A Polynomial Subspace Projection Approach for the Detection of Weak Voice Activity

Imperial College London

Vincent W. Neo, Stephan Weiss, Patrick A. Naylor

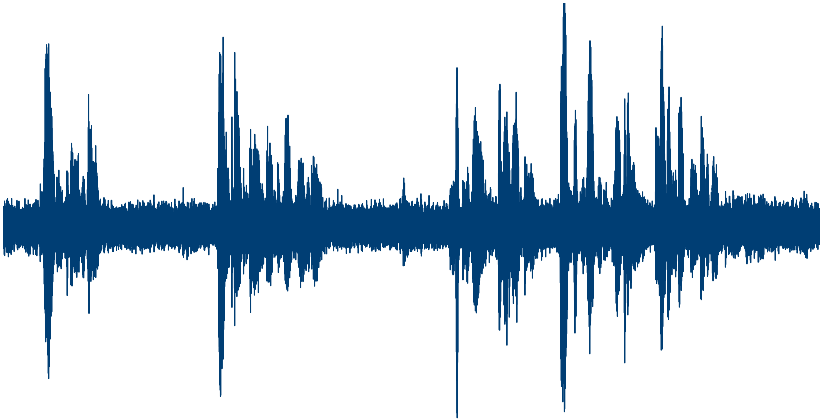
SSPD 2022

Outline

Introduction

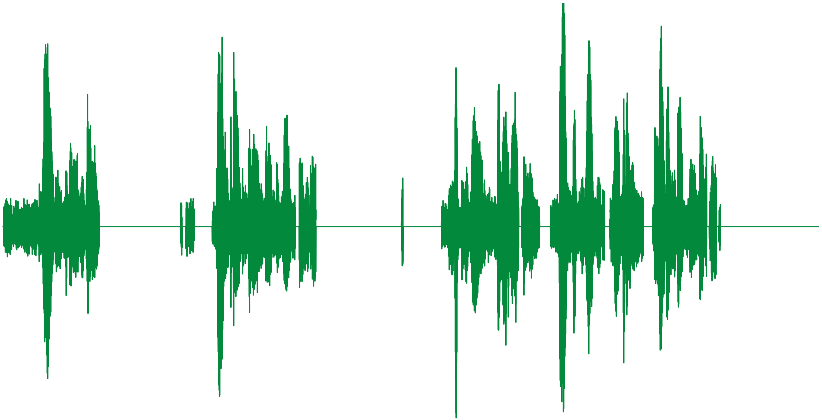
What is Voice Activity Detection (VAD)

Imperial College London



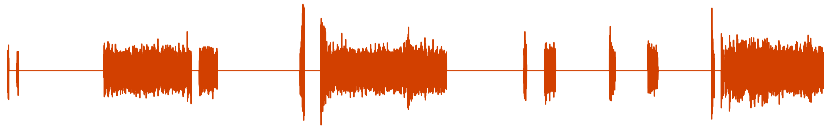
What is Voice Activity Detection (VAD)

Imperial College London



What is Voice Activity Detection (VAD)

Imperial College London



Motivation for VAD

Detection of voice activity is important for many applications:

- Speech enhancement in hearing aids, telecommunications
- Automatic speech recognition (ASR) systems
- Robot audition

Motivation for VAD

Detection of voice activity is important for many applications:

- Speech enhancement in hearing aids, telecommunications
- Automatic speech recognition (ASR) systems
- Robot audition

Main challenges:

- Background noise
- Interfering sources
- Reverberation

Existing Approaches to VAD

Statistical-based single channel methods [Sohn1999; ITU-T 2012; Gazor2003]

- Exploit differences in noise and speech distributions

Existing Approaches to VAD

Statistical-based single channel methods [Sohn1999; ITU-T 2012; Gazor2003]

- Exploit differences in noise and speech distributions

$\Rightarrow$ Challenging to measure signal statistics in very noisy environments

Existing Approaches to VAD

Statistical-based single channel methods [Sohn1999; ITU-T 2012; Gazor2003]

