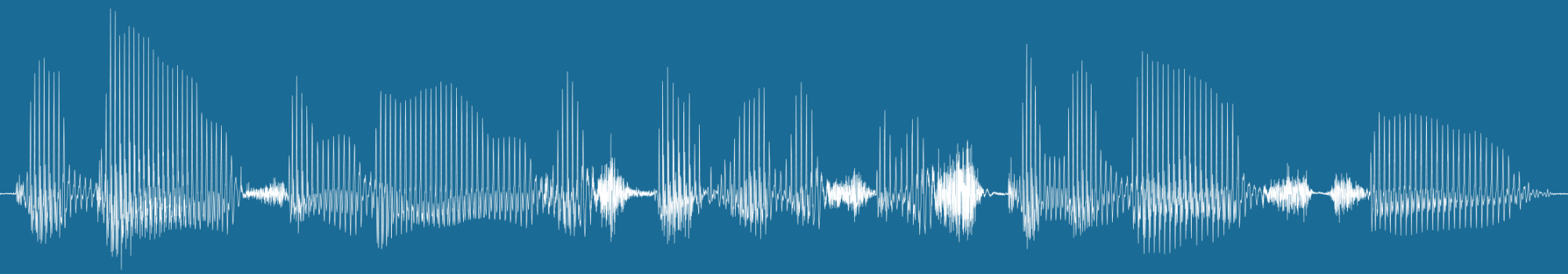
Imperial College London

UNIVERSITY OF Southampton

University of Strathclyde Glasgow



A POLYNOMIAL EIGENVALUE DECOMPOSITION MUSIC APPROACH FOR BROADBAND SOUND SOURCE LOCALIZATION

October, 2021

Aidan Hogg, Vincent Neo, Stephan Weiss, Christine Evers and Patrick Naylor

Electrical and Electronic Engineering, Imperial College London, UK
Electrical and Electronic Engineering, University of Strathclyde, Glasgow, UK
Electronics and Computer Science, University of Southampton, UK

Sound source localization is an important task for a multitude of applications, including robot audition and voice-controlled smart devices.

Direction of arrival estimates are essential in providing angular positional information for localization.

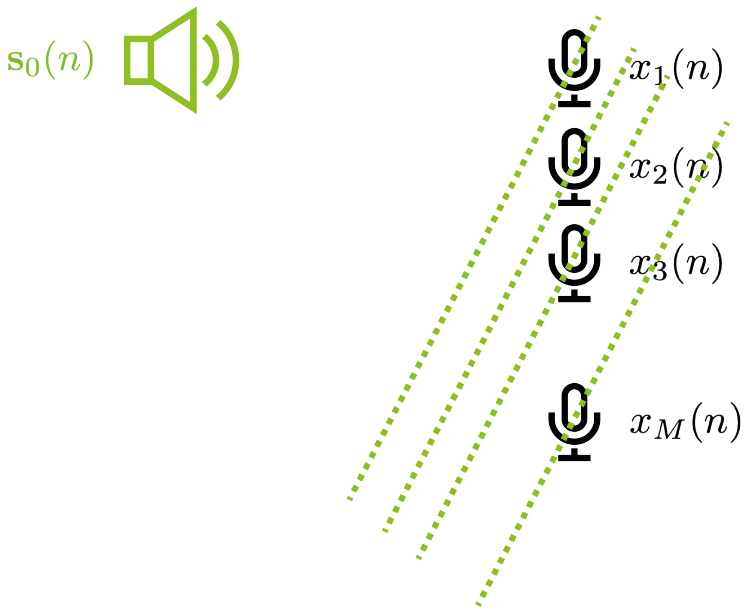
MUSIC ALGORITHM

Consider the scenario with a microphone array and a far-field sound source. The sound source signals arrive with delays expressed in a steering vector .

Data model:

$$\mathbf{x}(n) =$$

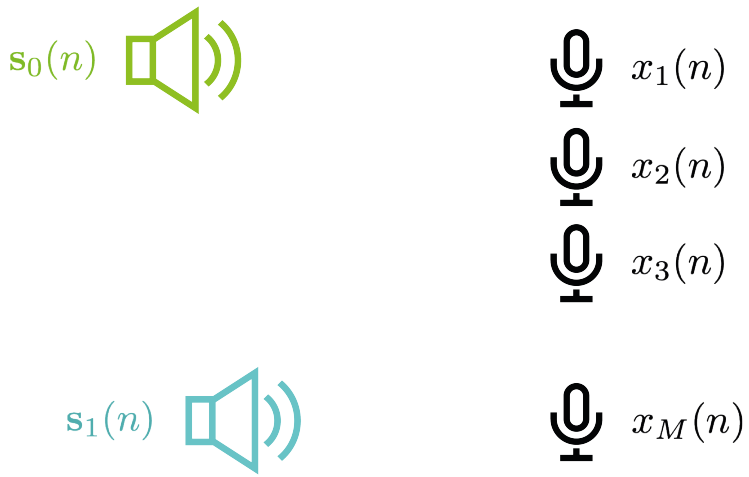
Consider the scenario with a microphone array and a far-field sound source. The sound source signals arrive with delays expressed in a steering vector $\mathbf{a}_0$.



Data model:

$$\mathbf{x}(n) = s_0(n) \cdot \mathbf{a}_0$$

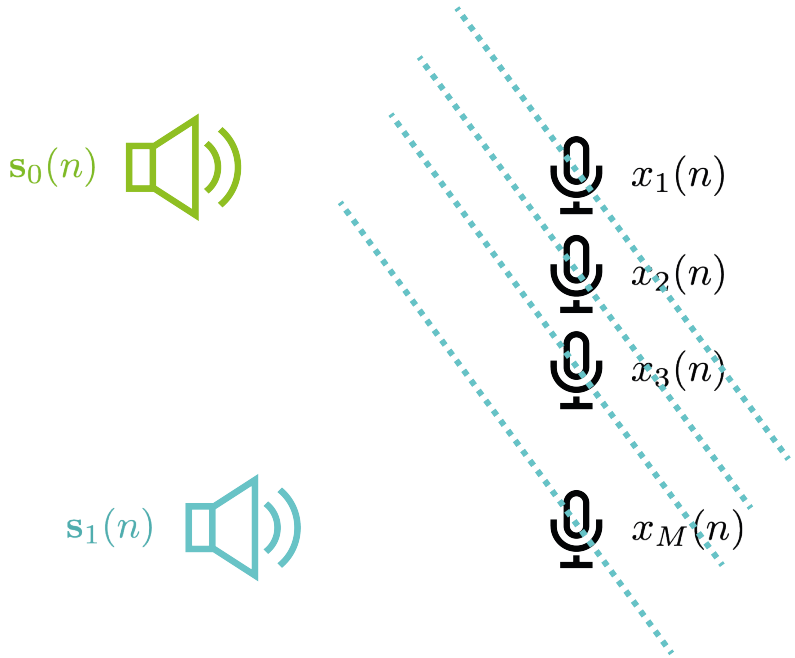
Consider the scenario with a microphone array and a far-field sound source. The sound source signals arrive with delays expressed in a steering vector $\mathbf{a}_0$.



Data model:

$$\mathbf{x}(n) = s_0(n) \cdot \mathbf{a}_0$$

Consider the scenario with a microphone array and a far-field sound source. The sound source signals arrive with delays expressed in a steering vector $\mathbf{a}_0$, $\mathbf{a}_1$.



