Energy-Based Voice Activity Detection Algorithm using Gaussian and Cauchy Kernels

Aminadabe dos S. P. Soares*, Wemerson D. Parreira†, Everton G. Souza*, Sérgio J. M. de Almeida*, Claudio M. Diniz*, Chiara D. Nascimento*, Matheus F. Stigger*
*Catholic University of Pelotas. Graduate Program on Electronics Engineering and Computing, Pelotas, Brazil
Email: aminadabe.soares@sou.ucpel.edu.br, everton.granemann@ucpel.edu.br, sergio.almeida@ucpel.edu.br, claudio.diniz@ucpel.edu.br, chiara.nascimento@ucpel.edu.br, matheustigger@outlook.com,
†Federal University of Santa Catarina, Florianópolis, Santa Catarina, Brazil
Email: wemerson.d.p@ufsc.br

***Abstract*—In this work we present a simple and robust Energy-Based Voice Activity Detection Algorithm using Kernel (KVAD). Taking advantage of kernel metrics, Gaussian and Cauchy kernels are used to classify acoustic signatures as speech and non-speech. As an evidence of the potentiality of KVAD algorithm, comparisons with existing energy-based algorithms are presented, showing better performance in adverse environments as low signal-to-noise ratio (SNR) and non-stationary noise**

Keywords*—*Voice Activity Detection, Energy, Gaussian Kernel, Cauchy Kernel, Low SNR.

I. Introduction

The process of distinguish speech and non-speech is called voice activity detection (VAD). VAD algorithms typically extract some feature from the input signal, usually determined during speech-absent periods, and compare with a threshold value. Generally, if the feature value exceeds the threshold, a classification rule is applied, assigning to the audio segment a binary operation ranking it as speech or non-speech. An accurate detection of the presence or absence of speech is challenging, in particular, when the speech signal is corrupted by background noise.

In the literature, VAD algorithms have been based on various features, as signal energy [1], [2], spectral entropy [3], hidden Markov Model [4], [5] and Neural Network [6]. One of the most traditional algorithms, is based on the level of energy and zero crossing rate [7]. These algorithms have the advantages of being simple, with no assumptions about noise characteristics but fail when the background noise increases. Furthermore, as pointed out by Lu et al [8], speech and noise statistical overlap each other when handle in traditional (euclidean) processing space. As the overlapping increases with noise level, it is suitable find a mapped signal subspace in which most of the speech information is kept while the noise information is discarded. To overcome these problems, detection methods based in kernel feature space has been popularized due its capability to increase data separability in low SNR environments. Progress has been achieved distinguishing speech spectrum from transients [9], as well as distinction of speech/non-speech using a likelihood ratio test [10].

978-1-5386-2311-4/18/ $ 31.00 ©2018 IEEE

However, none of the strategies achieved a global optimum, i.e., low computational complexity, suitable decision rules and adaptability to background noise. For this reason, we propose an alternative energy-based VAD algorithm using Kernel (KVAD). The detection strategy is based on a classification problem in Reproducing Kernel Hilbert Space (RKHS). Adopting the kernel trick, we can handle the inner products of the mapping functions through different kernel functions, even without knowing the mapping functions. Hence, the kernel metric is used to classify the speech and non-speech, as a simple binary problem. The algorithm performance is tested for three types of noise: white, pink and cafeteria babble, showing better SNR score when compared to existing approaches based on energy estimation. The paper is organized as it follows. The kernel function and KVAD algorithm are described in Section II. Section III contrasts the KVAD simulations results with others in the literature. In Section IV we conclude the work.

II. KVAD Algorithm

The proposed energy-based VAD algorithm using Kernel (KVAD) apply the classification criteria in RKHS. In general terms, the problem can be split into three main stages: i) selection of an appropriate feature; ii) choice of a metric as a measure of similarity between those features (data-points) and iii) selection of the classification algorithm. Fig. 1 presents the block diagram of the KVAD algorithm. Details are given in the following subsections.

