Multimodal Signal Processing on Manifolds

David Dov

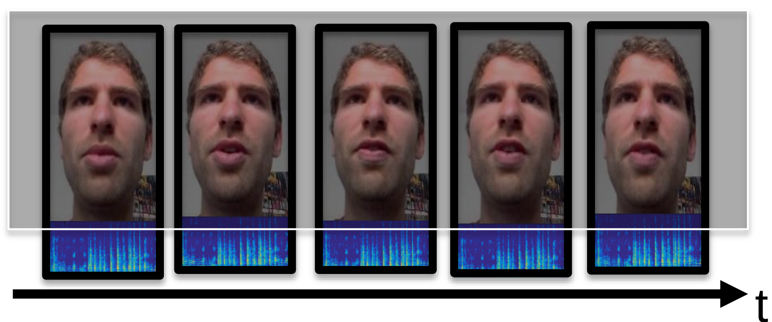
Supervised by Prof. Israel Cohen & Prof. Ronen Talmon

• Problem setup

Multimodal data:

❑ N pairs of samples (data points):

$\{\boldsymbol{v}_n, \boldsymbol{w}_n\}_1^N$, $\boldsymbol{v}_n \in \mathcal{R}^{L_v}, \boldsymbol{w}_n \in \mathcal{R}^{L_w}$

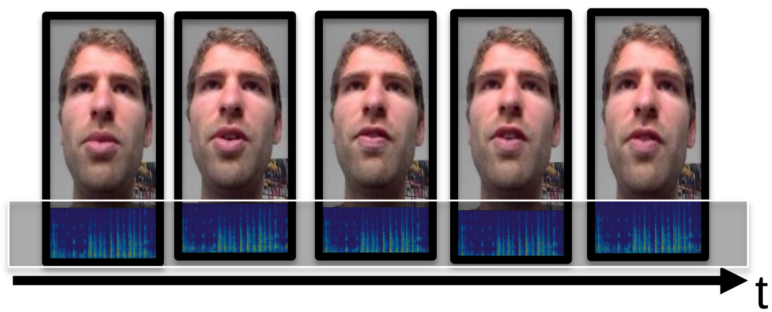


• Problem setup

Multimodal data:

❑ N pairs of samples (data points):

$$\{\boldsymbol{v}_n, \boldsymbol{w}_n\}_1^N \ , \boldsymbol{v}_n \in \mathcal{R}^{L_v}, \boldsymbol{w}_n \in \mathcal{R}^{L_w}$$



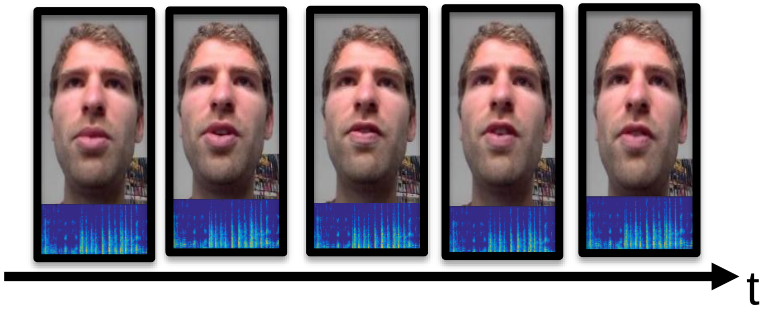
❑ The data points are aligned

• Problem setup

Different sources:

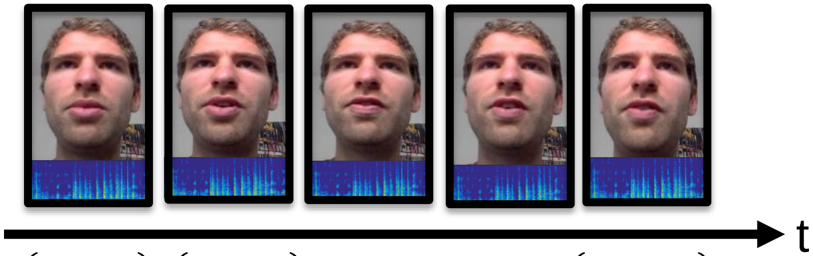
- ❑ Common
- ❑ Modality specific

$$\boldsymbol{v}_n = \boldsymbol{v}_n(\boldsymbol{x}, \boldsymbol{y})$$
$$\boldsymbol{w}_n = \boldsymbol{w}_n(\boldsymbol{x}, \boldsymbol{z})$$



- Example - sound source activity detection

❑ Given audio visual signals:



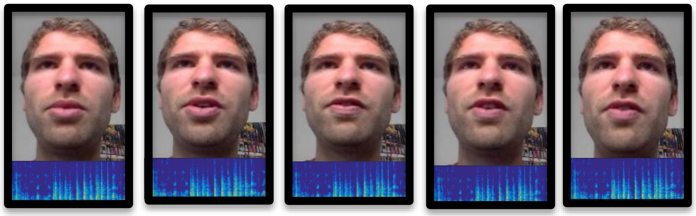
$(\boldsymbol{v}_1, \boldsymbol{w}_1), (\boldsymbol{v}_2, \boldsymbol{w}_2), \quad \dots \quad , (\boldsymbol{v}_N, \boldsymbol{w}_N)$

❑ Goal: for each frame, estimate the activity of the common source:

$$\mathbf{1}_n(\boldsymbol{x}) = \begin{cases} 1 & ; \quad n \in \mathcal{H}_1 \\ 0 & ; \quad n \in \mathcal{H}_0 \end{cases}$$

• Example - sound source activity detection

- Given audio visual signals:



$(\boldsymbol{v}_1, \boldsymbol{w}_1), (\boldsymbol{v}_2, \boldsymbol{w}_2), \quad \dots \quad, (\boldsymbol{v}_N, \boldsymbol{w}_N)$ t

- Special case: voice activity detection

- Challenge: structured modality-specific interferences
  - Head movements (we do no preprocessing)
  - Acoustic noises and transients

$$\boldsymbol{v}_n = \boldsymbol{v}_n(\boldsymbol{x}, \boldsymbol{y})$$
$$\boldsymbol{w}_n = \boldsymbol{w}_n(\boldsymbol{x}, \boldsymbol{z})$$

• Problem setup – cont'd

- ❑ Any type of modality
- ❑ Possibly, multiple modalities (more than two)
- ❑ Unsupervised setting – no labels
- ❑ The signals is the data
  - No external training datasets
- ❑ Online/batch

- Problem setup – cont'd

Goal:

- ❑Data fusion
- ❑Unified representation: $\{\boldsymbol{\phi}_n\}_{\mathbf{1}}^{\boldsymbol{N}} \in \mathcal{R}^L$

$$\{\boldsymbol{v}_n(\boldsymbol{x}, \boldsymbol{y}), \boldsymbol{w}_n(\boldsymbol{x}, \boldsymbol{z})\} \rightarrow \boldsymbol{\phi}_n(\boldsymbol{x})$$

- ❑Reduce the effect of structured interferences

- Related open questions

❑Limited availability of sensors over time

$$\boldsymbol{v}_n(\boldsymbol{x}, \boldsymbol{y}) \rightarrow \boldsymbol{\phi}_n(\boldsymbol{x})$$

- Do I need the data from all of the modalities?

❑Multi-modal correspondence

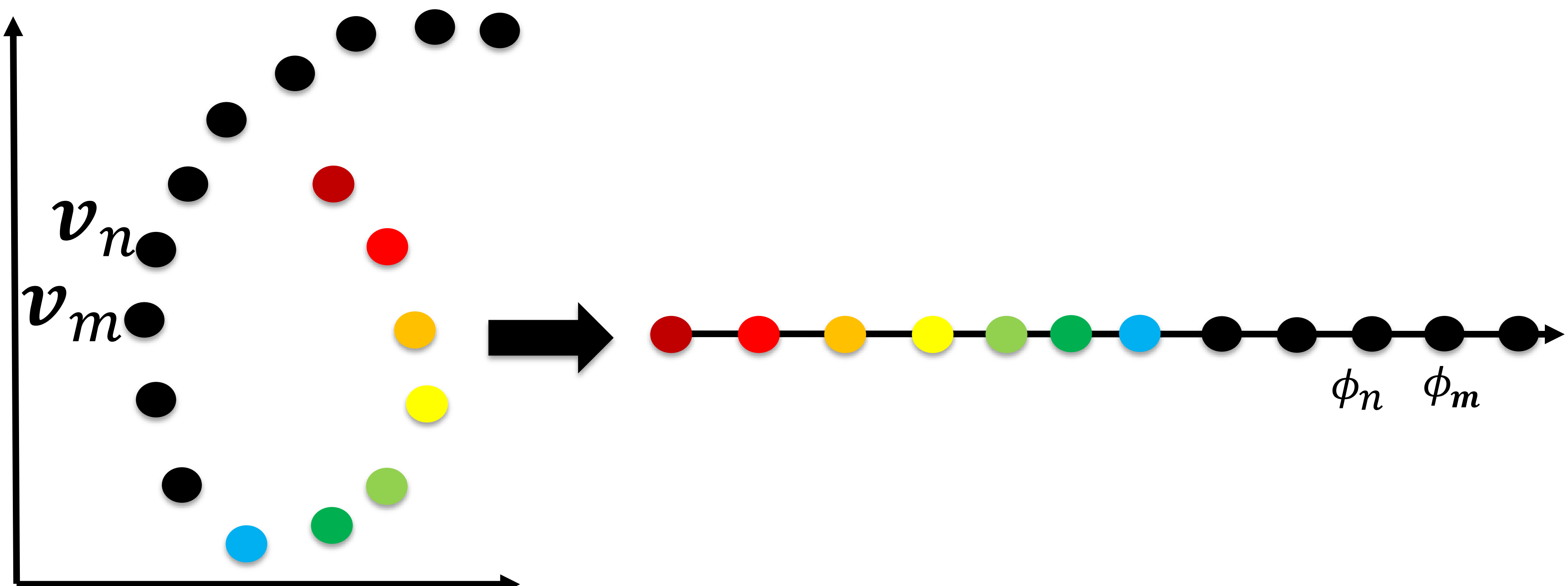
- "Correlation"

- Manifold learning

We take the kernel based geometric approach

Background - the single modal case

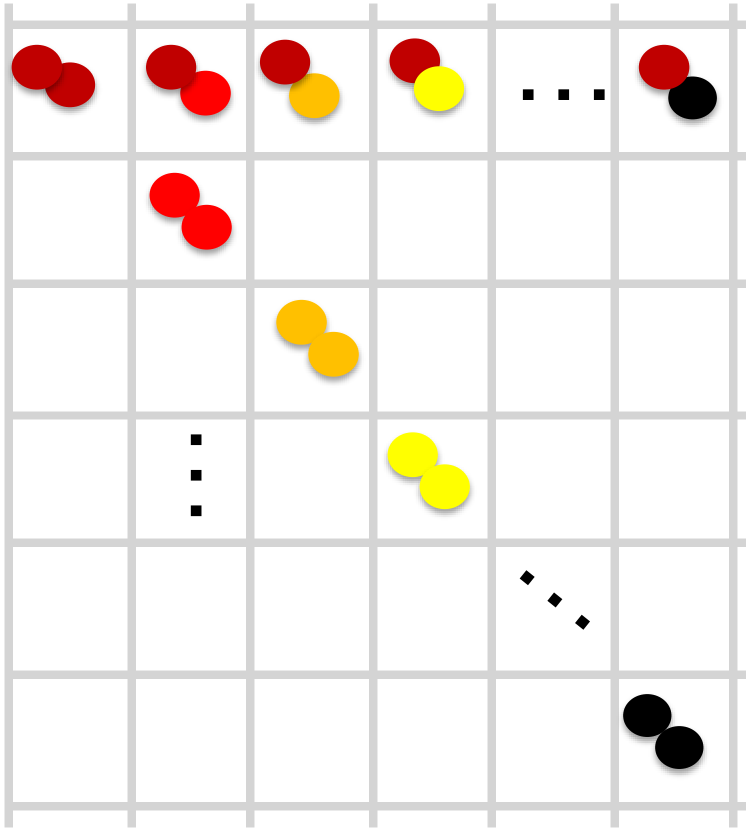
- ❑ Geometric assumption: low dimensional structure
- ❑ Goal: a representation that respects the geometric structure
  - ❑ Preserve local affinities

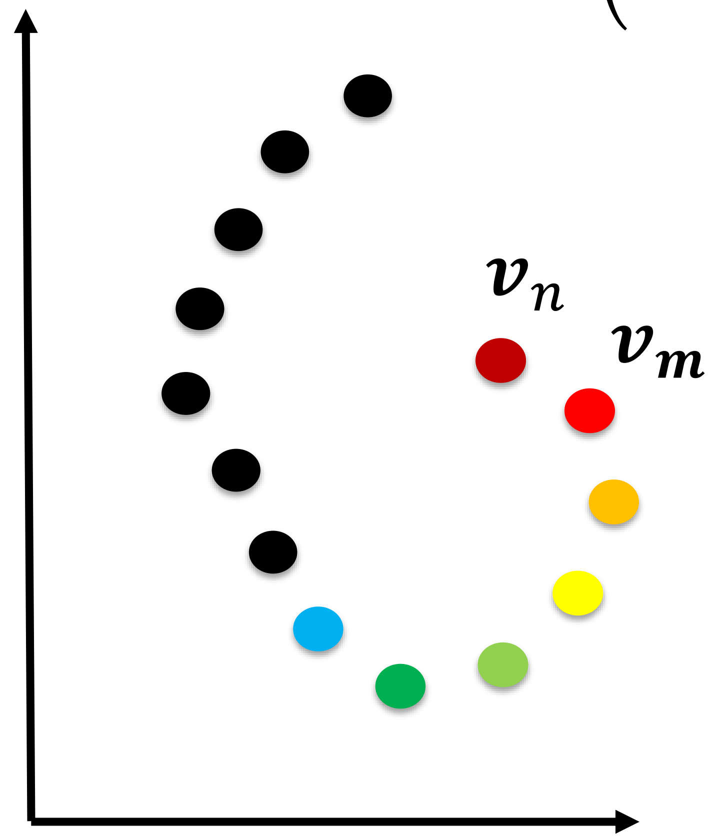


Manifold learning - the single modal case

Diffusion Maps (Coifman & Lafon 06):