Data model:

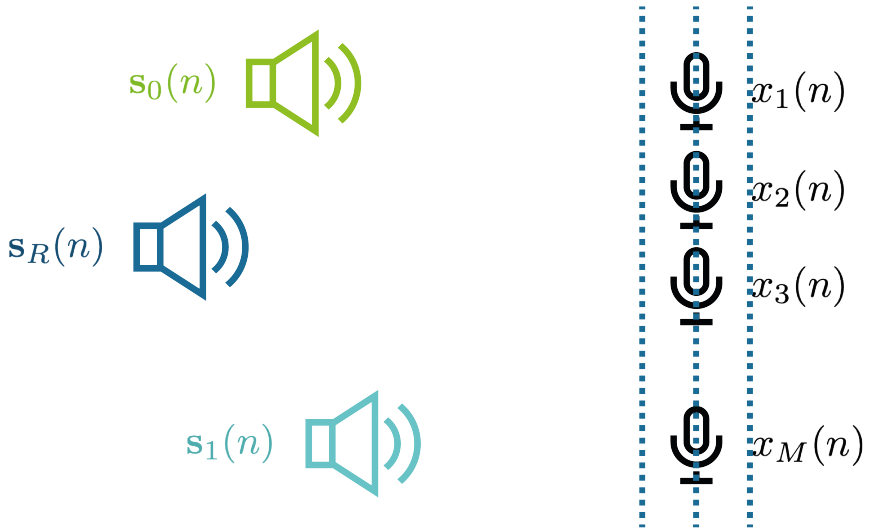
$$\mathbf{x}(n) = s_0(n) \cdot \mathbf{a}_0 + s_1(n) \cdot \mathbf{a}_1$$

Consider the scenario with a microphone array and a far-field sound source. The sound source signals arrive with delays expressed in a steering vector $\mathbf{a}_0$, $\mathbf{a}_1$.

Data model:

$$\mathbf{x}(n) = s_0(n) \cdot \mathbf{a}_0 + s_1(n) \cdot \mathbf{a}_1$$

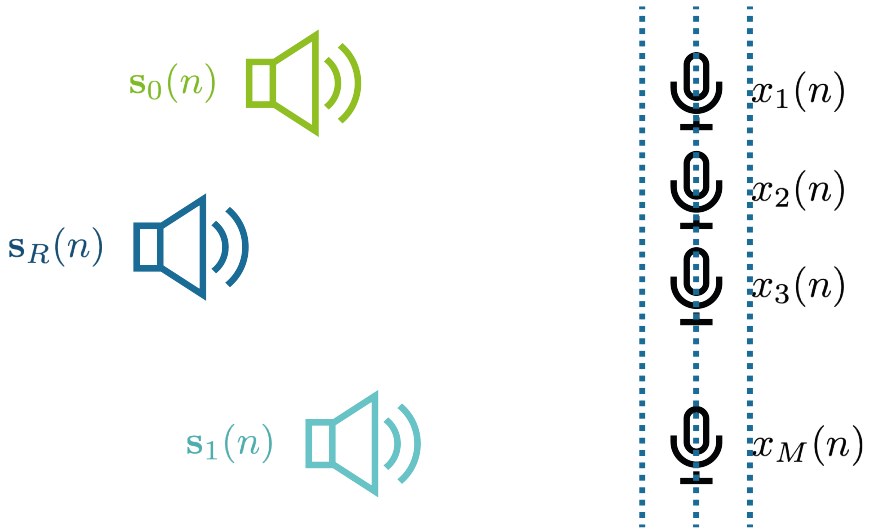
Consider the scenario with a microphone array and a far-field sound source. The sound source signals arrive with delays expressed in a steering vector $\mathbf{a}_0, \mathbf{a}_1, \cdots, \mathbf{a}_R$.



Data model:

$$\mathbf{x}(n) = s_0(n) \cdot \mathbf{a}_0 + s_1(n) \cdot \mathbf{a}_1 + \cdots + s_R(n) \cdot \mathbf{a}_R$$

Consider the scenario with a microphone array and a far-field sound source. The sound source signals arrive with delays expressed in a steering vector $\mathbf{a}_0, \mathbf{a}_1, \cdots, \mathbf{a}_R$.



Data model:

$$\mathbf{x}(n) = \sum_{r=0}^{R} s_r(n) \cdot \mathbf{a}_r$$

A signal $s(n)$ arriving at the array can be characterised by the delays of its wavefront (neglecting attenuation):

$$\begin{bmatrix} x_1(n) \\ x_2(n) \\ \vdots \\ x_M(n) \end{bmatrix} = \begin{bmatrix} s(n-\tau_1) \\ s(n-\tau_2) \\ \vdots \\ s(n-\tau_M) \end{bmatrix} = \begin{bmatrix} \delta(n-\tau_1) \\ \delta(n-\tau_2) \\ \vdots \\ \delta(n-\tau_M) \end{bmatrix} * s(n) \xrightarrow{\mathscr{Z}} \mathbf{a}_\theta(z) S(z) \quad (1)$$

If evaluated at a narrowband normalised angular frequency Ω_i, the time delays τ_m in the broadband steering vector $\mathbf{a}_\theta(z)$ collapse to phase shifts in the narrowband steering vector $\mathbf{a}_{\theta,\Omega_i}(z)$,

$$\mathbf{a}_{\theta,\Omega_i}(z) = \mathbf{a}_\theta(z)|_{z=e^{-j\Omega_i}} = \begin{bmatrix} e^{-j\tau_1\Omega_i} \\ e^{-j\tau_2\Omega_i} \\ \vdots \\ e^{-j\tau_M\Omega_i} \end{bmatrix}. \quad (2)$$

If the noisy and reverberant signal, $x_m(n)$, at the m-th microphone for discrete-time sample $n = 0, 1, \ldots, N$, is

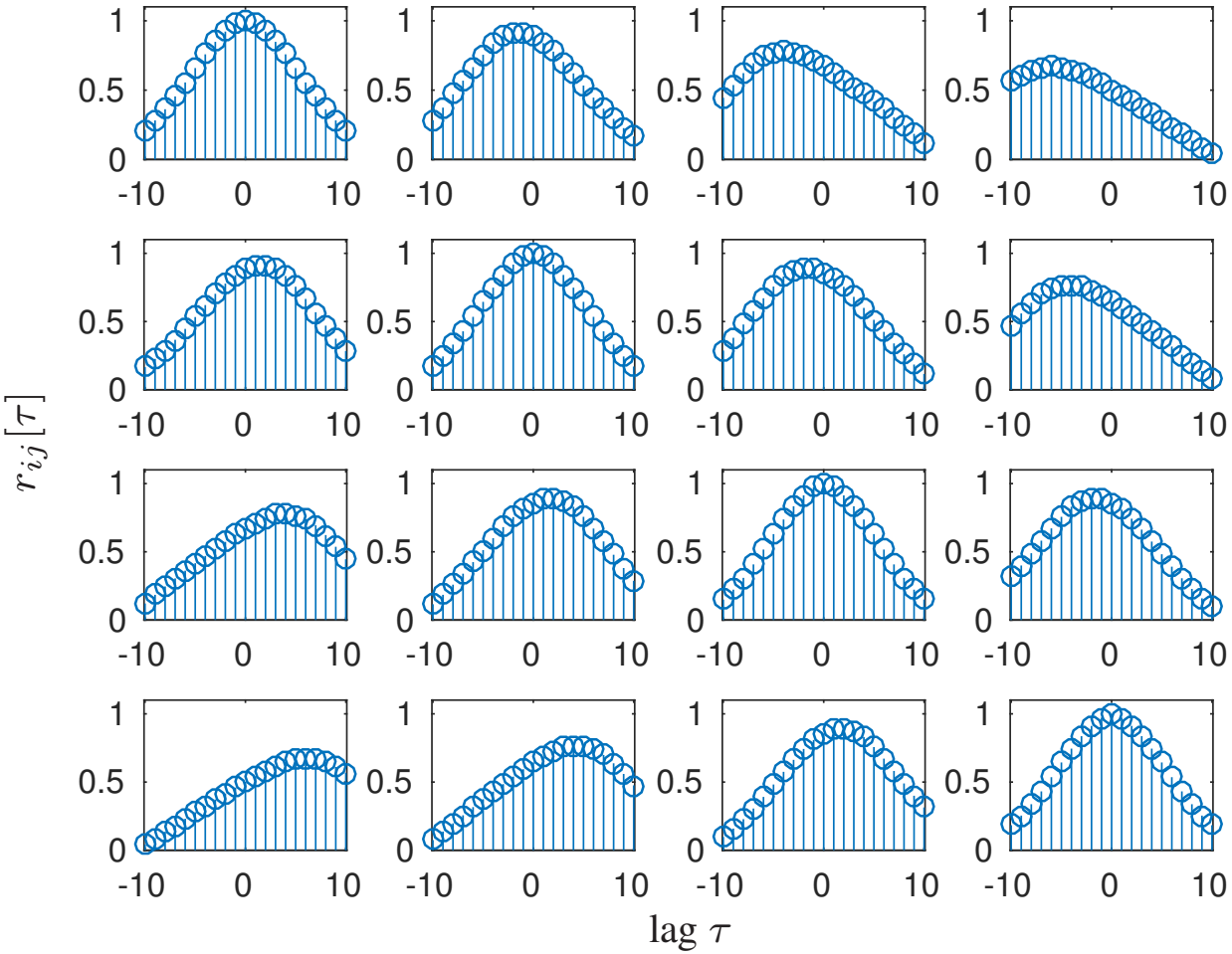
$$x_m(n) = \mathbf{h}_m^T \mathbf{s}_0(n) + v_m(n) \tag{3}$$

