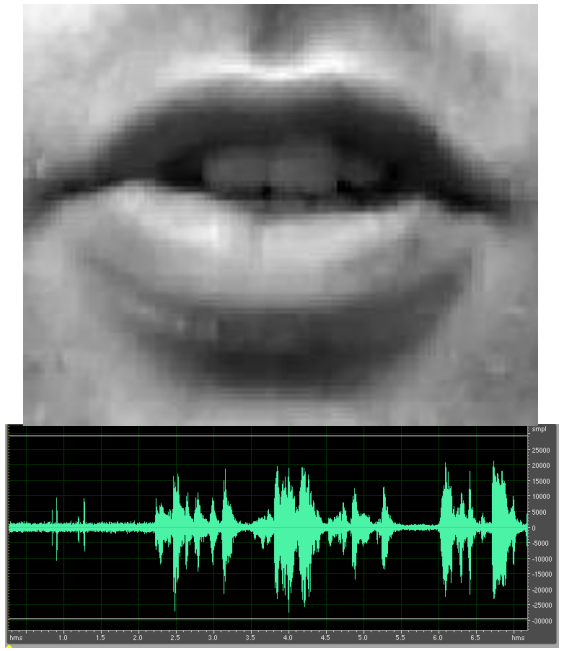
Audio-Visual Speech Processing Using Diffusion Maps and The Scattering Transform



- M. Sc. work by David Dov
- Supervised by: Prof. Israel Cohen

Outline

- Motivation and Related Work
- Audio-Visual Voice Activity Detection (VAD) Using Diffusion Maps
- VAD using the Scattering Transform
- Conclusions & Future Work

• Outline

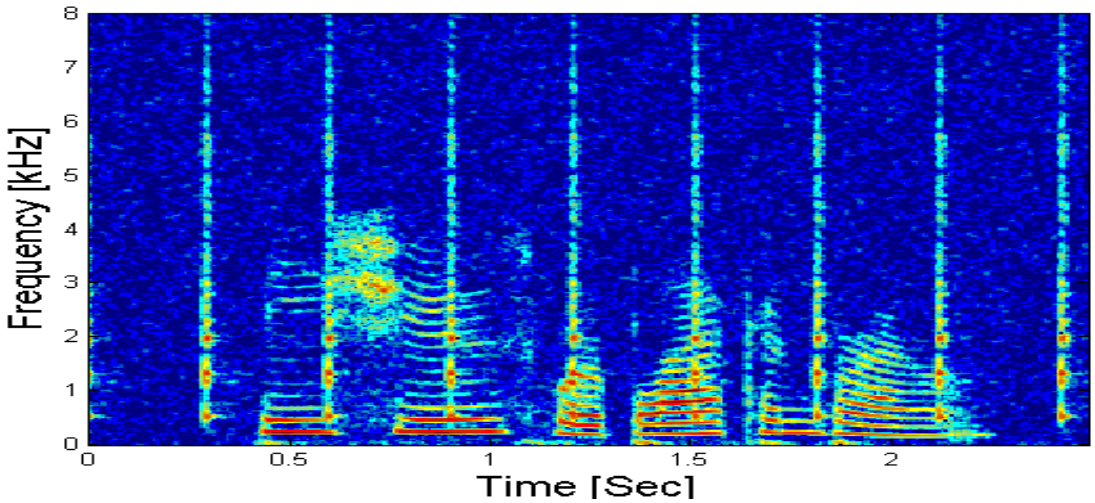
- *Motivation and Related Work*
- Audio-Visual Voice Activity Detection (VAD) Using Diffusion Maps
- VAD using the Scattering Transform
- Conclusions & Future Work

• Motivation

- Speech processing in noisy acoustic environments:
- The model: *speech* + *quasi-stationary noise* + *transients*

$$a[n] = a^s[n] + a^d[n] + a^t[n]$$

- Example:



- The challenges:
  - *Non-stationary noise*
  - *Transients*: short time interruptions, e.g., keyboard typing

• Motivation

- The common practice in the processing of noisy acoustic environments:
  - Exploit STFT domain: $A(j,k) = A^s(j,k) + A^d(j,k) + A^t(j,k)$
  - Assume: a statistical model for speech and the noise
  - Assume: the noise is slowly varying with respect to speech
- *The main limitation:*
  - Transients are poorly modeled by statistical models
  - Transients are not slowly varying with respect to speech

• Motivation

- One solution- incorporate the *visual signal*:
  - Available
  - Invariant to the acoustic environment

- Challenges:
  - High dimensional data
  - Non speech mouth movements (“transients”)
  - Large variability between speakers

• Motivation

- The goal:
  - Design signal representations where speech is separated from the interruptions
- Applications:
  - Classification
  - Speech Enhancement

• Motivation

- Application example: *Voice activity detection*
  - Speech/ non-speech classification
- Essential for:
  - Speech and speaker recognition
  - Speech coding
  - Speech enhancement
  - Dominant speaker identification

• Related Work

- State of the art *Voice Activity Detectors (VAD):*
  - Based on the log of the Likelihood Ratio:

$$\Lambda_j(k) = \log\left(\frac{p_r\left(A(j,k);\mathcal{H}_1\right)}{p_r\left(A(j,k);\mathcal{H}_0\right)}\right)$$

- Likelihood ratio estimation based on statistical models:
  - Gaussian [Sohn et al. 1999], [Ramirez et al. 2004]
  - Laplacian [Chang et al. 2003], [Shin et al. 2008]
  - Gamma [Shin et al. 2006]

• Related Work

- State of the art visual VADs:
  - Speech representation:
    - The shape of the lips: color, contours, keypoints
    - The dynamics of the lips: motion vectors
  - [Siatras et al. 2009]
    - Exploit *pixel intensity* values in the region of the mouth
    - Use a Gaussian model to model speech/non speech frames
  - [Minotto et al. 2013]:
    - Exploit a Gaussian model to *detect the lips*
    - Model the movement of lips using Hidden Markov Model

• The Gap

- Visual VADs have *an advantage* in noisy condition
- Audio VADs provide *near perfect* performance in quite environments
- *A bimodal VAD* may combine the advantages
- *The gap:*
  - Merely few bimodal VADs exist in the Literature
  - There is no unified framework for processing..

• Related work

- One such audio visual VAD:
  - [Tamura et al. 2010]:
  - Audio feature: log power coefficient
  - Video feature: the variance of the vertical part of the optical flow vectors

• Outline

- Motivation and Related Work
- *Audio-Visual VAD using Diffusion Maps*
- VAD Using The Scattering Transform
- Conclusions & Future Work

• Audio-Visual VAD using Diffusion Maps

Problem formulation

- The *audio* signal: $a[n] = a^s[n] + a^d[n] + a^t[n]$
- The *visual* signal: a front side clean video of a speaker
- Consider a training set:
  - Labeled
  - Includes: speech, noise and transients
- The goal- estimate the speech indicator: $\boxed{1_s(i) = \begin{cases} 1 & ; \quad i \in \mathcal{H}_1 \\ 0 & ; \quad i \in \mathcal{H}_0 \end{cases}}$
- Frame by frame processing $\{\mathbf{a}_i, \mathbf{v}_i\}$

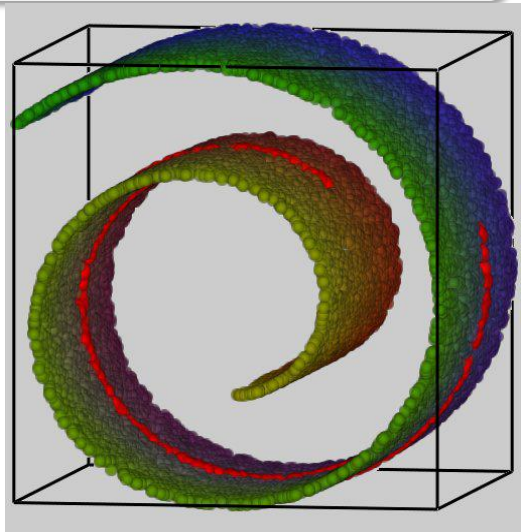
• Audio-Visual VAD using Diffusion Maps

- Assumption on the clean signal (speech)
  - Captured by different sensors (microphone and camera)
  - Has an *underling intrinsic geometric structure (based on the physics of speech pronunciation)*
- We seek for:
  - A data driven model
  - Invariant to the sensor
  - Reveal the underline structure
  - A low dimensional representation $\left\{\hat{\mathbf{a}}_i, \mathbf{v}_i\right\}$ where speech is separated from the non-speech part

• Audio-Visual VAD using Diffusion Maps

- ***Diffusion maps*** [Coifman & Lafon 2006]
  - Given a set of N data points of

    (high) dimension M: $\mathbf{f}_1, \mathbf{f}_2, .., \mathbf{f}_N \in \mathbb{R}^M$



