant Speech Separation with Visual Voice Activity Detection", *IEEE Transactions on Multimedia*, 2014. (in press)



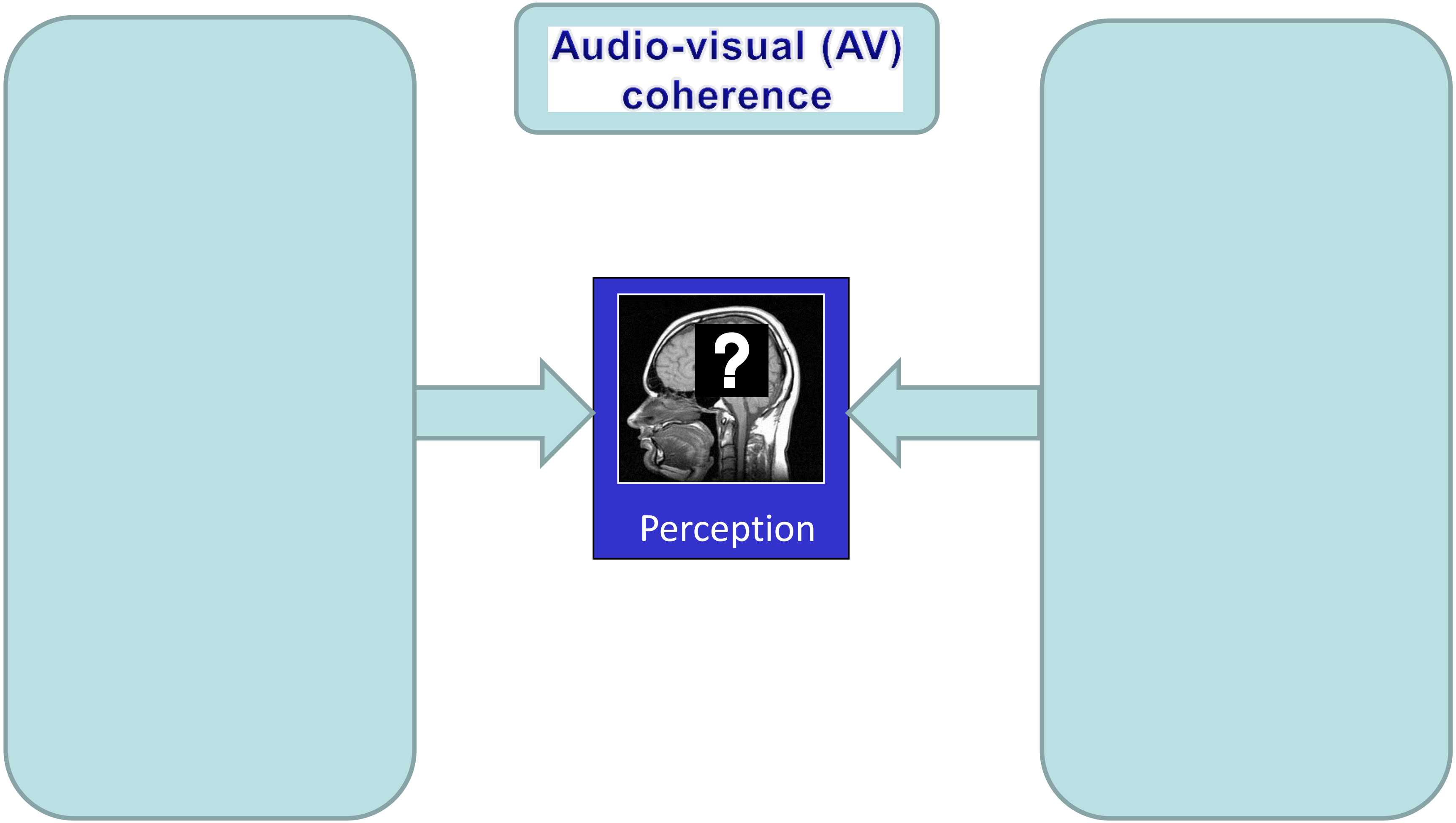
(a) Magnitude spectrum of source 1

(b) Magnitude spectrum of source 2

(c) Magnitude spectrum of source 1 estimate

(d) Magnitude spectrum of source 2 estimate

Why AV-BSS?----AV coherence



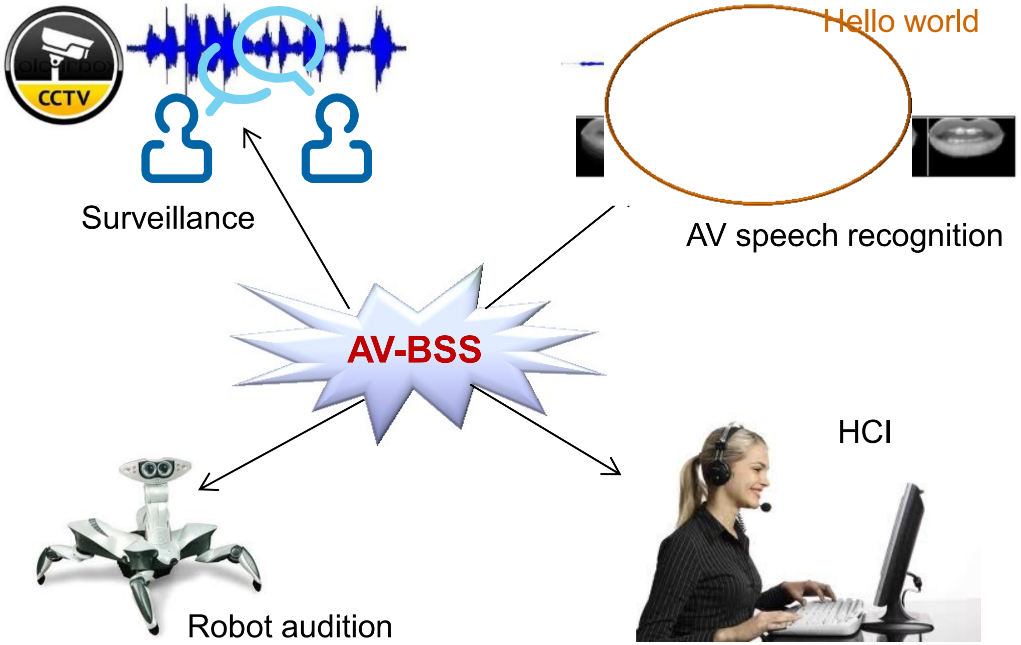
Audio-visual (AV) coherence

Perception

Why AV-BSS?

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

Potential applications



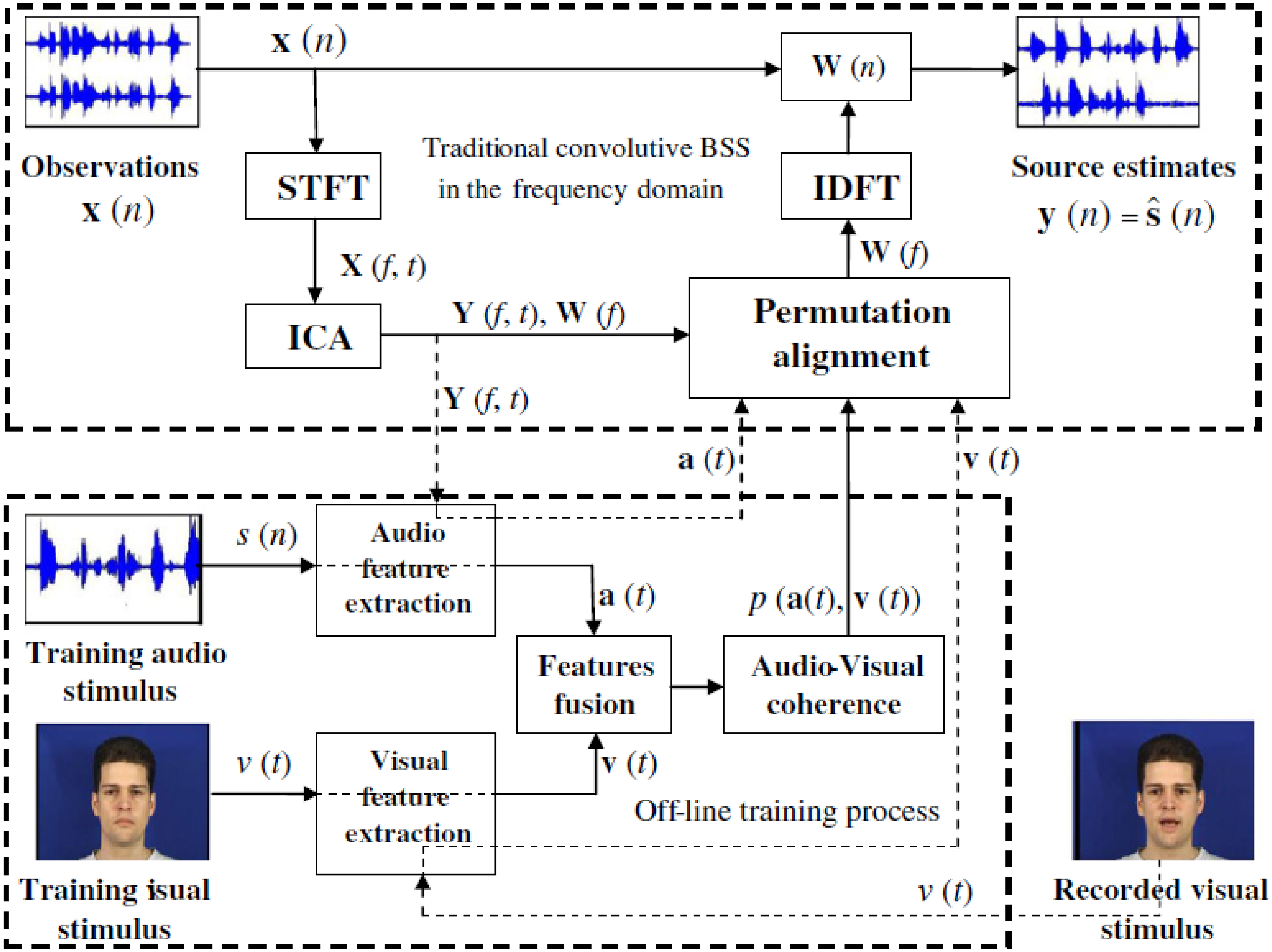
Objective

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

Key Challenges

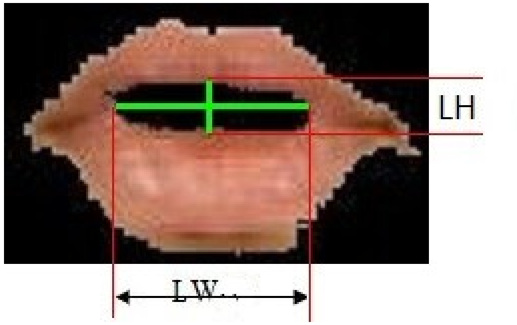
- Reliable **AV coherence modelling**
- **Bimodal differences** in size, dimensionality and sampling rates
- **Fusion of AV coherence** with audio-domain **BSS** methods

Visual Information to Resolve the Permutation Problem



Feature Extraction

- **Visual feature extraction**
  - **Internal lip Width and Height**
  - **2-Dimensional**



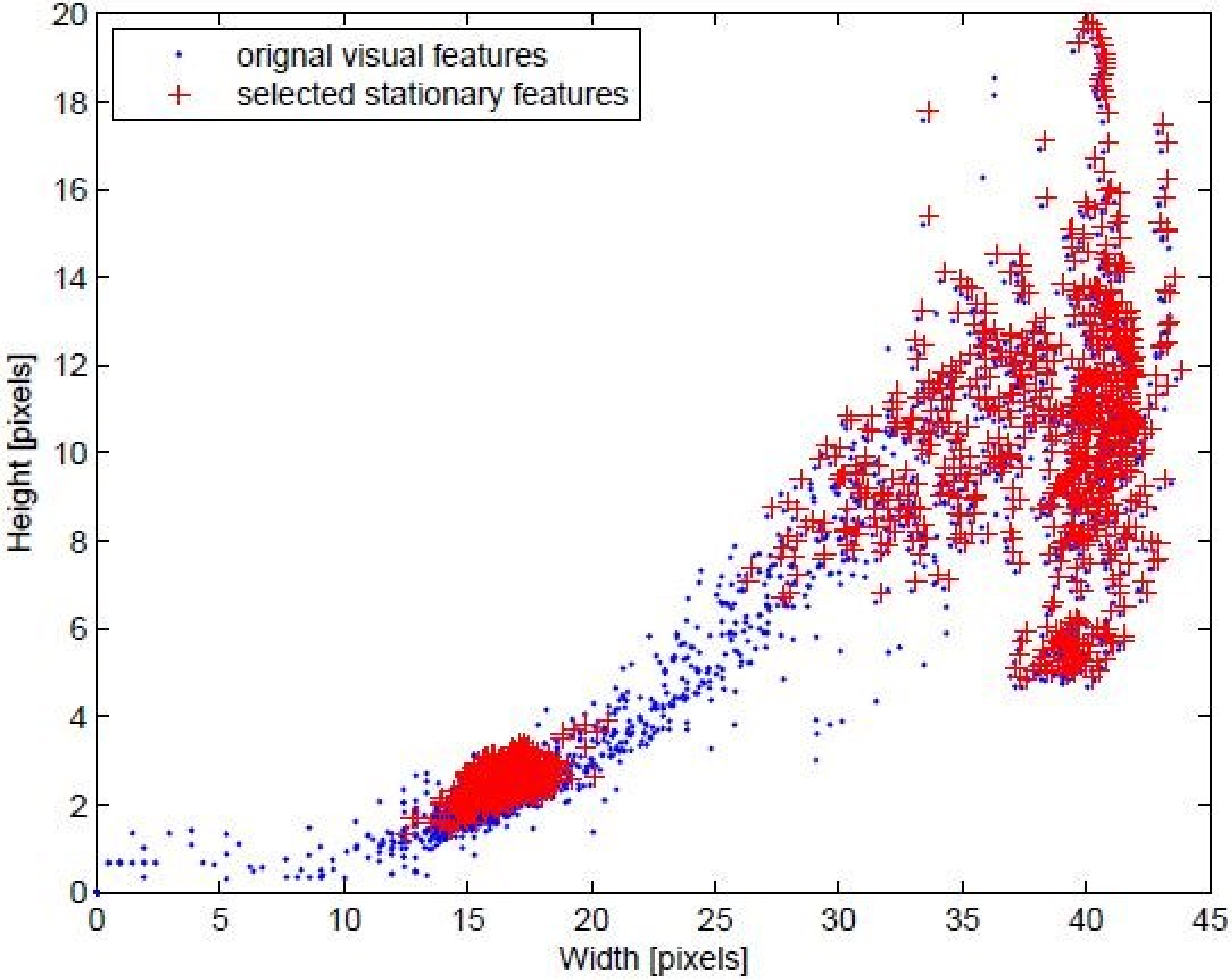
$$\mathbf{v}_{\mathrm{T}}(m) = [\mathrm{LW}(m), \mathrm{LH}(m)]^{T}$$