- ❑ Construct an affinity matrix $\boldsymbol{K} \in \mathbb{R}^{N \times N}$: $K(n,m) = \exp\left(-\frac{\|v_n - v_m\|^2}{\epsilon_v}\right)$





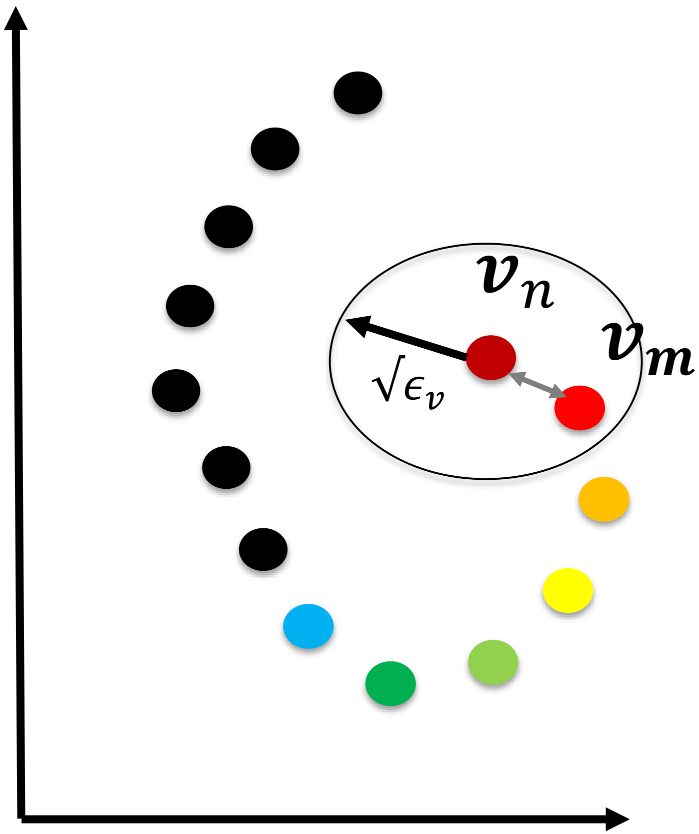
Manifold learning - the single modal case

Graph interpretation [Keller et al 10']

- Each point is a vertex
- The weights of the edges:

$$K_v(n,m) = \exp\left(-\frac{\|\boldsymbol{v}_n - \boldsymbol{v}_m\|^2}{\epsilon_v}\right)$$

- An *edge* exists between *similar* points
- $||\boldsymbol{v}_n - \boldsymbol{v}_m||^2 < \epsilon_v \rightarrow K_v(n,m) \neq 0$

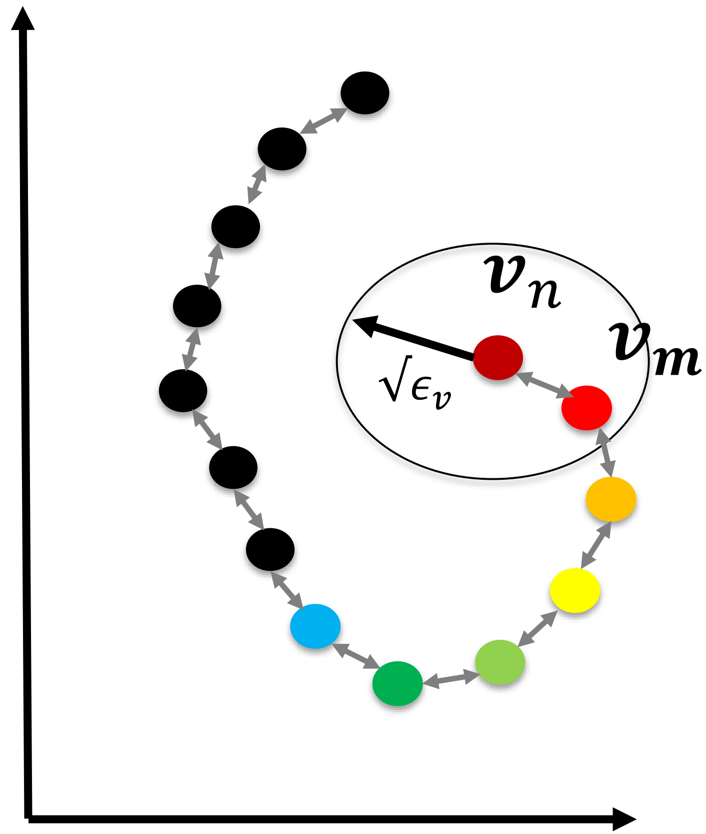


Manifold learning - the single modal case

Graph interpretation [Coifman & Lafon 06, Keller et al 10']

- ❑ Assumption: a single geometric structure
- ❑ A necessary condition: a connected graph
- ❑ In particular:
  - each point is connected

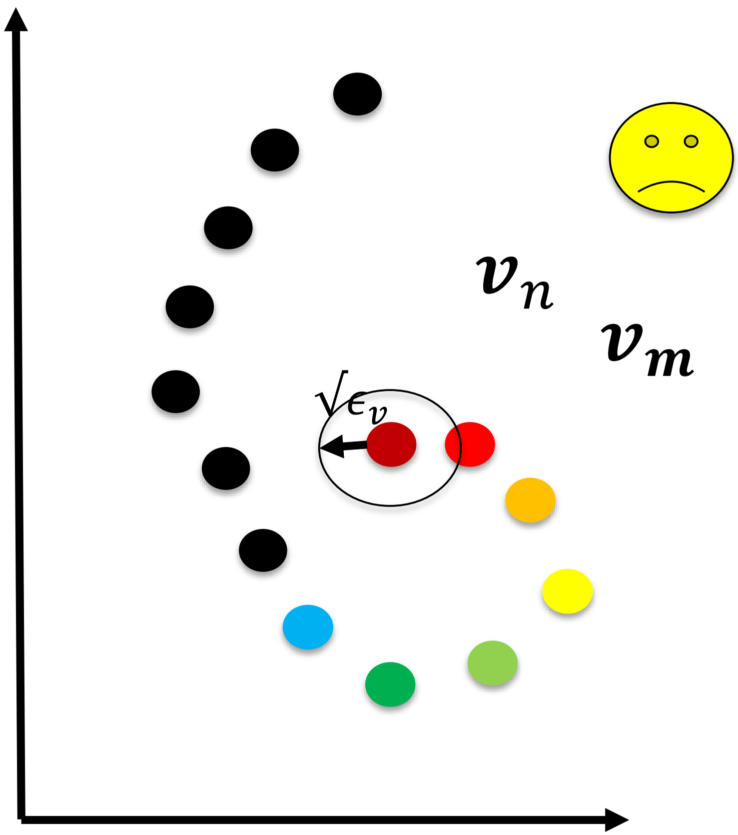
  (to at least one other point)



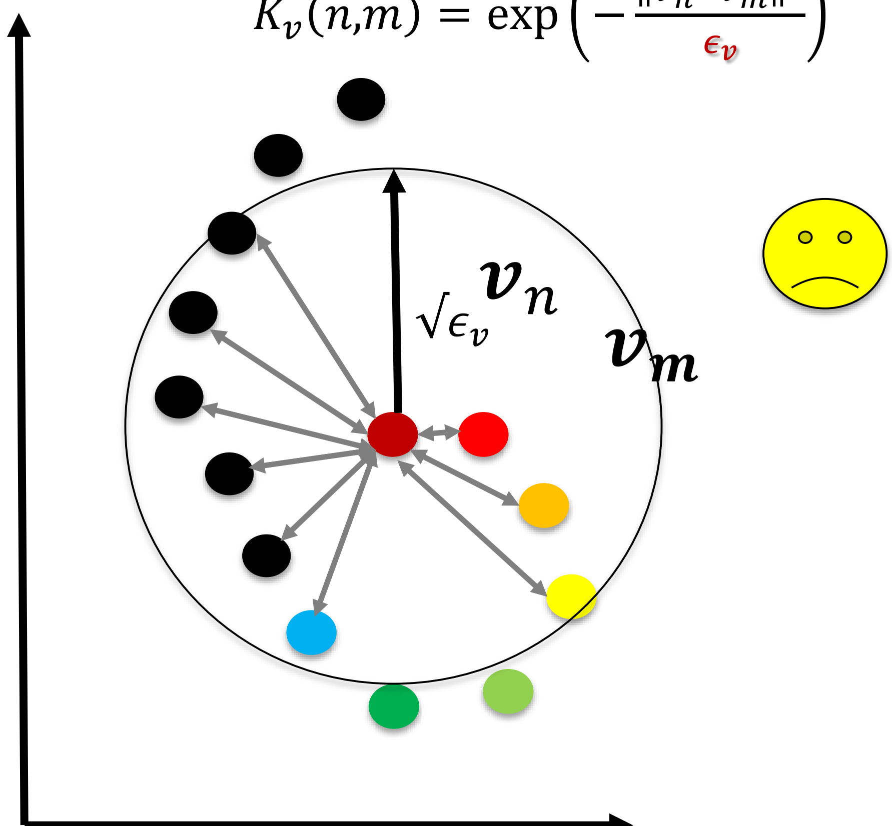
Manifold learning - the single modal case

- The tradeoff in kernel bandwidth (ϵ_v) selection ***trade-off***

$$K_v(n,m) = \exp\left(-\frac{\|v_n - v_m\|^2}{\epsilon_v}\right)$$



Too small kernel bandwidth
Disconnected graph

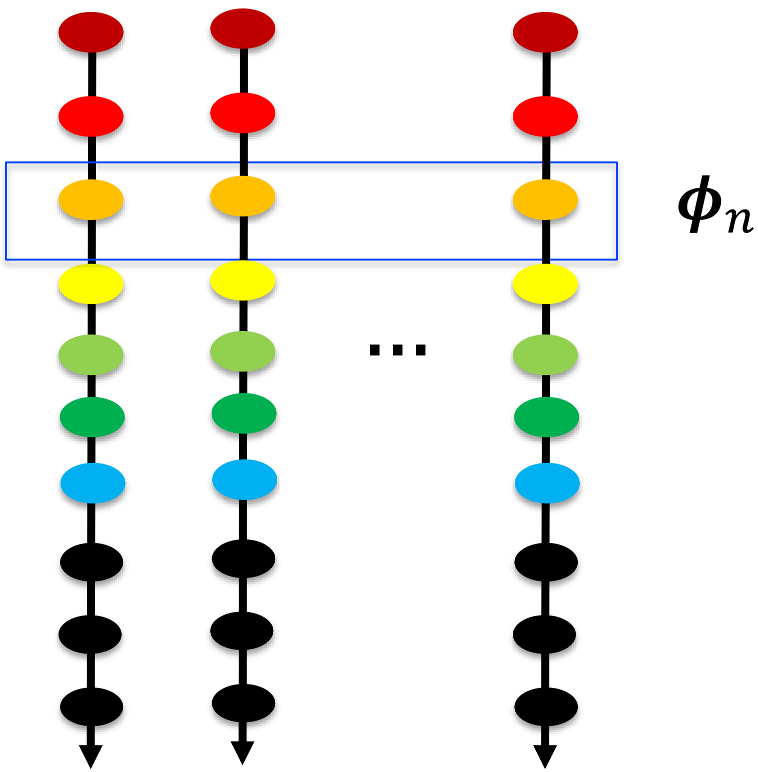


Too large kernel bandwidth
Wrong affinities

Manifold learning - the single modal case

Diffusion Maps (Coifman & Lafon 06):

- ❑ Row Normalize : $\boldsymbol{K} \rightarrow \boldsymbol{M} = \boldsymbol{D}^{-1}\boldsymbol{K}$
- ❑ Eigenvector decomposition of $\boldsymbol{M}$
- ❑ $\boldsymbol{\phi}_n$ is the nth row:



• Related studies – multimodal case

Kernel based approaches:

- ❑ Construct an affinity kernel $\boldsymbol{K}_v \in \mathbb{R}^{N \times N}$:

$$K_v(n,m) = \exp\left(-\frac{\|\boldsymbol{v}_n - \boldsymbol{v}_m\|^2}{\epsilon_v}\right)$$

- ❑ Combine the data:

$$\boldsymbol{K} = f(\boldsymbol{K}_v, \boldsymbol{K}_w)$$

[Wang 12', Lindenbaum et al. 15', Michaeli et al. 16', Vestner et al 17']

• Related studies – multimodal case

❑ Fusion by the product of kernels:

$$\boldsymbol{M} = \boldsymbol{M}_v \boldsymbol{M}_w$$

$\boldsymbol{M}_v, \boldsymbol{M}_w$ normalized versions (row stochastic) of $\boldsymbol{K}_v, \boldsymbol{K}_w$

❑ Analysis in [Lederman & Talmon 16', Talmon & Wu 18']:

- Representation according to common factors:

$$(\mathbf{v}_n(\boldsymbol{x}, \boldsymbol{y}), \mathbf{w}_n(\boldsymbol{x}, \boldsymbol{z})) \rightarrow \boldsymbol{\phi}_n(\boldsymbol{x})$$

- Alternating diffusion

- Limitations of the analysis

❑ What is the roll of the affinity kernel in the fusion process?

$$K_v(n,m) = \exp\left(-\frac{\|v_n - v_m\|^2}{\epsilon_v}\right)$$

❑ How to select the kernel bandwidths ϵ_v, ϵ_w?

❑ How the intensities of $\mathbf{x}, \mathbf{y}, \mathbf{z}$ ("SNR") effects the fusion?