- Exploit differences in noise and speech distributions

$\Rightarrow$ Challenging to measure signal statistics in very noisy environments

Machine learning-based methods [Google 2021; Zhang2016; Ivry2019]

- Speech feature extraction for classification

Existing Approaches to VAD

Statistical-based single channel methods [Sohn1999; ITU-T 2012; Gazor2003]

- Exploit differences in noise and speech distributions

$\Rightarrow$ Challenging to measure signal statistics in very noisy environments

Machine learning-based methods [Google 2021; Zhang2016; Ivry2019]

- Speech feature extraction for classification

$\Rightarrow$ Feature extraction becomes difficult in adverse acoustic environments

Existing Approaches to VAD

Statistical-based single channel methods [Sohn1999; ITU-T 2012; Gazor2003]

- Exploit differences in noise and speech distributions

$\Rightarrow$ Challenging to measure signal statistics in very noisy environments

Machine learning-based methods [Google 2021; Zhang2016; Ivry2019]

- Speech feature extraction for classification

$\Rightarrow$ Feature extraction becomes difficult in adverse acoustic environments

- Weak Transient Signal Detection Using PEVD [Weiss2021]
  - Exploits multichannel signal processing to amplify weak transient signals

Existing Approaches to VAD

Statistical-based single channel methods [Sohn1999; ITU-T 2012; Gazor2003]

- Exploit differences in noise and speech distributions

$\Rightarrow$ Challenging to measure signal statistics in very noisy environments

Machine learning-based methods [Google 2021; Zhang2016; Ivry2019]

- Speech feature extraction for classification

$\Rightarrow$ Feature extraction becomes difficult in adverse acoustic environments

- Weak Transient Signal Detection Using PEVD [Weiss2021]
  - Exploits multichannel signal processing to amplify weak transient signals

This Talk: PEVD-based Multichannel Preprocessing for VAD

Background

The received signal at the q-th sensor with time index n is

$$x_q(n) = \sum_{p=1}^{p} \mathbf{h}_{p,q}^T(n)\mathbf{s}_p(n)$$

where

- $\mathbf{h}_{p,q}(n)$ is the room impulse response from pth source to qth microphone modelled as a Jth order FIR filter,
- $\mathbf{s}_p(n)$ is the pth localized source signal.

The data vector collected from Q microphones:

$$\mathbf{x}(n) = [x_1(n), x_2(n), \ldots, x_Q(n)]^T \in \mathbb{R}^Q .$$

Assuming stationarity, the space-time covariance matrix is

$$\mathbf{R}(\tau) = \mathbb{E}[\mathbf{x}(n)\mathbf{x}^T(n-\tau)] \in \mathbb{R}^{Q\times Q} \ ,$$

where $(i,j)^{\text{th}}$ element is the correlation function $r_{ij}(\tau) = \mathbb{E}[x_i(n)x_j(n-\tau)]$ and τ is the time-shift.

Z-transform of $\mathbf{R}(\tau)$ is a para-Hermitian polynomial matrix

$$\mathscr{R}(z) = \sum_{\tau=-W}^{W} \mathbf{R}(\tau) z^{-\tau},$$

where $\mathbf{R}(\tau) \approx 0$ for $|\tau| > W$, calligraphic $\mathscr{R}$ for polynomial matrices and regular $\mathbf{R}$ for matrices.

The PEVD of $\mathcal{R}(z)$ is [Weiss2018a; Weiss2018b]

$$\mathcal{R}(z) = \mathcal{U}(z)\mathbf{\Lambda}(z)\mathcal{U}^P(z)\ , \tag{1}$$

where $\mathbf{\Lambda}(z), \mathcal{U}(z)$ contain the eigenvalues and eigenvectors and $\mathcal{R}^P(z) = \mathcal{R}^H(1/z^*)$.