To capture the temporal correlations of the speech signals at different microphones, we require a space-time covariance matrix

$$\boldsymbol{\mathcal{R}}_{\mathbf{xx}}(\tau) = \mathbb{E}\{\mathbf{x}(n)\mathbf{x}^T(n-\tau)\}, \tag{4}$$

where

$$\mathbf{x}^T(n) = [x_1(n), x_2(n) \cdots x_M(n)] \tag{5}$$

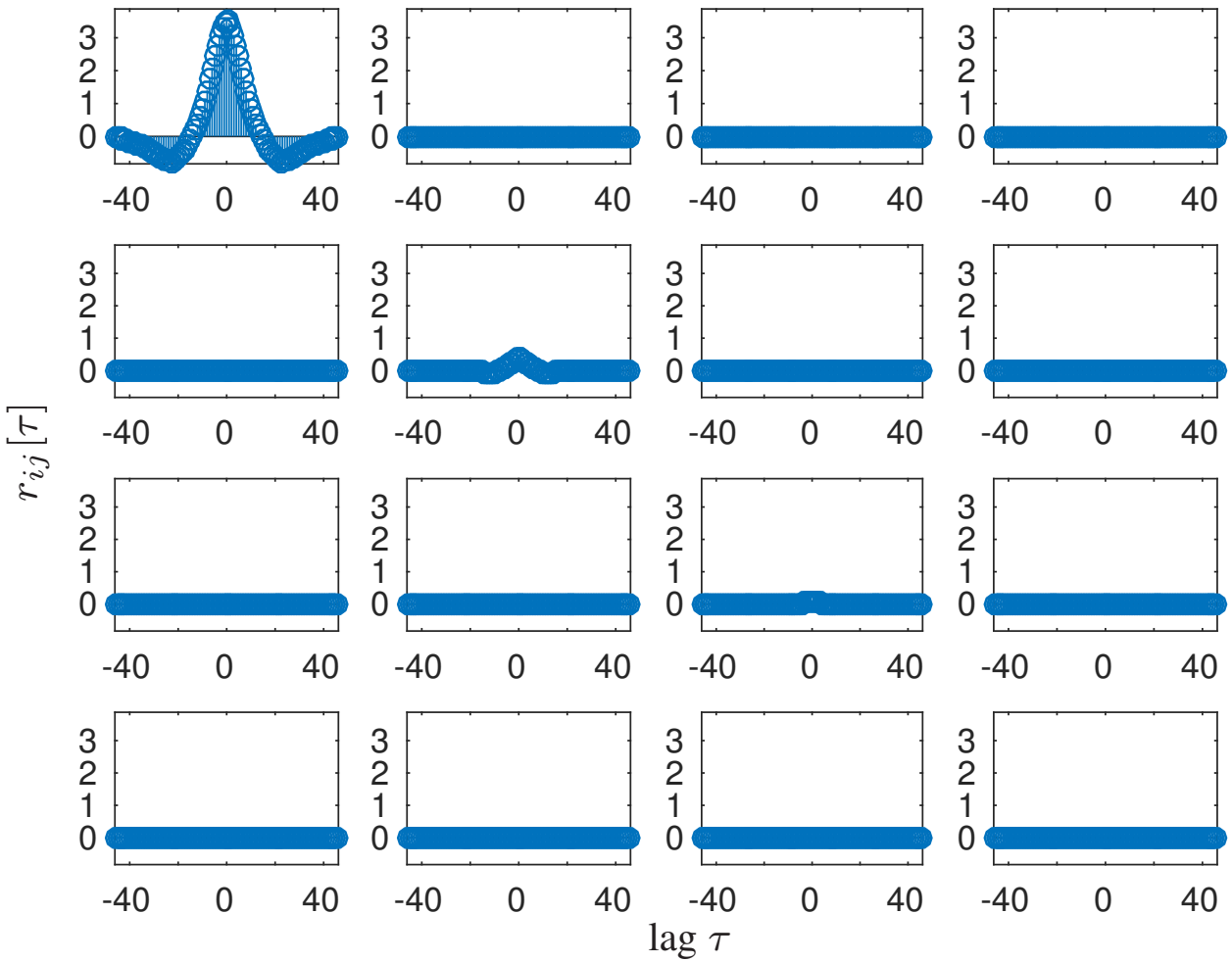


Space-time covariance matrix $\boldsymbol{\mathcal{R}}_{\mathbf{xx}}(z)$ on the interval $|\tau| \leq 10$.

Due to the signal model the PEVD of (4) can be partitioned by thresholding the eigenvalues to give

$$\tilde{\boldsymbol{\mathcal{R}}}_{\mathbf{xx}}(z) \approx \left[\boldsymbol{\mathcal{U}}_{\tilde{s}}(z) \middle| \boldsymbol{\mathcal{U}}_{\tilde{v}}(z)\right] \left[\begin{array}{c|c} \boldsymbol{\Lambda}_{\tilde{s}}(z) & \mathbf{0} \\ \hline \mathbf{0} & \boldsymbol{\Lambda}_{\tilde{v}}(z) \end{array}\right] \left[\begin{array}{c} \boldsymbol{\mathcal{U}}^{P}{}_{\tilde{s}}(z) \\ \hline \boldsymbol{\mathcal{U}}^{P}{}_{\tilde{v}}(z) \end{array}\right],$$

where $\{.\}_{\tilde{s}}$ and $\{.\}_{\tilde{v}}$ represent the orthogonal signal and noise subspace components.



Diagonalised matrix of eigenvectors produced by the SMD algorithm for the noise-free case with a direction of arrival (DoA) of 45°.