- Main contributions

❑ Graph theoretic analysis of the product of kernels:

$$\boldsymbol{M} = \boldsymbol{M}_v \boldsymbol{M}_w$$

❑ Improved fusion via proper selection of the kernel bandwidth

$$K_v(n,m) = \exp\left(-\frac{\|v_n - v_m\|^2}{\epsilon_v}\right)$$

❑ Address the task of sound source activity detection

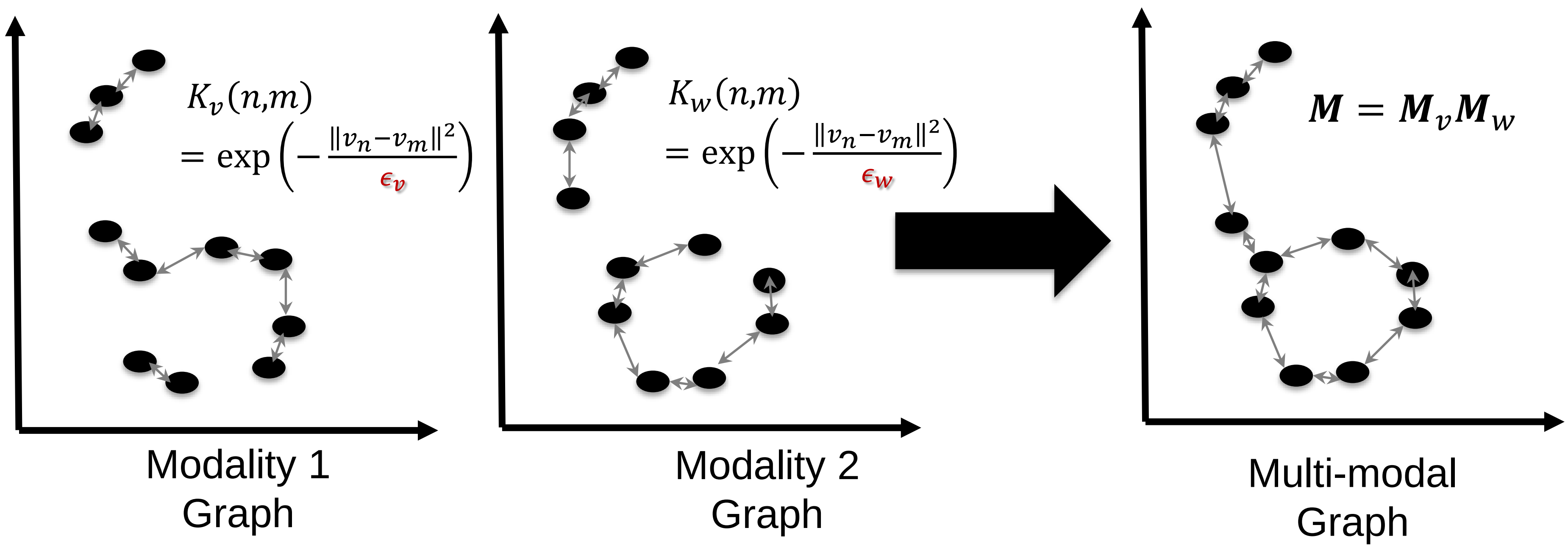
❑ Online setting and missing data

❑ The problem of multimodal correspondence

- Audio localization in video

• Proposed graph interpretation – multi-modal case

- The kernel product defines a ***multi-modal*** graph.
- Points n and m are connected if $\boldsymbol{M}_{n,m} \neq 0$



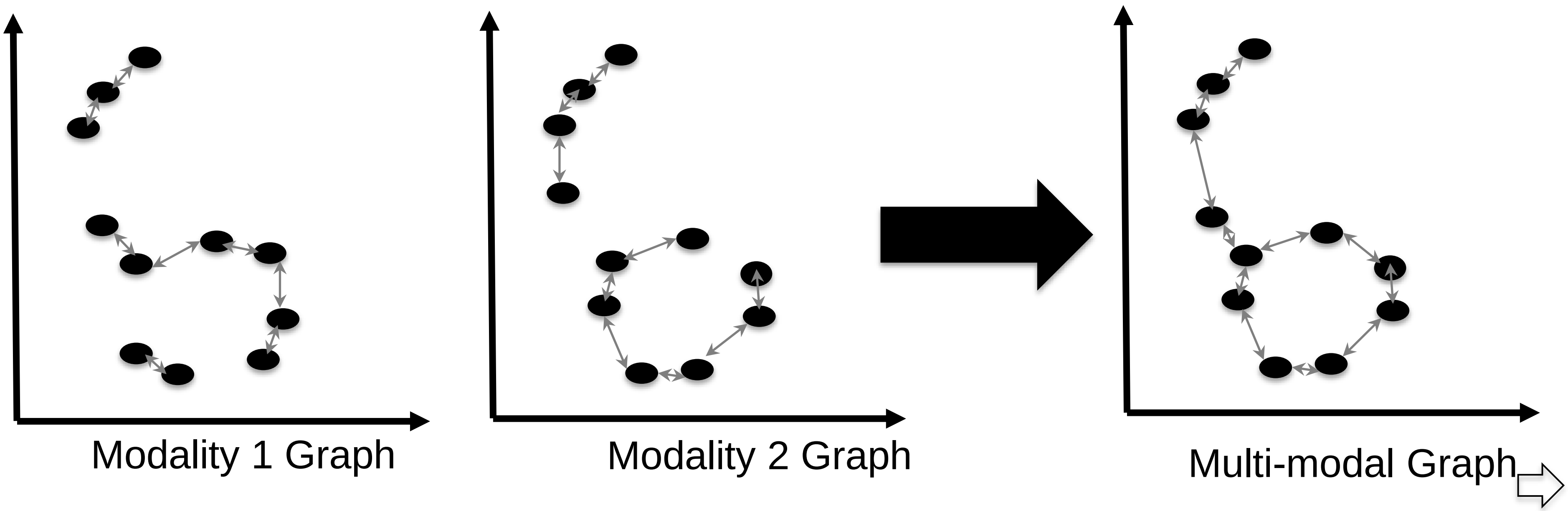
- Proposed graph interpretation – multi-modal case

Proposition1 [Dov, Talmon, and Cohen IEEE TSP 16']:

$\forall n,\ \exists m \neq n$ such that $M(n,m) \neq 0$ iff

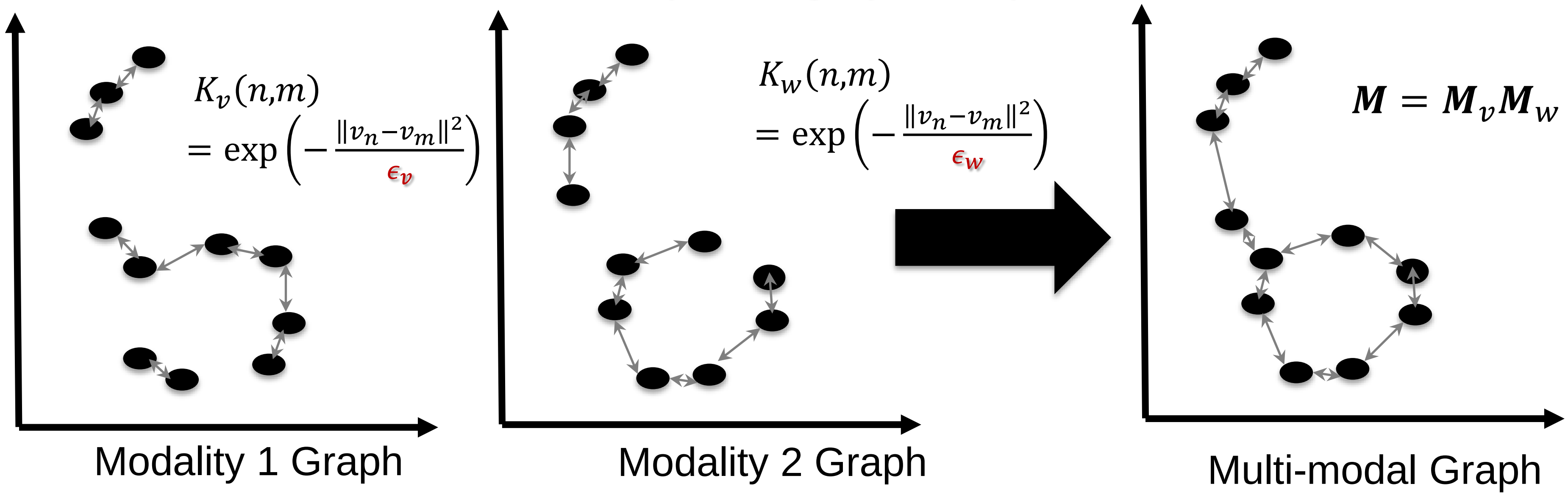
$\forall n,\ \exists m \neq n$ such that $M_v(n,m) \neq 0$ or $M_w(n,m) \neq 0$

- A point in the multi-modal graph is connected iff it is connected at least in one of the modalities



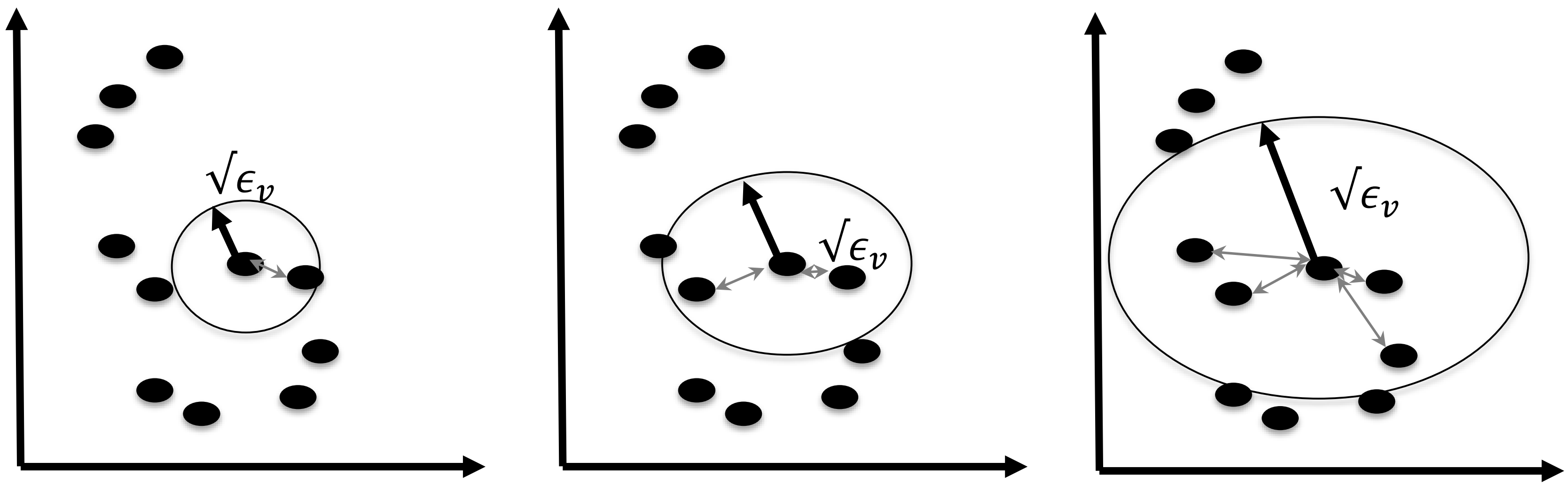
- Proposed graph interpretation – multi-modal case

  - ❑ The multi-modal graph may be connected even if the single-modal graphs are disconnected
  - ❑ Previous studies require the same connectivity as in the single modal case
  - ❑ ***The kernel bandwidth may be significantly reduced***



Proposed analysis of kernel bandwidth selection

- We relate between:
  - The kernel bandwidth
  - Average number of connections to each point



- Proposed analysis of kernel bandwidth selection

❑Assume a statistical model:

- The connectivity between a pair of points:

$$\mathbf{1}_v(n,m) = \begin{Bmatrix} 1 & w.p.\,p_v \\ 0 & \text{otherwise} \end{Bmatrix}$$

- IID
- Cross-modality independence

• Proposed analysis of kernel bandwidth selection

- We study the relation between the average number of connections in the single & multi-modal graphs

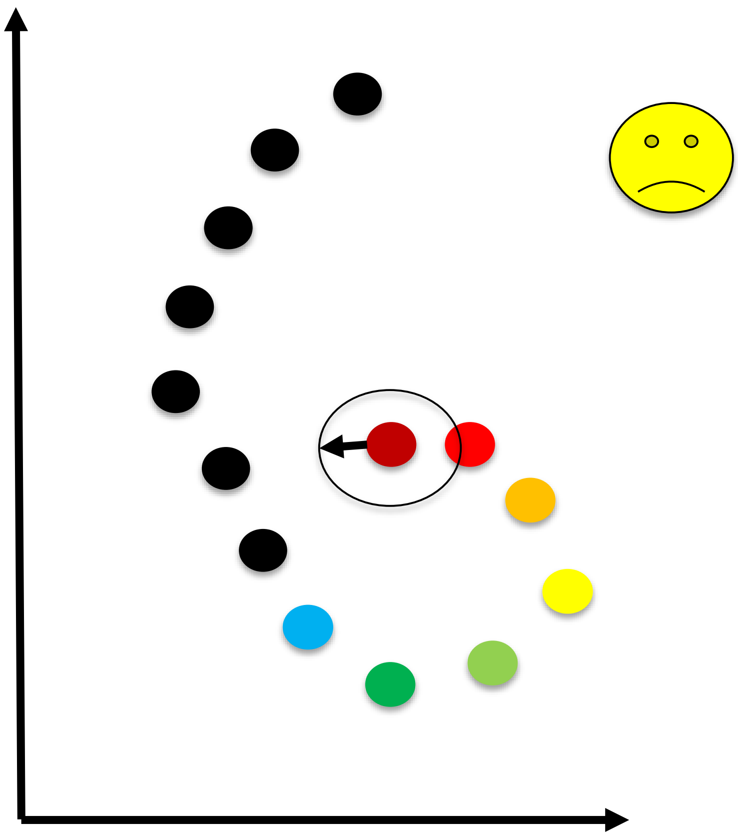
- Define the average number of connections:
  - S_v - modality 1
  - S_w - modality 2
  - S - multi-modal

Proposition 2 [Dov, Talmon, and Cohen IEEE TSP 16']:
the average number of connections in the multi-modal case: $S \xrightarrow[N\to\infty]{} S_v S_w$