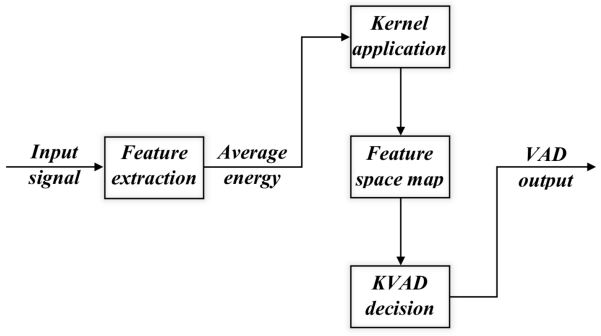

Fig. 1. Block diagram of the proposed KVAD algorithm.

A. Extraction of voice features

The first step in any automatic voice activity detection system is the extraction of acoustic feature from the signal [11], [12]. Most techniques assume that these features are statistically stationary over the interval of milliseconds. Thus, the speech signal is divided into short segments of fixed duration using a sliding window technique, which are then treated as frames of observation for features extraction. Let $u(i)$ be the ith sample of speech and M the number of samples in a frame. So the jth frame can be represented as:

$$\boldsymbol{u}_j = [u(jM), u(jM+1), \ldots, u(M(j+1)-1)]^T. \quad (1)$$

Then, the full-band energy of a speech signal within the jth frame is [13]:

$$\mathcal{E}_j = \frac{1}{M} \sum_{k=jM}^{M(j+1)-1} u^2(k). \quad (2)$$

Considering the total energy signal as the main feature, the KVAD algorithm proposed here will be restricted to the following characteristics:

- A1: Speech is locally stationary. Its spectral form do not change or change slightly over short periods of maximum 30 ms.
- A2: Energy of the speech signal is usually higher than background noise energy, otherwise, speech will be unintelligible. Besides, it is assumed that the initial samples ($\sim$ 100 ms) does not contain any speech.
- A3: Most part of the energy speech signal is concentrated in low frequency bands.

B. Kernel function, similarity and feature space map

In general, kernel functions are used to solve non-linear problems based in RKHS [14]. This space is known as the Features Space, $\mathcal{H}$. Let $\mathcal{H}$ be a Hilbert space of real-value functions on a domain $\mathcal{X}$, equipped with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and a real-valued bivariate function $\kappa(\boldsymbol{x}, \boldsymbol{y})$ on $\mathcal{X} \times \mathcal{X}$ [15]. The Feature Space allows a simpler classification of these elements than the original space [16]. Considering the patterns $\boldsymbol{x}$ and $\boldsymbol{y} \in \mathcal{X}$, e.g. a vector that describes the voiced and the unvoiced (noise) features (see Sec. II-A), the positive definite kernel function $\kappa(\cdot, \cdot)$ is given by [17]:

$$\kappa(\boldsymbol{x}, \boldsymbol{y}) = \langle \Phi(\boldsymbol{x}), \Phi(\boldsymbol{y}) \rangle \quad (3)$$

where Φ is the mapping function which represents a inner product in the feature space $\mathcal{H}$. Assuming $\mathcal{X}$ is a set of the voice features, then it is possible to map into an Euclidean space, $\mathbb{R}^p$, using the application Φ as (3). Thus, for the similarity analysis between $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathcal{X} \subset \mathbb{R}^p$, we use two positive definite kernels[1]:

1) The Gaussian kernel is defined by [17],

$$\kappa_G(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{2\xi^2}\right) \quad (4)$$

where $\|.\|$ denote the Euclidean distance and ξ is the kernel bandwidth ($\xi \neq 0$) .

2) The Cauchy kernel is derived from the Cauchy's distribution [18],

$$\kappa_C(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sigma^2}{\sigma^2 + \|\boldsymbol{x}-\boldsymbol{y}\|^2} \quad (5)$$

where $\|.\|$ denote the Euclidean distance and σ is the smoothing parameter.

As will be discussed later, the σ and ξ are also known as kernel parameters. The Gaussian (4) and Cauchy kernels (5) are unit-norm kernels, that is, $0 \leq \kappa(\boldsymbol{x}, \boldsymbol{y}) \leq 1$, for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$. Note that, if ξ in (4) and σ in (5) approaches to zero, then $\kappa(\boldsymbol{x}, \boldsymbol{y}) \approx 0$. In the other hand, if ξ and σ are close to infinite, then $\kappa(\boldsymbol{x}, \boldsymbol{y}) \approx 1$. For this reason, we should avoid to choose ξ and σ very close to zero.