- **Audio feature extraction**
  - **Mel-scale Frequency Cepstrum Coefficients (MFCCs)**
  - **Block processing (synchronize with each video frame)**
  - **L-dimensional**

$$\mathbf{a}_{\mathrm{T}}(m) = [a_{\mathrm{T}1}(m), \ldots, a_{\mathrm{T}L}(m)]^{T}$$

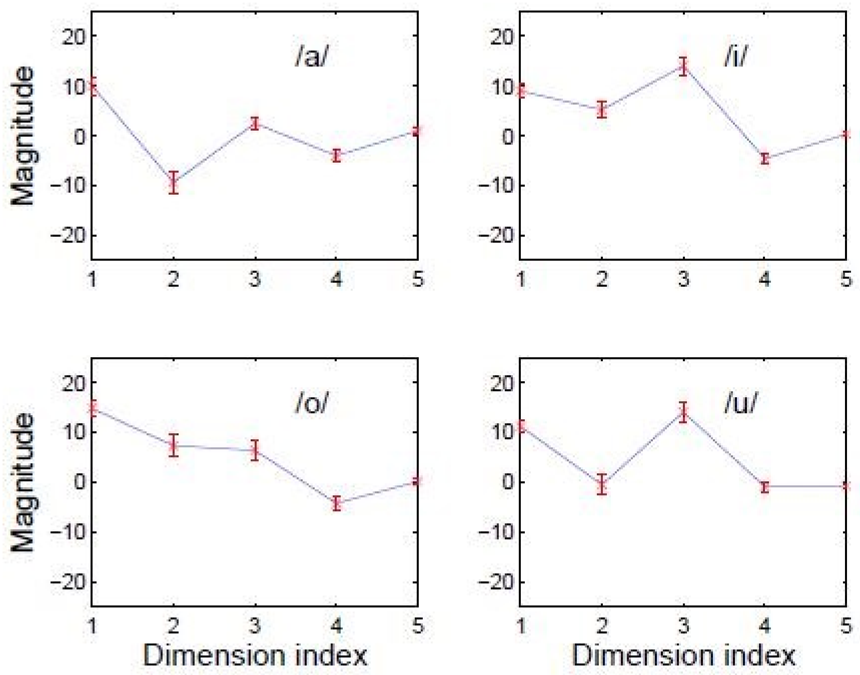
- **Audio-visual space-----Feature Selection**

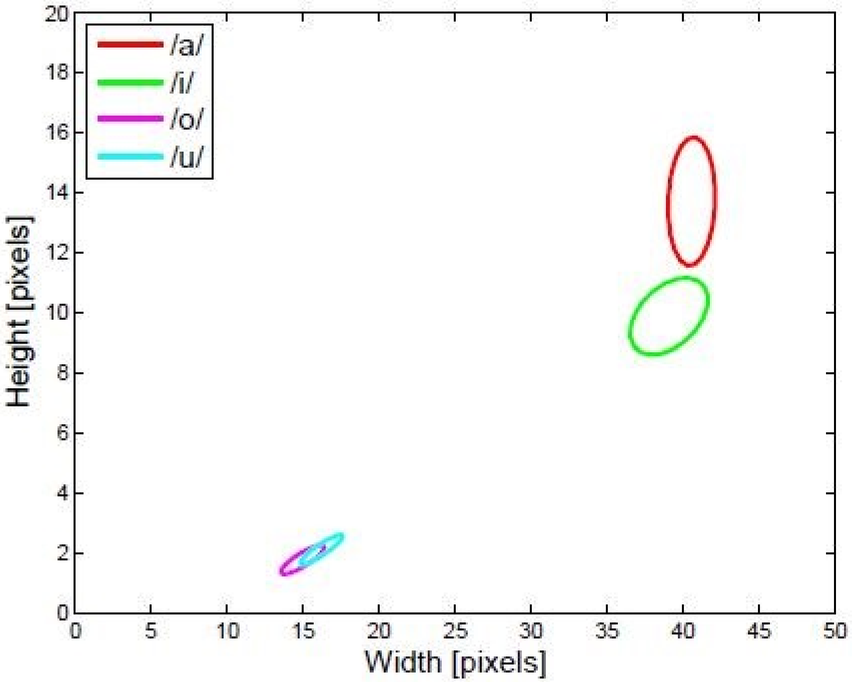
Robust AV Feature Selection



AV Coherence Modelling

$$p(\mathbf{a}(m), \mathbf{v}(m)) = p(\mathbf{u}(m)) = \sum_{d=1}^{D} w_d \mathcal{N}(\mathbf{u}(m) \mid \boldsymbol{\mu}_d, \boldsymbol{\Sigma}_d)$$



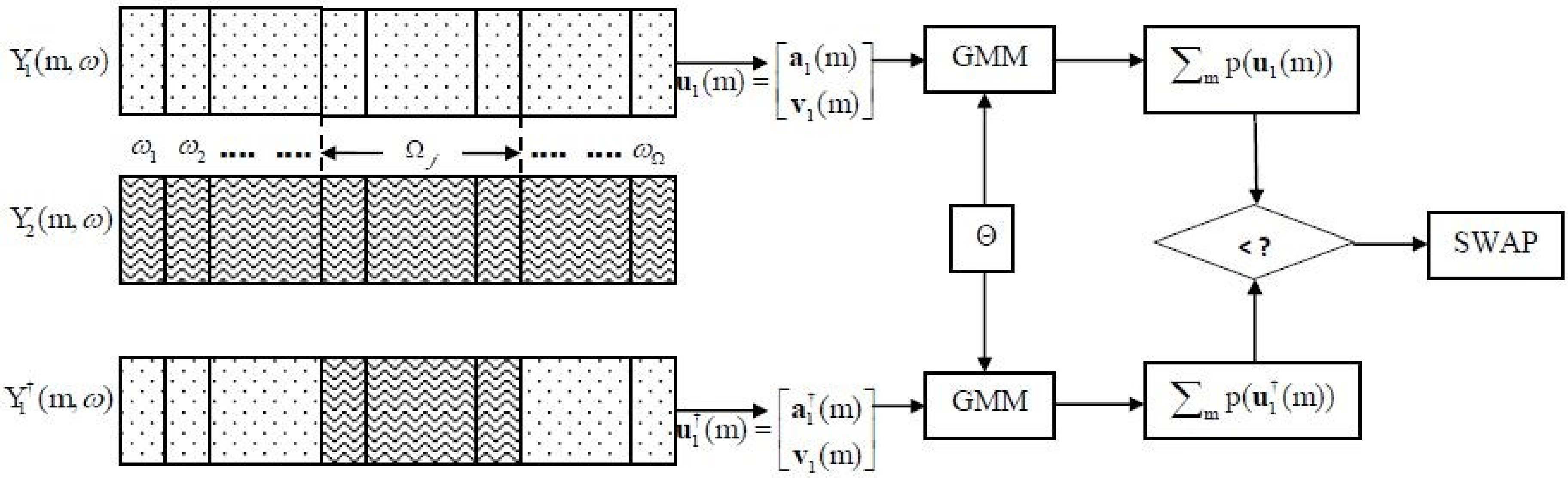


Resolution of the permutation problem

Objective

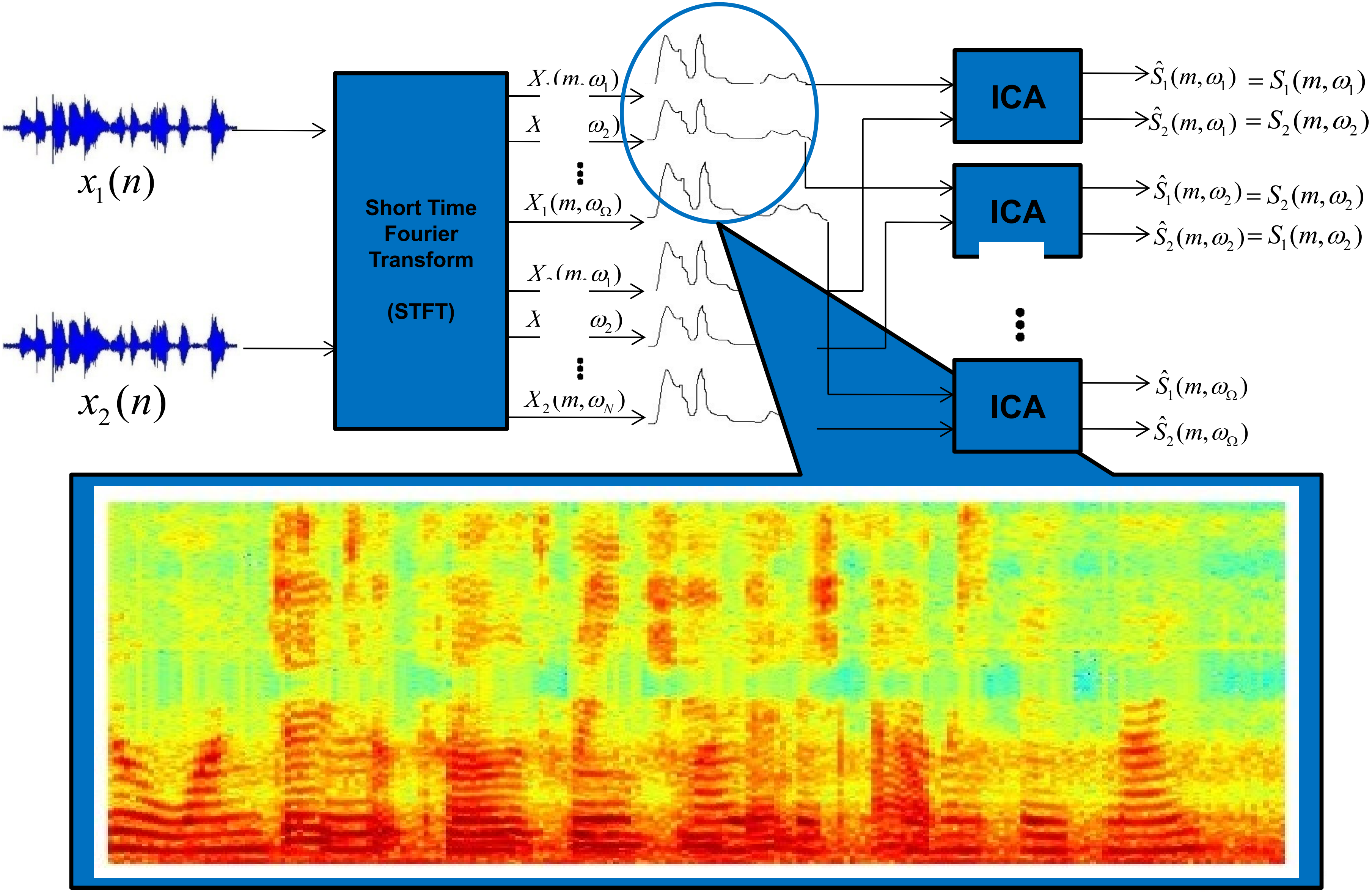
$$\hat{\mathbf{P}}(\omega) = \arg\max_{\mathbf{P}(\omega)} \sum_{m} \sum_{k=1}^{K} p(\mathbf{u}_k(m))$$

Solution: An iterative sorting scheme



$Y_1(m,\omega)$
ω_1 ω_2 ....
Ω_j
.... ω_Ω
$Y_2(m,\omega)$
$Y_1^{\dagger}(m,\omega)$
$\mathbf{u}_1(m) = \begin{bmatrix} \mathbf{a}_1(m) \\ \mathbf{v}_1(m) \end{bmatrix}$
GMM
$\sum_m p(\mathbf{u}_1(m))$
Θ
< ?
SWAP
$\mathbf{u}_1^{\dagger}(m) = \begin{bmatrix} \mathbf{a}_1^{\dagger}(m) \\ \mathbf{v}_1(m) \end{bmatrix}$
GMM
$\sum_m p(\mathbf{u}_1^{\dagger}(m))$

FD-BSS using ICA



$x_1(n)$

$x_2(n)$

Short Time Fourier Transform

(STFT)