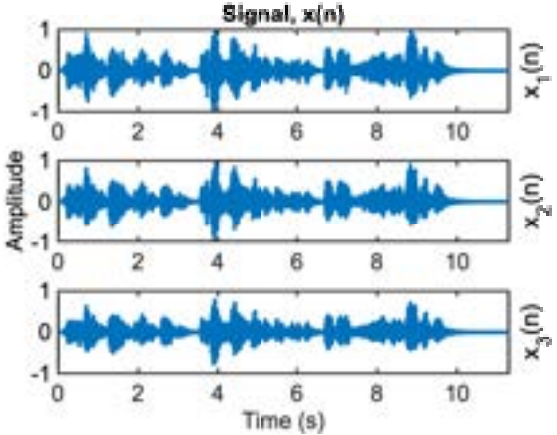
Subspace decomposition using PEVD:

$$\mathcal{R}(z) = \begin{bmatrix} \mathcal{U}_s(z) & \mathcal{U}_\perp(z) \end{bmatrix} \begin{bmatrix} \mathbf{\Lambda}_s(z) & \mathbf{0} \\ \mathbf{0} & \mathbf{\Lambda}_{\bar{s}}(z) \end{bmatrix} \begin{bmatrix} \mathcal{U}_s^P(z) \\ \mathcal{U}_\perp^P(z) \end{bmatrix}, \tag{2}$$

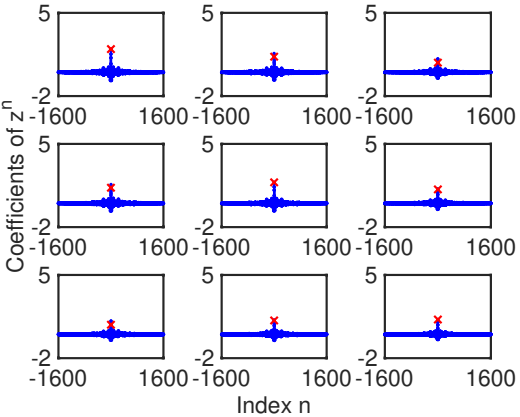
associated with signal, $\{\cdot\}_s$ and orthogonal complement, $\{\cdot\}_\perp$ subspaces.

Example: Polynomial Matrix from ST-Covariance

Imperial College London



Received signals $\mathbf{x}(n)$.

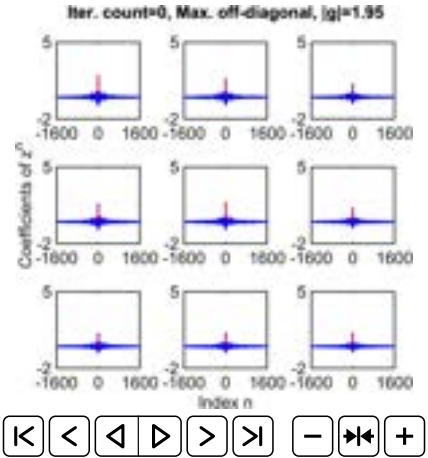


Space-time polynomial matrix, $\mathcal{R}(z)$.

Example: PEVD Algorithm

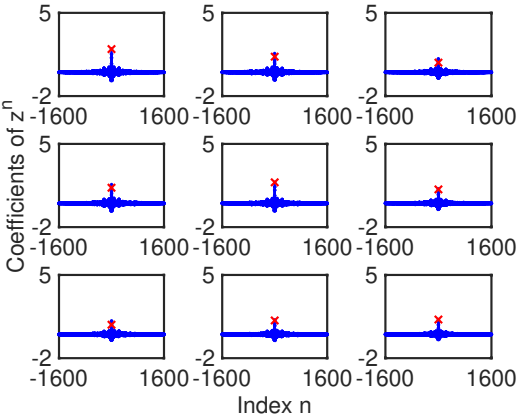
Imperial College London

Algorithm converges when $|g| < 1.68 \times 10^{-2}$

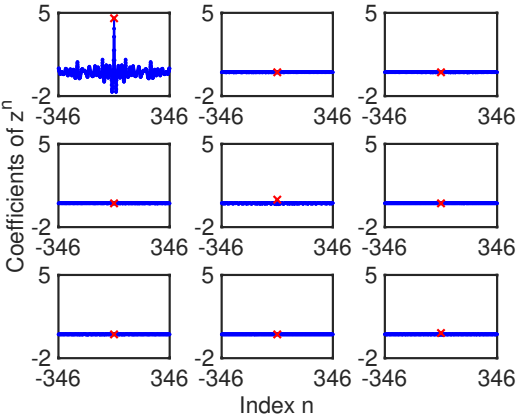


Example: PEVD Algorithm Outputs

Imperial College London



Space-time polynomial matrix, $\mathcal{R}(z)$.