Swiss roll [Takahashi et al. 2009]

  - Assume: the data points lie on a low dimensional manifold
  - Embed the data into a low dimensional domain
  - Provide ***a new low dimensional representation***

• Audio-Visual VAD using Diffusion Maps

- ***Diffusion maps:***
  - Define a similarity kernel function:

$$k\left(\mathbf{f}_i,\mathbf{f}_j\right)=\exp\left(-\frac{\|\mathbf{f}_i-\mathbf{f}_j\|^2}{\varepsilon}\right)$$

  - Remove the effect of the density:

$$k_d\left(\mathbf{f}_i,\mathbf{f}_j\right)=\frac{k\left(\mathbf{f}_i,\mathbf{f}_j\right)}{d(\mathbf{f}_i)d(\mathbf{f}_j)}\ ;\ d\left(\mathbf{f}_i\right)=\sum_{\mathbf{f}_l}k\left(\mathbf{f}_i,\mathbf{f}_l\right)$$

  - Construct a graph:

$$\{\mathbf{f}_i\}-nodes\ ;\ k_d(i,j)-edges$$

• Audio-Visual VAD using Diffusion Maps

– Define a row stochastic Markov Matrix:

$$\mathbf{M} \in \mathbb{R}^{N \times N} \;\; ; \;\; \mathbf{M}_{i,j} = \frac{k_d\left(\mathbf{f}_i, \mathbf{f}_j\right)}{s\left(\mathbf{f}_j\right)} \;\; ; \;\; s\left(\mathbf{f}_j\right) = \sum_{\mathbf{f}_l} k_d\left(\mathbf{f}_l, \mathbf{f}_j\right)$$

– Eigenvalue decomposition: $\left\{\lambda_k, \phi_k\right\}_{k=1}^{K}$

$$\mathbf{\Phi} \equiv \left(\lambda_1\phi_1, \lambda_2\phi_2, \ldots, \lambda_K\phi_K\right)$$

– The diffusion maps:

$$\hat{\mathbf{f}}_i = \left(\mathbf{\Phi}_{i,1}, \mathbf{\Phi}_{i,2}, \ldots, \mathbf{\Phi}_{i,K}\right) \quad \boxed{K \ll M}$$

$$\hat{\mathbf{f}}_1, \hat{\mathbf{f}}_2, .., \hat{\mathbf{f}}_N \in \mathbb{R}^K$$

• Audio-Visual VAD using Diffusion Maps

- The *diffusion distance*:

$$\|\hat{\mathbf{f}}_i - \hat{\mathbf{f}}_j\|^2 \approx D^2(\mathbf{f}_i, \mathbf{f}_j) = \sum_{k=1}^{N} \frac{\left(\mathbf{M}_{i,k} - \mathbf{M}_{j,k}\right)^2}{\nu_0(k)} \;;\; \nu_0(k) = \frac{s(\mathbf{f}_k)}{\sum_{\mathbf{f}_l} s(\mathbf{f}_l)}$$

  - $\nu_0(k)$: the unique stationary distribution
  - Short distances =
    - Highly connected points on the graph

• Audio-Visual VAD using Diffusion Maps

- Related work:
  - Audio-Visual Group Recognition using Diffusion Maps [Keller et al. 2010]
  - Voice Activity Detection in Presence of Transient Noise using Spectral Clustering [Mousazadeh & Cohen 2013]

• Proposed Approach

- The main limitations of diffusion maps:
  - The local similarity is based on the l_2 distance:

$$k\left(\mathbf{f}_i, \mathbf{f}_j\right) = \exp\left(-\frac{\|\mathbf{f}_i - \mathbf{f}_j\|^2}{\varepsilon}\right)$$

  - Even locally:
    - Speech frames may be close to transients frames (for audio)
    - Speech frames may be distant from other speech frames of a different speaker (for video)
- Diffusion maps reveal the “noisy manifold”!

• Proposed Approach

- We propose: $k\left(\tilde{\mathbf{f}}_i, \tilde{\mathbf{f}}_j\right) = \exp\left(-\frac{\|\tilde{\mathbf{f}}_i - \tilde{\mathbf{f}}_j\|^2}{\varepsilon}\right)$
- Namely, *define a domain* where:
  - Speech frames are (locally) *highly similar*

• Audio-Visual VAD using Diffusion Maps

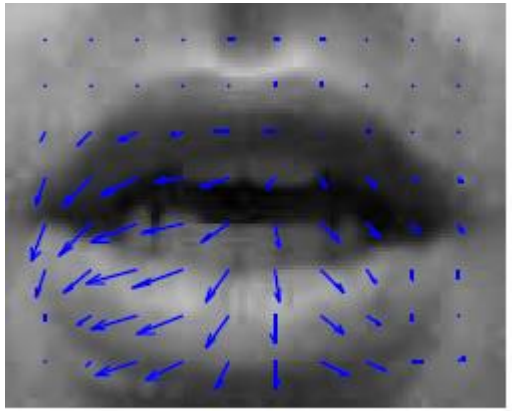
- High dimensional representation: $\tilde{\mathbf{f}}_1, \tilde{\mathbf{f}}_2, .., \tilde{\mathbf{f}}_N \in \mathbb{R}^{\tilde{M}}$
- *Audio signal* [Mousazadeh & Cohen 2013]:
  - Based on Mel-Frequency Cepstral Coefficients (MFCC)
  - Include noise estimation

$$\tilde{\mathbf{a}}_i = \begin{bmatrix} w_\Lambda(i-1)\cdot\mathbf{a}_{i-1}^{MFCC} \\ w_\Lambda(i)\cdot\mathbf{a}_i^{MFCC} \\ w_\Lambda(i+1)\cdot\mathbf{a}_{i+1}^{MFCC} \end{bmatrix} ; \; w_\Lambda(i) = 1 - e^{-\frac{\Lambda_i}{\epsilon}}$$

  - The concatenation over time- reduce the effect of transients

Audio-Visual VAD using Diffusion Maps

- *Visual signal:*
  - Based on motion vectors
  - Invariant to different speakers



$$\mathbf{v}_i = \begin{bmatrix} \mathbf{v}_{i-1}^{MV} & \mathbf{v}_i^{MV} & \mathbf{v}_{i+1}^{MV} \end{bmatrix}^T$$

- $\mathbf{v}_i^{MV}$: The abs of the motion vectors

• Audio-Visual VAD using Diffusion Maps

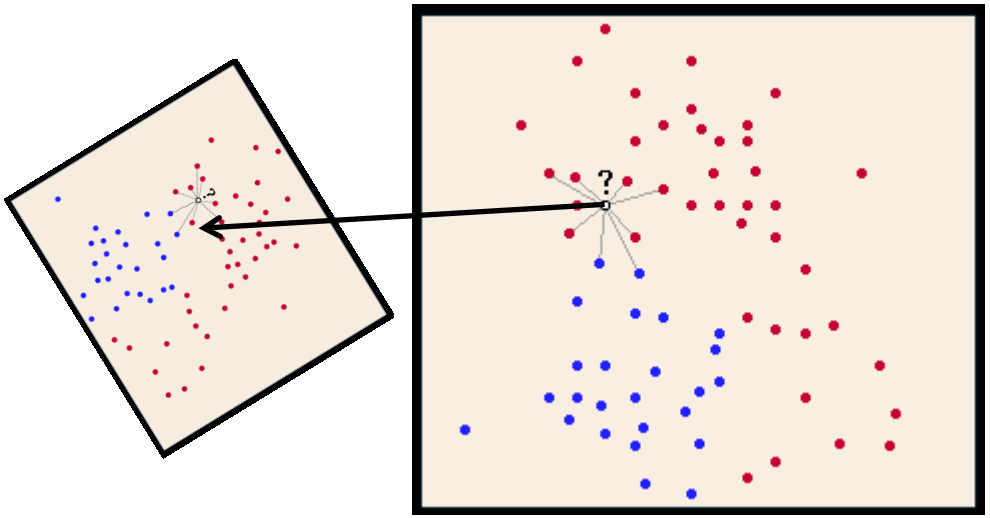
- ***The proposed new representation:*** $\hat{\mathbf{f}}_1, \hat{\mathbf{f}}_2, .., \hat{\mathbf{f}}_N \in \mathbb{R}^K$
- ***Incorporates the advantages*** of the specifically designed features and the diffusion maps:
  - Separates speech and non-speech frames
  - Invariant to different speakers
  - Robust to noise
  - Unit-less and adequate to data merging
  - Low dimensional
  - Preserves local distances

• Audio-Visual VAD using Diffusion Maps

- Implementation:
  - Use a training set to construct diffusion maps
  - Extend for ***sequential processing*** using local similarities

$$\phi_k' = \mathbf{w}^T \mathbf{\Phi}(:,k)$$

$$\hat{\mathbf{f}}_m = \left(\phi_1', \phi_2', \ldots, \phi_K'\right)$$



[H Lohninger, Fundamentals of Statistics]

