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
Speech processing in noisy acoustic environments:

The model: \textit{speech} + \textit{quasi-stationary noise} + \textit{transients}

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

Example:

The challenges:

- \textit{Non-stationary noise}
- \textit{Transients}: short time interruptions, e.g., keyboard typing
A-V Speech Processing Using Diffusion Maps & The Scattering Transform

• 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
- **State of the art Voice Activity Detectors (VAD):**
  
  - Based on the log of the Likelihood Ratio:
    
    \[
    \Lambda_j(k) = \log \left( \frac{p_r \left( A(j, k); H_1 \right)}{p_r \left( A(j, k); H_0 \right)} \right)
    \]

- 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
A-V Speech Processing Using Diffusion Maps & The Scattering Transform

• 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**

- **Audio** signal: \( a[n] = a^s[n] + a^d[n] + a^t[n] \)
- **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:
  \[
  1_s(i) = \begin{cases} 
  1 & ; \ i \in \mathcal{H}_1 \\
  0 & ; \ i \in \mathcal{H}_0 
  \end{cases}
  \]
- Frame by frame processing \( \{a_i, v_i\} \)
• 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 $\{\hat{a}_i, v_i\}$ 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 
    \[(high) \text{ dimension} \quad M: \quad f_1, f_2, \ldots, f_N \in \mathbb{R}^M\]
  - Assume: the data points lie on a low dimensional manifold
  - Embed the data into a low dimensional domain
  - Provide a new low dimensional representation

Swiss roll [Takahashi et al. 2009]
Diffusion maps:

- Define a similarity kernel function:
  \[ k(f_i, f_j) = \exp\left( -\frac{\|f_i - f_j\|^2}{\varepsilon} \right) \]

- Remove the effect of the density:
  \[ k_d(f_i, f_j) = \frac{k(f_i, f_j)}{d(f_i)d(f_j)} \quad ; \quad d(f_i) = \sum_{f_i} k(f_i, f_i) \]

- Construct a graph:
  \[ \{f_i\} - nodes \quad ; \quad k_d(i, j) - edges \]
Audio-Visual VAD using Diffusion Maps

- Define a row stochastic Markov Matrix:

\[
M \in \mathbb{R}^{N \times N} ; \quad M_{i,j} = \frac{k_d(f_i, f_j)}{s(f_j)} ; \quad s(f_j) = \sum_{i} k_d(f_i, f_j)
\]

- Eigenvalue decomposition: \( \{ \lambda_k, \phi_k \}_{k=1}^K \)

\[
\Phi \equiv (\lambda_1 \phi_1, \lambda_2 \phi_2, ..., \lambda_K \phi_K)
\]

- The diffusion maps:

\[
\hat{f}_i = (\Phi_{i,1}, \Phi_{i,2}, ..., \Phi_{i,K}) \quad K \ll M
\]

\[
\hat{f}_1, \hat{f}_2, .., \hat{f}_N \in \mathbb{R}^K
\]
• **Audio-Visual VAD using Diffusion Maps**

**The *diffusion distance***:

\[
\| \hat{f}_i - \hat{f}_j \|^2 \approx D^2(f_i, f_j) = \sum_{k=1}^{N} \left( \frac{M_{i,k} - M_{j,k}}{\nu_0(k)} \right)^2 ; \quad \nu_0(k) = \frac{s(f_k)}{\sum_{f_l} s(f_l)}
\]

– \( \nu_0(k) \): 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]

• Proposed Approach

- The main limitations of diffusion maps:
  - The local similarity is based on the $l_2$ distance:
    \[
    k(f_i, f_j) = \exp\left(-\frac{||f_i - f_j||^2}{\varepsilon}\right)
    \]
  - 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:  
  \[ k(\tilde{f}_i, \tilde{f}_j) = \exp \left( - \frac{\| \tilde{f}_i - \tilde{f}_j \|^2}{\varepsilon} \right) \]

- Namely, **define a domain** where:
  