C. KVAD algorithm design guidelines

The previous analysis allow us now to establish some guidelines. Suppose the desired goal is to obtain a comparison between the similarity measure (induced by kernel) and τ_0, threshold. The following procedure could be applied.

1) Define a kernel function and kernel parameter(s);
2) Set a threshold $\tau_0 \in (0, 1)$.
3) Apply a feature extractor in the first samples of the voice signal. This sequence, $\boldsymbol{u}_0$, of short unit of time ($\sim$ 10 ms) will be the reference frame, named $\boldsymbol{f}_0$, whose speech is absent.
4) Apply a feature extractor in the next frames, $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N$, named $\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_N$, where $N+1$ is the number of the digital signal frames to be tested.
5) Compute $\kappa(\boldsymbol{f}_0, \boldsymbol{f}_j)$, with $1 \leq j \leq N$.
6) Make a decision, if $\kappa(\boldsymbol{f}_0, \boldsymbol{f}_j) \leq \tau_0$, with $1 \leq j \leq N$ then the jth frame has voice activity, so mark it as a *speech* frame, otherwise mark it as *non-speech* frame.

III. Methods, Results and Performance Evaluation

A. Identification measures

In order to establish a comparative criterion between different methods, we obtained a Gold Standard (GS) by a hand-marked process using Praat software [19]. Differently of Soleimani et al. [20] we have used three statistical quality estimators: the Mean Square Error (MSE) and the False Identification Rate for speech (FIRs) and non-speech samples (FIRns). The idea behind was to check the correspondence between GS, our KVAD algorithm and other two related energy-based algorithms: Voice Activity Detection with Zero-Crossing Rate (VAD-ZC) and Voice Activity Detection with Adaptive Scaling Factor (VAD-ASF). For the calculation we defined MSE, J_{mse}, as:

$$J_{\text{mse}} = \frac{1}{N} \sum_{k=1}^{N} \|l_k - \hat{l}_k\|^2 \quad (6)$$

where N is the number of samples of the signal, l_k and $\hat{l}_k$ are the labels for the kth sample from GS and VAD algorithms, respectively. It worth mentioning that l_k and $\hat{l}_k$ are zero/one for all non-speech/speech samples. By way of clarification, we computed $10\log(J_{mse})$ instead of J_{mse} (in Table I) to increase some orders of magnitude of J_{mse}. The FIRs, which

[1] A number of different terms are used for positive definite kernels, such as Mercer kernel, admissible kernel, Support Vector kernel and covariance function [17].

represents the detection of voice samples where there was no voice activity with respect to GS, was calculated as follows:

$$\Gamma_{\mathrm{s}} = \frac{C_{\mathrm{s}}}{(N - N_{\mathrm{noise}})} \tag{7}$$

where Γ_{s} give the false identification rate of the speech samples, with C_{s} computing the number of false identification occurrence of the speech samples and N_{noise} the number of the non-speech samples. In contrast, FIRns computes the detection of non-speech where there was voice activity:

$$\Gamma_{\mathrm{ns}} = \frac{C_{\mathrm{ns}}}{(N_{\mathrm{noise}})} \tag{8}$$

where Γ_{ns} is the false identification rate of the non-speech samples (FIRns) and C_{ns} is the occurrence of false identifications.

B. Kernel parameter selection

As discussed in Sec. II-B, when kernel functions are used as measure of similarity, the kernel parameter plays a special role in VAD process. It scales appropriately high distances, between the reference frame and the others frames, into short distances whereas greater distances than the kernel parameter become negligible. Fig. 2 presents the simulation results for the Gaussian-KVAD (G-KVAD) and the Cauchy-KVAD (C-KVAD) algorithms when using the Gaussian kernel (4) and Cauchy kernel (5), respectively. For select the best kernel parameters, we ran out simulations for ξ, σ in the range $[0; 5 \times 10^{-3}]$, searching for the ones which minimize the MSE. For the sake of simplicity, we adopted a fixed threshold $\tau_0 = 0.5$ in this search. The results are depicted in Fig. 2.