• Audio-Visual VAD using Diffusion Maps

- Speech presence estimation:
  - The measure for voice activity P is based on:
    - Gaussian Mixture Model (GMM)

$$\Gamma_i = \min\left(\frac{p_r(\hat{\mathbf{f}}_i;\mathcal{H}_1)}{p_r(\hat{\mathbf{f}}_i;\mathcal{H}_0)}, \Gamma_{\max}\right)$$

$$P^S(\mathbf{f}_i) = \frac{1}{(2L^S+1)\cdot\Gamma_{\max}} \sum_{r=-L^S}^{L^S} \Gamma_{i+r}$$

    - The variability of the data over time

$$P^{US}(\mathbf{f}_i) = \frac{\min\left(\sum_{l=1}^{L^{US}} D(\tilde{\mathbf{f}}_i, \tilde{\mathbf{f}}_{i-l}), \sum_{l=1}^{L^{US}} D(\tilde{\mathbf{f}}_i, \tilde{\mathbf{f}}_{i+l})\right)}{L^{US}\cdot D_{\max}}$$

• Audio-Visual VAD using Diffusion Maps

- Merge the modalities:

$$P^B(\mathbf{a}_i, \mathbf{v}_i) = \alpha P(\mathbf{a}_i) + (1-\alpha) P(\mathbf{v}_i)$$

  – $\alpha \in [0,1]$: can be set according to the quality of the signals

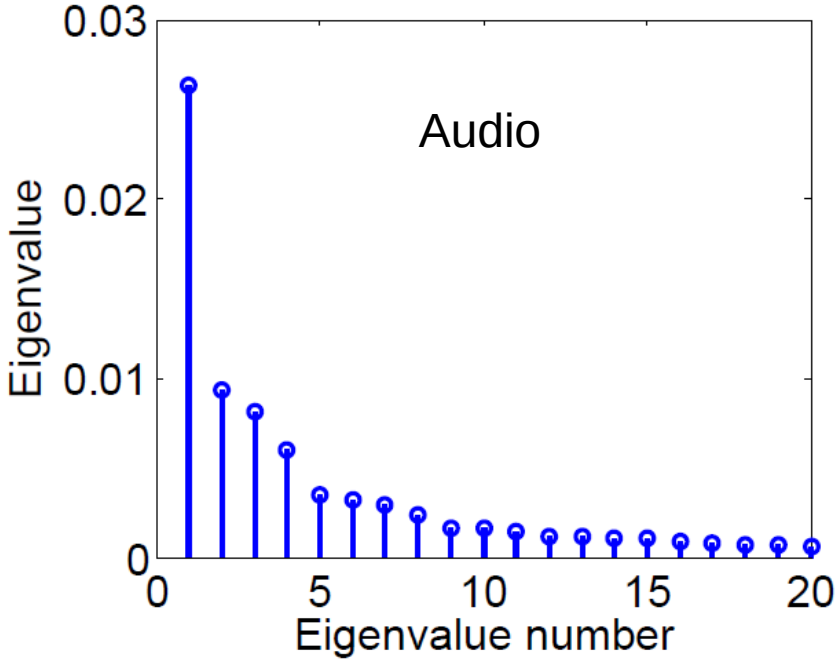
- Speech presence indicator:

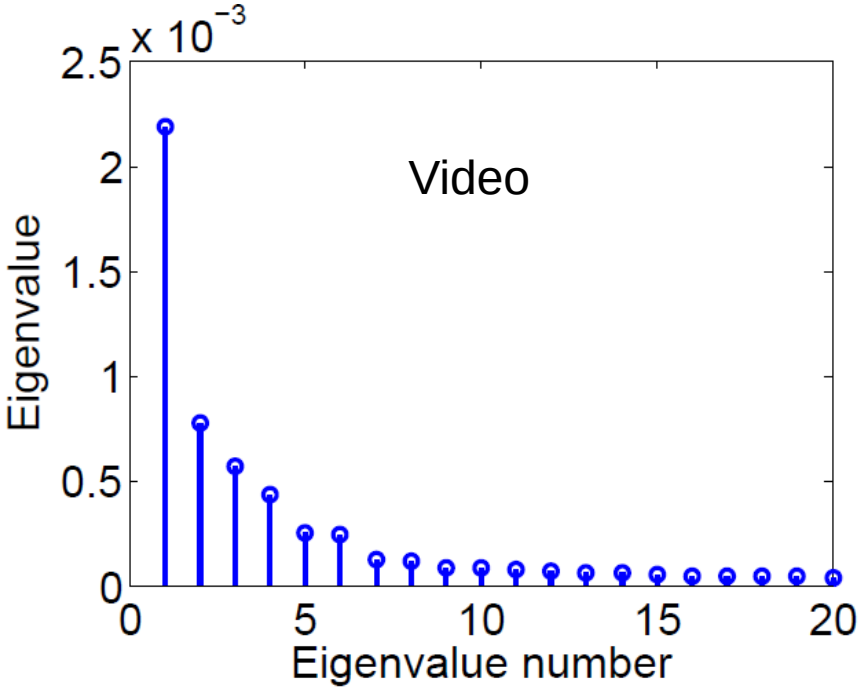
$$\hat{1}_{\mathbf{s}}(i) = \begin{cases} 1 & ; \quad P^B(\mathbf{a}_i, \mathbf{v}_i) > \tau \\ 0 & ; \quad otherwise \end{cases}$$

• Audio-Visual VAD using Diffusion Maps

• *Experimental results*

– The spectrum of the eigenvalues:



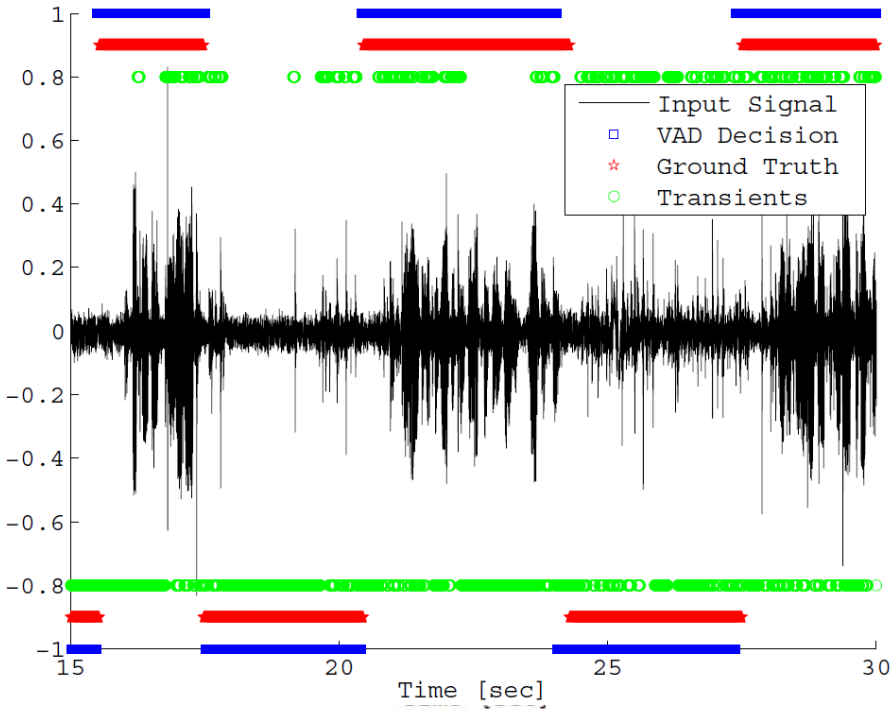


– The dimension of the features- audio: $M = 72$, Video: $M = 297$

– Set $K = 4$ for *the dimension* of diffusion maps

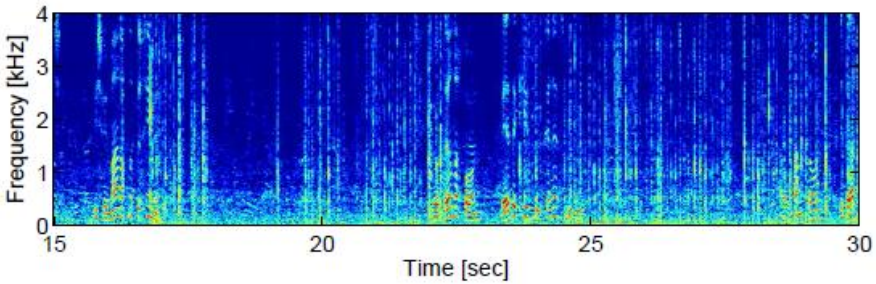
• Audio-Visual VAD using Diffusion Maps

- An example of a typical voice activity detection:



(a)

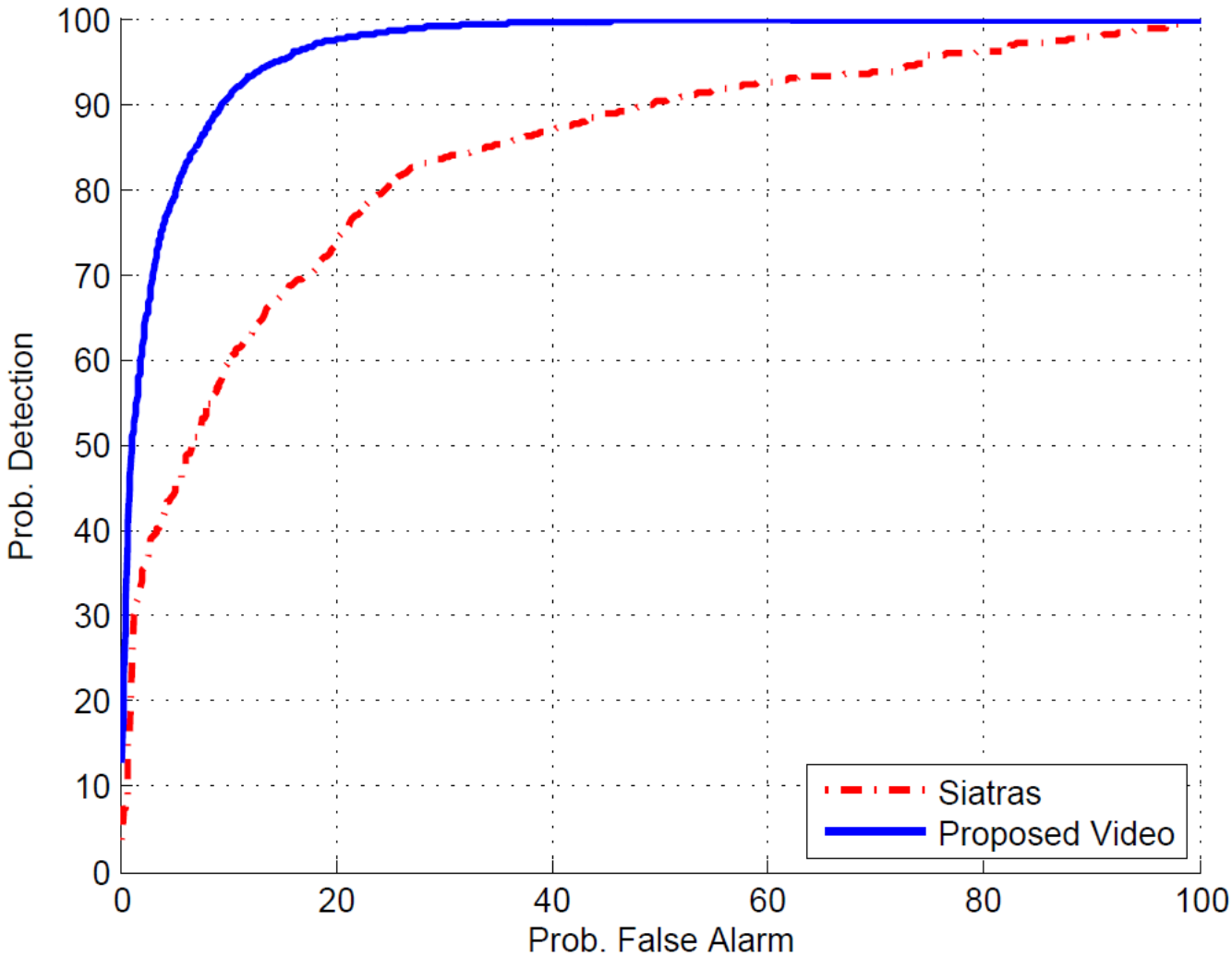
- 10 [dB] SNR babble noise and keyboard typing transient



(b)

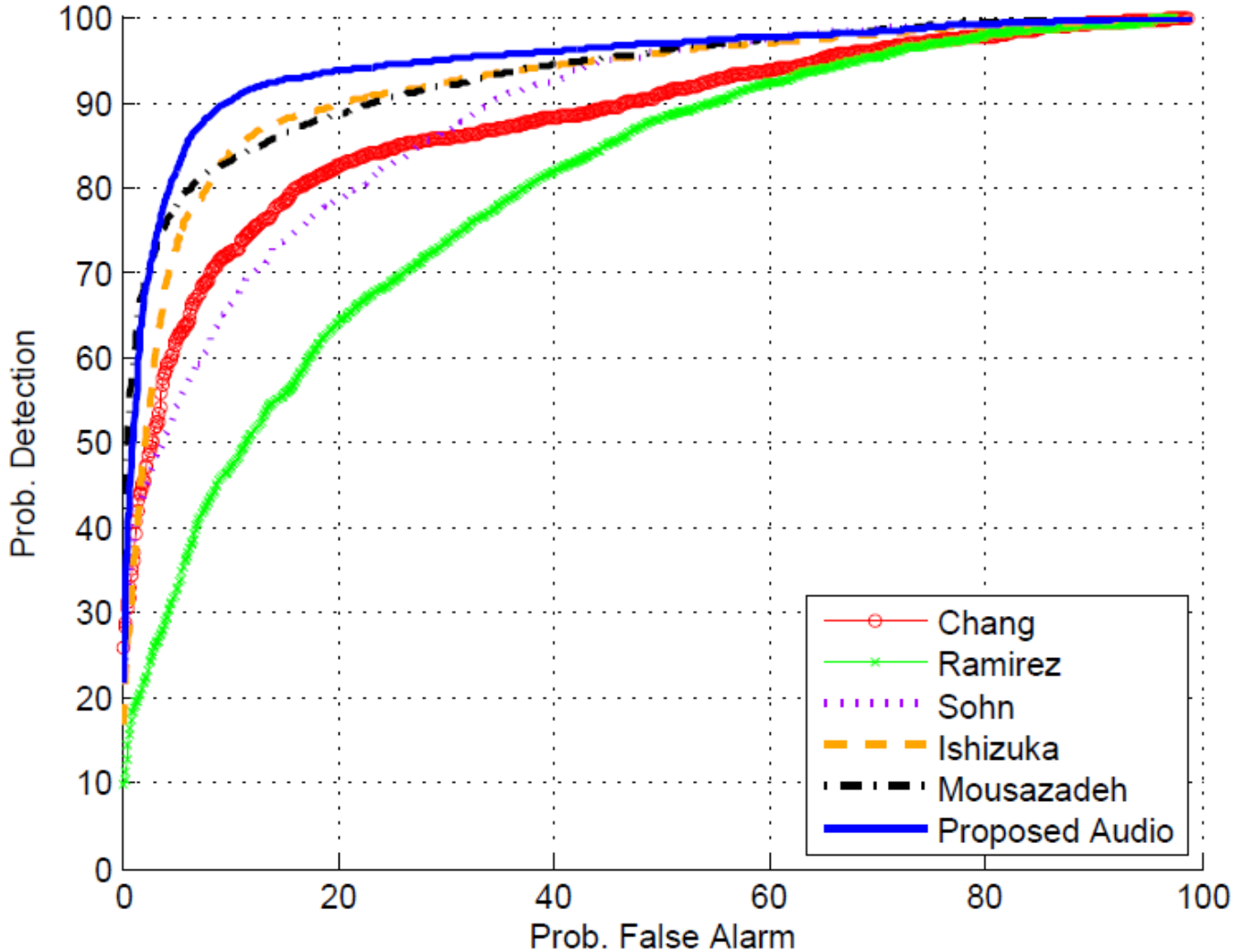
• Audio-Visual VAD using Diffusion Maps

- VAD evaluation: (Correct detections rates: Siatras: 77.5%, Proposed Video: 90.6%)