  – Speech frames are (locally) **highly similar**
• Audio-Visual VAD using Diffusion Maps

- High dimensional representation: \( \tilde{\mathbf{f}}_1, \tilde{\mathbf{f}}_2, \ldots, \tilde{\mathbf{f}}_n \in \mathbb{R}^\tilde{M} \)

- **Audio signal** [Mousazadeh & Cohen 2013]:
  - Based on Mel-Frequency Cepstral Coefficients (MFCC)
  - Include noise estimation

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

- 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:* \( \hat{f}_1, \hat{f}_2, \ldots, \hat{f}_N \in \mathbb{R}^K \)

*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

\[ \phi_k' = w^T \Phi(:, k) \]

\[ \hat{f}_m = (\phi_1', \phi_2', ..., \phi_K') \]

[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)

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

\[
P^S(f_i) = \frac{1}{(2L^S + 1) \cdot \Gamma_{\max}} \sum_{r=-L^S}^{L^S} \Gamma_{i+r}
\]

• The variability of the data over time

\[
P^{US}(f_i) = \frac{\min \left( \sum_{l=1}^{L^{US}} D(\tilde{f}_i, \tilde{f}_{i-l}), \sum_{l=1}^{L^{US}} D(\tilde{f}_i, \tilde{f}_{i+l}) \right)}{L^{US} \cdot D_{\max}}
\]
- **Audio-Visual VAD using Diffusion Maps**

  - **Merge the modalities:**
    
    $$ P^B(a_i, v_i) = \alpha P(a_i) + (1 - \alpha) P(v_i) $$
    
    - $\alpha \in [0,1]$ : can be set according to the quality of the signals

  - **Speech presence indicator:**
    
    $$ \hat{1}_s(i) = \begin{cases} 1 & ; \quad P^B(a_i, v_i) > \tau \\ 0 & ; \quad \text{otherwise} \end{cases} $$
Audio-Visual VAD using Diffusion Maps

**Experimental results**

- The spectrum of the eigenvalues:

  ![Audio spectrum](image1)

  ![Video spectrum](image2)

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

- Set $K = 4$ for *the dimension* of diffusion maps
A-V Speech Processing Using Diffusion Maps & The Scattering Transform

- Audio-Visual VAD using Diffusion Maps

- An example of a typical voice activity detection:

- 10 [dB] SNR babble noise and keyboard typing transient
• 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 (%)

<table>
<thead>
<tr>
<th></th>
<th>Babble (10dB SNR)</th>
<th>Musical (10dB SNR)</th>
<th>Colored (5dB SNR)</th>
<th>Musical (0dB SNR)</th>
<th>Babble (15dB SNR)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Keyboard</strong></td>
<td>82</td>
<td>88.4</td>
<td>87</td>
<td>74.6</td>
<td>88.2</td>
</tr>
<tr>
<td><strong>Hammering</strong></td>
<td>72.2</td>
<td>85.5</td>
<td>82.6</td>
<td>75.1</td>
<td>70.6</td>
</tr>
<tr>
<td><strong>Hammering</strong></td>
<td>79.4</td>
<td>87.5</td>
<td>84</td>
<td>80.7</td>
<td>88.4</td>
</tr>
<tr>
<td><strong>Sohn</strong></td>
<td>87.3</td>
<td>84.7</td>
<td>88.9</td>
<td>64.7</td>
<td>91.5</td>
</tr>
<tr>
<td><strong>Ishizuka</strong></td>
<td>86.7</td>
<td>84.7</td>
<td>84.1</td>
<td>87.6</td>
<td>91</td>
</tr>
<tr>
<td><strong>Mousazadeh</strong></td>
<td><strong>90.1</strong></td>
<td><strong>91.5</strong></td>
<td><strong>90.5</strong></td>
<td><strong>88.7</strong></td>
<td><strong>92.4</strong></td>
</tr>
<tr>
<td><strong>Proposed Audio</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Chang**