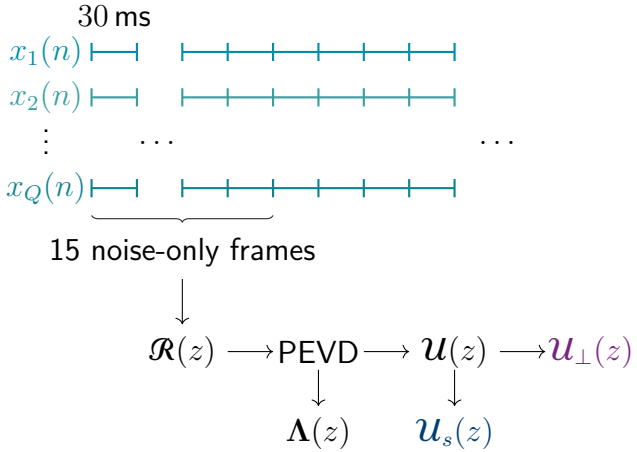
Eigenvalue polynomial matrix, $\mathbf{\Lambda}(z)$.

PEVD Algorithms

Iterative PEVD algorithms approximating (1) include:

- Second-order Sequential Best Rotation (SBR2) [McWhirter2007]
- Sequential Matrix Diagonalization (SMD) [Redif2015]
- Householder PEVD [Neo2019]
- Fixed-order approximate PEVD [Tkacenko 2010]
- Multiple-shift SBR2/SMD [Wang2015; Corr2014b]
- Causality-constrained Multiple-shift SMD [Corr2014a]

PEVD Preprocessor for VAD


30 ms
$x_1(n)$
$x_2(n)$
$x_Q(n)$
15 noise-only frames
$\mathcal{R}(z)$
PEVD
$\mathcal{U}(z)$
$\mathcal{U}_\perp(z)$
$\mathbf{\Lambda}(z)$
$\mathcal{U}_s(z)$

For L estimated signal components, $\mathcal{U}_s(z) \in \mathbb{C}^{Q\times L}$ and $\mathcal{U}_\perp(z) \in \mathbb{C}^{Q\times(Q-L)}$,

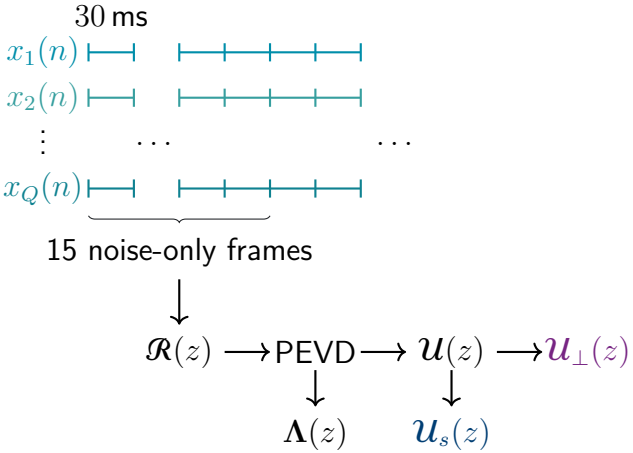
$$\mathcal{U}_s(z)\mathcal{U}_s^P(z) + \mathcal{U}_\perp(z)\mathcal{U}_\perp^P(z) = \mathbf{I}\,.$$

The component associated with $\mathcal{U}_\perp(z) \bullet\!\!-\!\!\circ\, \mathbf{U}(n)$ can be recovered using