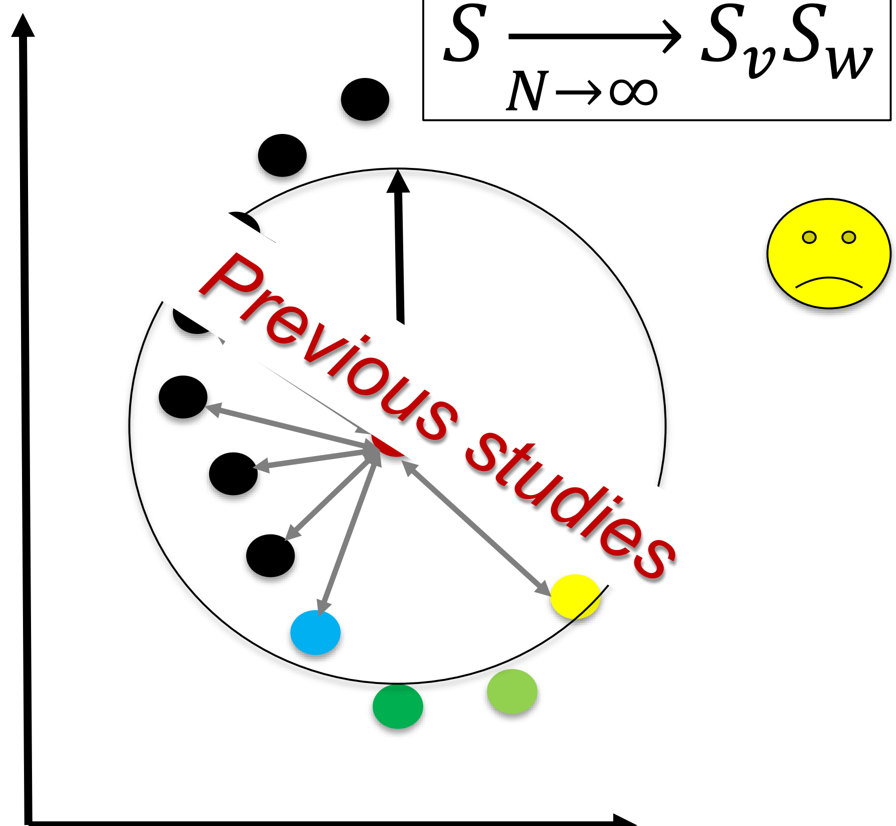
The tradeoff

- ❑ The tradeoff in kernel bandwidth (ϵ_v) selection ***trade-off***

$$S \xrightarrow[N \to \infty]{} S_v S_w$$



Too small kernel bandwidth
Disconnected graph



Too large kernel bandwidth
Wrong affinities

- Proposed algorithm for kernel bandwidth selection

Algorithm outline:

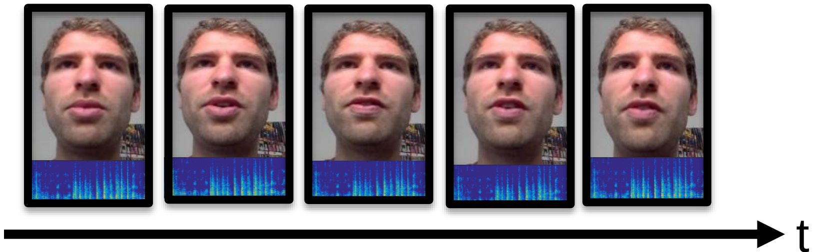
- ❑ Select the kernel bandwidth ϵ_v as in the single-modal case
- ❑ Estimate the average number of connections $\delta = S_v$:
  - $\hat{\delta} = (N-1)\hat{p}_v = \frac{1}{N}\sum_m \sum_{n \neq m} K_v(n,m)$
- ❑ Reduce the kernel bandwidth until:

$$\delta^{\mathrm{AD}} = \sqrt{\hat{\delta}}$$

via an iterative search

- Sound source activity detection

❑ Given audio visual signals:



$(\boldsymbol{v}_1, \boldsymbol{w}_1), (\boldsymbol{v}_2, \boldsymbol{w}_2), \quad \dots \quad , (\boldsymbol{v}_N, \boldsymbol{w}_N)$

❑ Goal: for each frame, estimate the activity of the common source:

$$\mathbf{1}_n(\boldsymbol{x}) = \begin{cases} 1 & ; \quad n \in \mathcal{H}_1 \\ 0 & ; \quad n \in \mathcal{H}_0 \end{cases}$$

- Proposed algorithm for sound source activity detection

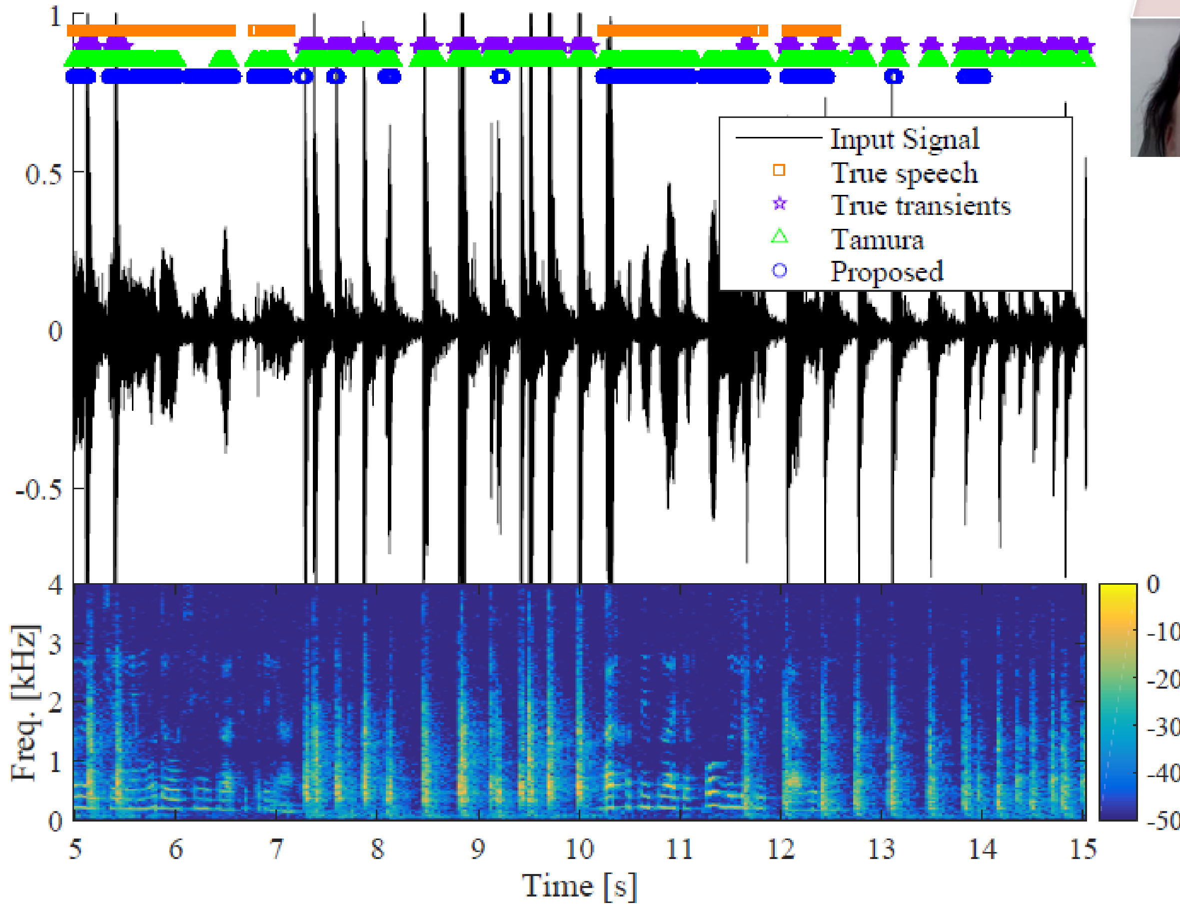
Proposed algorithm outline:

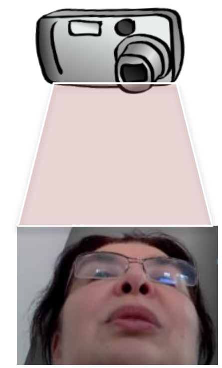
- ❑ Construct the *improved* affinity kernels: $\boldsymbol{M}_v$ and $\boldsymbol{M}_w$
- ❑ Fuse the modalities: $\boldsymbol{M} = \boldsymbol{M}_v \boldsymbol{M}_w$
- ❑ Use the leading eigenvector $\boldsymbol{\phi}_1 \in R^N$
- ❑ Activity indicator:

$$\hat{1}_n(\boldsymbol{x}) = \begin{Bmatrix} 1 & ; & \phi_1(n) > \tau \\ 0 & ; & \text{otherwise} \end{Bmatrix}$$

- Experimental Results

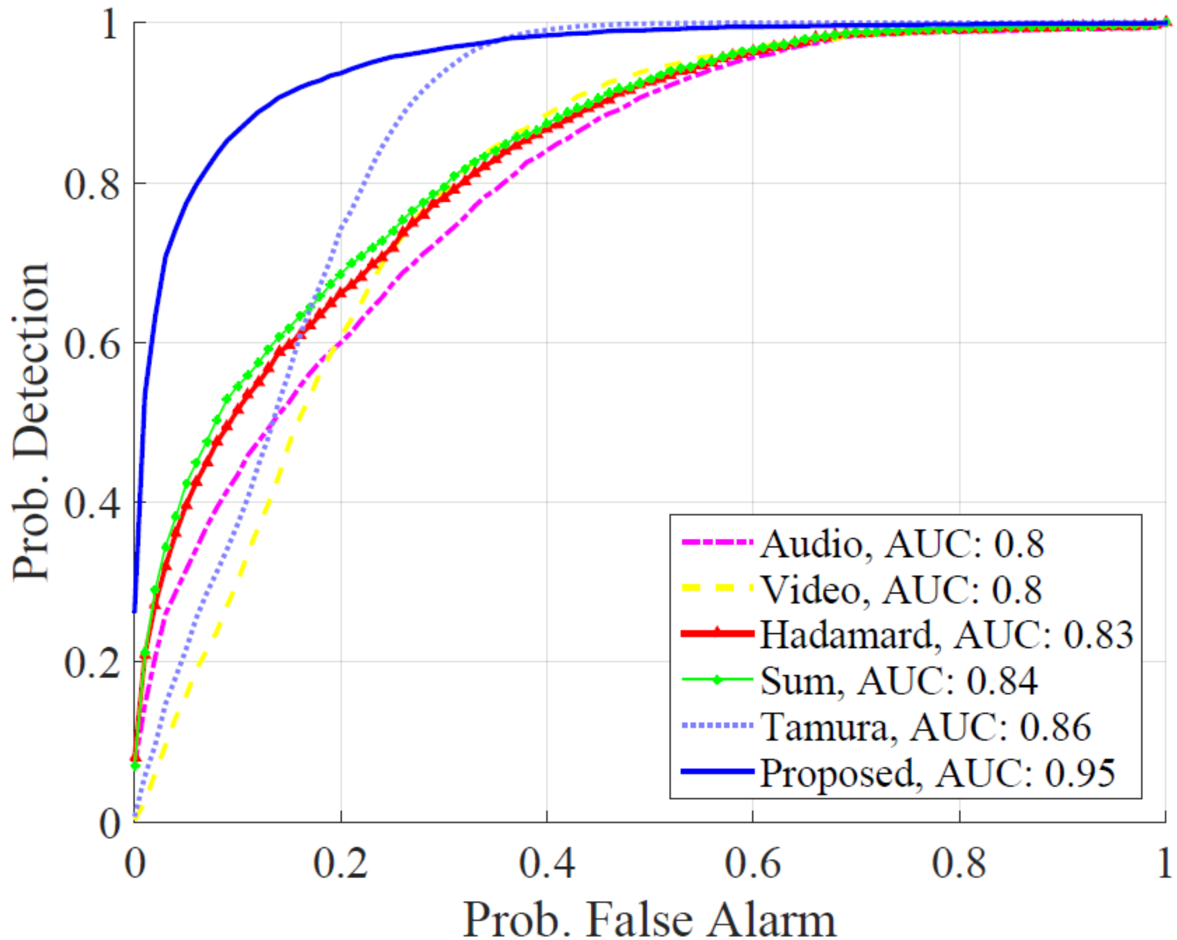
❑ Voice activity detection. Transient type: hammering