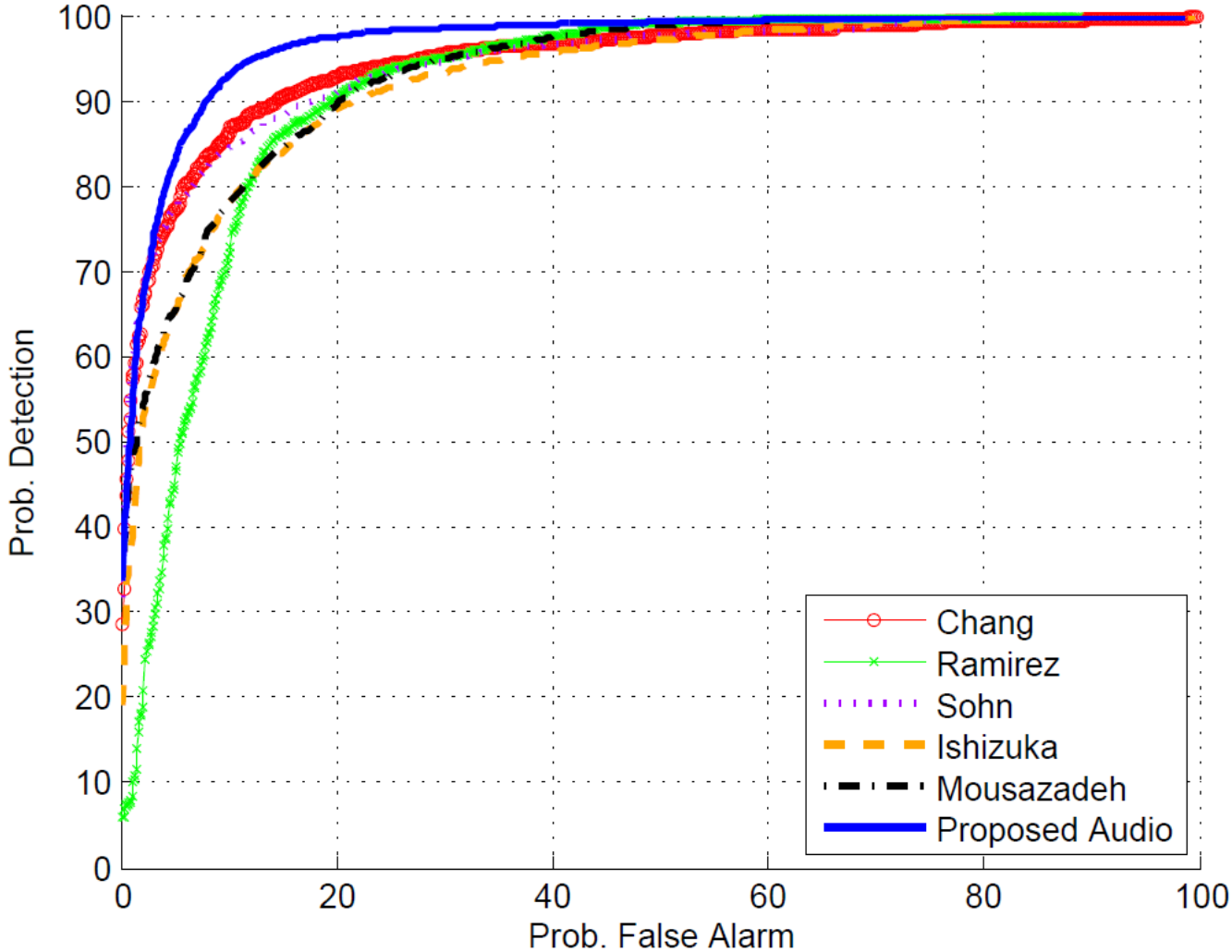
• Audio-Visual VAD using Diffusion Maps

- VAD evaluation: ***babble noise*** with 10 [dB] SNR and a ***keyboard typing transient***