$X_1(m,\omega_1)$

$X_1(m,\omega_2)$

$X_1(m,\omega_\Omega)$

$X_2(m,\omega_1)$

$X_2(m,\omega_2)$

$X_2(m,\omega_N)$

ICA

ICA

ICA

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

$\hat{S}_2(m,\omega_1) = S_2(m,\omega_2)$

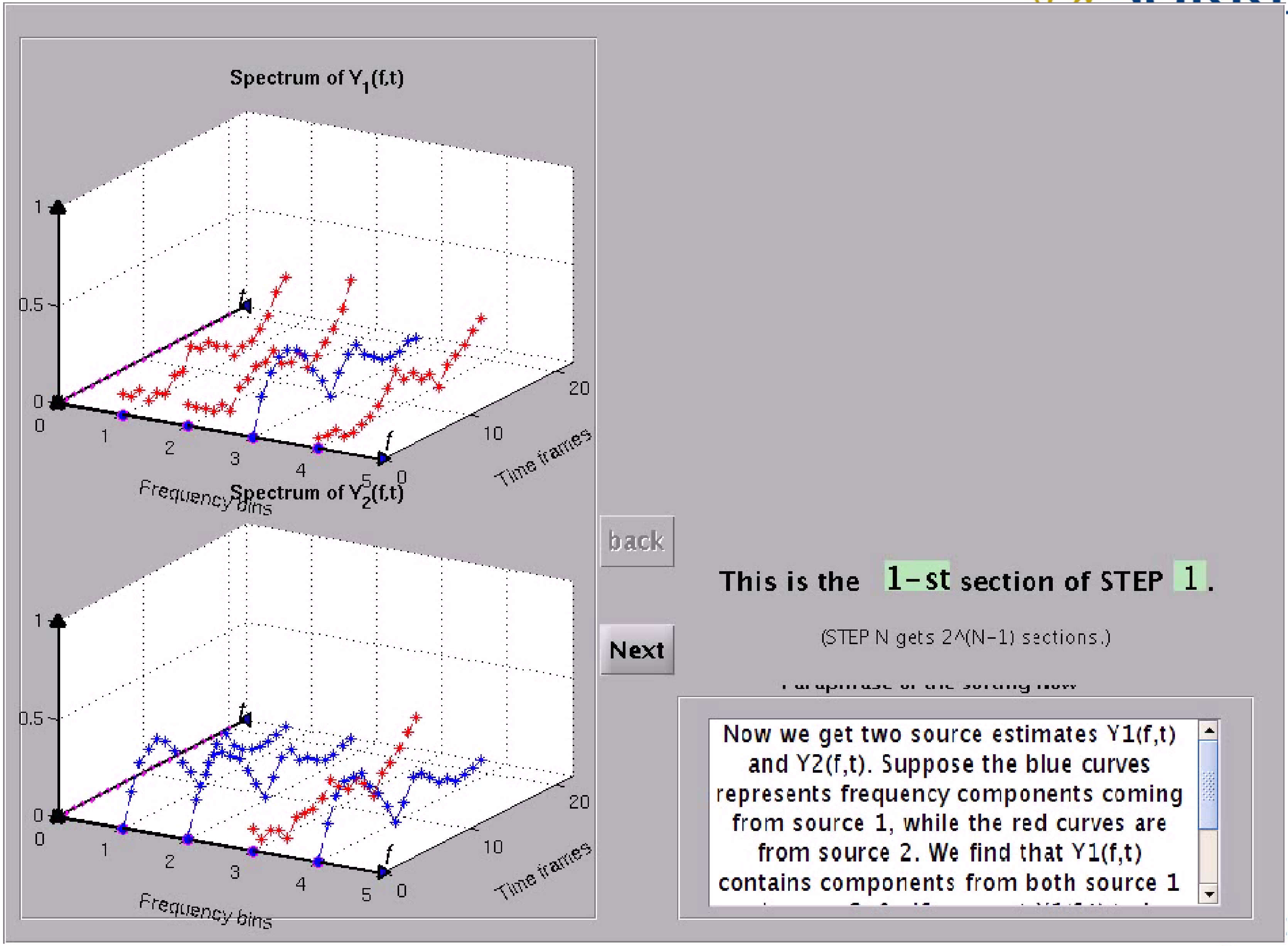
$\hat{S}_1(m,\omega_2) = S_2(m,\omega_2)$

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

$\hat{S}_1(m,\omega_\Omega)$

$\hat{S}_2(m,\omega_\Omega)$

Resolution of the permutation problem



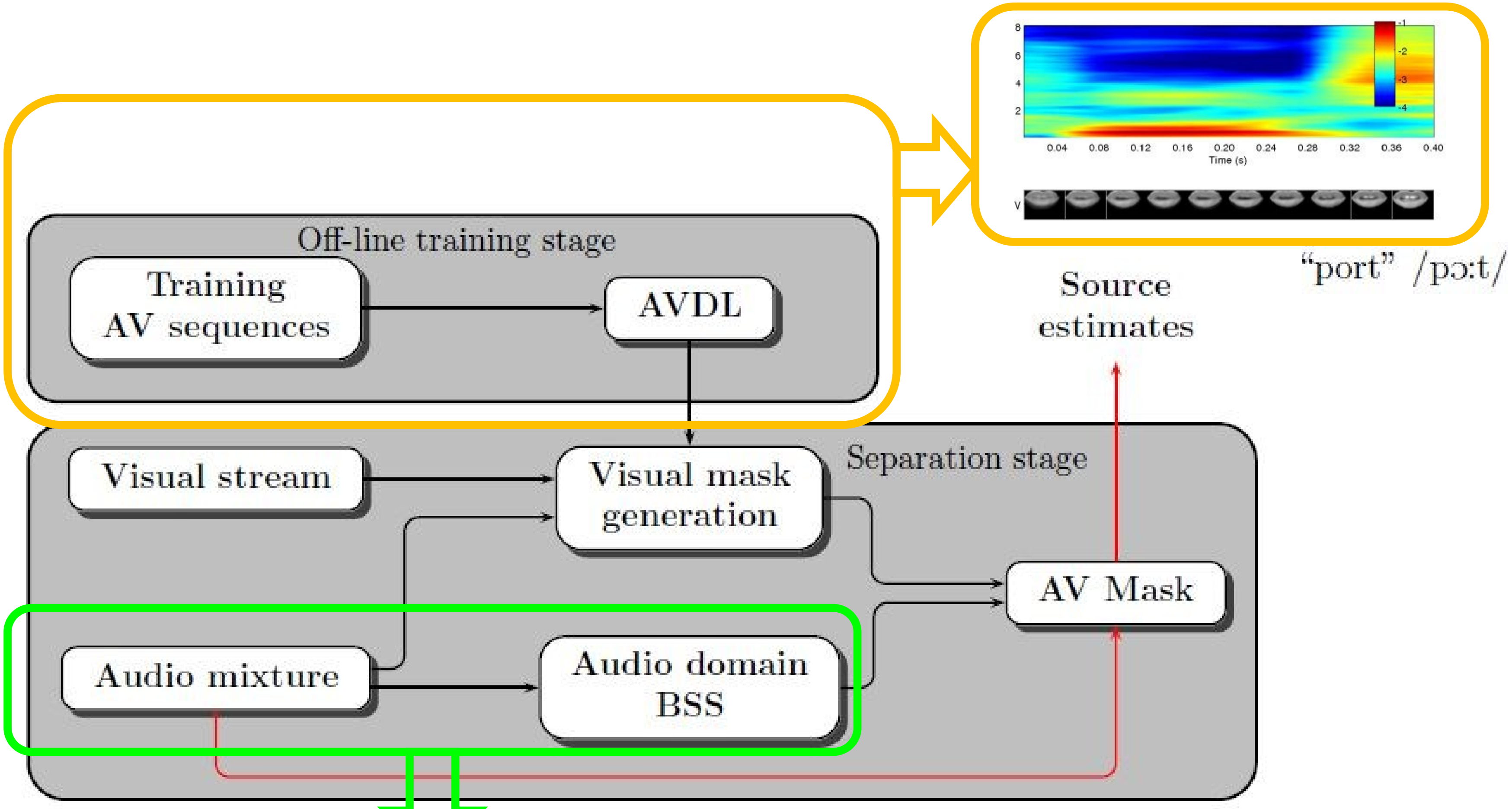
This is the 1-st section of STEP 1.

(STEP N gets 2^(N-1) sections.)

Paraphrase of the sorting now

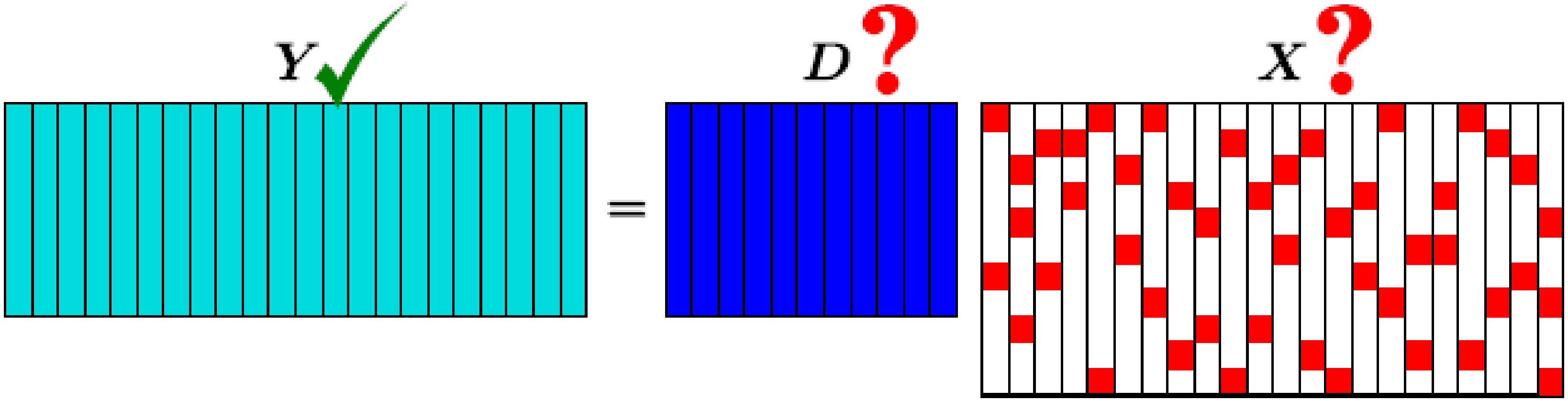
Now we get two source estimates Y1(f,t) and Y2(f,t). Suppose the blue curves represents frequency components coming from source 1, while the red curves are from source 2. We find that Y1(f,t) contains components from both source 1

AVDL based BSS



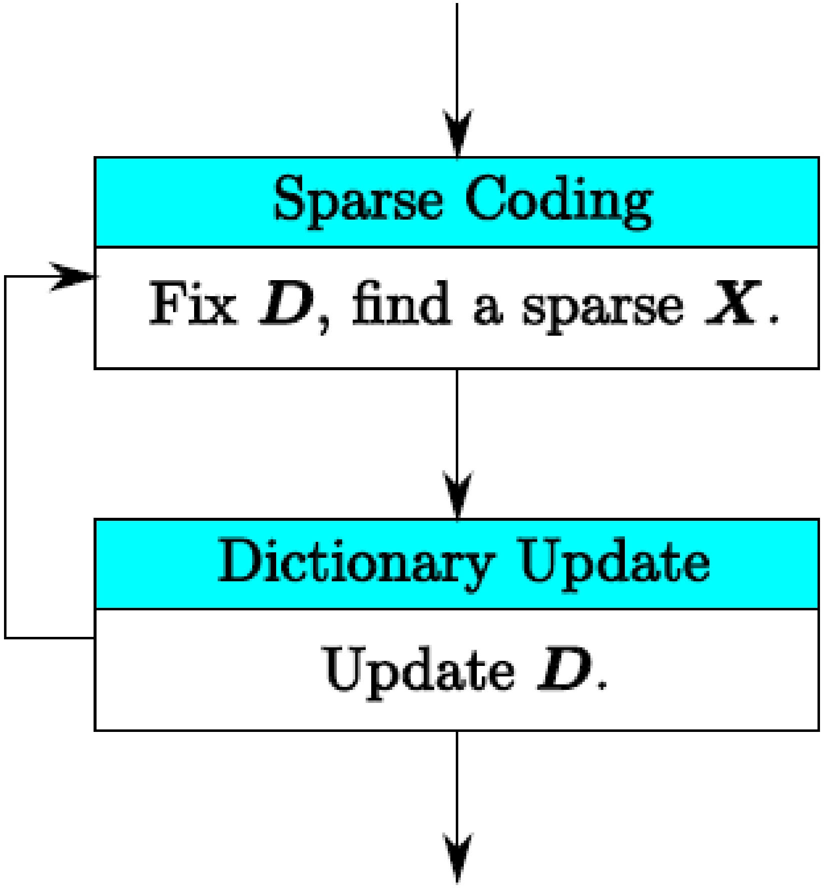
TF masking, Mandel et al. 2010.

Dictionary learning



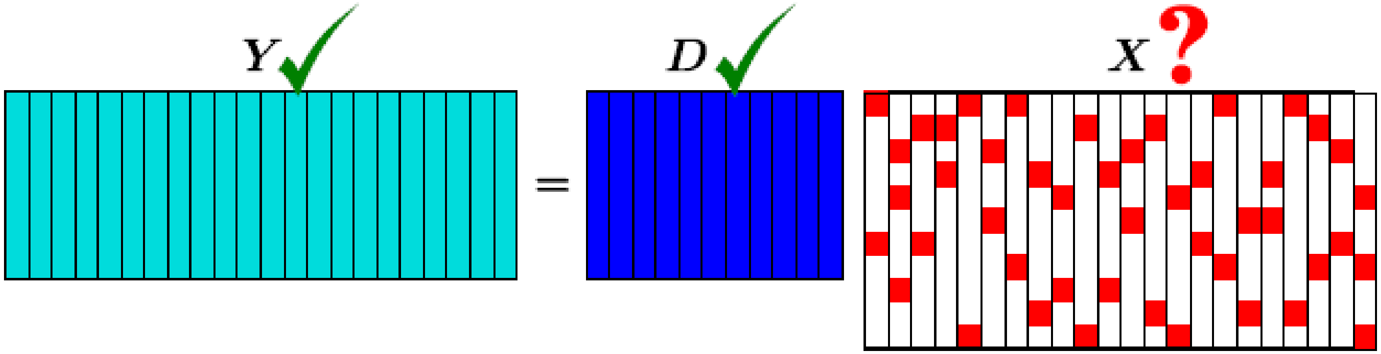
Figures taken from ICASSP 2013 Tutorial 11, by Dai, Maihe and Wang. Likewise for next four pages. Acknowledgement to Wei Dai for making these figures.