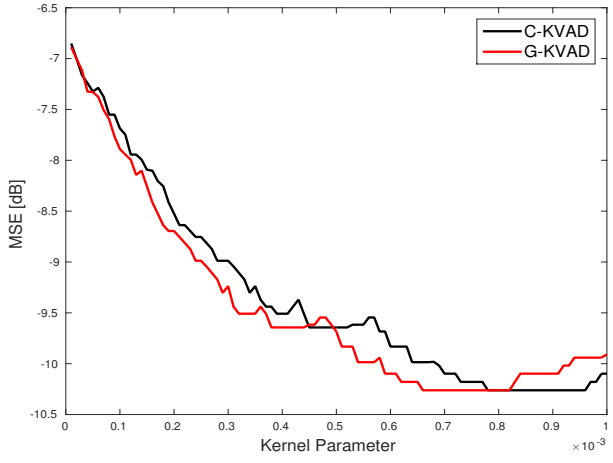

Fig. 2. MSE behavior of the proposed G-KVAD (red) and C-KVAD (black) algorithms.

As we can see, Fig. 2 shows a global minimum (picture was truncated for 10^{-3} in x-axis, for better visualization) for $\xi \in [0.7 \times 10^{-3}; 0.8 \times 10^{-3}]$ for G-KVAD and $\sigma \in [0.8 \times 10^{-3}; 0.9 \times 10^{-3}]$ for the C-KVAD algorithm.

C. Comparison with referenced energy-based VAD algorithms

After establishing the appropriated kernel parameters, the performance of G-KVAD and C-KVAD was tested against two others VAD algorithms: the Energy-Based Voice Activity Detection with Zero-Crossing Rate (VAD-ZC) [1] and the Energy-Based Voice Activity Detection with Adaptive Scaling Factor (VAD-ASF) [2]. For the analysis we use speech samples from a Brazilian database of 8 kHz, 16 bits, with frame length of 10 ms, corresponding to 80 samples in time domain, with approximately 5.54 s total time length each. To verify the accuracy of the proposed algorithms, they were compared statistically in terms of MSE, FIRs and FIRns. For all the tests, the speech samples were corrupted by a stationary (Gaussian white and pink) and non-stationary (babble) noises. Tables I, II and III summarize the results. For each one of the statistical estimators we performed Monte Carlo simulations over 50 runs, with SNR values ranging from 0 to 40 dB. The kernel parameters adopted were $\xi = 0.7 \times 10^{-3}$ for the G-KVAD and $\sigma = 0.8 \times 10^{-3}$ for the C-KVAD, according to the estimations obtained in Sec. III-B. The results displayed on the tables are averaged values over these 50 runs. In general, KVAD algorithms obtained better performance than VAD-ASF and VAD-ZC in the most of low SNR environments. As we can see in Table I, the MSE values are smaller for KVAD, at least two orders of magnitude, for all types of noise.

TABLE I. COMPARISON OF THE AVERAGED VALUES OF MSE (DB) FOR PINK, WHITE AND BABBLE NOISES

	SNR - White Noise			
Algorithm	0 dB	10 dB	20 dB	40 dB
VAD-ZC	-2.049	-2.049	-2.489	-10.242
VAD-ASF	-2.081	-3.785	-5.776	-8.655
G-KVAD	-6.192	-10.074	-10.248	-10.262
C-KVAD	-6.247	-10.091	-10.233	-10.262
	SNR - Pink Noise			
Algorithm	0 dB	10 dB	20 dB	40 dB
VAD-ZC	-2.049	-2.049	-2.498	-10.223
VAD-ASF	-2.325	-4.030	-5.969	-7.776
G-KVAD	-4.843	-8.938	-10.177	-10.262
C-KVAD	-4.780	-8.945	-10.164	-10.262
	SNR - Babble Noise			
Algorithm	0 dB	10 dB	20 dB	40 dB
VAD-ZC	-2.049	-2.049	-3.585	-10.262
VAD-ASF	-2.336	-3.514	-5.151	-5.652
G-KVAD	-4.862	-9.095	-10.179	-10.262
C-KVAD	-4.798	-8.911	-10.104	-10.262

Table II and Table III display the percentage of FIRs and FIRns for all the algorithms. Table II depicts some cases where KVAD gets higher indices of errors than VAD-ASF, for example, at 10 dB for white and pink noise and 0 dB for babble noise. This is a consequence of VAD-ASF adaptive threshold, which treats each voice segment adaptively, different from KVAD whose threshold is static. On the other hand, the percentage of FIRns for VAD-ZC and VAD-ASF are around of 60% for low SNR values, as we can see in Table III, indicating that although there was a good classification in the voice segments for FIRs, there was loss of information. In addition, VAD-ASF and VAD-ZC were not even able to classify some stretch of voice activity considering low SNR, denoted by (-). In this way, the KVAD algorithms withhold more regularity with respect to errors: around of 4% and 16% for FIRs and FIRns, respectively, except for some special cases with very low SNR.