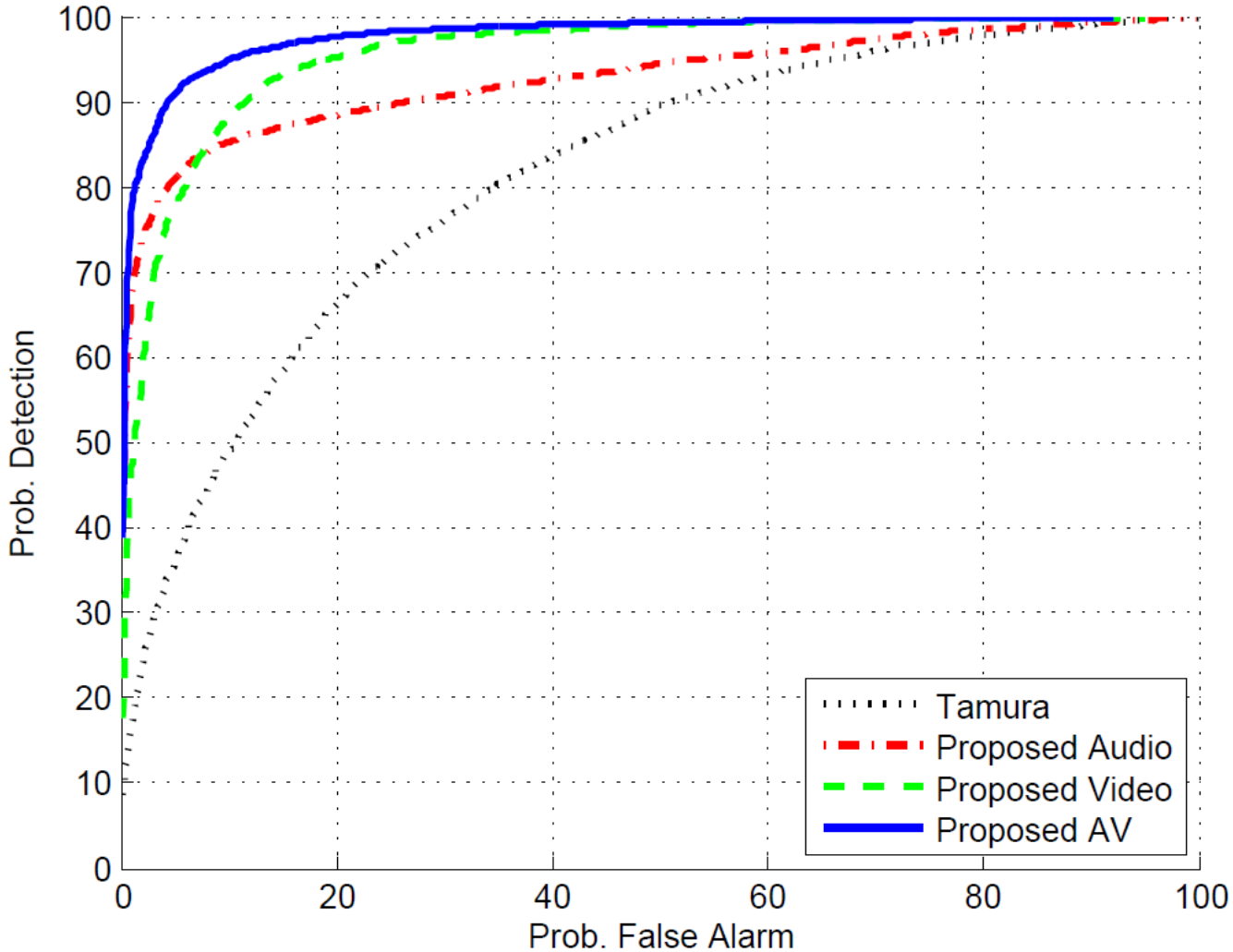
• Audio-Visual VAD using Diffusion Maps

- VAD evaluation: musical noise with 10 [dB] SNR and a *hammering transient*



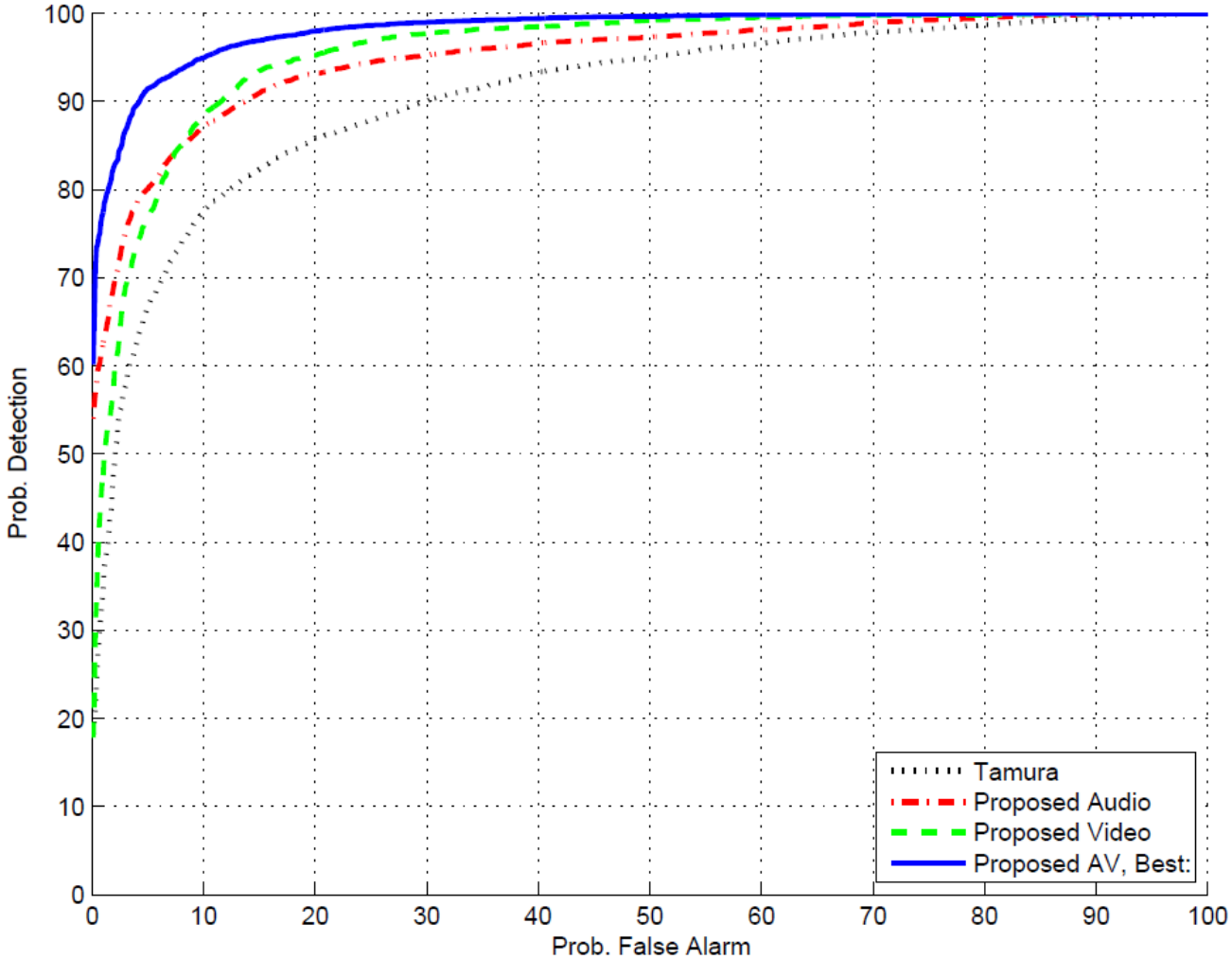
• Audio-Visual VAD using Diffusion Maps

- VAD evaluation: ***babble noise*** with 10 [dB] SNR and a ***keyboard typing transient***



• Audio-Visual VAD using Diffusion Maps

- VAD evaluation: musical noise with 10 [dB] SNR and a *hammering transient*



• Audio-Visual VAD using Diffusion Maps

- Audio VAD evaluation: correct decision (%)

	Babble (10dB SNR) Keyboard	Musical (10dB SNR) Hammering	Colored (5dB SNR) Hammering	Musical (0dB SNR) Keyboard	Babble (15dB SNR) Scissors
Chang	82	88.4	87	74.6	88.2
Ramirez	72.2	85.5	82.6	75.1	70.6
Sohn	79.4	87.5	84	80.7	88.4
Ishizuka	87.3	84.7	88.9	64.7	91.5
Mousazadeh	86.7	84.7	84.1	87.6	91
Proposed Audio	**90.1**	**91.5**	**90.5**	**88.7**	**92.4**

• Audio-Visual VAD using Diffusion Maps

- Audio-visual VAD evaluation:

	Babble (10dB SNR) Keyboard	Musical (10dB SNR) Hammering	Colored (5dB SNR) Hammering	Musical (0dB SNR) Keyboard	Babble (15dB SNR) Scissors
Tamura	73.6	83.8	83.9	73.8	81.2
Audio	87.7	89.9	87.8	86.5	90.2
Video	89.6	89.6	89.6	89.6	89.6
P^US AV	89.3	89.8	89.6	89.5	85.8
P^S AV	92.3	94.2	92.1	92	93.8
Proposed AV	**92.9**	**94.5**	**92.8**	**92.9**	**94.6**

• Outline

- Motivation and Related Work
- Audio-Visual Voice Activity Detection (VAD) Using Diffusion Maps
- *VAD using the Scattering Transform*
- Conclusions & Future Work

• VAD Using The Scattering Transform

- Focus on the audio signal
- *Open problem:*
  - Voice activity detection in *presence of transients*
- Benchmark: [Mousazadeh & Cohen 2013]

• Audio-Visual VAD using Diffusion Maps

- *Problem formulation*
- The input signal: $y(\tau) = x(\tau) + d(\tau) + z(\tau)$
- Consider a training set:
  - Labeled
  - Includes: speech, noise and transients
- The goal- estimate the speech indicator: $\boxed{1_s(t) = \begin{cases} 1 & ; \quad t \in \mathcal{H}_1 \\ 0 & ; \quad t \in \mathcal{H}_0 \end{cases}}$
- Frame by frame processing,
  - Frame y_t is given by: $y(\tau); \tau \in [t - T/2, t + T/2]$

• VAD Using The Scattering Transform

- Recall: The Mel Frequency Spectrogram:

$$\mathbf{y}_t^M(\lambda) = \frac{1}{2\pi}\int |Y(t,\omega)|^2 |\Psi_\lambda(\omega)|^2 \, d\omega \approx |y \star \psi_\lambda|^2 \star |\phi|^2 (t)$$

  - Smooth the spectrum
  - Invariant to small translation and deformation

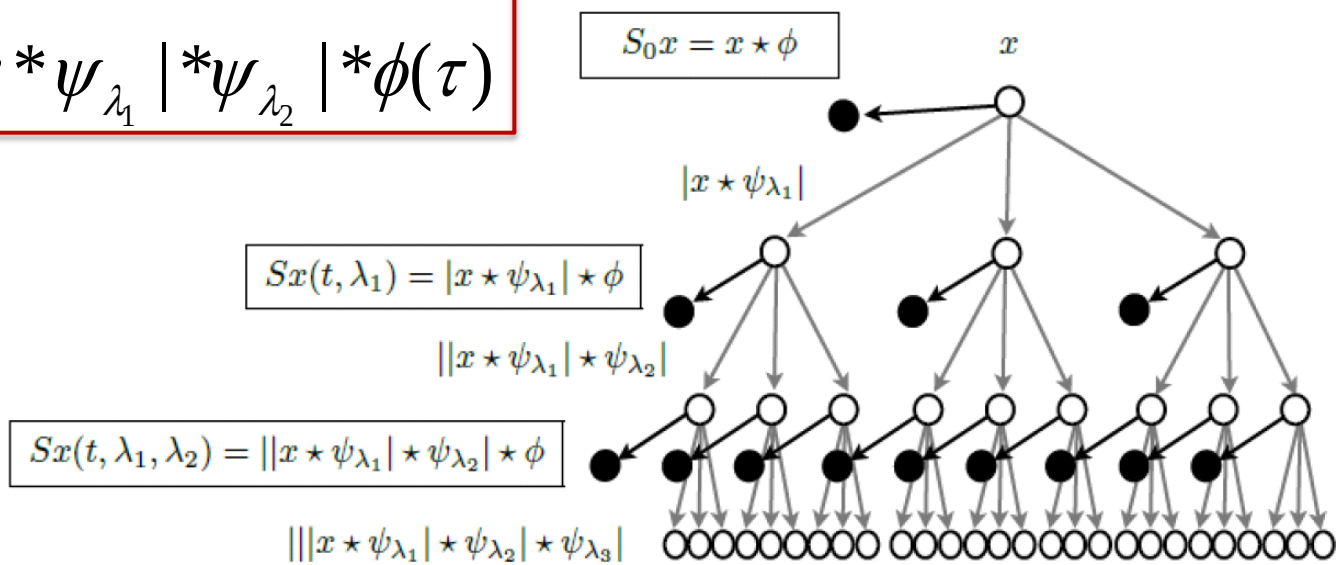
- Suffer from information loss: [Anden & Mallat 2013]
  - Smooth over the frequency = smooth over the time
  - Removes fine scale information: transients

• VAD Using The Scattering Transform

- ***The Scattering Transform*** [Anden & Mallat 2013]
  - Based on a cascade of: Wavelet convolutions & modulus operators
  - The 1st & the 2nd orders:

$$S_1(\tau,\lambda_1) = |\, y * \psi_{\lambda_1} \,| * \phi(\tau)$$

$$S_2(\tau,\lambda_1,\lambda_2) = \|\, y * \psi_{\lambda_1} \,| * \psi_{\lambda_2} \,| * \phi(\tau)$$



• VAD Using The Scattering Transform

- The first order coefficients of the transform ≈ the Mel Frequency Spectrogram
- *The lost information:*
  - Recovered using higher order coefficients
- Proposed representation:

$$\mathbf{y}_t = w_t \mathbf{y}_t^S \;\; ; \;\; w_t = 1 - e^{-\frac{\Lambda_t}{\epsilon}}$$