A two-stage procedure



W. Dai, T. Xu, and W. Wang, "Simultaneous Codeword Optimisation (SimCO) for Dictionary Update and Learning", *IEEE Transactions on Signal Processing*, vol. 60, no. 12, pp. 6340-6353, 2012.

Sparse coding (approximation)



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

Greedy algorithms:

- OMP Y. Pati, et al. 1993; J. Tropp 2004
- 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

$$\begin{aligned}\Omega &= \{(i,j):\ \boldsymbol{X}_{i,j} \neq 0\}, \\ \mathcal{X}_\Omega &= \{\boldsymbol{X}:\ \boldsymbol{X}_{i,j} = 0,\ \forall (i,j) \in \Omega^c\}.\end{aligned}$$

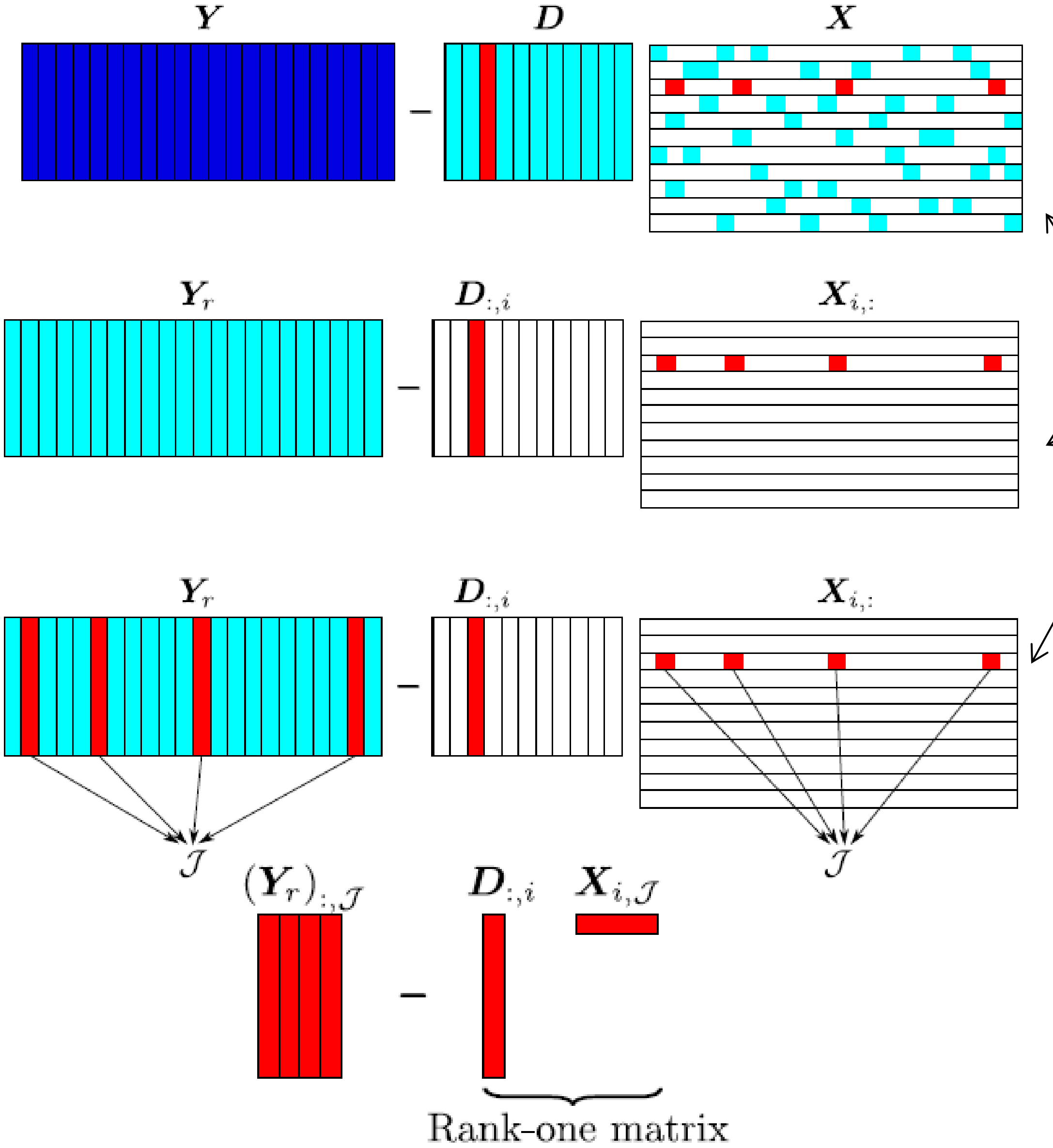
  - Unit norm codewords

$$\mathcal{D} = \left\{\boldsymbol{D}:\ \|\boldsymbol{D}_{:,j}\|_2 = 1,\ \forall j \in [d]\right\}.$$

- Dictionary Update:

$$\min_{\boldsymbol{D}\in\mathcal{D},\ \boldsymbol{X}\in\mathcal{X}_\Omega} \|\boldsymbol{Y} - \boldsymbol{D}\boldsymbol{X}\|_F^2 .$$

Dictionary update: K-SVD algorithm



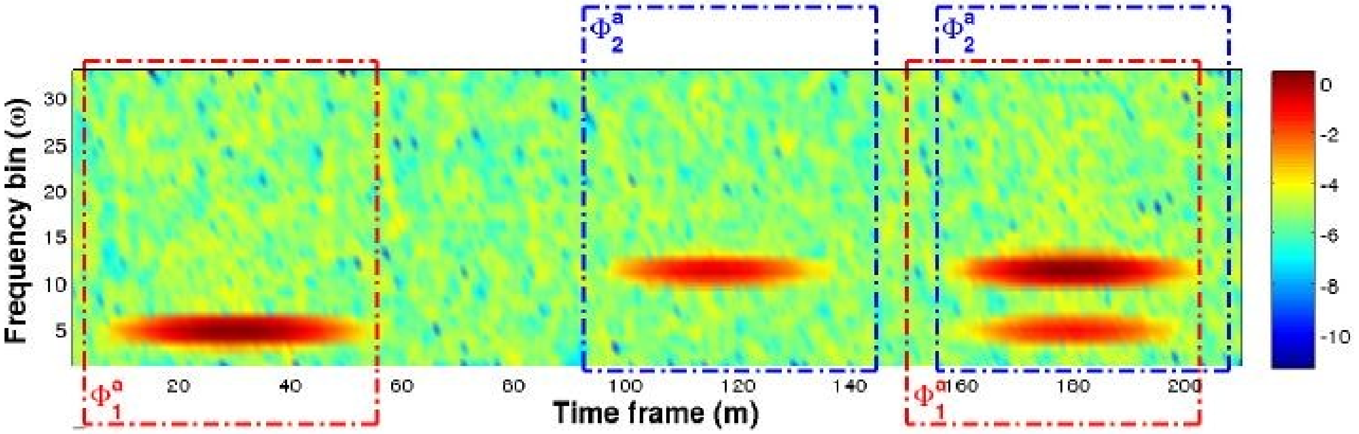
$$
\begin{aligned}
&\|\boldsymbol{Y} - \boldsymbol{D}\boldsymbol{X}\|^2 \\
&= \|\boldsymbol{Y} - \boldsymbol{D}_{:,j\neq i}\boldsymbol{X}_{j\neq i,:} - \boldsymbol{D}_{:,i}\boldsymbol{X}_{i,:}\|^2 \\
&= \|\boldsymbol{Y}_r - \boldsymbol{D}_{:,i}\boldsymbol{X}_{i,:}\|^2 \\
&= \left\|(\boldsymbol{Y}_r)_{:,\mathcal{J}} - \boldsymbol{D}_{:,i}\boldsymbol{X}_{i,\mathcal{J}}\right\|^2 + c
\end{aligned}
$$

Audio-visual dictionary learning: a generative model

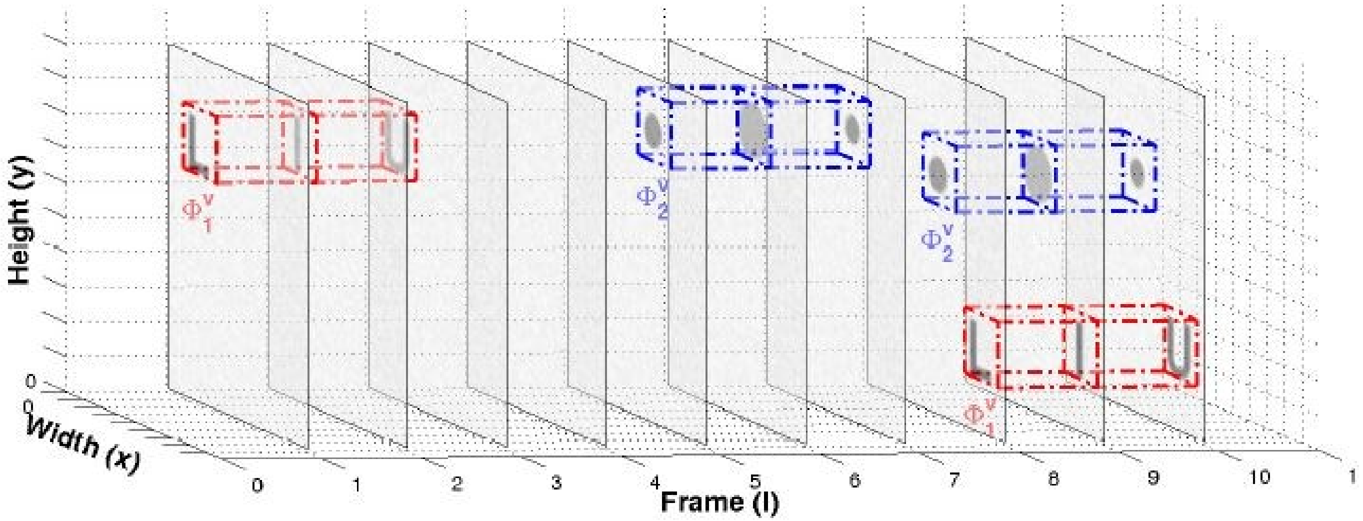
$$\begin{pmatrix}\psi^a(m)\\ \psi^v(y,x,l)\end{pmatrix} \approx \begin{pmatrix}\hat{\psi}^a(m)\\ \hat{\psi}^v(y,x,l)\end{pmatrix} = \sum_{d=1}^{D}\begin{pmatrix}\sum_{\breve{m}=1}^{M_s} c_{d\breve{m}}\phi_d^a(m-\breve{m})\\ \sum_{\breve{y}=1,\breve{x}=1,\breve{l}=1}^{Y_s,X_s,L_s} b_{d\breve{y}\breve{x}\breve{l}}\phi_d^v(y-\breve{y},x-\breve{x},l-\breve{l})\end{pmatrix}$$

Q. Liu, W. Wang, P. Jackson, M. Barnard, J. Kittler, and J.A. Chambers, “Source separation of convolutive and noisy mixtures using audio-visual dictionary learning and probabilistic time-frequency masking", IEEE Transactions on Signal Processing, vol. 61, no. 22, pp. 5520-5535, 2013.