- Experimental Results

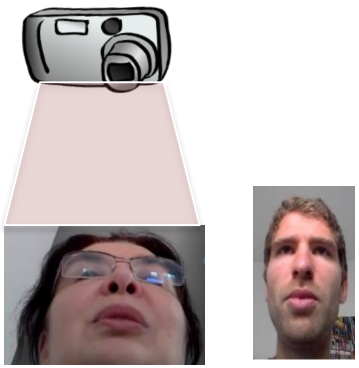
❑ ROC curves. Transient type: hammering

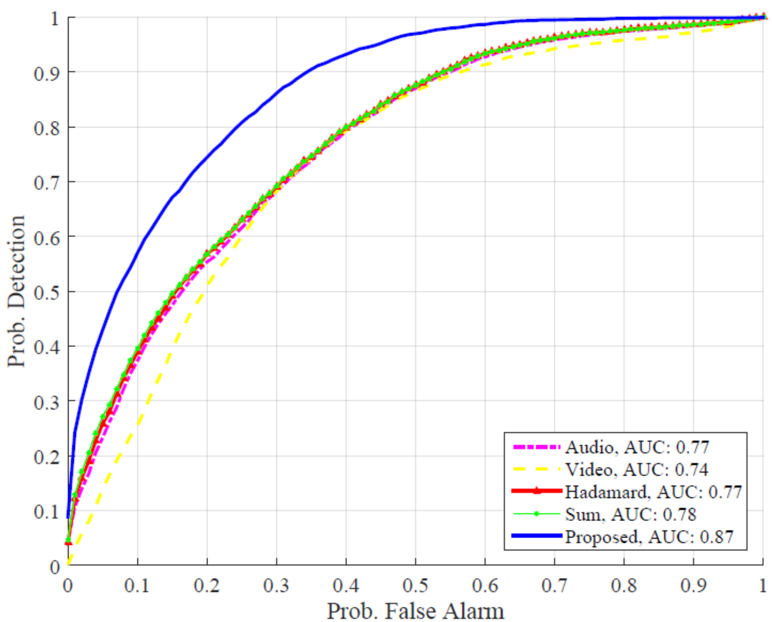


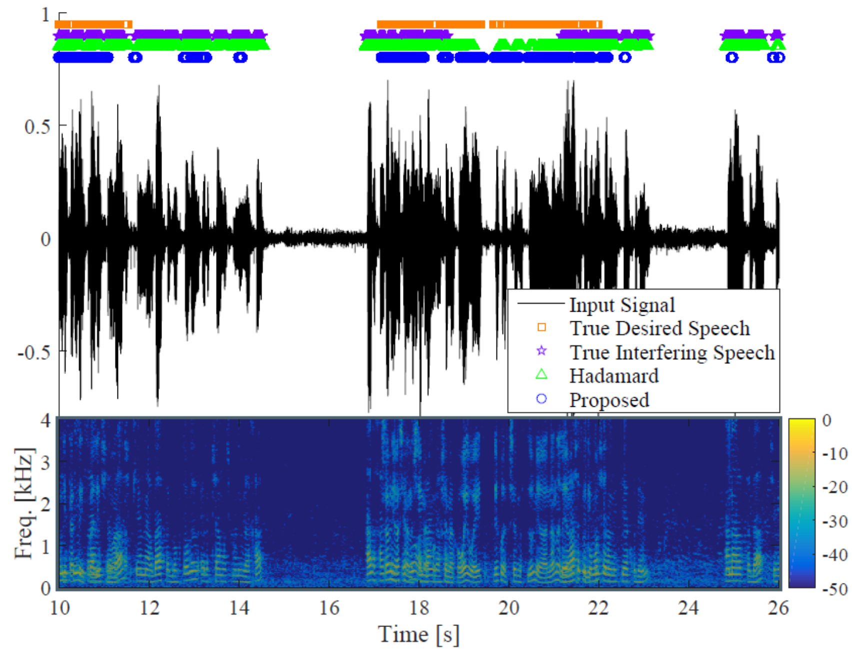
Multimodal Signal Processing on Manifolds

• Application – desired speaker activity detection

- ❑ Interfering source: speech of another speaker
- ❑ Challenge: same acoustic characteristics to the desired and the interfering sources







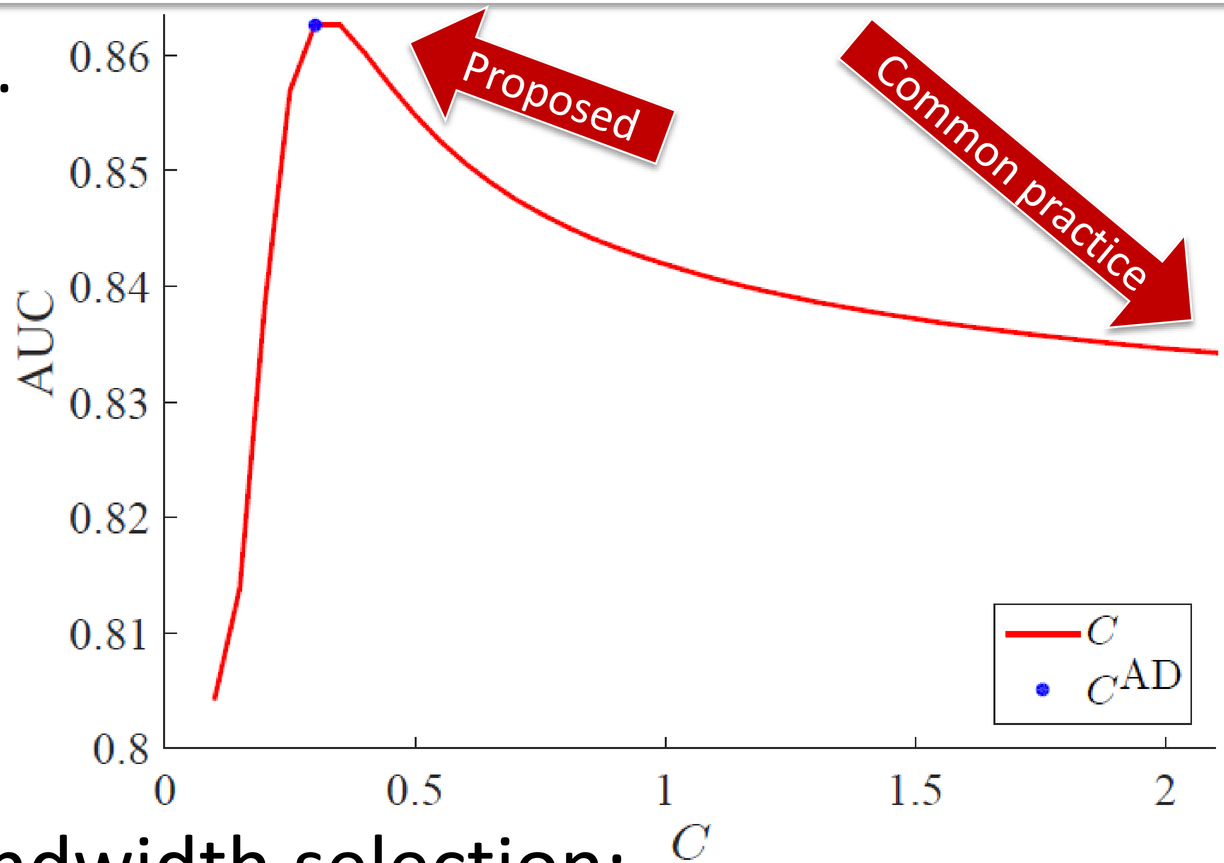
[Dov, Talmon, and Cohen, ICSEE 16']

- Experimental results: voice activity detection

❑ ROC curves.

Babble noise

-5 dB SNR

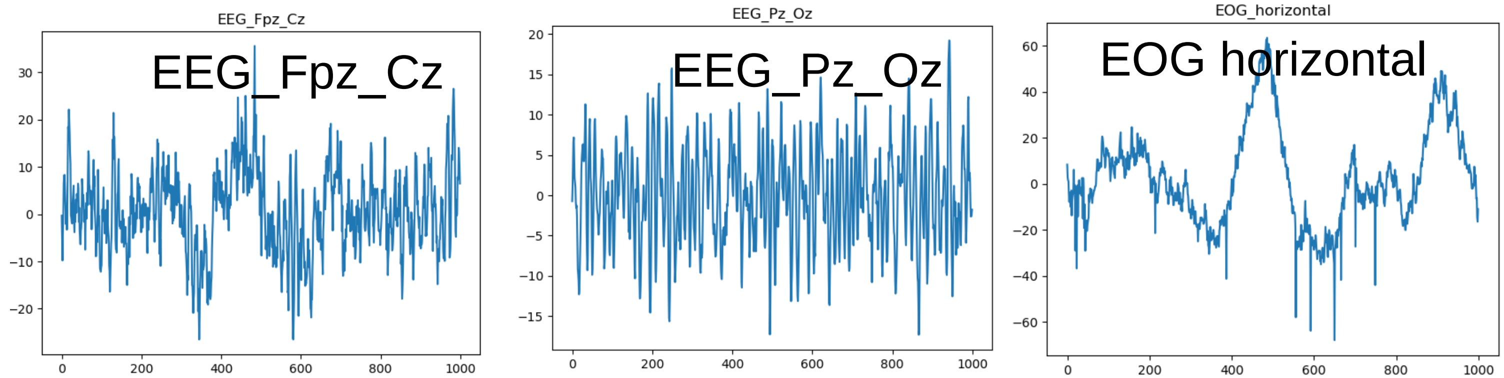


❑ Kernel bandwidth selection:

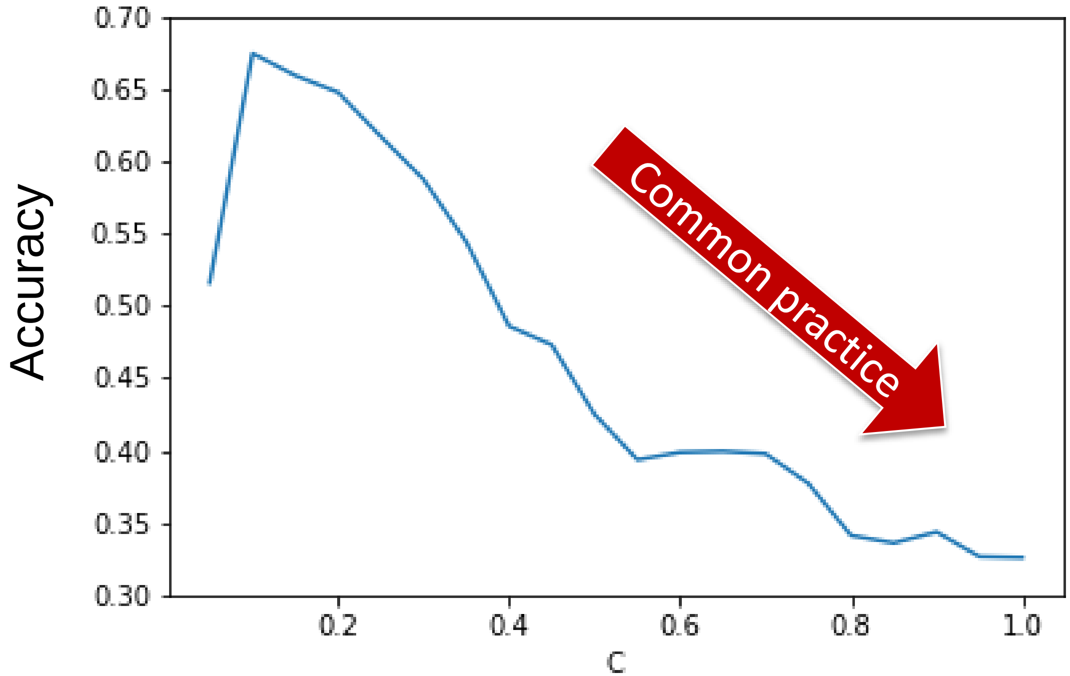
$$\epsilon_v = C \cdot \max_m [\min_n (||\boldsymbol{v}_n - \boldsymbol{v}_m||^2)]$$

• Sleep stages classification [joint with Jonas Laake]

- ❑ Modalities:



- ❑ Classes: REM, shallow sleep (NREM 1,2), deep sleep (NREM 3,4)

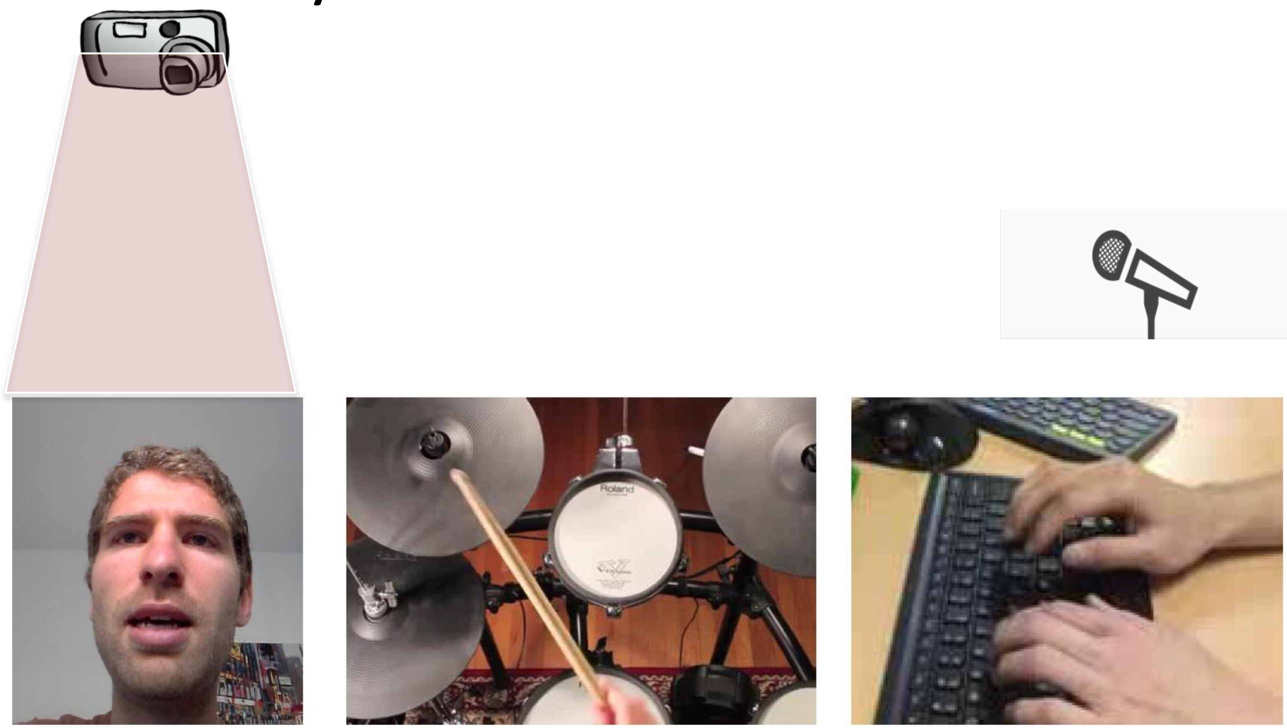


- Extending the fusion problem

❑ Online

❑ Limited availability of the sensors

❑ Sound scene analysis



- Fusion in an online setting

❑ A short calibration set: $\{\boldsymbol{v}_r(\boldsymbol{x},\boldsymbol{y}),\boldsymbol{w}_r(\boldsymbol{x},\boldsymbol{z})\}_{\mathbf{1}}^{R}$

Goal:

❑ Unified representation: $\boldsymbol{\phi}_n \in \mathcal{R}^L$

$$\boldsymbol{v}_n(\boldsymbol{x},\boldsymbol{y}) \rightarrow \boldsymbol{\phi}_n(\boldsymbol{x})$$

❑ Reduce the effect of structured interferences

• Out of sample extension

Single-modal approach:

❑ Obtain a representation using the reference set:

$$(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \dots, \boldsymbol{\phi}_L)$$

❑ Online extension (Nystrom method) [Fowlkes 04]:

$$\phi_j(n) = \frac{1}{\lambda_j} \sum_{r=1}^{R} M_v\,(n, r)\phi_j(r)$$

• Out of sample extension

Multi-modal approach:

- ❑ Obtain a representation using the reference set:

$$(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \dots, \boldsymbol{\phi}_L)$$

- ❑ Online extension :

$$\phi_j(n) = \frac{1}{\lambda_j} \sum_{r=1}^{R} M(n,r)\phi_j(r)$$

$$= \frac{1}{\lambda_j} \sum_{m=1}^{R} M_v(n,m)\, f(m)$$

$$f(m) \triangleq \sum_{r=1}^{R} M_w(m,r)\phi_j(r)$$

• Extending the fusion problem

We take ***advantage*** of the ***limitation*** of the extension and show:

- ❑ Multimodal geometric structure ***can*** be learned from a short "calibration" set
- ❑ The common source can be extract from ***one modality***:

$$v_n(x,y) \rightarrow \phi_n(x)$$

- ❑ Challenging interfering sources such as ***speech*** are reduced

- Proposed algorithm for sound scene analysis

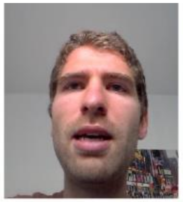
- ❑ Point a video camera to a particular source of interest
- ❑ Construct the multimodal representation via $\boldsymbol{M} = \boldsymbol{M}_v \boldsymbol{M}_w$
- ❑ Extend the representation to new frames using *one* modality

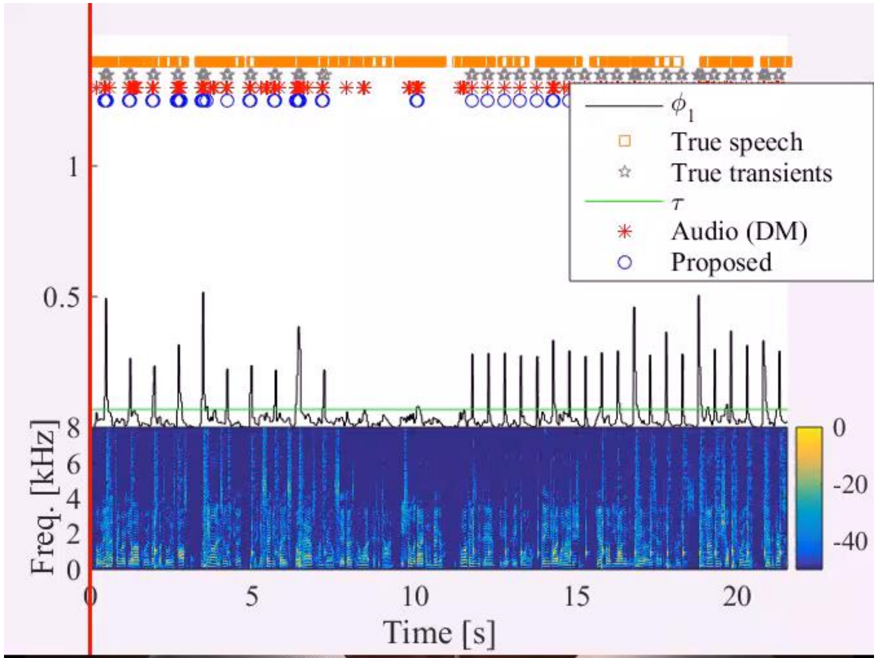
[Dov, Talmon, Cohen ACM/IEEE TASLP 17']

• Sound source activity detection

Results:

- ❑ Source of interest:
  - Drums beats
- ❑ Interfering source:
  - Speech

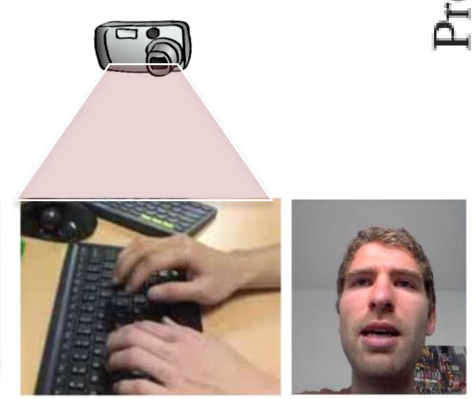


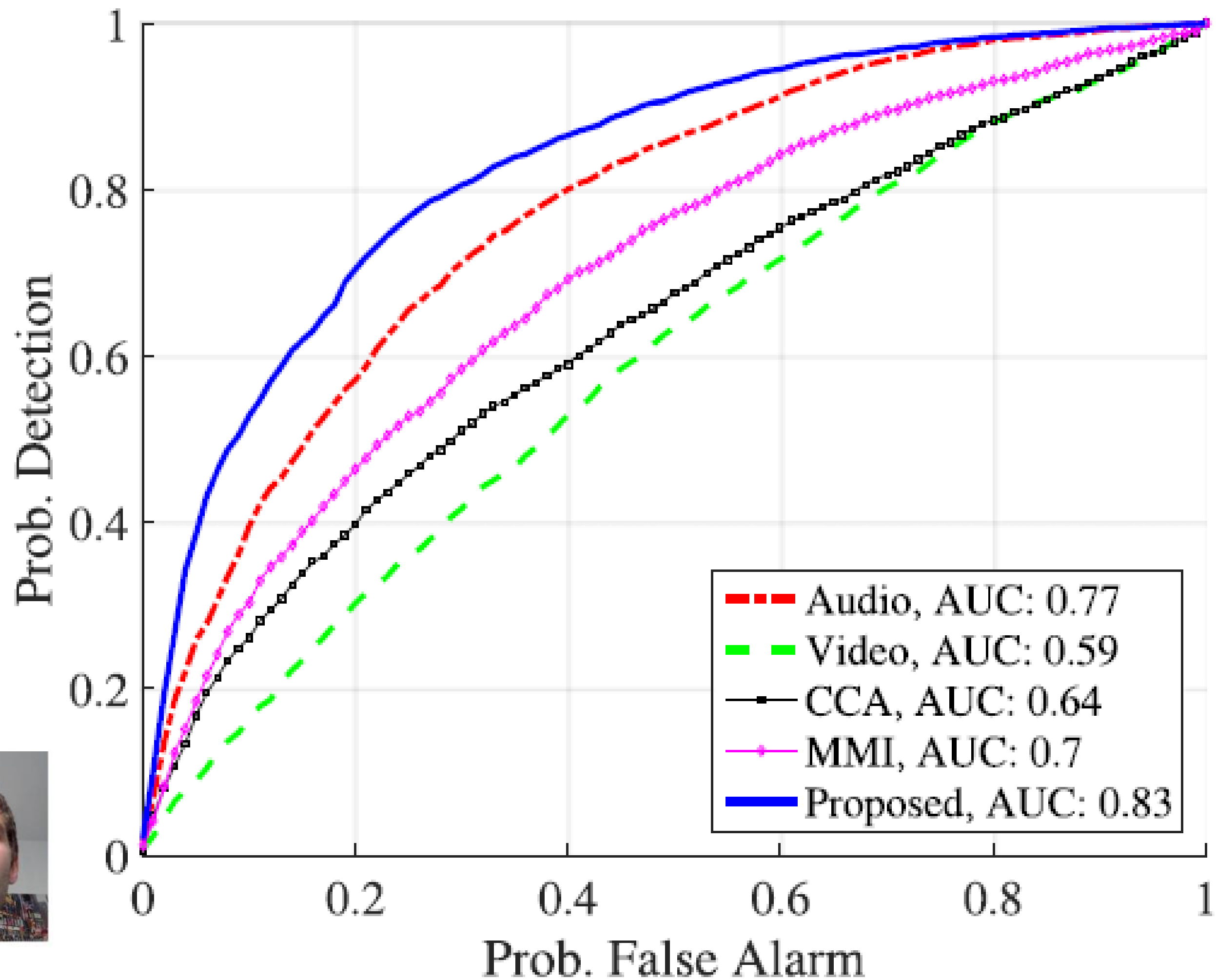


• Sound source activity detection

Results:

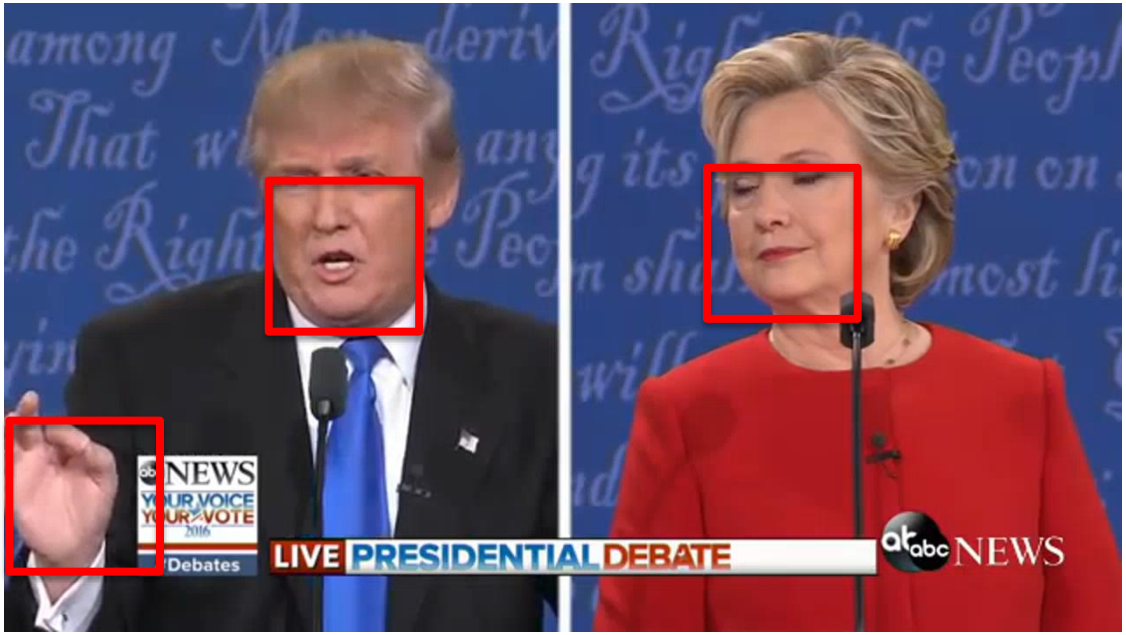
- ❑ Source of interest:
  - Keyboard-taps
- ❑ Interfering event:
  - Speech





Measuring multimodal correspondence

- ❑ Example1: audio localization in video



- ❑ Which part of the video corresponds more to the audio?

• Why the problem is challenging – example 2

❑ Very “simple” case:

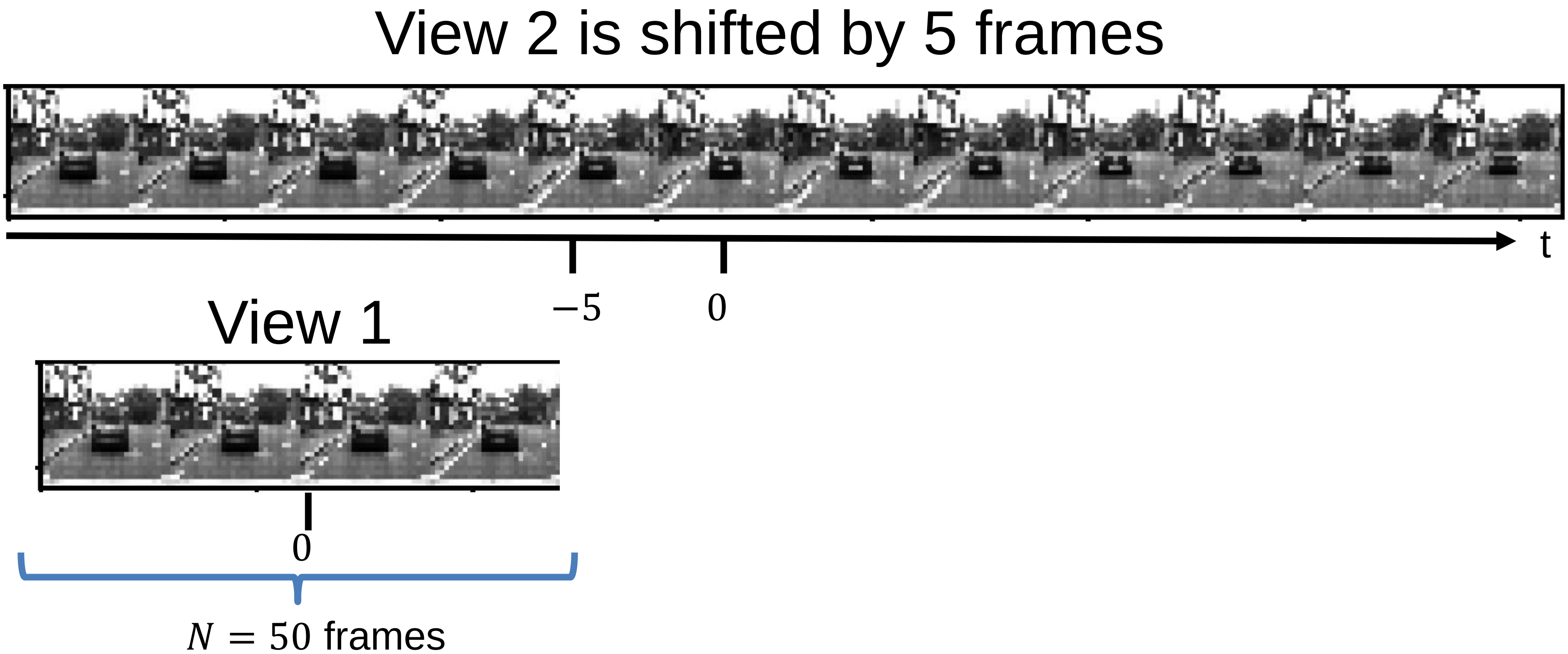
- Multi-view (not multi-modal)
- Almost the same view