IV. CONCLUSION

In this work, an Energy-Based Voice Activity Detection Algorithm using Gaussian and Cauchy kernels was proposed.

TABLE II. COMPARISON OF FIRS (%) FOR PINK, WHITE AND BABBLE NOISES

Algorithm	SNR - White Noise 0 dB	10 dB	20 dB	40 dB
VAD-ZC	-	-	0	4.00
VAD-ASF	-	2.57	5.36	8.01
G-KVAD	23.70	4.52	4.05	4.01
C-KVAD	23.53	4.62	4.10	4.01
Algorithm	SNR - Pink Noise 0 dB	10 dB	20 dB	40 dB
VAD-ZC	-	-	0	4.01
VAD-ASF	23.73	7.68	8.01	9.24
G-KVAD	32.63	9.06	4.29	4.01
C-KVAD	33.06	9.49	4.31	4.01
Algorithm	SNR - Babble Noise 0 dB	10 dB	20 dB	40 dB
VAD-ZC	-	-	0.02	4.00
VAD-ASF	13.21	6.22	6.55	9.90
G-KVAD	32.39	8.84	4.30	4.00
C-KVAD	32.79	9.88	4.59	4.00

(-) denote an undefined value.

TABLE III. COMPARISON OF FIRNS (%) FOR PINK, WHITE AND BABBLE NOISES

Algorithm	SNR - White Noise 0 dB	10 dB	20 dB	40 dB
VAD-ZC	62.39	62.39	59.98	16.85
VAD-ASF	64.56	55.20	42.93	23.29
G-KVAD	23.58	17.09	16.78	16.77
C-KVAD	23.72	16.92	16.79	16.77
Algorithm	SNR - Pink Noise 0 dB	10 dB	20 dB	40 dB
VAD-ZC	62.39	62.39	59.93	16.91
VAD-ASF	63.69	53.79	40.73	27.41
G-KVAD	36.57	18.39	16.86	16.77
C-KVAD	38.37	17.81	16.89	16.77
Algorithm	SNR - Babble Noise 0 dB	10 dB	20 dB	40 dB
VAD-ZC	62.39	62.39	53.81	16.77
VAD-ASF	63.41	56.93	47.11	43.54
G-KVAD	34.01	17.71	16.84	16.77
C-KVAD	35.05	17.65	16.89	16.77

The behavior of KVAD algorithm, for both kernel functions, was tested for different types of noise, stationary and non-stationary, and low SNR levels. The KVAD algorithm was effective balancing the false identification rate of speech through the reducing the false identification rate of non-speech. The performance of the algorithm was compared with other algorithms in the literature, i.e., VAD-ZC and VAD-ASF, showing more regularity in the processes of voice identification and lower MSE.

REFERENCES

[1] R. G. Bachu et al. "Voiced/Unvoiced Decision for Speech Signals Based on Zero-Crossing Rate and Energy". *Advanced Techniques in Computing Sciences and Software Engineering*, pp 279-282, 2010.

[2] K. Sakhnov and E. Veterletskaya and B. Simak. Approach for Energy-Based Voice Detector with Adaptive Scaling Factor. *IAENG – International Journ. of Comp. Science*, 2009.

[3] S.A. McClellan, J.D. Gibson, "Spectral entropy: An alternative indicator for rate allocation", IEEE Int. Conf. on Acoustics, Speech, Signal Processing, Adelaide, Australia, p. 201–204, Apr. 1994.

[4] H. Veisi and H. Sameti, "Hidden-Markov-model-based voice activity detector with high speech detection rate for speech enhancement", IET Signal Process., vol. 6, iss. 1, pp. 54 – 63, 2012.