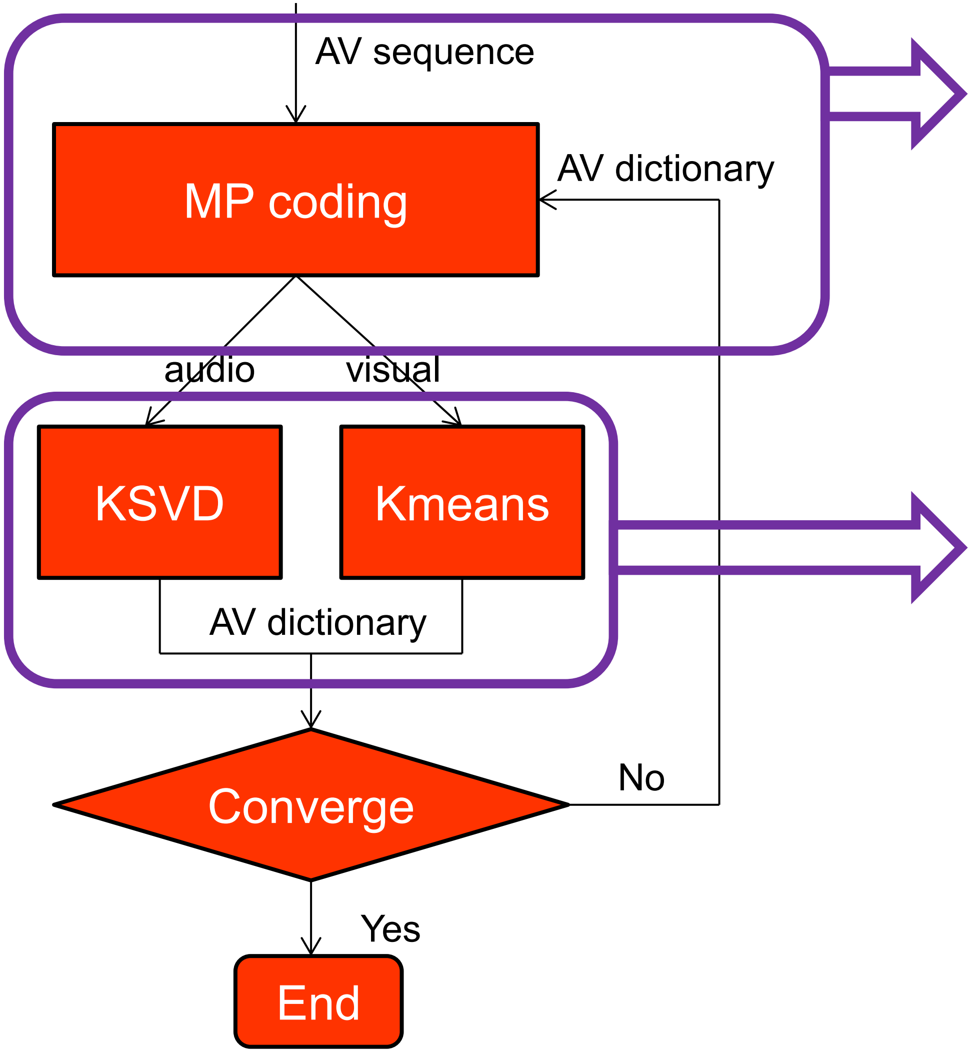
Sparse assumption of AVDL



(a) Audio stream ψ^a



Flow of the AVDL



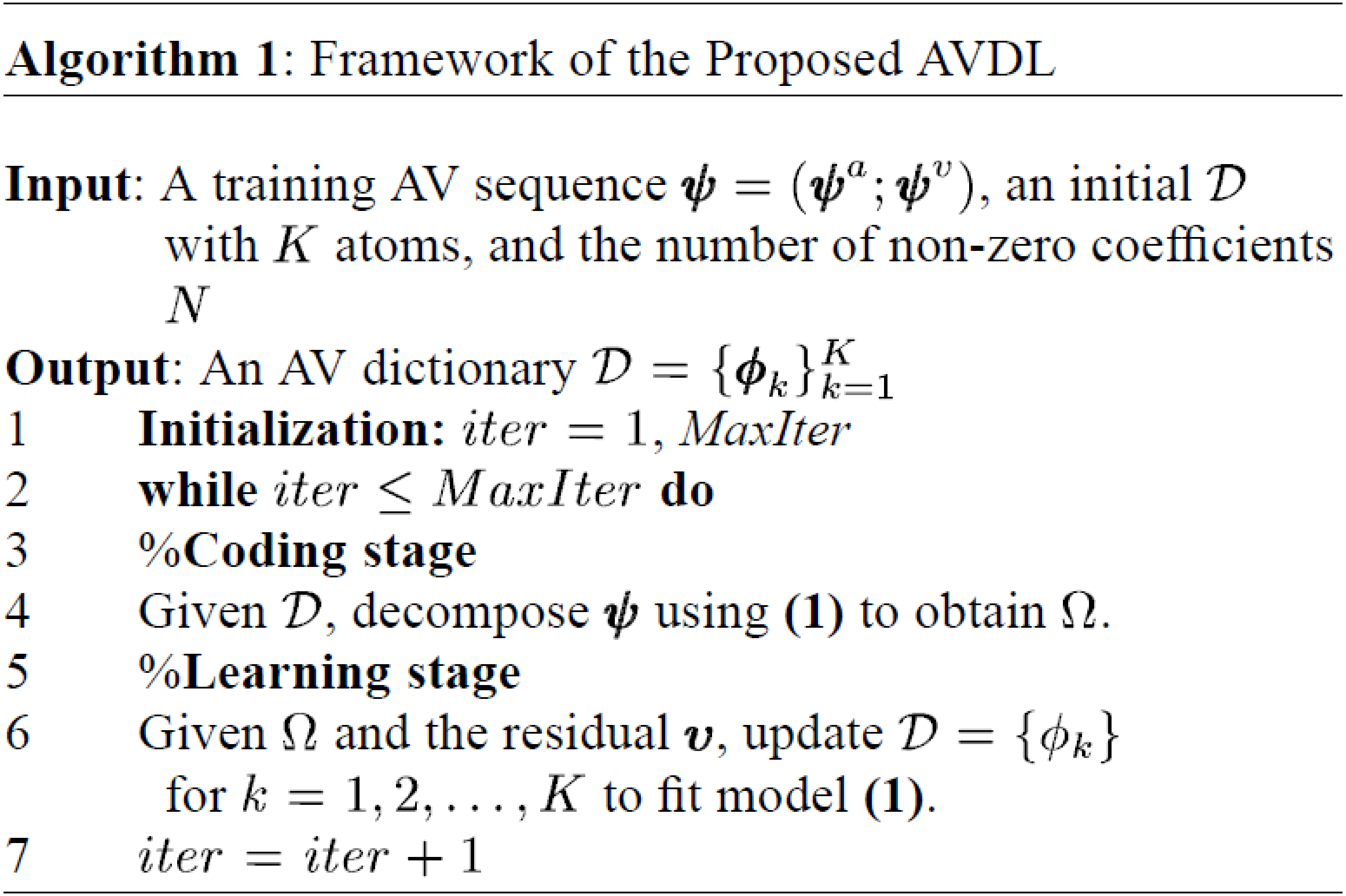
The coding process relies on the matching criterion, how well an atom fits the signal in the MP algorithm

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 $\boldsymbol{\psi} = (\boldsymbol{\psi}^a; \boldsymbol{\psi}^v)$, an initial $\mathcal{D}$
       with $K$ atoms, and the number of non-zero coefficients
       $N$
Output: An AV dictionary $\mathcal{D} = \{\boldsymbol{\phi}_k\}_{k=1}^{K}$
Initialization: $iter = 1$, $MaxIter$
while $iter \leq MaxIter$ do
    %Coding stage
    Given $\mathcal{D}$, decompose $\boldsymbol{\psi}$ using (1) to obtain $\Omega$.
    %Learning stage
    Given $\Omega$ and the residual $\boldsymbol{\upsilon}$, update $\mathcal{D} = \{\phi_k\}$
      for $k = 1, 2, \ldots, K$ to fit model (1).
    $iter = iter + 1$
```

The coding process

$$J^{av}(\bar{\boldsymbol{v}}_{\breve{y}\breve{x}\breve{l}\breve{m}}, \boldsymbol{\phi}_k) = J^a(\bar{\boldsymbol{v}}^a_{\breve{m}}, \boldsymbol{\phi}^a_k) J^v(\bar{\boldsymbol{v}}^v_{\breve{y}\breve{x}\breve{l}}, \boldsymbol{\phi}^v_k),$$

$$J^a_{\mathrm{Mon}} = |\langle \bar{\boldsymbol{v}}^a_{\breve{m}}, \boldsymbol{\phi}^a_k \rangle|$$

$$J^v(\bar{\boldsymbol{v}}^v_{\breve{y}\breve{x}\breve{l}}, \boldsymbol{\phi}^v_k) = \exp\left\{ \frac{-1}{YXL} \left\| \bar{\boldsymbol{v}}^v_{\breve{y}\breve{x}\breve{l}} - \boldsymbol{\phi}^v_k \right\|_1 \right\}.$$

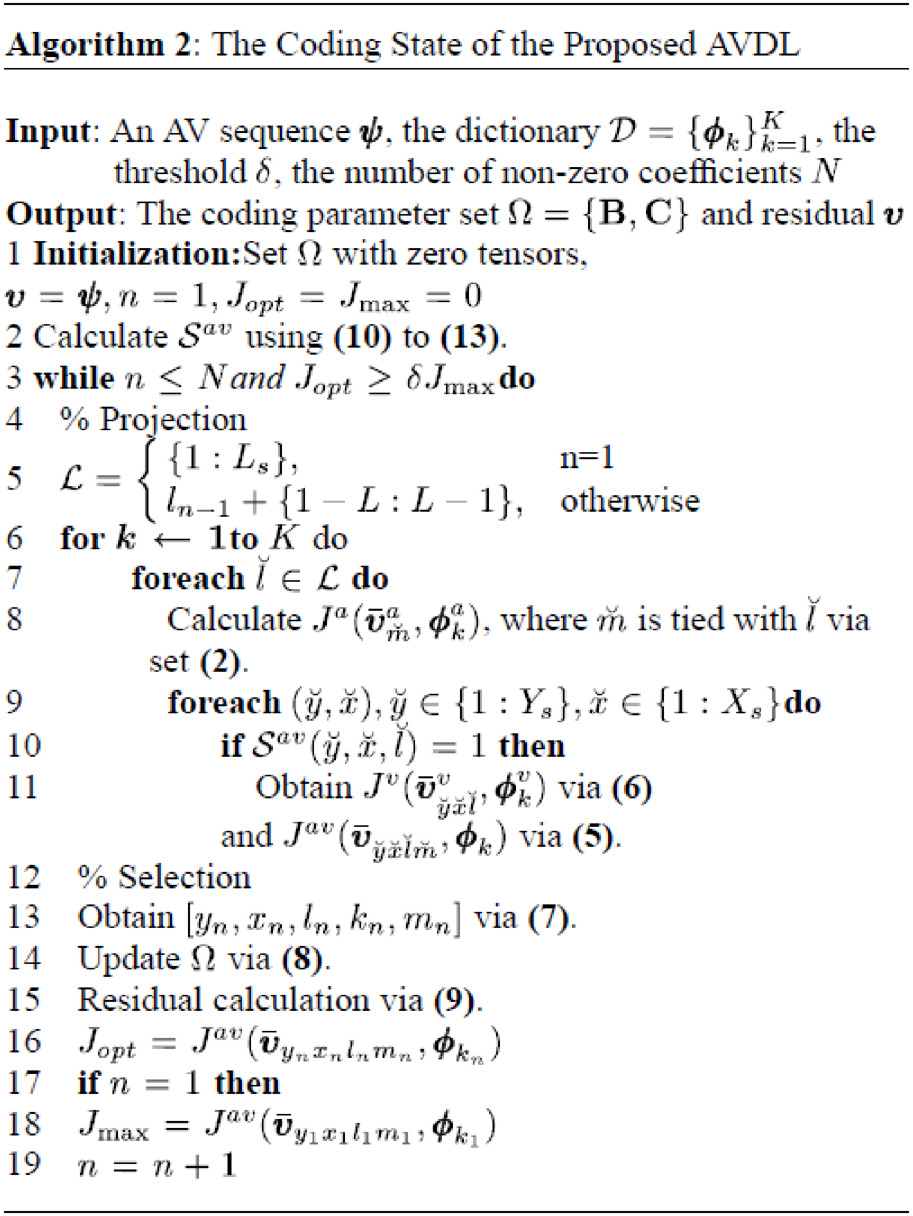
$$[k_n, y_n, x_n, l_n, m_n] = \arg\max_{[k,\breve{y},\breve{x},\breve{l},\breve{m}]} J^{av}(\bar{\boldsymbol{v}}_{\breve{y}\breve{x}\breve{l}\breve{m}}, \boldsymbol{\phi}_k),$$

$$B(k_n, y_n, x_n, l_n) = 1$$

$$C(k_n, m_n) = J^a(\bar{\boldsymbol{v}}^a_{m_n}, \boldsymbol{\phi}^a_{k_n}).$$

$$\bar{\boldsymbol{v}}^a_{l_n} \leftarrow \bar{\boldsymbol{v}}^a_{l_n} - C(k_n, l_n)\boldsymbol{\phi}^a_{k_n}.$$

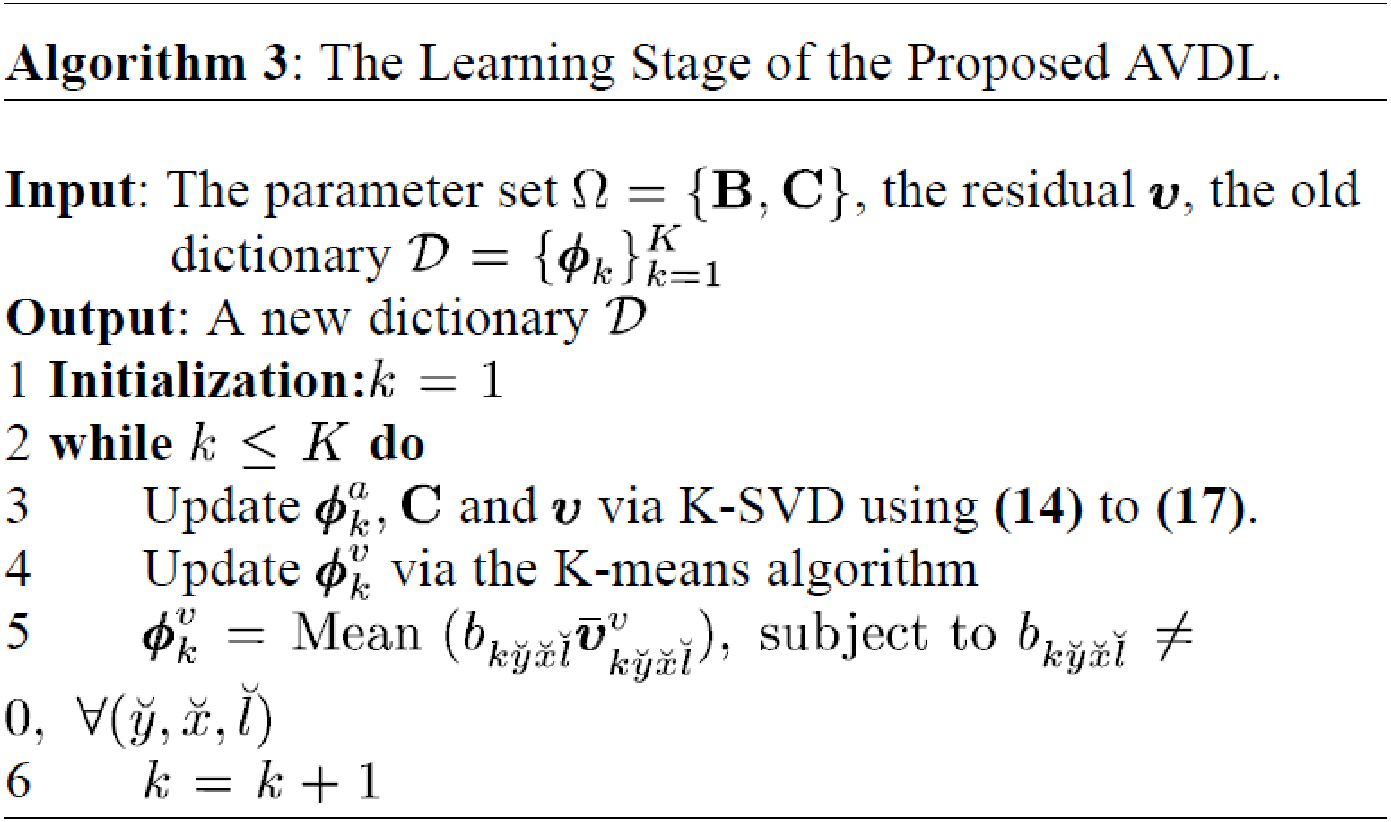
The coding process (algorithm)



Algorithm 2: The Coding State of the Proposed AVDL