Generalising from MUSIC, the following quantity is defined

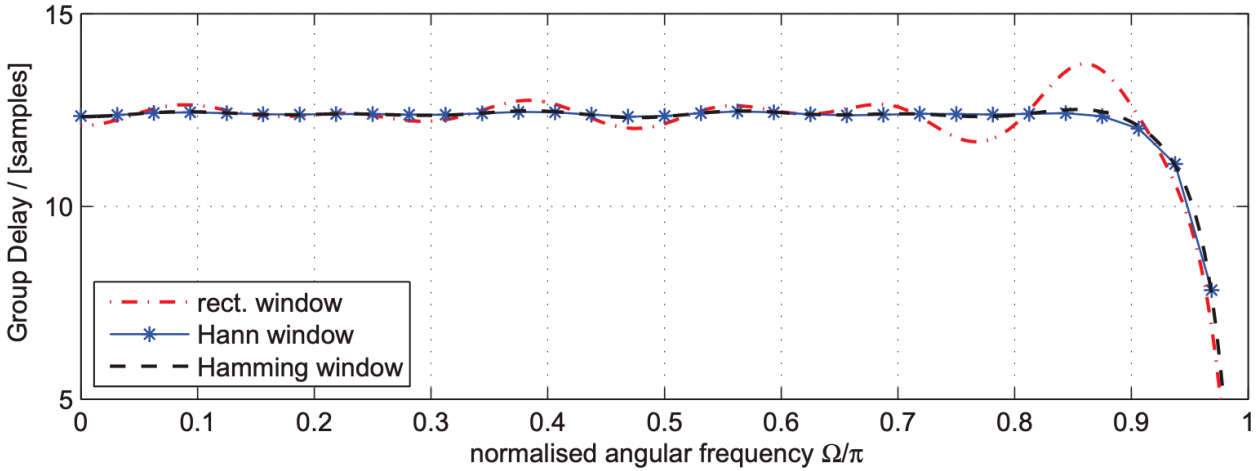
$$\Gamma_\theta(z) = \mathbf{a}_\theta^P(z)\boldsymbol{\mathcal{U}}_v(z)\boldsymbol{\mathcal{U}}_v^P(z)\mathbf{a}_\theta(z), \tag{6}$$

which is used to compute the pseudo-spectrogram for Spatio-Spectral P-MUSIC,

$$\mathbf{P}_{SSP-MU}(\theta,\Omega) = \left.\frac{1}{\Gamma_\theta(z)}\right|_{z=e^{-j\Omega}}, \tag{7}$$

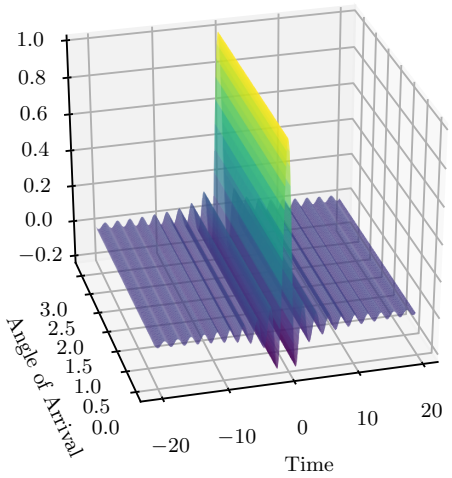
where frequency Ω is obtained by evaluating z on the unit circle.

CONTRIBUTIONS

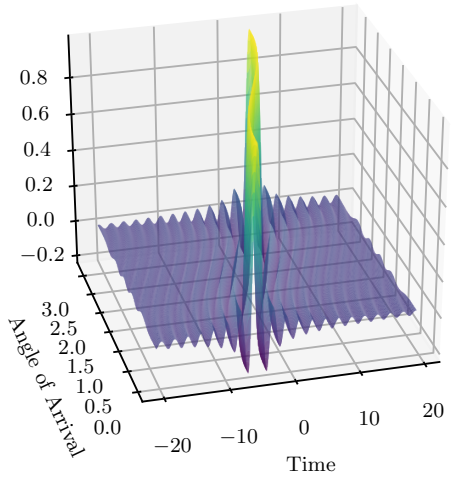


(Group delay response for tapered windows & rectangular window for fractional delay τ = 2.4)

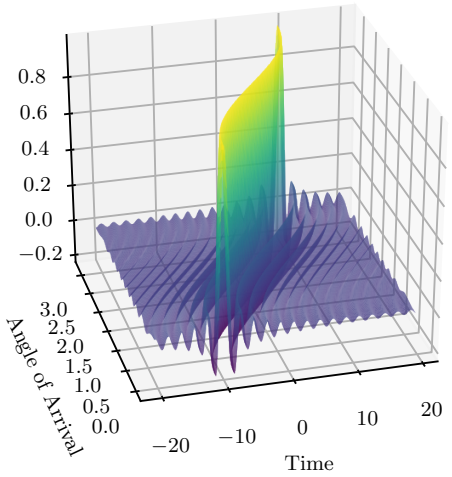
- A broadband steering vector can be realised by sampling a sinc function which leads to an infinite impulse response
- truncation results in inaccuracies towards $\frac{f_s}{2}$ (Alrmah, 2015)
- using a tapered window to create a finite length filter dramatically improves the accuracy (Selva, IEEE TSP 2008)