[5] Y-W. Tan, W-J. Liu, W. Jiang and H. Zheng, "Hybrid SVM/HMM Architectures for Statistical Model-based Voice Activity Detection", *International Joint Conference on Neural Networks (IJCNN)*, Jul. 6-11, Beijing, China, 2014.

[6] M. Farsinejad, M. Mohammadi, B. Nasersharif and A. Akbari, "A Model-based Voice Activity Detection Algorithm using probabilistic neural networks", Proceedings of 14th Asia-Pacific Conference on Communications (APCC 2008), 2008.

[7] J.C. Junqua, B. Reaves and B. Mak, "A study of endpoint detection algorithms in adverse conditions: Incidence on a DTW and HMM recognizers", Proceed. of Eurospeech, Genova, Italy, pp. 1371–1374, Sep. 24–26, 1991.

[8] X. Lu, M. Unoki, R. Isotani, H. Kawai, S. Nakamura, "Regularization in a reproducing kernel hubert space for robust voice activity detection", Signal Processing (ICSP), 2010 IEEE 10th International Conference, 585–588. 2010.

[9] D. Dov, R. Talmon and I. Cohen. "Kernel Method for Voice Activity Detection in the Presence of Transients", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, Issue 12, Dec. 2016.

[10] D. K. Kim, J-H. Chang, "Statistical voice activity detection in kernel space", *J. Accoust. Soc. Am.*, 132 (4), 2012.

[11] K. Ishizuka, T. Nakatani, M. Fujimoto and N. Miyazaki, "Noise robust voice activity detection based on periodic based to aperiodic component ratio", Speech Communication, vol. 52, no. 1, pp. 41–60, 2010.

[12] N. Cho and E. Kim, "Enhanced voice activity detection using acoustic event detection and classification", IEEE Trans. on Consumer Electronics, vol. 57, no. 1, pp. 196–202, 2011.

[13] L. R. Rabiner and R. W. Schafer. *Introduction to Digital Speech Processing, Foundations and Trends in Signal Processing*. Boston: Now Publishers Inc., 2007.

[14] W. D. Parreira, J. C. M. Bermudez, C. Richard and J-Y. Tourneret. "Stochastic Behavior Analysis of the Gaussian Kernel Least-Mean-Square Algorithm". *IEEE Trans. on Signal Process.*, vol. 60, no. 5, pp. 2208 – 2222, 2012.

[15] N. Aronszajn, "Theory of Reproducing Kernels", *Trans. Amer. Math. Soc*., vol. 68, pp. 337–404, 1950.

[16] R. Herbrich, *Learnig kernel classifiers. Theory and Algorithms*, The MIT Press, Cambridge, MA, 2002.

[17] B. Schölkopf and A. J. Smola, *Learning with Kernels - Support Vector Machines, Regularization, Optimization, and Beyond*. London: MIT Press, 2002.

[18] J. Basak, "A Least Square Kernel Machine With Box Constraints". Journal of Pattern Recognition Research, vol. 1, pp. 38–51, 2010.

[19] P. Boersma. Praat, a system for doing phonetics by computer. Glot International 5:9/10, 341–345, 2001.

[20] S. A. Soleimani and S. M. Ahadi, "Voice activity detection based on combination of multiple features using linear/kernel discriminant analysis", *Proc. 3rd Int. Conf. Inf. Commun. Technol.: From Theory to Applicat.* , pp. 1–5, 2008.

[21] R. V. Prasad, R. V. et al. "Voice Activity Detection for VoIP - An Information Theoretic Approach". In: IEEE Global Telecommunications Conference (IEEE GLOBECOM), pp. 1–6, San Francisco, 2006.

[22] J. Sohn and N. Kim, "Statistical model-based voice activity detection", IEEE Signal Process Letter, 6(1), pp. 1–3, 1999.

[23] P. C. Loizou, "Speech Enhancement - Theory and Practice", Second edition, CRC Press, 2013.

[24] E. A. P. Habets, I. Cohen and S. Gannot. "Generating nonstationary multisensor signals under a spatial coherence constraint", *Jour. of the Acoust. Soc. of America*, vol. 124, Issue 5, pp. 2911–2917, Nov. 2008.