```
Input: An AV sequence $\boldsymbol{\psi}$, the dictionary $\mathcal{D} = \{\boldsymbol{\phi}_k\}_{k=1}^{K}$, the
       threshold $\delta$, the number of non-zero coefficients $N$
Output: The coding parameter set $\Omega = \{\mathbf{B}, \mathbf{C}\}$ and residual $\boldsymbol{v}$
Initialization: Set $\Omega$ with zero tensors,
$\boldsymbol{v} = \boldsymbol{\psi}, n = 1, J_{opt} = J_{\max} = 0$
Calculate $\mathcal{S}^{av}$ using (10) to (13).
while $n \leq N$ and $J_{opt} \geq \delta J_{\max}$ do
    % Projection
    $\mathcal{L} = \begin{cases} \{1 : L_s\}, & \text{n=1} \\ l_{n-1} + \{1 - L : L - 1\}, & \text{otherwise} \end{cases}$
    for $k \leftarrow 1$ to $K$ do
        foreach $\breve{l} \in \mathcal{L}$ do
            Calculate $J^a(\bar{\boldsymbol{v}}^a_{\breve{m}}, \boldsymbol{\phi}^a_k)$, where $\breve{m}$ is tied with $\breve{l}$ via
            set (2).
            foreach $(\breve{y}, \breve{x}), \breve{y} \in \{1 : Y_s\}, \breve{x} \in \{1 : X_s\}$ do
                if $\mathcal{S}^{av}(\breve{y}, \breve{x}, \breve{l}) = 1$ then
                    Obtain $J^v(\bar{\boldsymbol{v}}^v_{\breve{y}\breve{x}\breve{l}}, \boldsymbol{\phi}^v_k)$ via (6)
                and $J^{av}(\bar{\boldsymbol{v}}_{\breve{y}\breve{x}\breve{l}\breve{m}}, \boldsymbol{\phi}_k)$ via (5).
    % Selection
    Obtain $[y_n, x_n, l_n, k_n, m_n]$ via (7).
    Update $\Omega$ via (8).
    Residual calculation via (9).
    $J_{opt} = J^{av}(\bar{\boldsymbol{v}}_{y_n x_n l_n m_n}, \boldsymbol{\phi}_{k_n})$
    if $n = 1$ then
    $J_{\max} = J^{av}(\bar{\boldsymbol{v}}_{y_1 x_1 l_1 m_1}, \boldsymbol{\phi}_{k_1})$
    $n = n + 1$
```

The learning stage