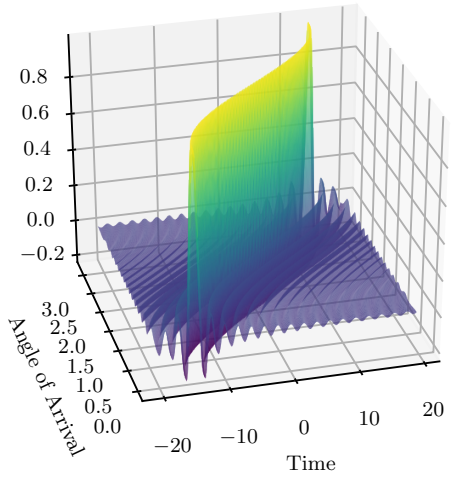
Broadband Steering Vector for m=0



Broadband Steering Vector for m=1



Broadband Steering Vector for m=2

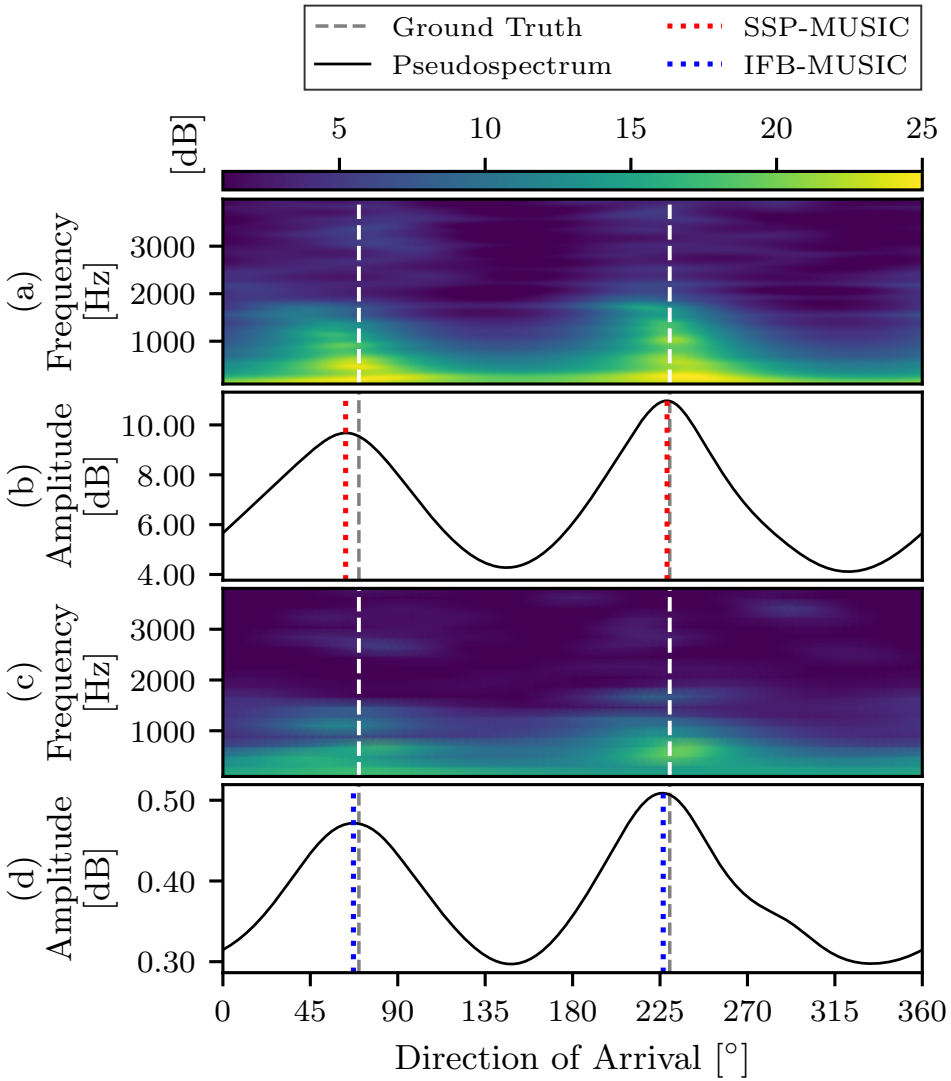


Broadband Steering Vector for m=3

We modify SSP-MUSIC to only include the direct-path to reduce the impact of reverberation on the direction of arrival (DoA) estimation.

If microphone array geometry is known, the largest direct-path delay can inform the choice of the temporal lag support, W.

$$\tilde{\boldsymbol{\mathcal{R}}}_{\mathbf{xx}}(z) \approx \sum_{\tau=-W}^{W} \mathbf{R}_{\mathbf{xx}}(\tau) z^{-\tau}, \tag{8}$$



Illustrative example of a 100 ms frame from Exp-2 (SNR: -10 dB, T60: 0.25 s).

(a) pseudo-spectrogram of SSP-MUSIC,

(b) pseudo-spectrum of SSP-MUSIC,

(c) pseudo-spectrogram of IFB-MUSIC,

(d) pseudo-spectrum of IFB-MUSIC.

RESULTS

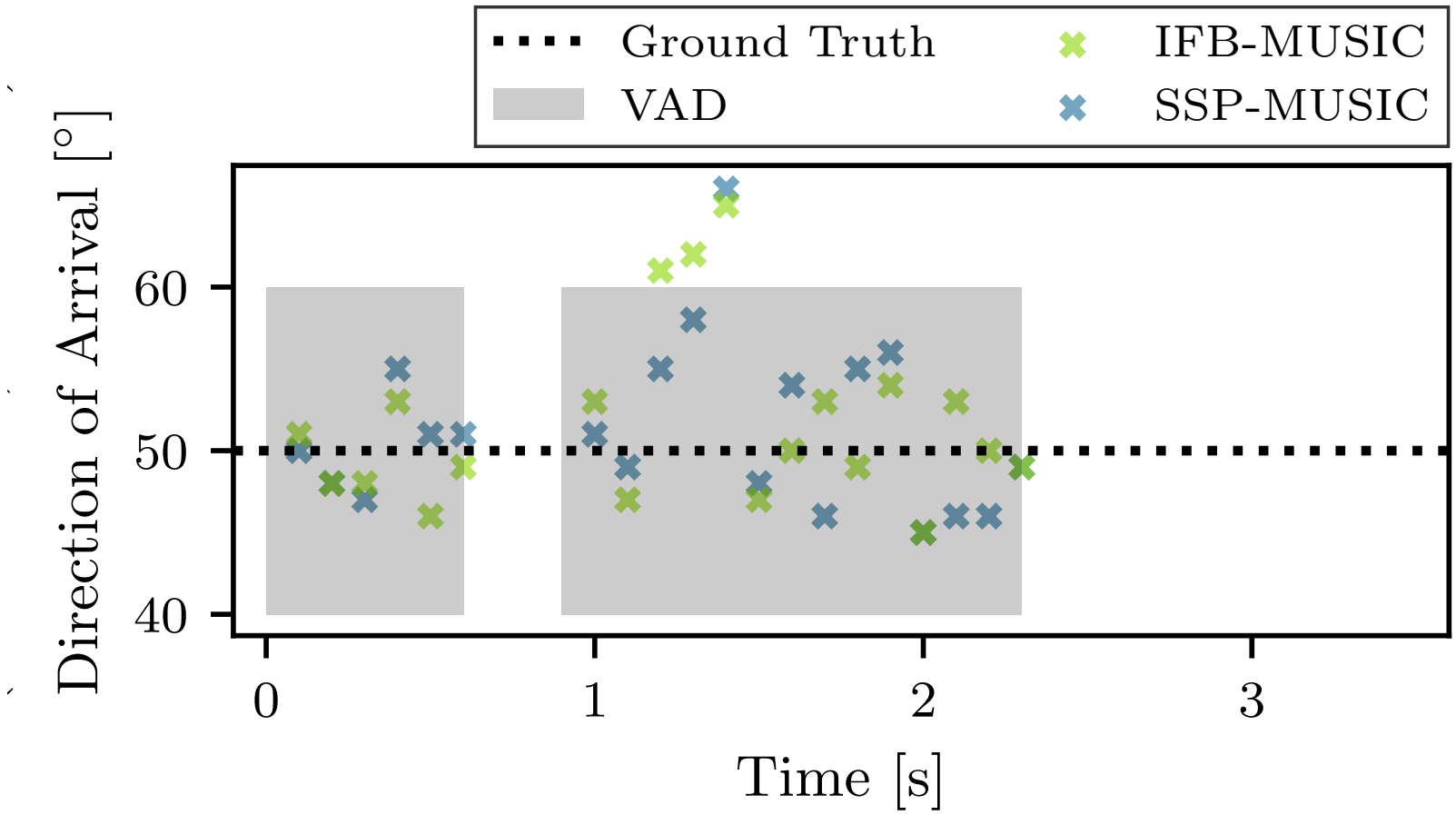
A '**HIT**' is when a sound source (speaker) has been detected once within a $\pm 15^{\circ}$ collar applied around the ground-truth azimuth.

A '**MISS**' is when a speaker change has not been detected within this collar and a FA is when a detection falls outside of a ground-truth azimuth collar.