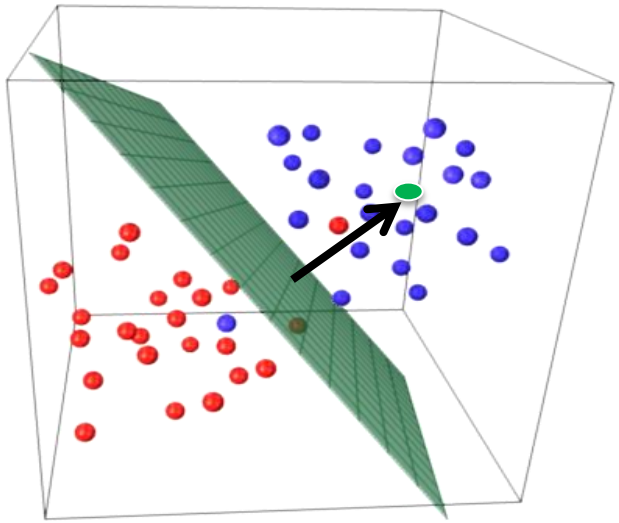
• VAD Using The Scattering Transform

- Speech presence estimation:
  - *Continuous* measure based on SVM classifier
  - Exploit the distance to a hyper plane: $L_t = \frac{<\mathbf{y}_t, \mathbf{n}> + b}{\|\mathbf{n}\|}$
  - Normalize and average over time:

$$P_t = \frac{1}{2J+1} \sum_{j=t-JT}^{t+JT} \tilde{L}_j$$

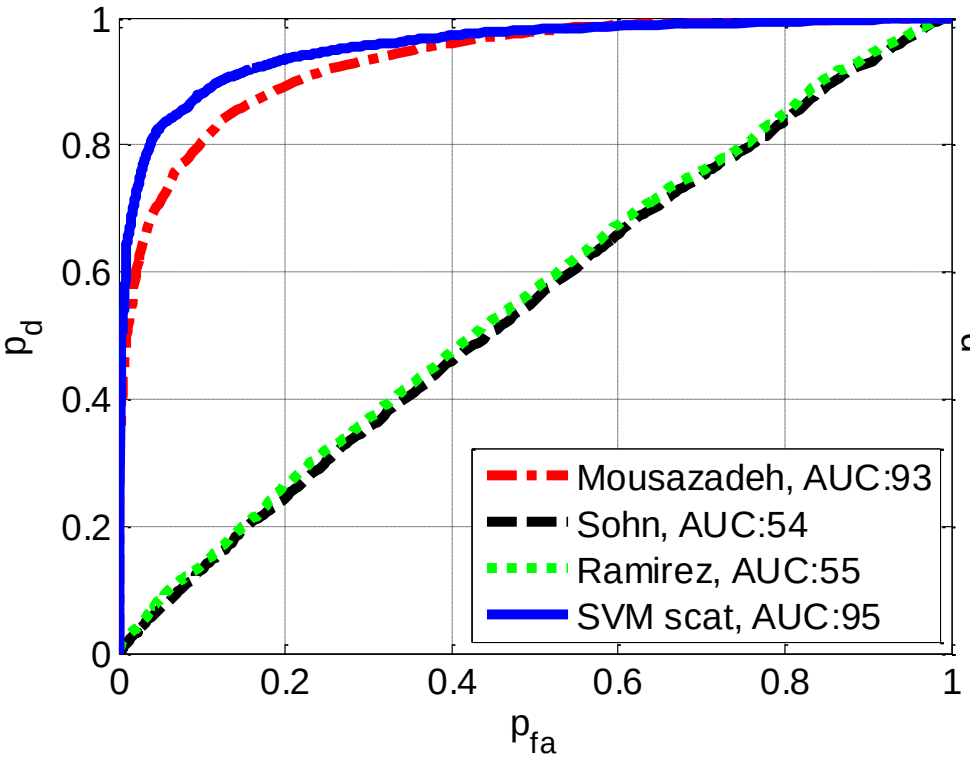
  - Speech indicator:

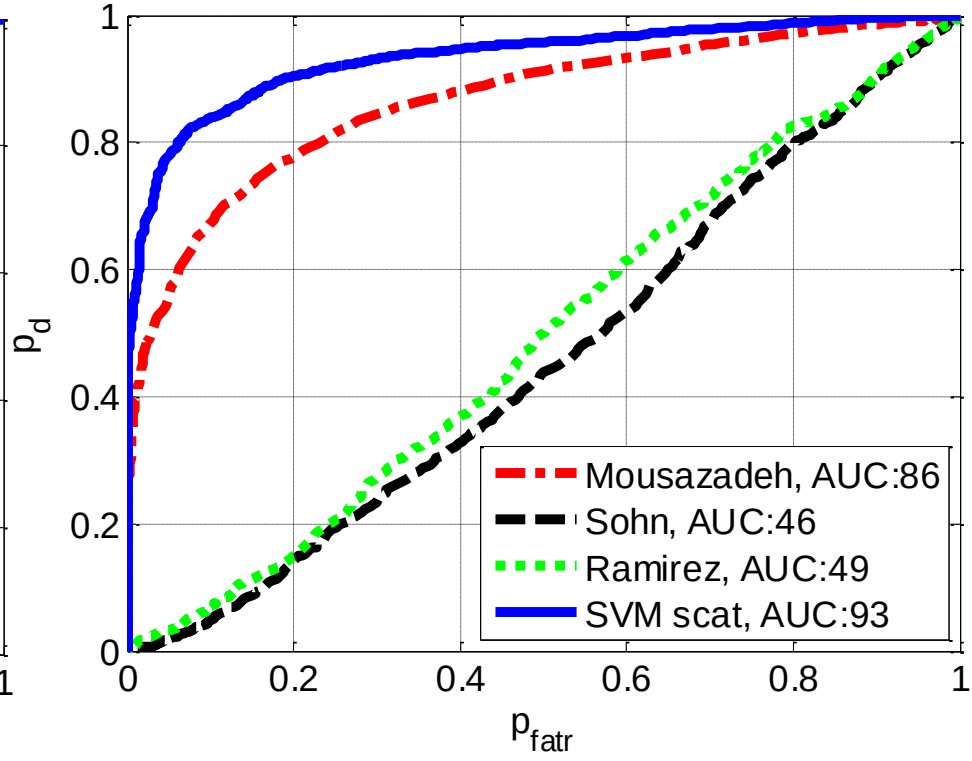
$$\boxed{\hat{1}_{\mathbf{s}}(t) = \begin{cases} 1 & ; \quad P_t > \alpha \\ 0 & ; \quad \text{otherwise} \end{cases}}$$



• VAD Using The Scattering Transform

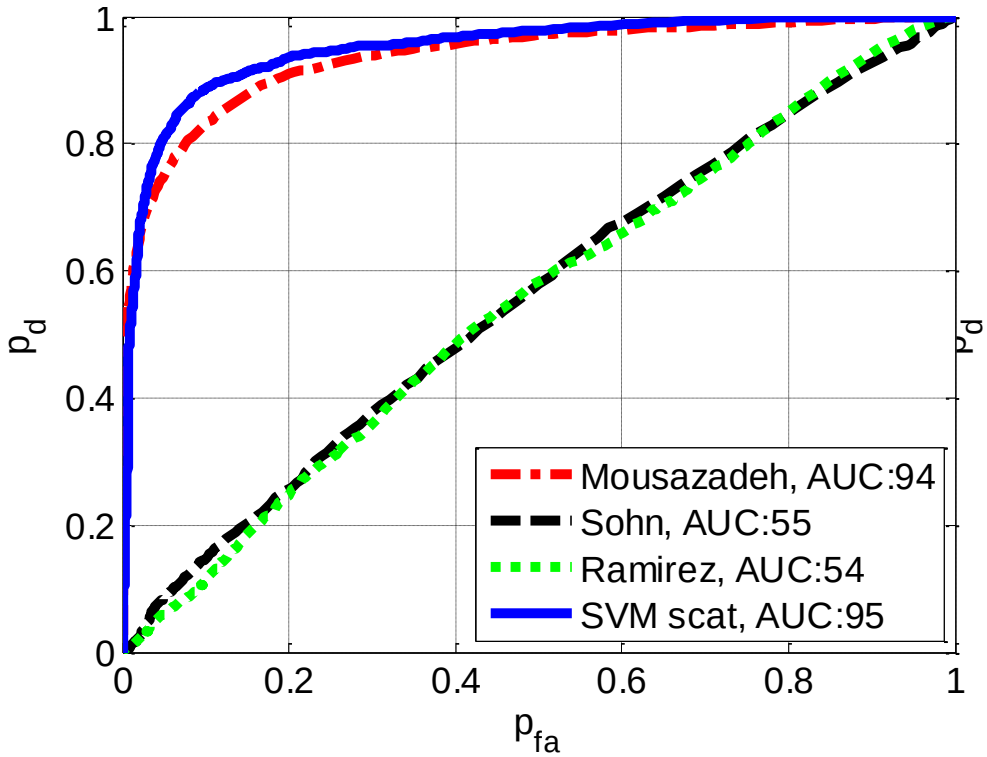
- VAD evaluation: 0 [dB] SNR Gaussian noise and a keyboard typing transient