Algorithm 3: The Learning Stage of the Proposed AVDL.

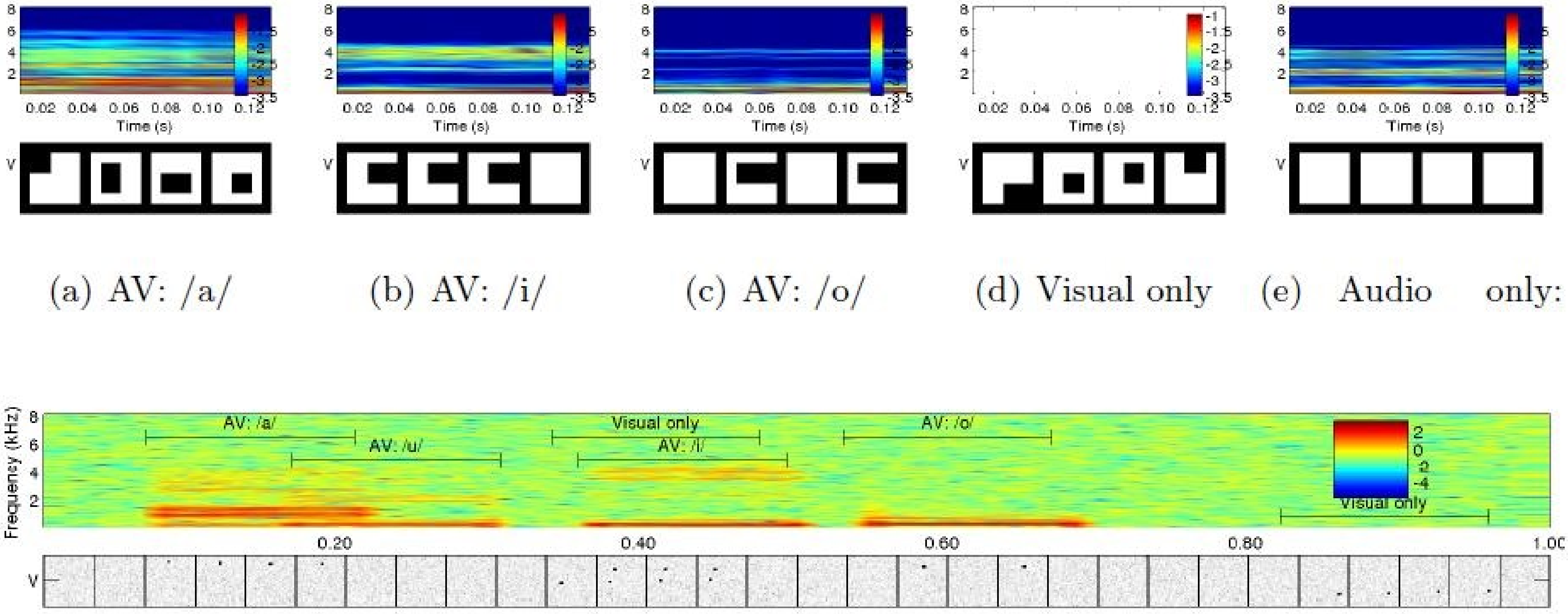
```
Input: The parameter set $\Omega = \{\mathbf{B}, \mathbf{C}\}$, the residual $\boldsymbol{\upsilon}$, the old
       dictionary $\mathcal{D} = \{\boldsymbol{\phi}_k\}_{k=1}^{K}$
Output: A new dictionary $\mathcal{D}$
Initialization: $k = 1$
while $k \leq K$ do
    Update $\boldsymbol{\phi}_k^a$, $\mathbf{C}$ and $\boldsymbol{\upsilon}$ via K-SVD using (14) to (17).
    Update $\boldsymbol{\phi}_k^v$ via the K-means algorithm
    $\boldsymbol{\phi}_k^v = \text{Mean}\,(b_{k\breve{y}\breve{x}\breve{l}}\bar{\boldsymbol{\upsilon}}^v_{k\breve{y}\breve{x}\breve{l}})$, subject to $b_{k\breve{y}\breve{x}\breve{l}} \neq 0, \ \forall(\breve{y}, \breve{x}, \breve{l})$
    $k = k + 1$
```

$$\bar{\boldsymbol{\upsilon}}^a_{\breve{m}} \leftarrow \bar{\boldsymbol{\upsilon}}^a_{\breve{m}} + c_{k\breve{m}}\boldsymbol{\phi}^a_k, \ \forall \breve{m}. \qquad \boldsymbol{\phi}^a_k \leftarrow \mathbf{ivec}(\mathbf{u}_k|\boldsymbol{\phi}^a_k).$$

$$\Upsilon_k \approx \lambda_k \mathbf{u}_k \mathbf{v}_k^T, \qquad \bar{\boldsymbol{\upsilon}}^a_{\breve{m}} \leftarrow \bar{\boldsymbol{\upsilon}}^a_{\breve{m}} - c_{k\breve{m}}\boldsymbol{\phi}^a_k, \ \forall \breve{m}.$$

AVDL evaluations

Synthetic data

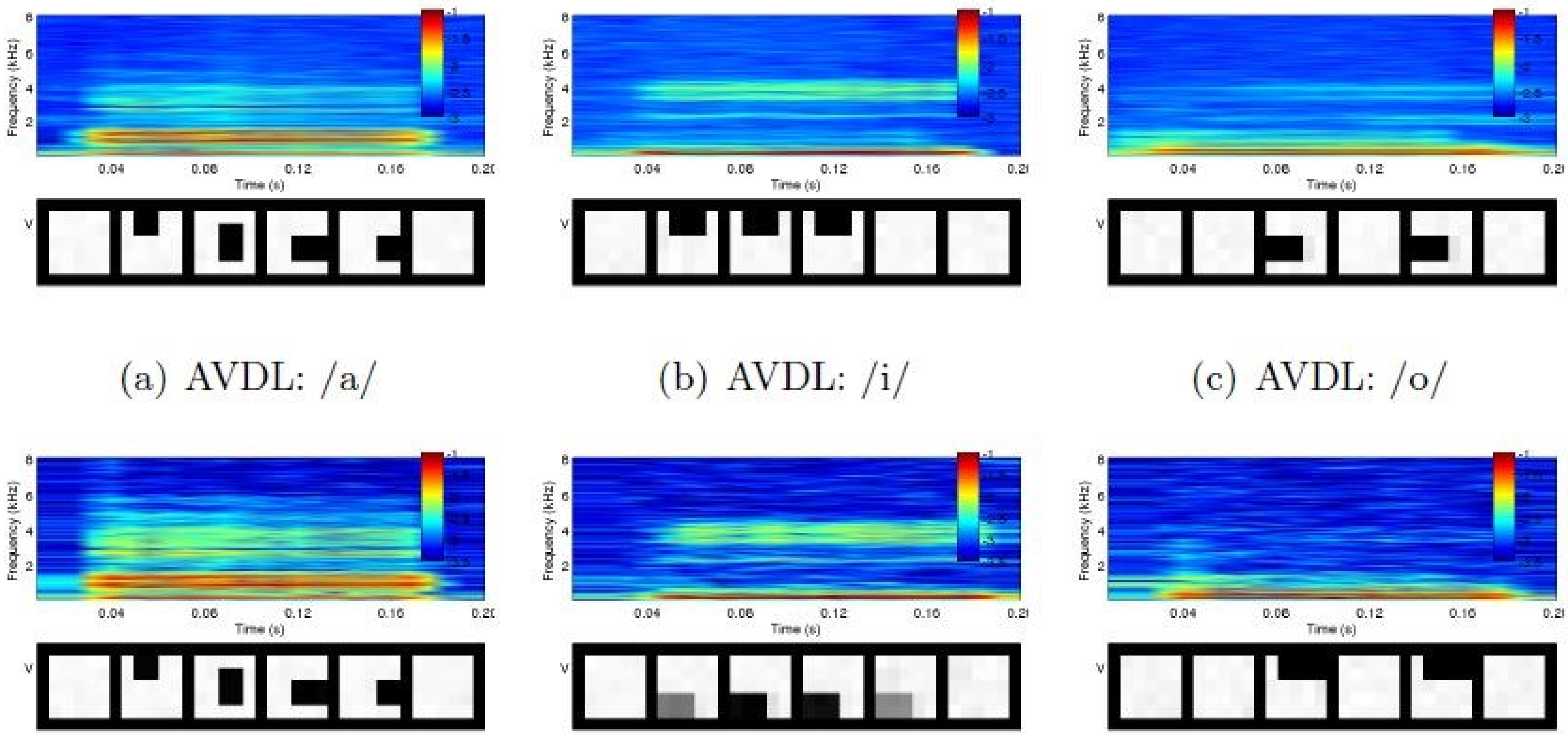


(a) AV: /a/ (b) AV: /i/ (c) AV: /o/ (d) Visual only (e) Audio only:

(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/

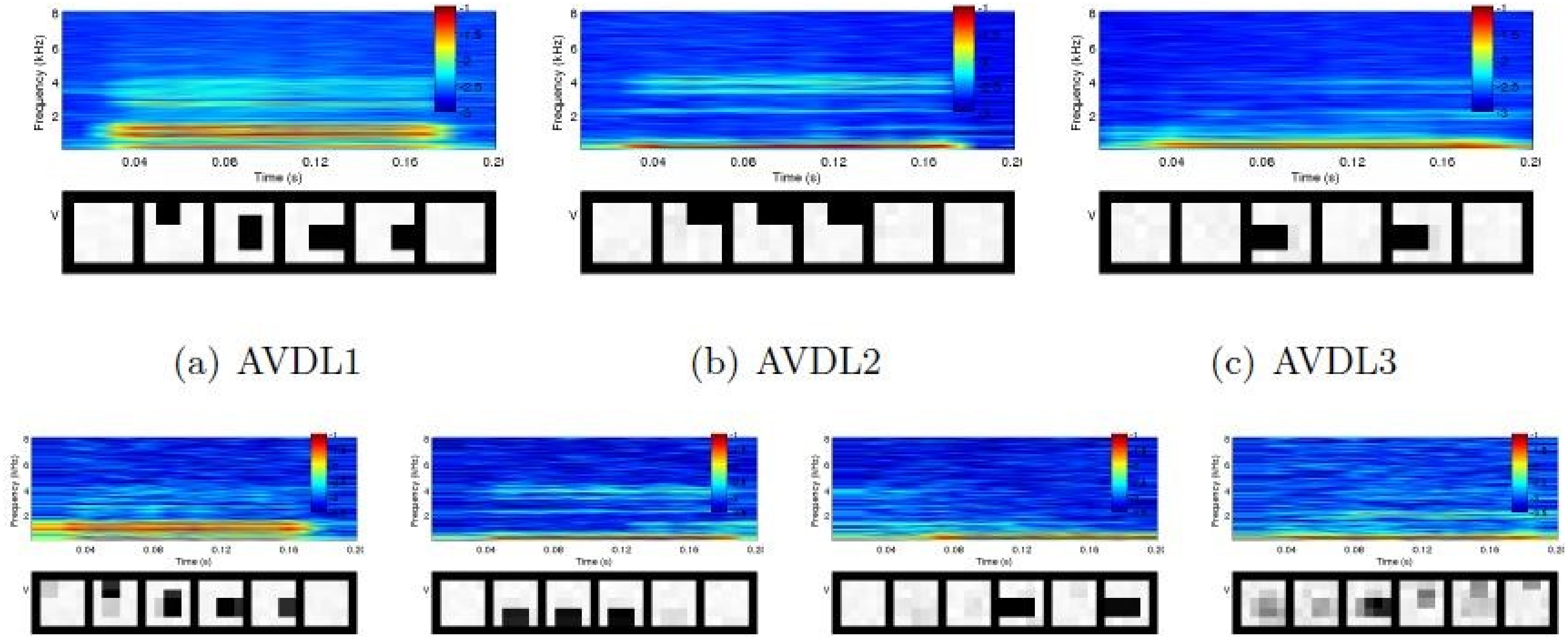
(d) Monaci: /a/

(e) Monaci: /i/

(f) Monaci: /o/

AVDL evaluations

Convolutive noise added



(a) AVDL1

(b) AVDL2

(c) AVDL3

(d) Monaci1

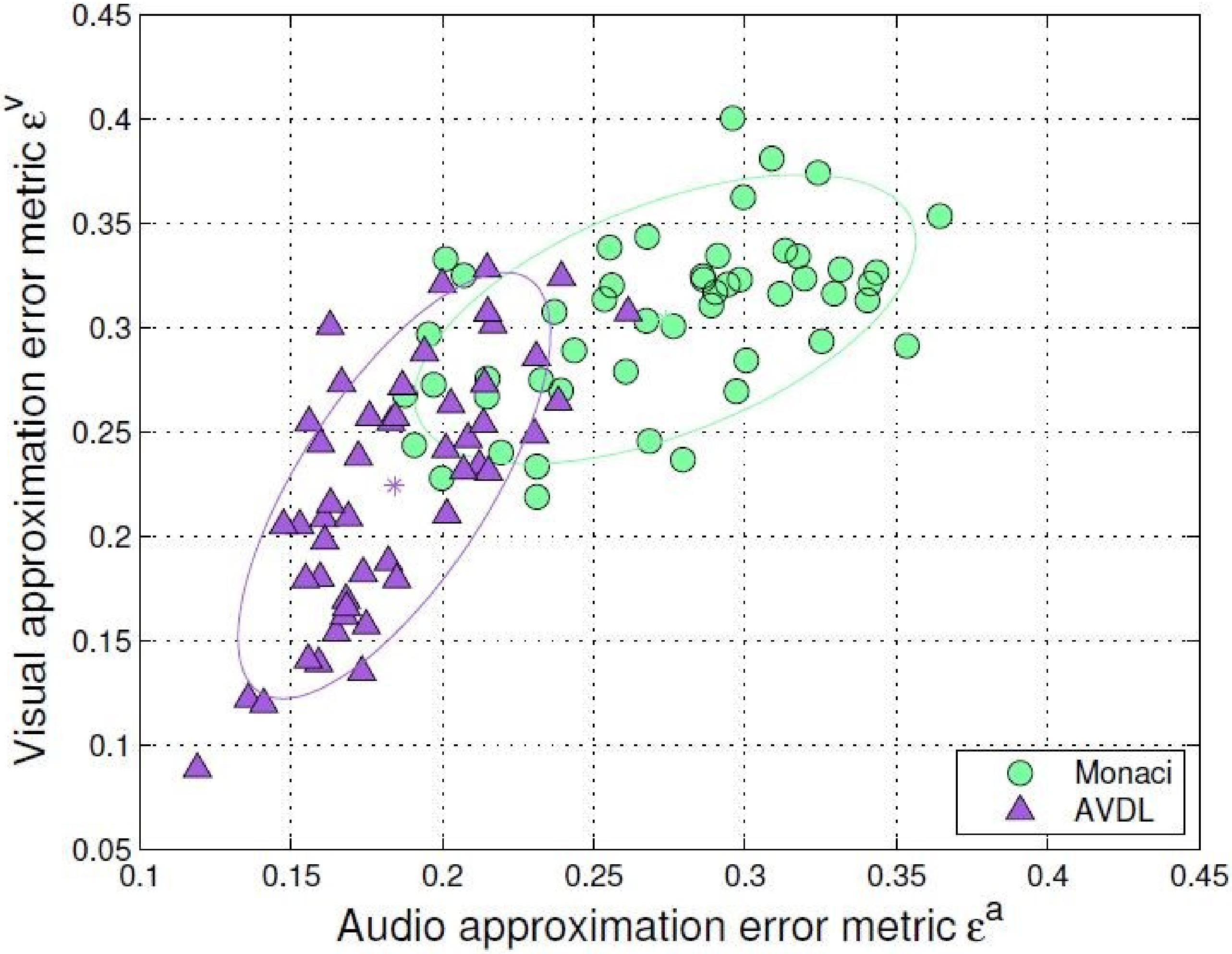