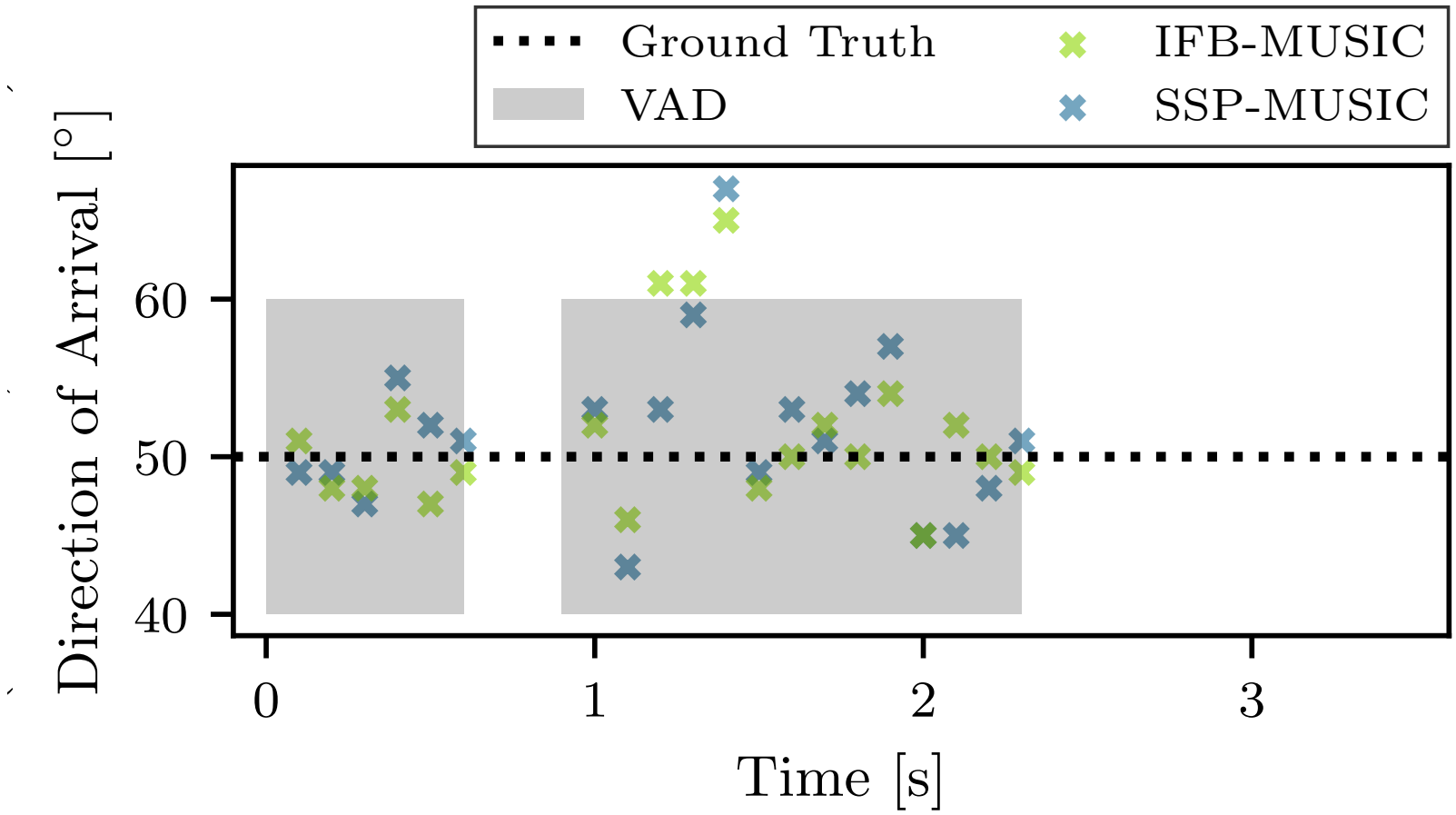
The **Hit Rate** and **False Alarm Rate** are, therefore, defined as,

$$\text{Hit Rate} = \frac{\text{HITs}}{\text{HITs} + \text{MISSs}}\%, \quad \text{False Alarm Rate} = \frac{\text{FAs}}{\text{HITs} + \text{FAs}}\% \tag{9}$$