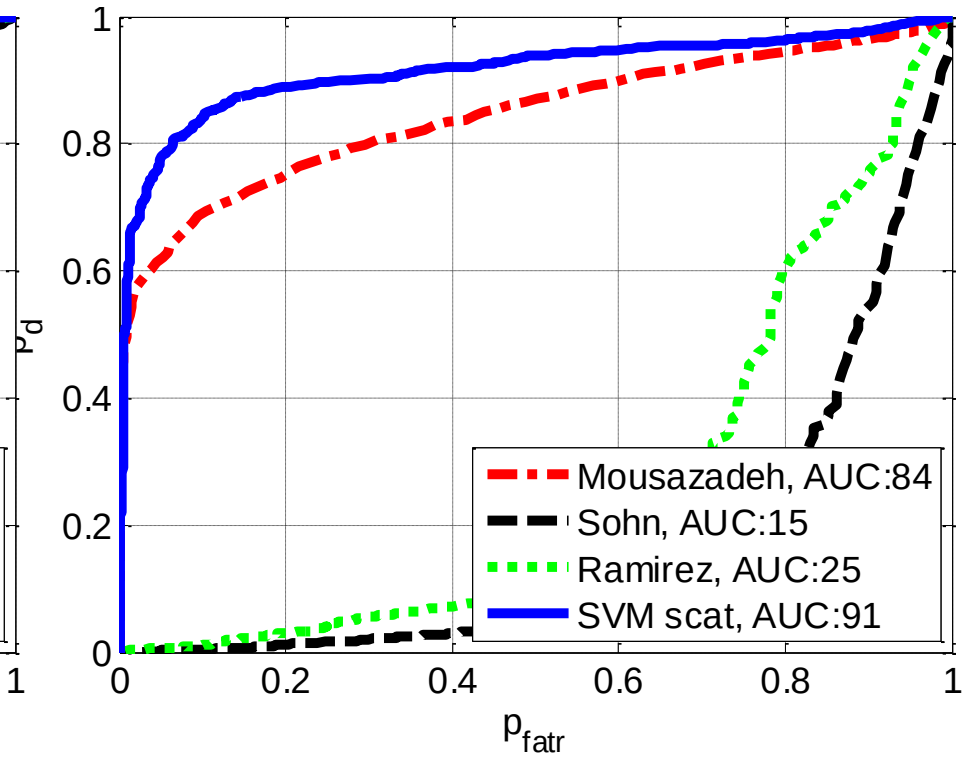




• VAD Using The Scattering Transform

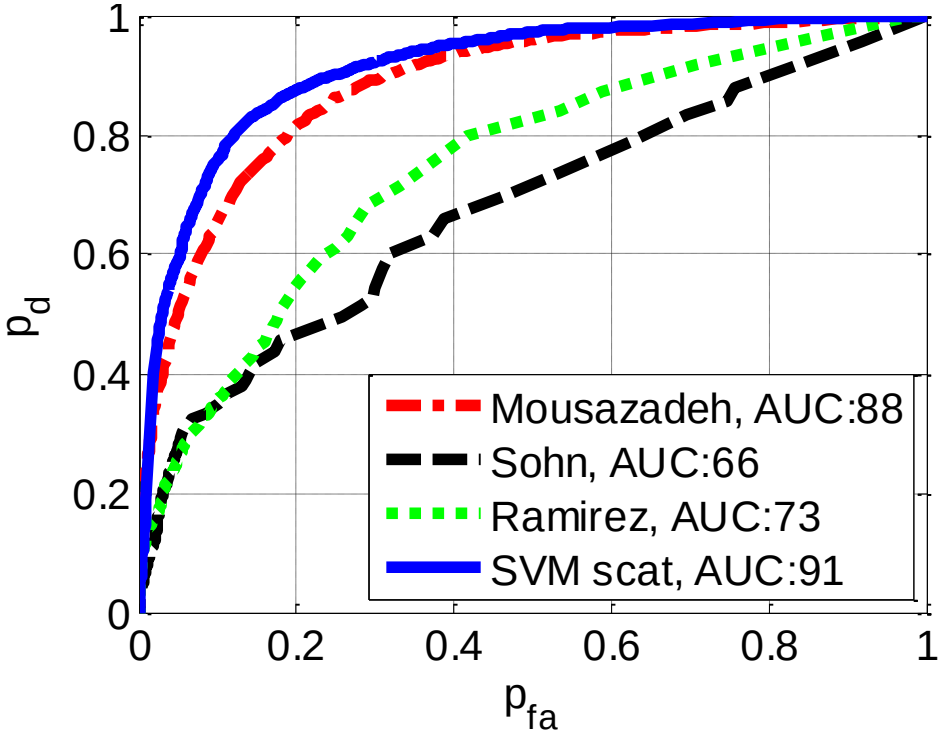
- VAD evaluation: 0 [dB] SNR colored Gaussian noise and a scissors transient

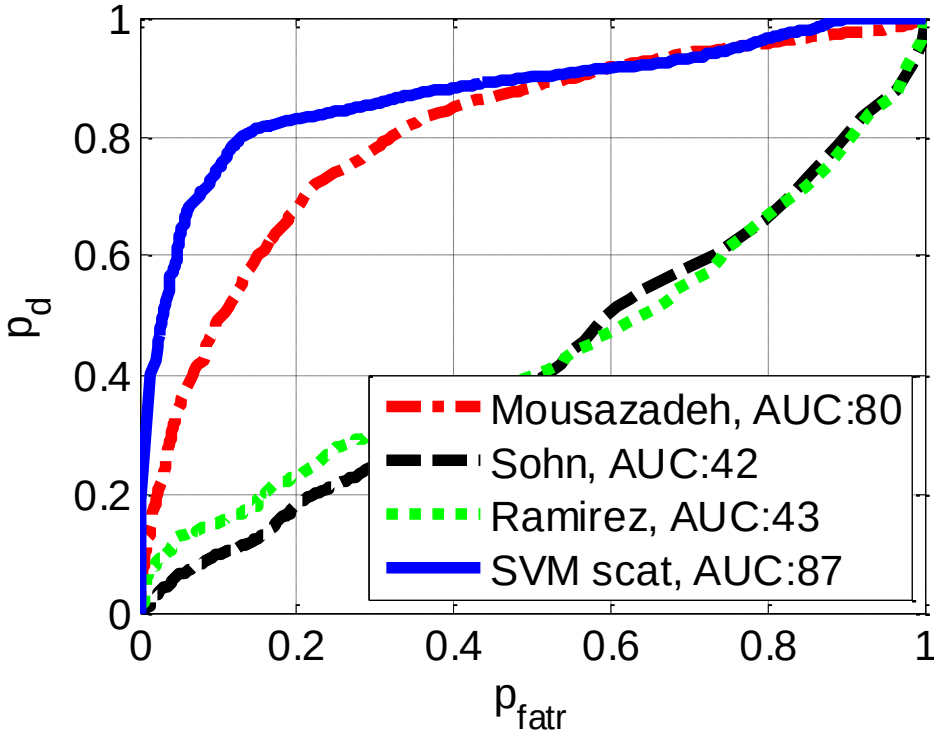




• VAD Using The Scattering Transform

- VAD evaluation: 0 [dB] SNR babble noise and a door nocks transient





• Outline

- Motivation and Related Work
- Audio-Visual Voice Activity Detection (VAD) Using Diffusion Maps
- VAD using the Scattering Transform
- *Conclusions & Future Work*

Conclusions

- We proposed Audio-Visual VAD:
  - A new representation:
    - *Compact*
    - Separates speech to non-speech frames
    - *Merge data* captured by different types of sensors
  - Incorporating audio & visual data is beneficial for VAD
  - The algorithm outperforms state of the art VADs

• Conclusions

- VAD Using The Scattering Transform
  - New method for VAD
  - High orders of the scattering transform separate speech and transient
  - We proposed a new supervised measure for voice activity
  - Outperforms state of the art VADs

• Future work

- Extend the proposed approach to other applications
  - Speech enhancement
  - Speech recognition in a noisy environment
- Develop adaptive merging schemes
- Incorporate transient estimation

Future work

- Theoretical questions:
  - What “feature space” is suitable for data driven learning and how to find it?
  - What is a good similarity kernel and how to design an application specific kernel
    - E.g for speaker recognition
  - How accurate is an out of sample extension?
- Design signal processing tools that:
  - Reveal the mutual structure of multimodal signals
  - Based on specifically designed metrics:
    - Based on different variability rates
    - Metric learning

- The End

Thank you!