(e) Monaci2

(f) Monaci3

(g) Monaci4

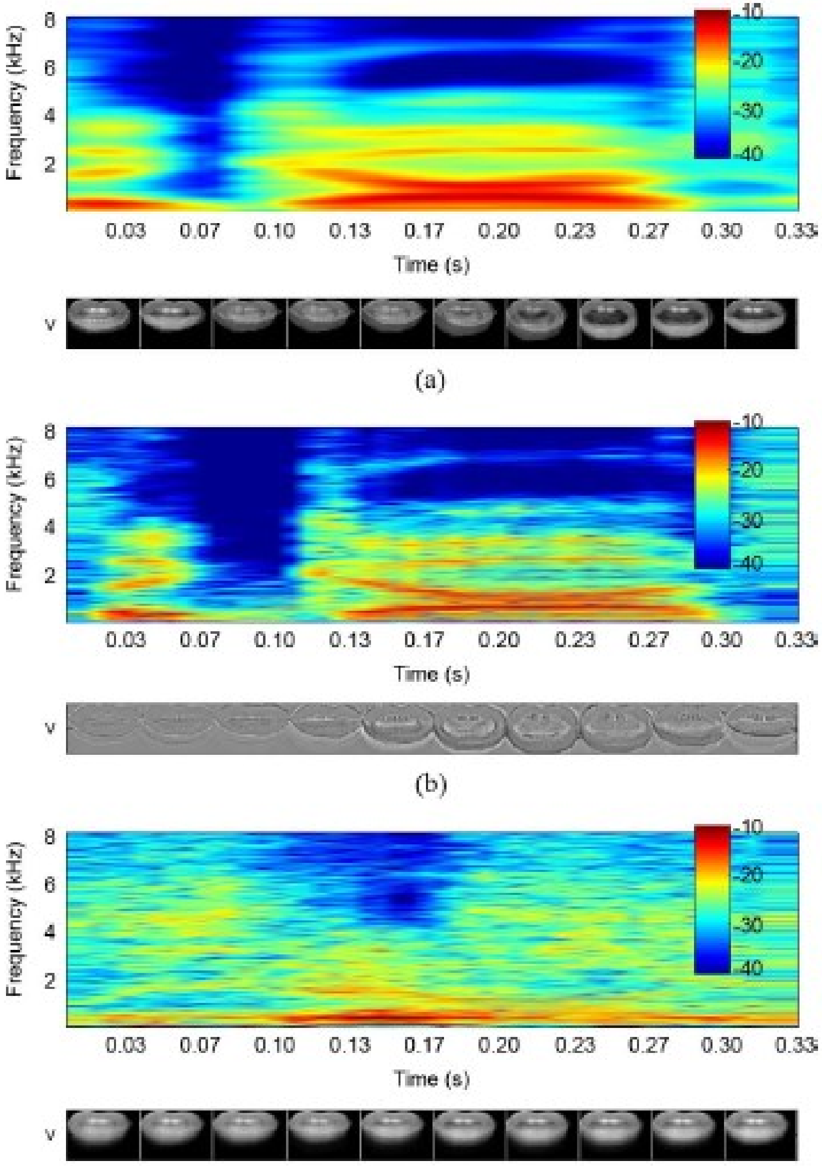
AVDL evaluations

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



(a)

(b)

(c)

AV mask fusion for AVDL-BSS

$$\mathcal{M}^{av}(m,\omega) = \boxed{\mathcal{M}^{a}(m,\omega)}^{r\boxed{(\mathcal{M}^{v}(m,\omega))}}$$

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

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

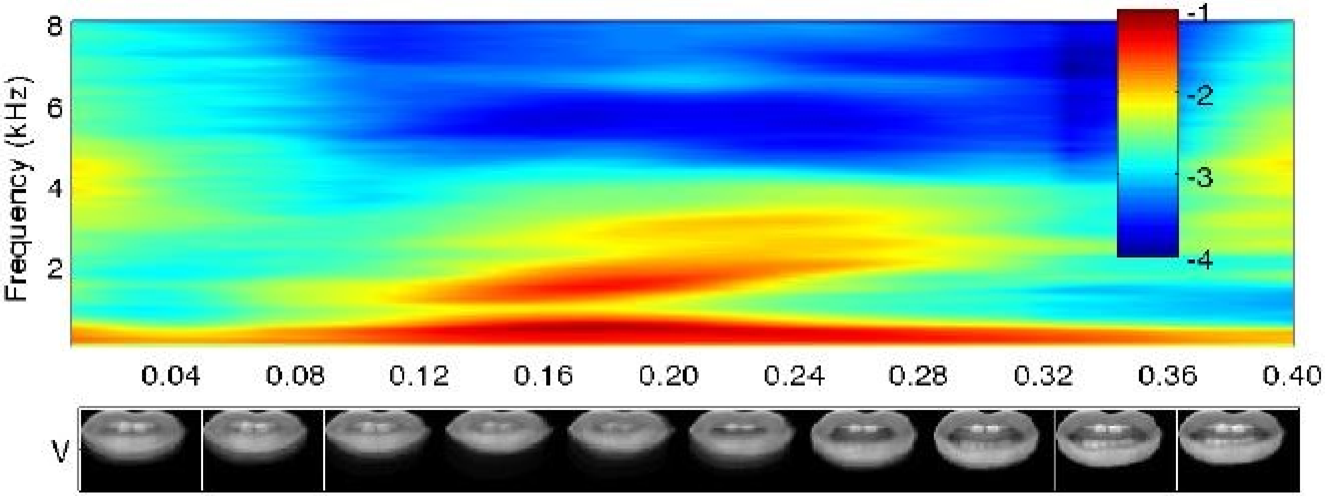
Q. Liu, W. Wang, P. Jackson, M. Barnard, J. Kittler, and J.A. Chambers, “Source separation of convolutive and noisy mixtures using audio-visual dictionary learning and probabilistic time-frequency masking", IEEE Transactions on Signal Processing, vol. 61, no. 22, pp. 5520-5535, 2013.

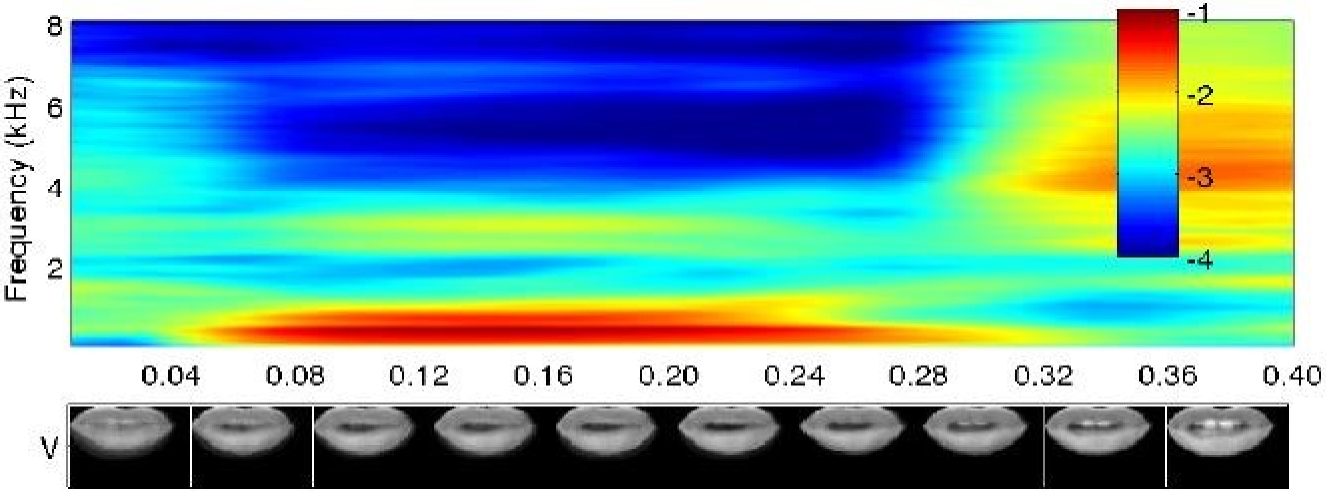
AVDL evaluations

Long Speech

Sheerman-Chase et al. LILiR Twotalk database 2011

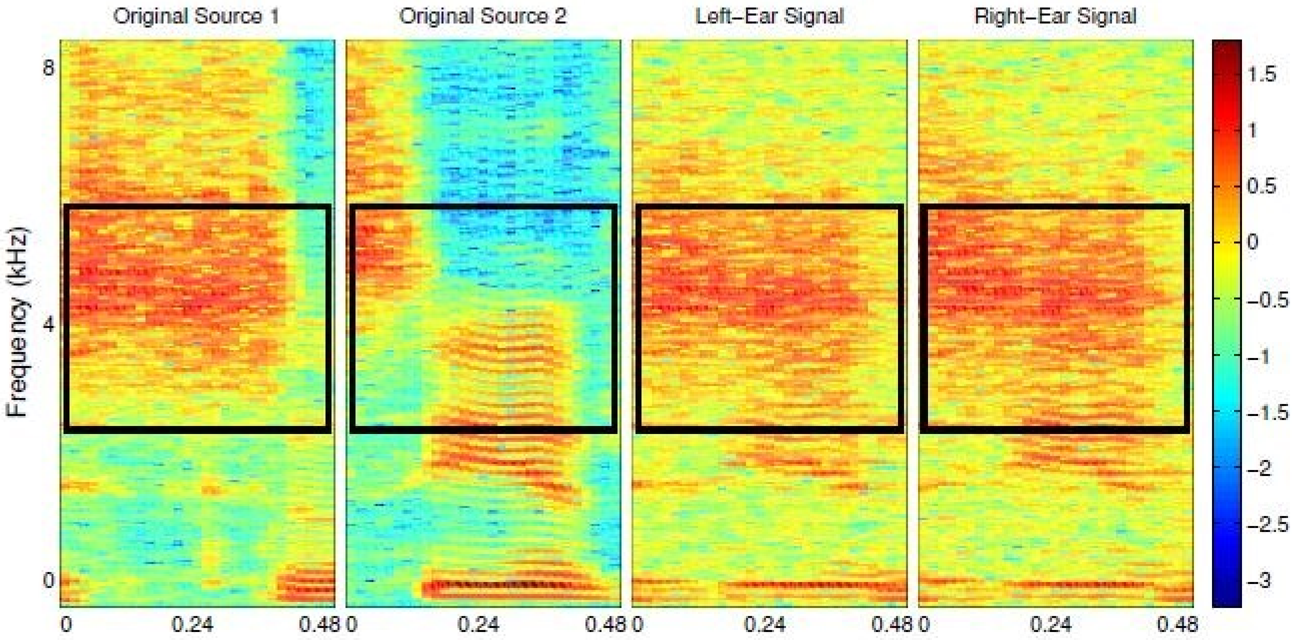
Lip tracking,
Ong et al. 2008

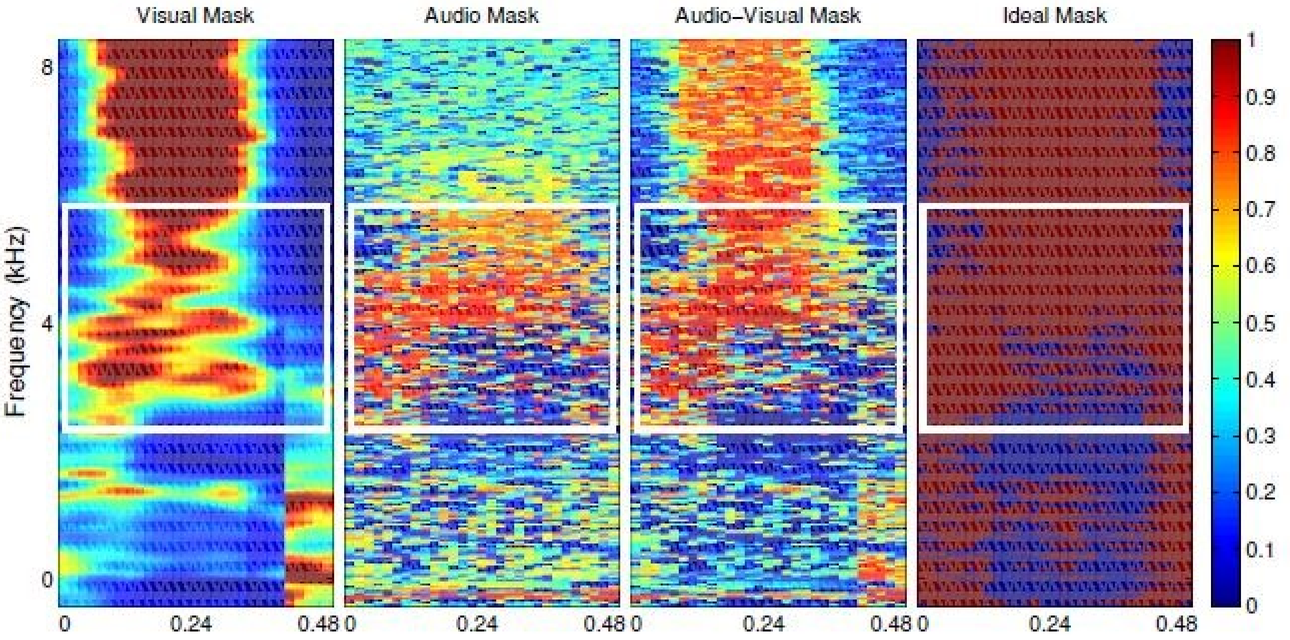




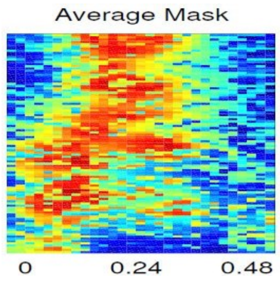
The first AV atom represents the utterance "**marine**" /mᵊri:n/ while the second one denotes the utterance "**port**" /pᵓ:t/.

Demonstration of TF mask fusion in AVDL-BSS

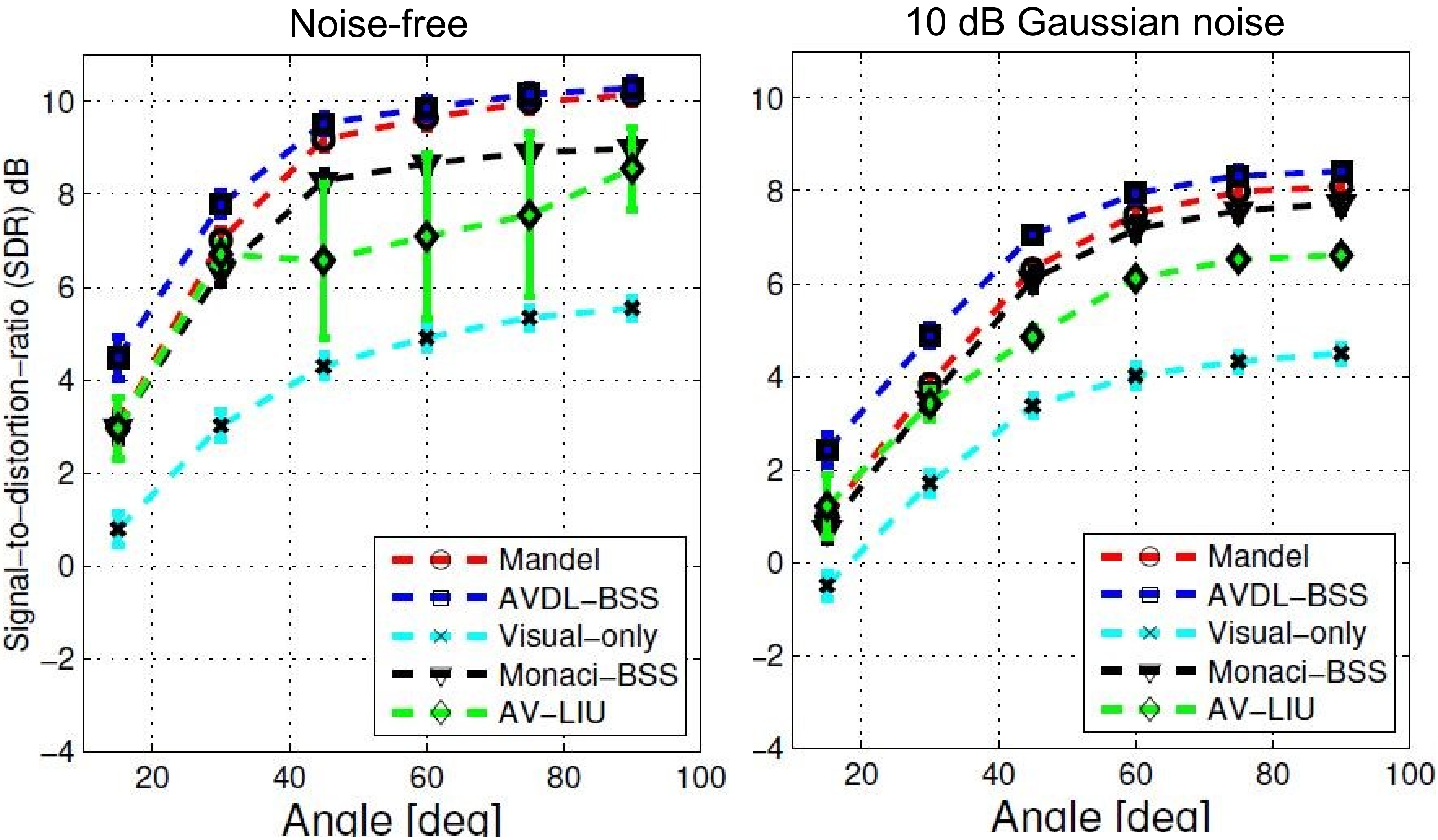




Why do we choose the power law combination, instead of, e.g., a linear combination?



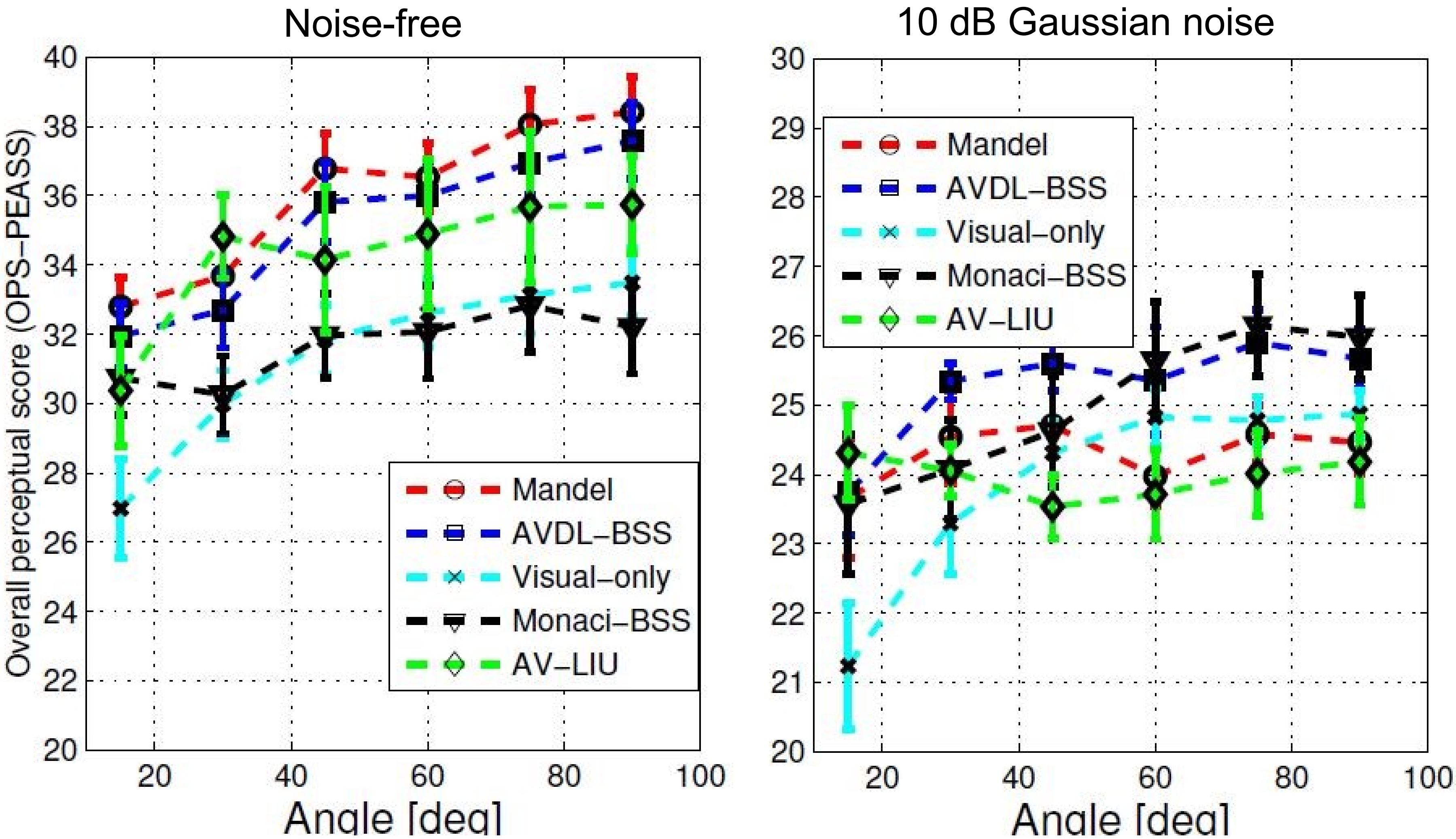
AVDL-BSS evaluations----SDR



Noise-free

10 dB Gaussian noise

AVDL-BSS evaluations----OPS-PEASS



Some examples

	Mixture	Ideal	Mandel	AV-LIU	AVDL-BSS	Rivet	AVMP-BSS
A	🔊	🔊	🔊	🔊	🔊	🔊	🔊
B	🔊	🔊	🔊	🔊	🔊	🔊	🔊
C	🔊	🔊	🔊	🔊	🔊	🔊	🔊
D	🔊	🔊	🔊	🔊	🔊	🔊	🔊

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