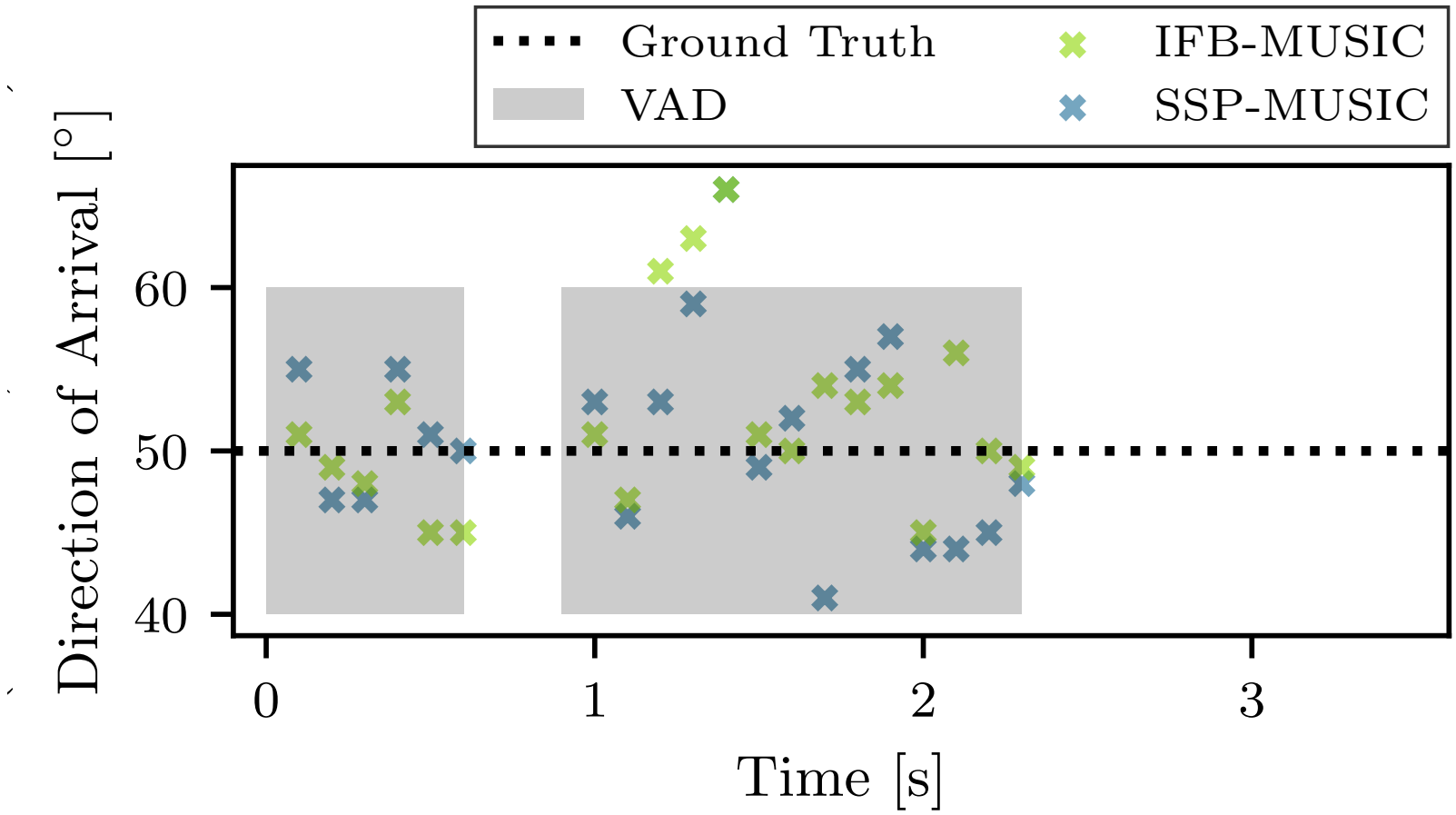
SNR: 25 dB, T60: 0.25 s



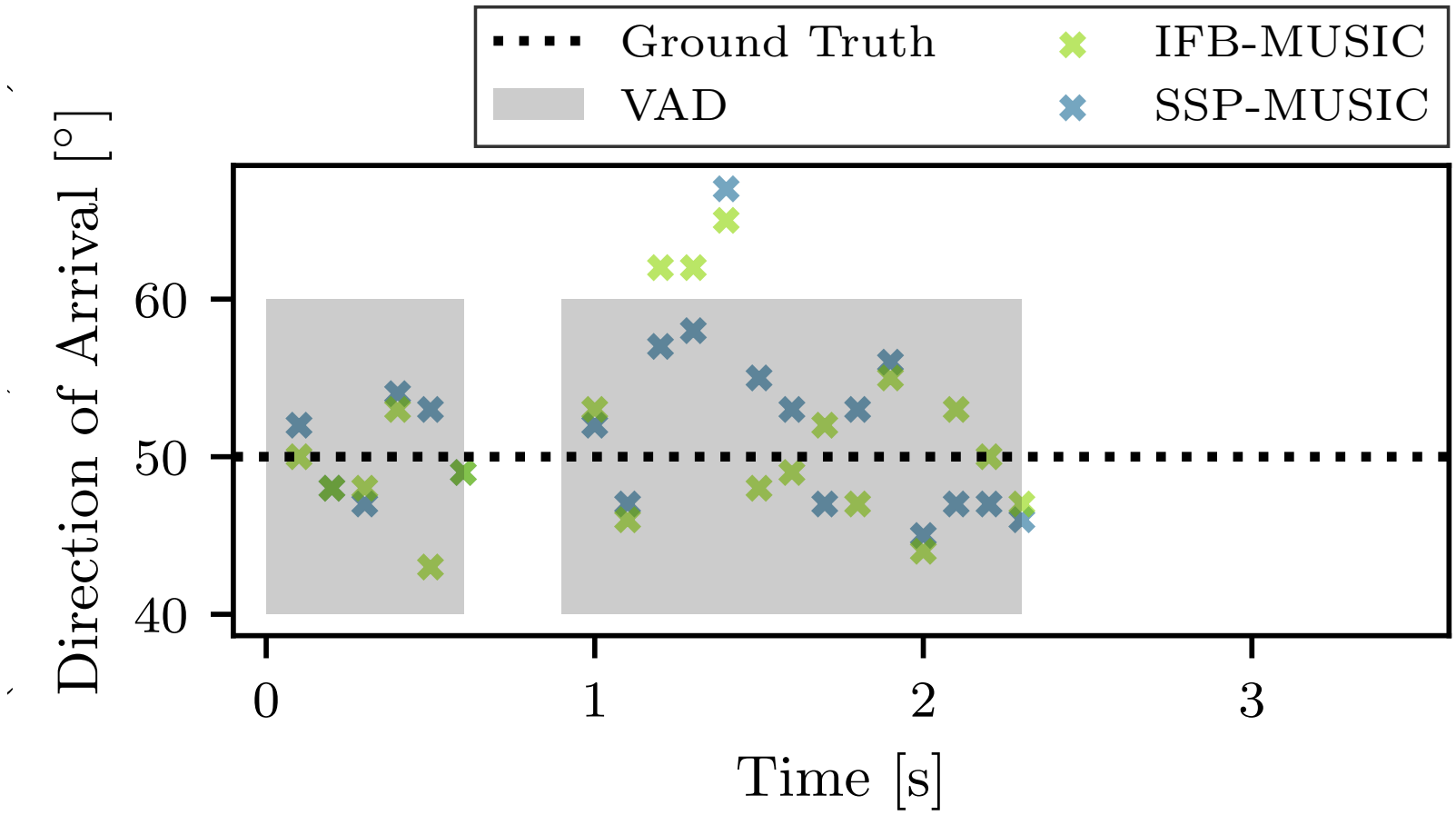
SNR: 20 dB, T60: 0.25 s



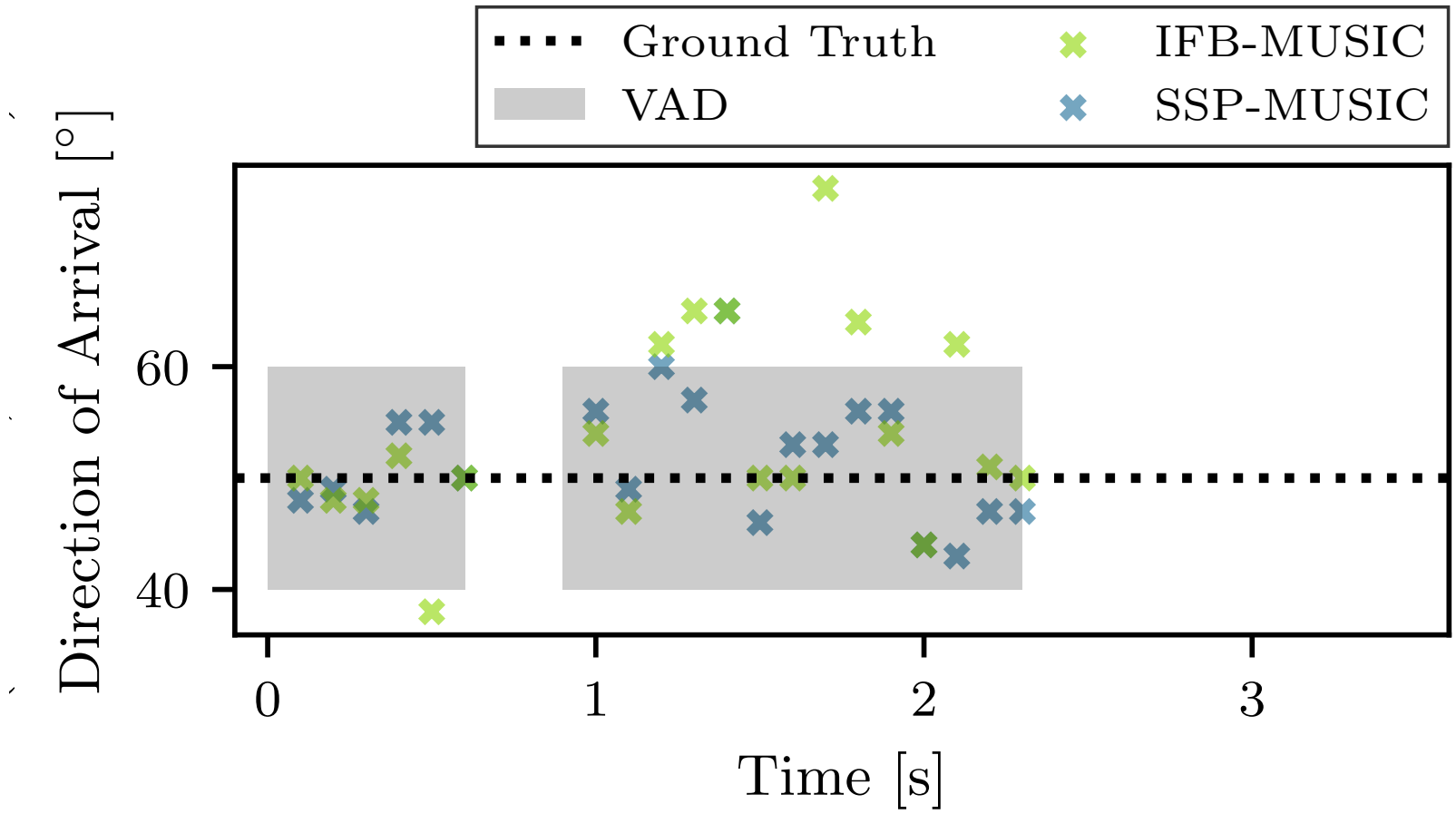
