Voice Activity Detection (VAD) in Presence of Transient Noise

By: Tal Rozenfeld & Tal Toledano
Supervisor: Saman Mousazadeh
The aim of the project

Examine and improve a VAD based on spectral clustering method to overcome the limitation of statistical model based VADs in presence of transient noise.
Statistical Model Based VAD-results

- Detection results for only speech signal:

- Detection results for only transient noise signal:

- The transient noise is detected as speech
Spectral Clustering

**DATA**
\{x_1, x_2 \ldots x_N\}

**Choose a weight function** \( W(i,j) \)
& build the similarity matrix \( W \)

**Eigenvectors**

**K-means**

- Choose a weight function
  i.e \( W(x_i, x_j) = \exp \left( -\frac{|x_i - x_j|^2}{2\sigma^2} \right) \)
- Construct a similarity matrix which express the similarity between two data points.

- Compute the first \( K \) eigenvectors of \( W \) : \( u_1, u_2 \ldots u_k \)
- Let \( U_{[N \times K]} \) be the matrix containing the vectors \( u_1, u_2 \ldots u_k \) as columns.
- Let \( y_i \) be the vector corresponding to the i-th row of \( U \)
- Cluster the points \( y_i \) with k-means algorithm into \( k \) clusters
Spectral Clustering – applied to VAD

• In practical we classify the input signal to 4 groups:
  • Speech + stationary noise
  • Transient noise + stationary noise
  • Speech + Transient noise + stationary noise
  • Stationary noise

• We expand the number of clusters because a frame consisting of transient noise is more similar to a speech frame than a noise-only frame. If we define two clusters, the K-means algorithm will cluster the frames consisting of only transient and stationary noise as a speech frame and not as a noise frame.
Spectral Clustering – applied to VAD

- The novel VAD algorithm using spectral clustering is a supervised learning one. Which means it is divided into two steps:
  - Learning step – in which training data is used to adjust the parameters of the algorithm.
  - Testing step – which uses the adjusted parameters for clustering unlabeled data.
Flowchart – Spectral Clustering VAD Learning algorithm

X=S+Nt+Ns
Signal=speech+stationary noise+transient noise

\[ \text{Log(|STFT|)} \]

manually build the indicator matrix C

Normalize W & find it’s first k eigenvectors

Build the similarity matrix W

Choose weight function & find optimized $\alpha$

Using K-means algorithm to cluster the data
Flowchart – Spectral Clustering VAD
Testing (real time) algorithm

1. \( \tilde{W} = \begin{bmatrix} W & B \\ B^T & A \end{bmatrix} \) when \( W \) is symmetric matrix and \( W = U \Lambda U^T \)

2. The Nystrom extension gives \( \tilde{U} = \begin{bmatrix} U \\ B^T U \Lambda^{-1} \end{bmatrix} \)

3. \( d_{speech} - \gamma \geq d_{non\ speech} + \gamma. \)
   \( d \) is distance function, \( -1 < tr < 1 \), \( \gamma = (tr) \ast d \)
spectral clustering vs. statistical model

Dynamic detection range - statistical VS. spectral clustering

- Statistical model
- Spectral clustering

speech

Transient noise
Separation of frequency ranges

- From the frequency span of the signal we can see that there is more difference between frames of transient noise and frames of speech in the lowest range of frequency.
- Therefore we decided to try the proposed VAD algorithm on different ranges of frequencies.
- We expected to reach better results on lower frequencies ranges.

![Diagram showing speech only, transient noise only, and transient noise & speech categories with high and low frequencies]
Results

- Different frequency span usage comparison of \( P_d \) vs. \( P_f \) dynamic range in different SNRs:

  \[ \text{SNR}=0 \]

  ![SNR=0](image)

  \[ \text{SNR}=5 \]

  ![SNR=5](image)

  \[ \text{SNR}=10 \]

  ![SNR=10](image)
In some cases, statistical VAD is superior to the spectral clustering. Thus possible performance improvement can be achieved by cascading both solution:

- Good separation of speech and transient from stationary noise
- Good separation of speech from transient