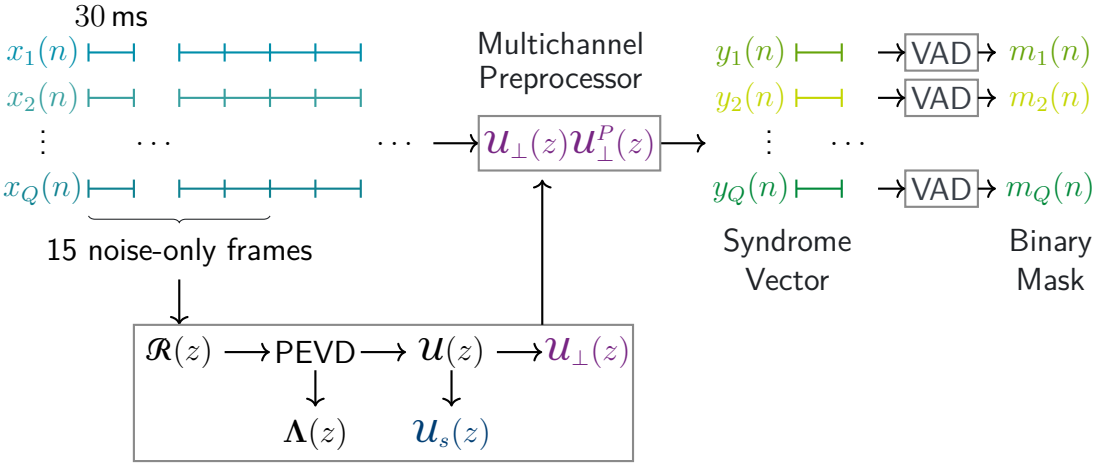
$$\mathbf{y}(n) = \sum_k \sum_m \mathbf{U}_\perp(k)\,\mathbf{U}_\perp^H(k-m)\,\mathbf{x}(n-m)\,.$$

This is equivalent to $\mathbf{x}(n)$ with the $\mathcal{U}_s(z)$ component removed.

Syndrome Signal Generation



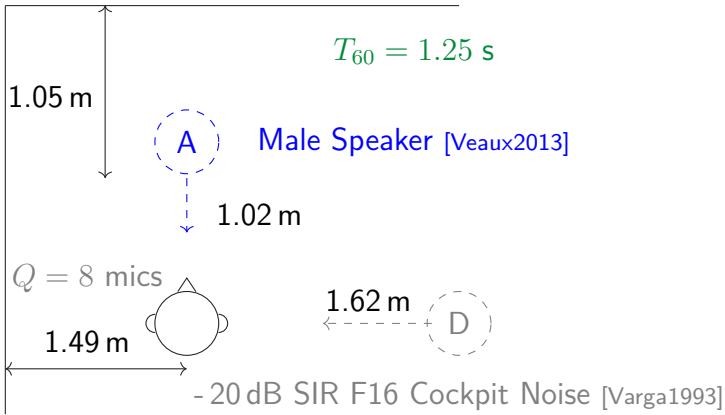
Syndrome Signal Generation



Experiment and Results

Setup: Male Speaker in a Measured Room [Kayser2009]

Imperial College London



VAD Evaluation

Comparative algorithms:

1. Sohn [Sohn1999]
2. WebRTC [Google 2021] : G0, G3 (Least to most aggressive)
3. Proposed (PEVD+Sohn): R1, R2, R5, R7 (different rank estimates)

VAD Evaluation

Comparative algorithms:

1. Sohn [Sohn1999]
2. WebRTC [Google 2021] : G0, G3 (Least to most aggressive)
3. Proposed (PEVD+Sohn): R1, R2, R5, R7 (different rank estimates)

Evaluation measures [Tharwat 2018] :

- Label evaluation metrics
  - Correct labels: True Positive (TP), True Negative (TN)
  - Wrong labels: False Positive (FP), False Negative (FN)
- Overall scores: F1, Balanced Accuracy (BACC)

VAD Evaluation

Comparative algorithms:

1. Sohn [Sohn1999]
2. WebRTC [Google 2021] : G0, G3 (Least to most aggressive)
3. Proposed (PEVD+Sohn): R1, R2, R5, R7 (different rank estimates)

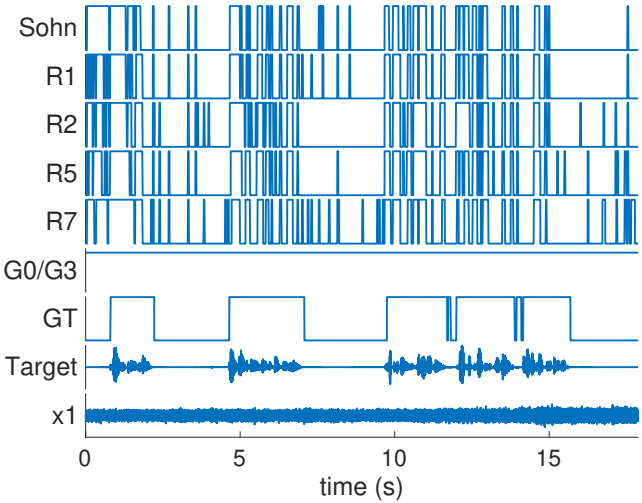
Evaluation measures [Tharwat 2018] :

- Label evaluation metrics
  - Correct labels: True Positive (TP), True Negative (TN)
  - Wrong labels: False Positive (FP), False Negative (FN)
- Overall scores: F1, Balanced Accuracy (BACC)

$\Longrightarrow$ Focus on first microphone in the results.

VAD Performance for -20 dB SIR F16 Cockpit Noise

Imperial College London



VAD Performance for -20 dB SIR F16 Cockpit Noise

Imperial College London

Method	TP	TN	FP	FN	F1	BACC
Sohn	130	241	38	185	0.538	0.638
R1	136	249	30	179	0.565	0.662
R2	158	244	35	157	0.622	**0.688**
R5	148	247	32	167	0.598	0.678
R7	136	224	55	179	0.538	0.617
G0	315	0	279	0	**0.693**	0.500
G3	315	0	279	0	**0.693**	0.500

VAD Performance for -20 dB SIR F16 Cockpit Noise

Imperial College London

Method	TP	TN	FP	FN	F1	BACC
Sohn	130	241	38	185	0.538	0.638
R1	136	249	30	179	0.565	0.662
R2	158	244	35	157	0.622	**0.688**
R5	148	247	32	167	0.598	0.678
R7	136	224	55	179	0.538	0.617
G0	315	0	279	0	**0.693**	0.500
G3	315	0	279	0	**0.693**	0.500

Other results in the paper:

- Since G0, G3 always predict the presence of speech, F1 scores significantly decrease when the speech segment is short.
- Tested on destroyer noise at various SIR from -30 dB to 20 dB.

Conclusion

Conclusion

- PEVD-based multi-microphone preprocessing for VAD
  - Characterize the ambient acoustics using PEVD to generate multichannel syndrome signals, which are microphone signals without ambient acoustics
  - Apply single channel VAD to each microphone

- Performance of proposed PEVD-based approach
  - Almost always improves F1 and BACC scores over the single channel method even in adverse environments, i.e. -30 dB SIR
  - Consistent performance unaffected by length of speech segments

References

Corr, J., K. Thompson, S. Weiss, J. G. McWhirter, and I. K. Proudler (2014a). "Causality-constrained multiple shift sequential matrix diagonalisation for parahermitian matrices". In: *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, pp. 1277–1281.

Corr, J., K. Thompson, S. Weiss, J. G. McWhirter, S. Redif, and I. K. Proudler (2014b). "Multiple shift maximum element sequential matrix diagonalisation for parahermitian matrices". In: *Proc. IEEE/SP Workshop on Statistical Signal Process.* Pp. 844–848.

Gazor, S. and Wei Zhang (Sept. 2003). "A soft voice activity detector based on a Laplacian-Gaussian model". In: *IEEE Trans. Speech Audio Process.* 11.5, pp. 498–505.