View 1 View 2 View 2 shifted

• Why the problem is challenging

❑ Application: synchronization

- Measure cross-correlation

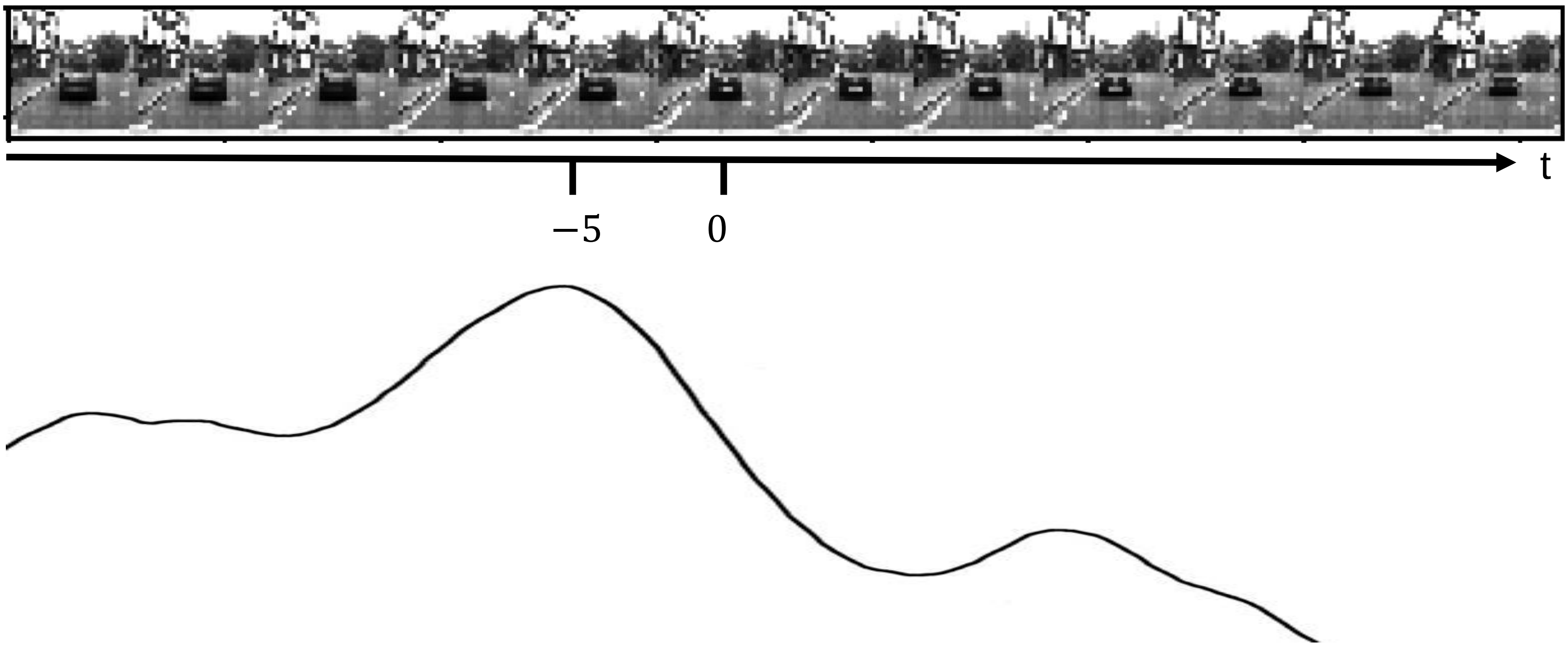


- Why the problem is challenging

❑ Application: synchronization

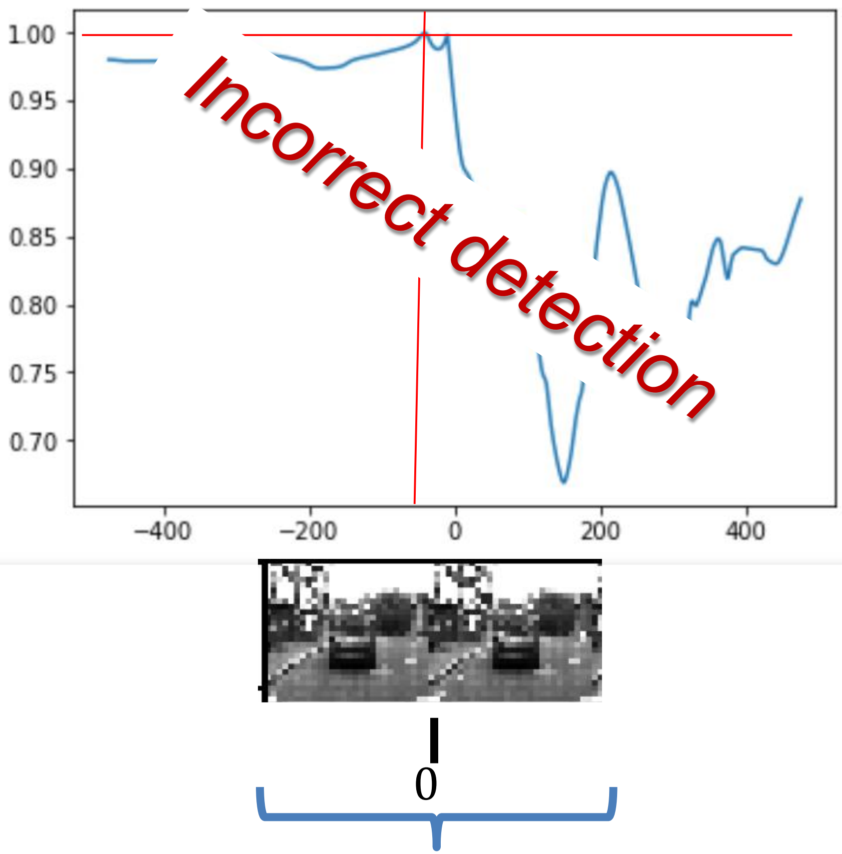
  - Measure cross-correlation

View 2 is shifted by 5 frames

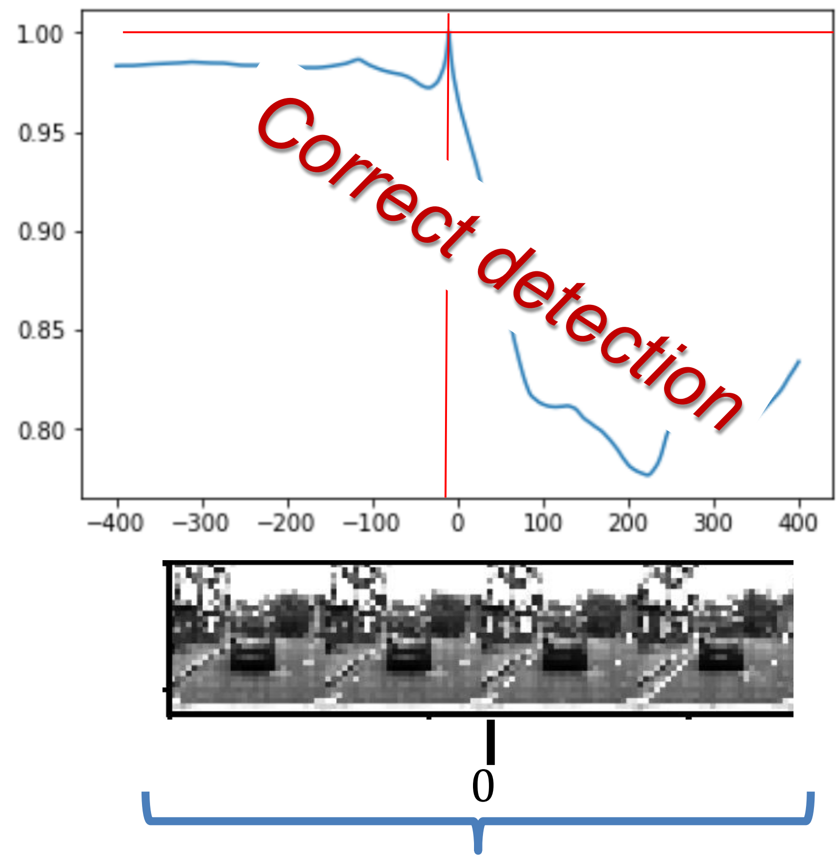


• Why the problem is challenging

❑ Apply cross-correlation to find the shift:



$N = 50$ frames



$N = 200$ frames

- Proposed measure of multimodal correspondence

❑ Trace of the kernel product:

$$\mathrm{Tr}\{\boldsymbol{M}\}$$

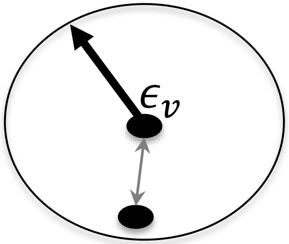
- Proposed graph interpretation

Graph interpretation of:

$$\mathrm{Tr}\{\boldsymbol{M}\}$$

❑ Recall the graph interpretation of the affinity kernel:

$$K_v(n,m) = \exp\left(-\frac{\|\mathbf{v}_n - \mathbf{v}_m\|^2}{\epsilon_v}\right)$$



❑ The statistical model for the connectivity:

$$\mathbf{1}_v(n,m) = \begin{Bmatrix} 1 & w.p.\, p_v \\ 0 & \text{otherwise} \end{Bmatrix}$$

• Proposed graph interpretation

- Consider the extreme cases
  - The modalities are *uncorrelated (UC)*
  - The modalities are *fully correlated (C)*
- Assume:

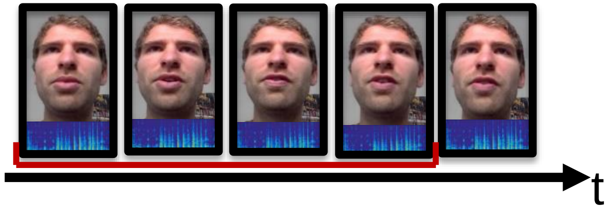
$$p_v = p_w \triangleq p \in (0,1)$$

Proposition 1 [Dov, Talmon, Cohen IEEE TSP 18']:

$$E^{\mathbf{UC}}\{\mathrm{Tr}\{\boldsymbol{M}\}\} = p \cdot E^{\mathbf{C}}\{\mathrm{Tr}\{\boldsymbol{M}\}\} < E^{\mathbf{C}}\{\mathrm{Tr}\{\boldsymbol{M}\}\}$$

• Measuring multimodal correspondence

Fast online update of the proposed measure



Proposition 2 [Dov, Talmon, Cohen IEEE TSP 18']:

$$\mathrm{Tr}\left\{\mathbf{M}\right\} = \mathrm{Tr}\left\{\mathbf{D}_v^{-1}\mathbf{D}_w^{-1}\mathbf{K}_v\mathbf{K}_w\right\} = \mathrm{Tr}\left\{\mathbf{DK}\right\} \triangleq \sum_{n=1}^{N} D(n,n)K(n,n)$$

$$\boldsymbol{D} \triangleq \boldsymbol{D}_v^{-1}\boldsymbol{D}_w^{-1}, \boldsymbol{K} \triangleq \boldsymbol{K}_v\boldsymbol{K}_w$$

- Measuring multimodal correspondence

❑ Fast online update of $K(n,n)$:

$$\widetilde{K}(n,n) = K(n,n)$$
$$-K_v(n,1)K_w(n,1)$$

- $\widetilde{K}(n,n)$ is the updated kernel

$\boldsymbol{K}$

Proposed measure

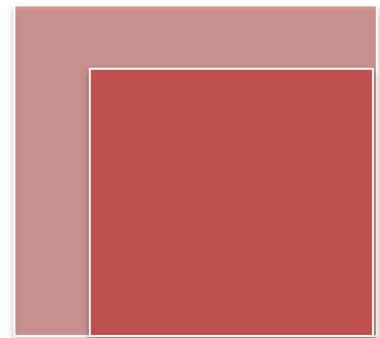
$$\mathrm{Tr}\left\{\mathbf{M}\right\} = \sum_{n=1}^{N} D(n,n)K(n,n)$$

• Measuring multimodal correspondence

❑ Fast online update of $K(n,n)$:

$$\widetilde{K}(n,n) = K(n,n) - K_v(n,1)K_w(n,1)$$

K



- $\widetilde{K}(n,n)$ is the updated kernel

Proposed measure

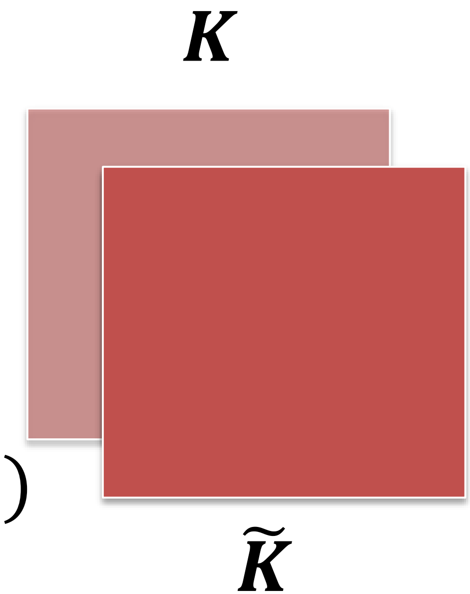
$$\mathrm{Tr}\{\mathbf{M}\} = \sum_{n=1}^{N} D(n,n)K(n,n)$$

• Measuring multimodal correspondence

❑ Fast online update of $K(n,n)$:

$$\widetilde{K}(n,n) = K(n,n)$$
$$-K_v(n,1)K_w(n,1)$$
$$+K_v(n,N+1)K_w(n,N+1)$$

- $\widetilde{K}(n,n)$ is the updated kernel



❑ Complexity:

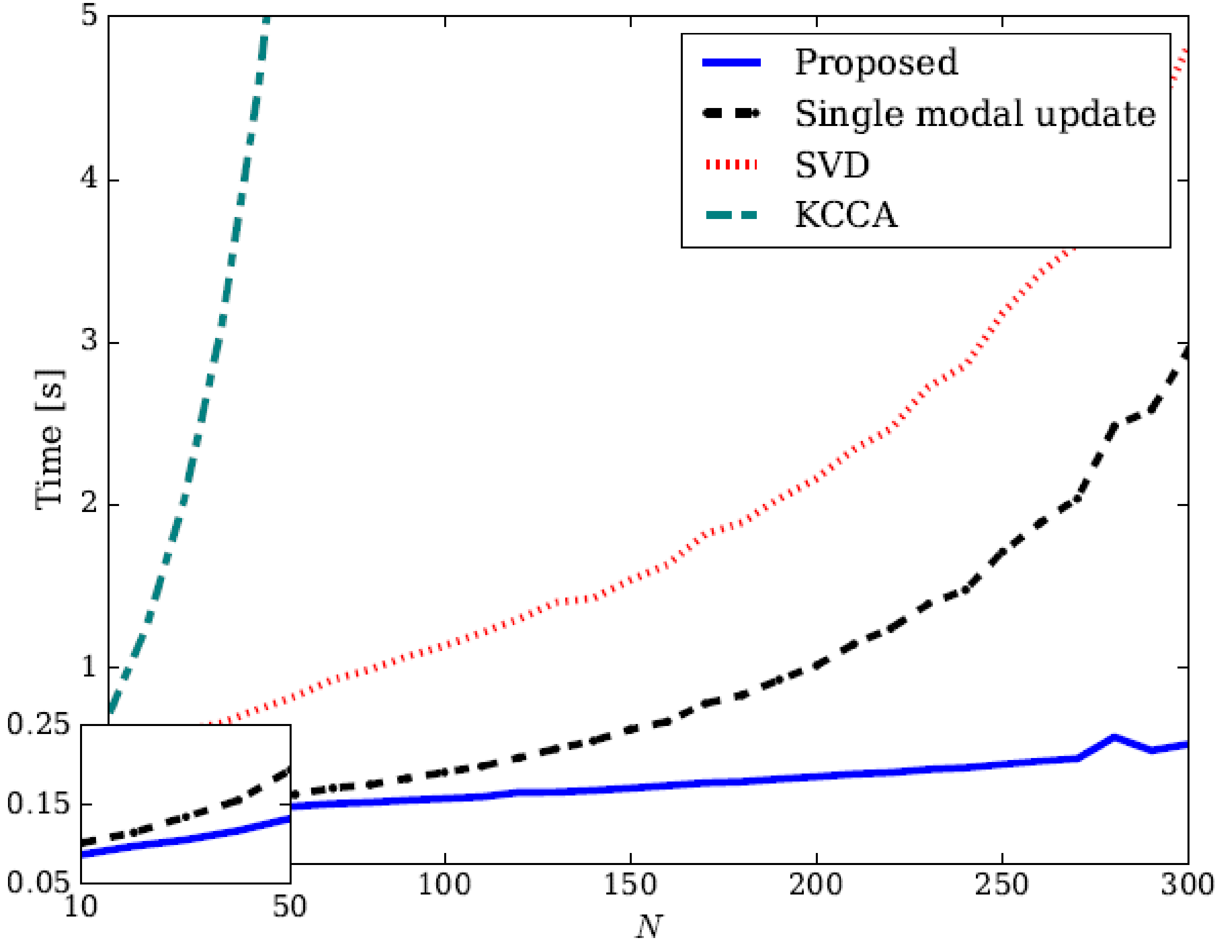
- $O(N)$
- No matrix product ($> O(N^2)$)

Proposed measure

$$\mathrm{Tr}\left\{\mathbf{M}\right\} = \sum_{n=1}^{N} D(n,n)K(n,n)$$

• Measuring multimodal correspondence

Runtime simulations:



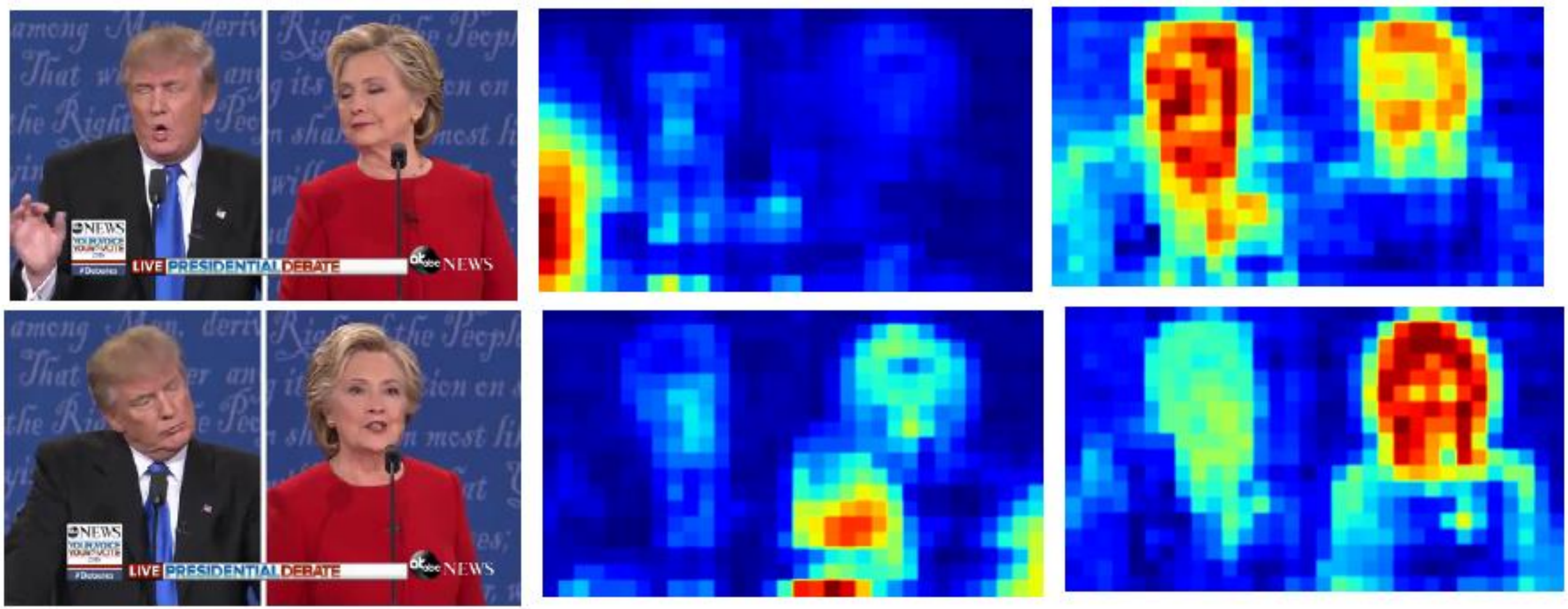
Measuring multimodal correspondence

- ❑ Example: audio localization in video

- ❑ Which part of the video corresponds more to the audio?

- Measuring multimodal correspondence

Audio localization in video



Motion in video

Proposed

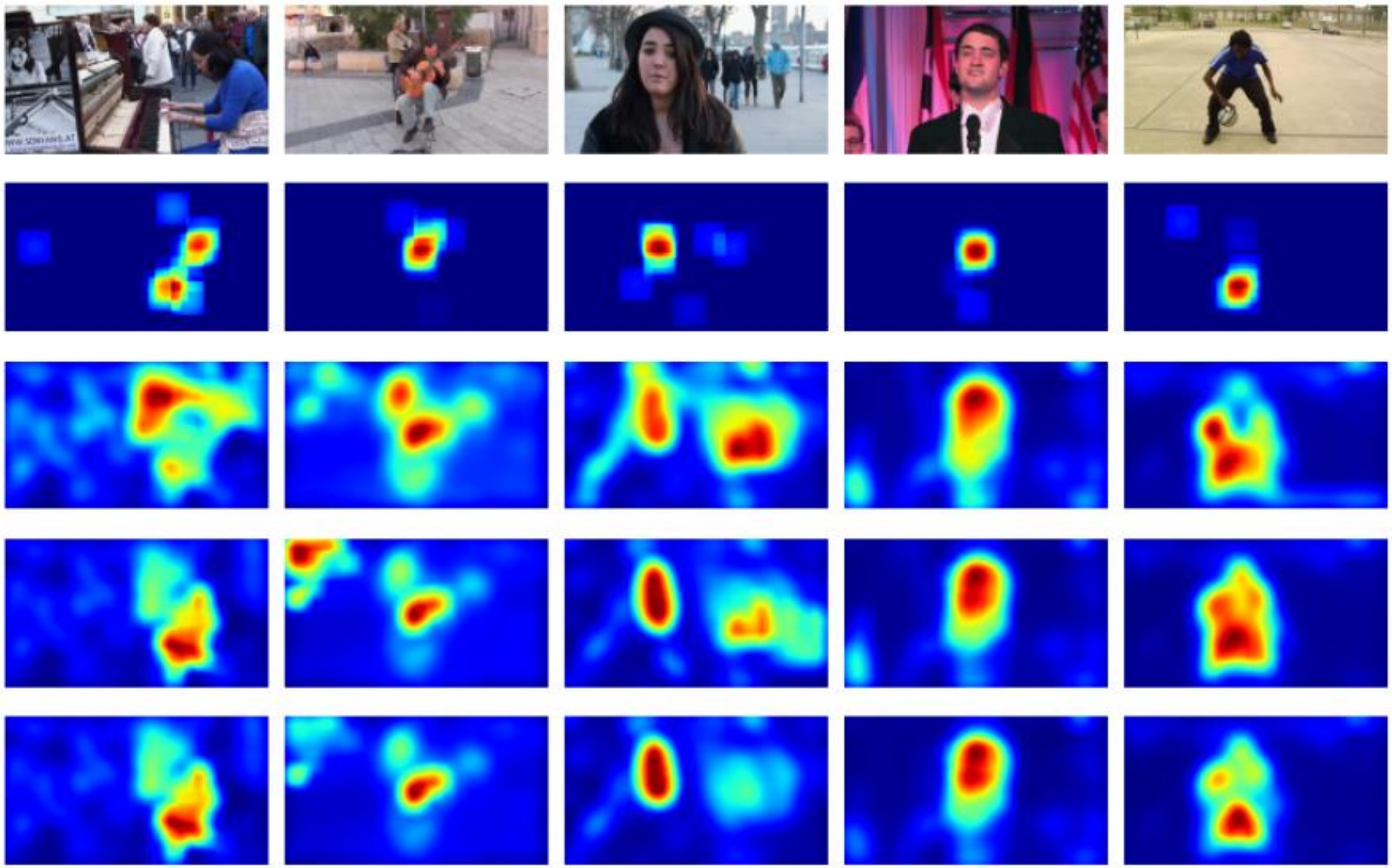
- Measuring multimodal correspondence

Eye fixation prediction

❑ Find the salient regions in the video

- Measuring multimodal correspondence

Eye fixation prediction



- True gaze
- [Min 16’]
- KCCA
- Proposed

- Measuring multimodal correspondence

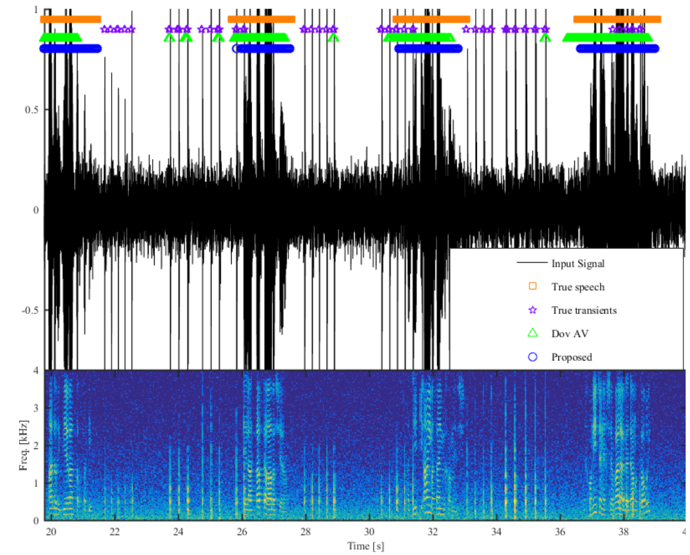
Eye fixation prediction

Algorithm	sAUC	CC	NSS
Video only	0.7292	0.3612	1.4295
KCCA	0.7628	0.4362	1.7904
Empirical HSIC	0.7530	0.4197	1.7229
Zhang et al. 2016	0.7235	0.3725	1.4667
Izadinia et al. 2013	0.6915	0.3519	1.5165
Min et al. 2016	0.7556	0.4182	1.6941
Proposed	**0.7660**	**0.4432**	**1.8309**

• Note about neural networks

- ❑ Transient reducing autoencoders + RNN for audio-visual VAD

[Ariav, Dov, and Cohen, Signal Processing, 18’]



- ❑ Synchronization in audio-visual recordings

[Aides, Dov, and Aronowitz, ICASSP 2018]

Passphrase	[12] (l_2)	S_{DTW} (proposed)
My voice...	0.7	1.68
Please verify...	3	2.84
Average	1.85	2.26

Passphrase	[12] (l_2)	S_{DTW} (proposed)
My voice...	32.71	2.98
Please verify...	31.07	4.39
Average	31.89	3.69

• Conclusions

- ❑ Kernel based multi-modal fusion
  - Missing data
  - Structured interferences
  - No labels and external training datasets
- ❑ *Insights* via discrete analysis using *graph* theory:
  - Relation between connectivity within and between modalities
  - Not as in the single modal case
- ❑ *Challenging* audio-visual tasks

- Future work

- ❑ Learning modality specific (vs common) factors
- ❑ Measuring an "SNR" style ratio between modality specific to common factors
- ❑ Sensor selection
- ❑ Going beyond 2 modalities
  - ❑ Fusion
  - ❑ Missing sensors
- ❑ Sensor reliability and liveness

- The End

Thank you!