**Ramirez**

**Sohn**

**Ishizuka**

**Mousazadeh**

**Proposed Audio**
## Audio-Visual VAD using Diffusion Maps

### Audio-visual VAD evaluation:

<table>
<thead>
<tr>
<th></th>
<th>Babble (10dB SNR)</th>
<th>Musical (10dB SNR)</th>
<th>Colored (5dB SNR)</th>
<th>Musical (0dB SNR)</th>
<th>Babble (15dB SNR)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Keyboard</td>
<td>Hammering</td>
<td>Hammering</td>
<td>Keyboard</td>
<td>Scissors</td>
</tr>
<tr>
<td>Tamura</td>
<td>73.6</td>
<td>83.8</td>
<td>83.9</td>
<td>73.8</td>
<td>81.2</td>
</tr>
<tr>
<td>Audio</td>
<td>87.7</td>
<td>89.9</td>
<td>87.8</td>
<td>86.5</td>
<td>90.2</td>
</tr>
<tr>
<td>Video</td>
<td>89.6</td>
<td>89.6</td>
<td>89.6</td>
<td>89.6</td>
<td>89.6</td>
</tr>
<tr>
<td>$P^{US\ AV}$</td>
<td>89.3</td>
<td>89.8</td>
<td>89.6</td>
<td>89.5</td>
<td>85.8</td>
</tr>
<tr>
<td>$P^{S\ AV}$</td>
<td>92.3</td>
<td>94.2</td>
<td>92.1</td>
<td>92</td>
<td>93.8</td>
</tr>
<tr>
<td>Proposed AV</td>
<td><strong>92.9</strong></td>
<td><strong>94.5</strong></td>
<td><strong>92.8</strong></td>
<td><strong>92.9</strong></td>
<td><strong>94.6</strong></td>
</tr>
</tbody>
</table>
• 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]
**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:
  \[
  1_s(t) = \begin{cases} 
  1 & ; \ t \in \mathcal{H}_1 \\
  0 & ; \ t \in \mathcal{H}_0 
  \end{cases}
  \]

- Frame by frame processing,
  - Frame \( y_t \) is given by: \( y(\tau); \tau \in [t - T/2, t + T/2] \)
A-V Speech Processing Using Diffusion Maps & The Scattering Transform

• VAD Using The Scattering Transform

Recall: The Mel Frequency Spectrogram:

\[ y^M_t (\lambda) = \frac{1}{2\pi} \int |Y(t, \omega)|^2 |\Psi_{\lambda}(\omega)|^2 \, d\omega \approx |y * \psi_{\lambda}|^2 * |\phi|^2 (t) \]

- 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
The Scattering Transform [Anden & Mallat 2013]

- Based on a cascade of: Wavelet convolutions & modulus operators
- The 1\textsuperscript{st} & the 2\textsuperscript{nd} orders:

\[
S_1(\tau, \lambda_1) = |y \ast \psi_{\lambda_1} | \ast \phi(\tau)
\]
\[
S_2(\tau, \lambda_1, \lambda_2) = || y \ast \psi_{\lambda_1} | \ast \psi_{\lambda_2} | \ast \phi(\tau)
\]
• VAD Using The Scattering Transform

• The first order coefficients of the transform \( \approx \) the Mel Frequency Spectrogram

• **The lost information:**
  
  – Recovered using higher order coefficients

• Proposed representation:

\[
y_t = w_t y^S_t \quad ; \quad w_t = 1 - e^{-\frac{\Lambda_t}{\epsilon}}
\]
• VAD Using The Scattering Transform

Speech presence estimation:

- **Continuous** measure based on SVM classifier

- Exploit the distance to a hyper plane: 
  \[ L_t = \frac{<y_t, n> + b}{||n||} \]

- Normalize and average over time:
  \[ P_t = \frac{1}{2J + 1} \sum_{j=t-JT}^{t+JT} \tilde{L}_j \]

- Speech indicator:
  \[ \hat{s}_s(t) = \begin{cases} 1 & ; P_t > \alpha \\ 0 & ; \text{otherwise} \end{cases} \]
• 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
Thank you!