Google (2021). *WebRTC Voice Activity Detector*.

ITU-T (Nov. 2003). *Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs*. Recommendation P.862. Int. Telecommun. Union (ITU-T).

ITU-T (June 2012). *Recommendation G.729 Annex B: a silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications*. Recommendation. Int. Telecommun. Union (ITU-T).

Ivry, A., B. Berdugo, and I. Cohen (May 2019). "Voice activity detection for transient noisy environment based on diffusion nets". In: *IEEE J. Sel. Topics Signal Process.* 13.2, pp. 254–264.

Kayser, H., S. D. Ewert, J. Anemüller, T. Rohdenburg, V. Hohmann, and B. Kollmeier (July 2009). "Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses". In: *EURASIP J. on Advances in Signal Process.* 2009.1, p. 298605.

References

McWhirter, J. G., P. D. Baxter, T. Cooper, S. Redif, and J. Foster (May 2007). "An EVD algorithm for para-hermitian polynomial matrices". In: *IEEE Trans. Signal Process.* 55.5, pp. 2158–2169.

Neo, V. W. and P. A. Naylor (2019). "Second order sequential best rotation algorithm with Householder transformation for polynomial matrix eigenvalue decomposition". In: *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, pp. 8043–8047.

Redif, S., S. Weiss, and J. G. McWhirter (Jan. 2015). "Sequential matrix diagonalisation algorithms for polynomial EVD of parahermitian matrices". In: *IEEE Trans. Signal Process.* 63.1, pp. 81–89.

Scheibler, Robin, Eric Bezzam, and Ivan Dokmanić (Apr. 2018). "Pyroomacoustics: a python package for audio room simulation and array processing algorithms". In: *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, pp. 351–355.

Sohn, Jongseo, Nam Soo Kim, and Wonyong Sung (1999). "A statistical model-based voice activity detection". In: *IEEE Signal Process. Lett.* 6.1, pp. 1–3.

Tharwat, A. (Aug. 2018). "Classification assessment methods". In: *Applied Computing and Informatics* 17.1, pp. 168–192.

Tkacenko, A. (2010). "Approximate eigenvalue decomposition of para-hermitian systems through successive FIR paraunitary transformations". In: *Proc. IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, pp. 4074–4077.

Varga, A. and H. J. M. Steeneken (July 1993). "Assessment for automatic speech recognition II: NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems". In: *Speech Commun.* 3.3, pp. 247–251.

References

Veaux, C., J. Yamagishi, and S. King (Nov. 2013). "The voice bank corpus design, collection and data analysis of a large regional accent speech database". In: *Conf. Asian Spoken Language Research and Evaluation*.

Wang, Z., J. G. McWhirter, J. Corr, and S. Weiss (2015). "Multiple shift second order sequential best rotation algorithm for polynomial matrix EVD". In: *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, pp. 844–848.

Weiss, S., C. Delaosa, J. Matthews, I. K. Proudler, and B. A. Jackson (Sept. 2021). "Detection of weak transient signals using a broadband subspace approach". In: *Sensor Signal Process. for Defence Conf. (SSPD)*.

Weiss, S., J. Pestana, and I. K. Proudler (May 2018a). "On the existence and uniqueness of the eigenvalue decomposition of a parahermitian matrix". In: *IEEE Trans. Signal Process.* 66.10, pp. 2659–2672.

Weiss, S., J. Pestana, I. K. Proudler, and F. K. Coutts (Dec. 2018b). "Corrections to "On the Existence and Uniqueness of the Eigenvalue Decomposition of a Parahermitian Matrix"". In: *IEEE Trans. Signal Process.* 66.23, pp. 6325–6327.

Zhang, Z.-L. and D. Wang (Feb. 2016). "Boosting contextual information for deep neural network based voice activity detection". In: *IEEE/ACM Trans. Audio, Speech, Language Process.* 24.2, pp. 252–264.

Thank you

Listening Examples: https://vwn09.github.io/research/pevd-vad