SNR: 15 dB, T60: 0.25 s



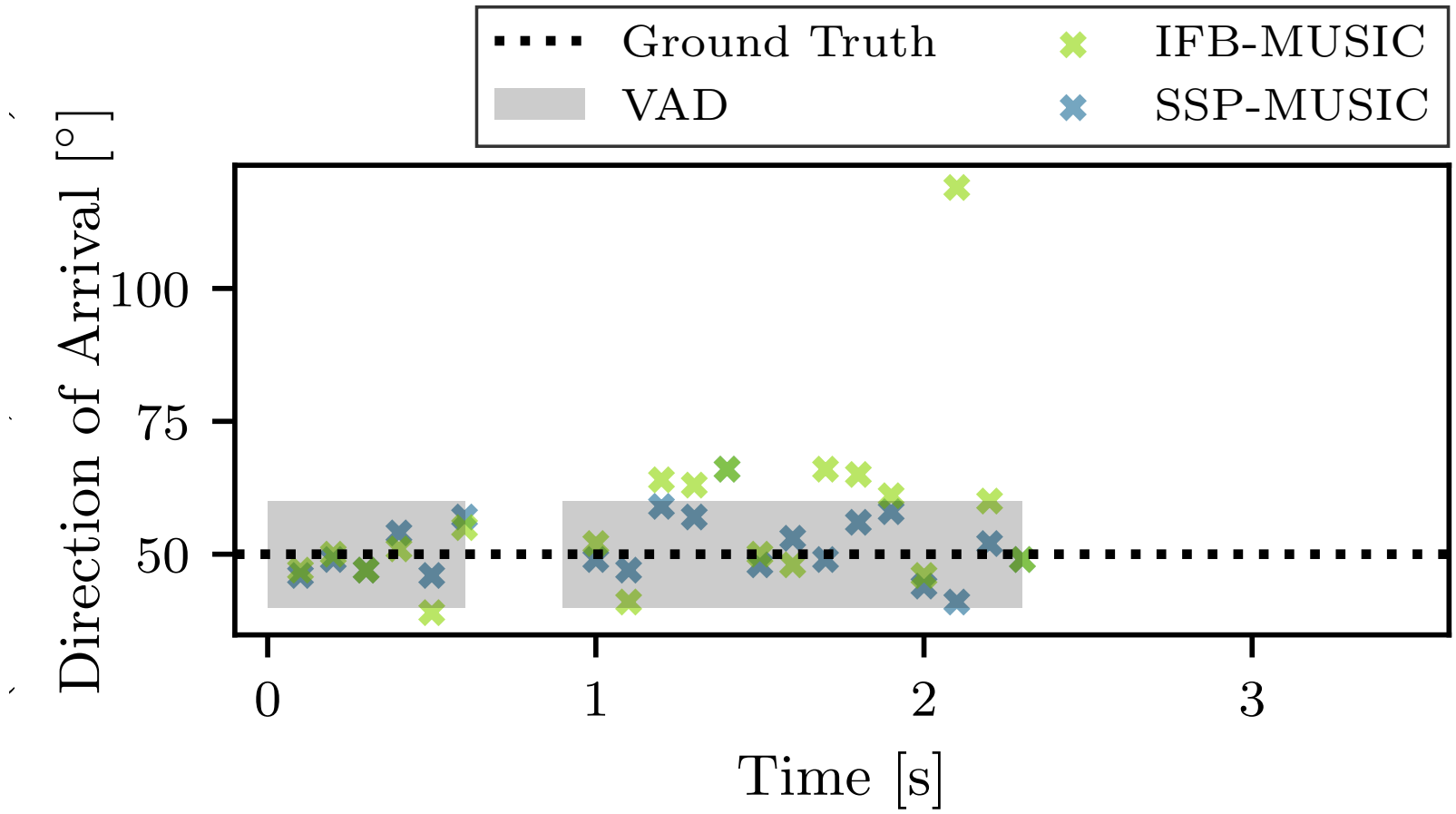
SNR: 10 dB, T60: 0.25 s



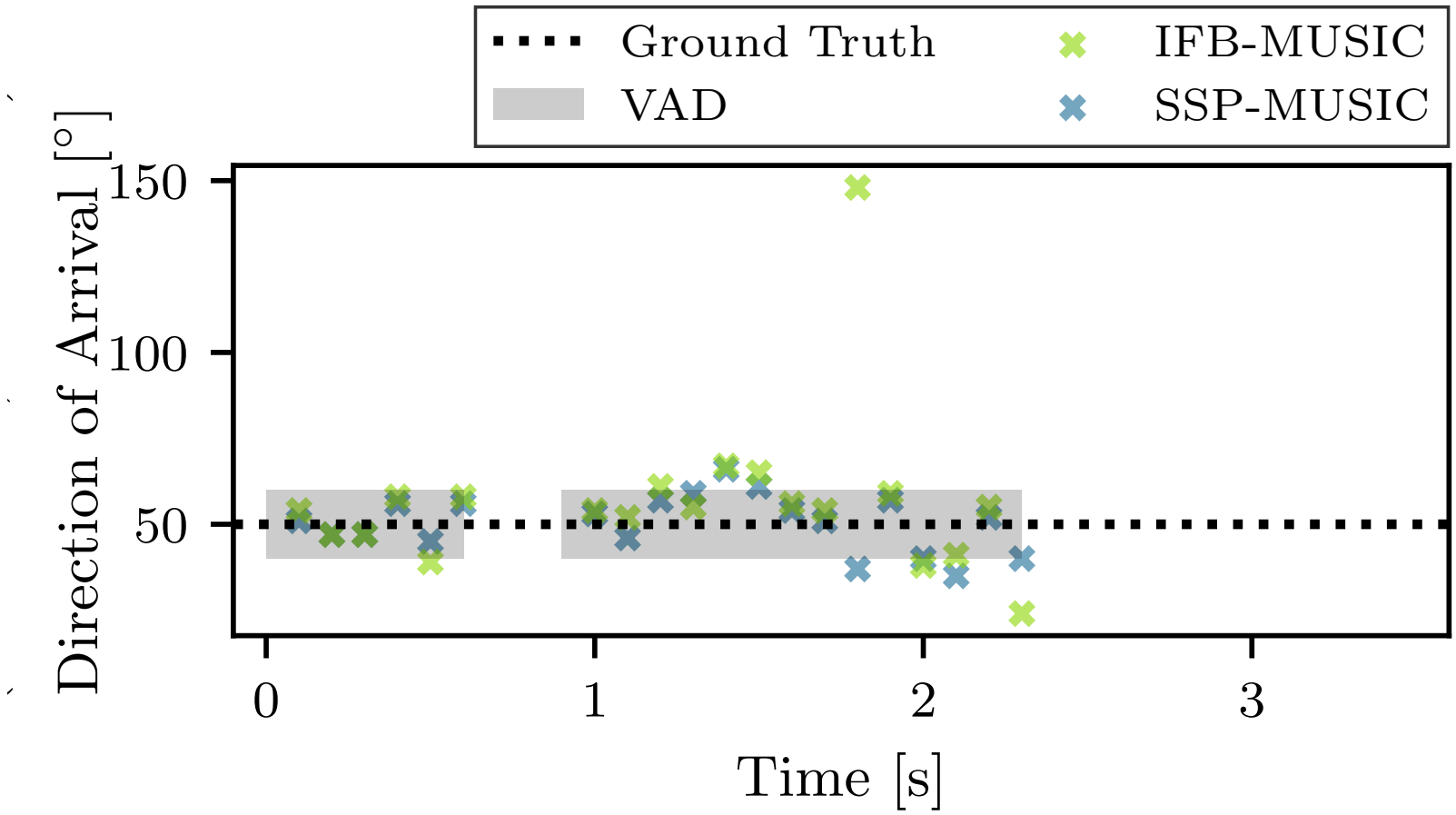
SNR: 5 dB, T60: 0.25 s



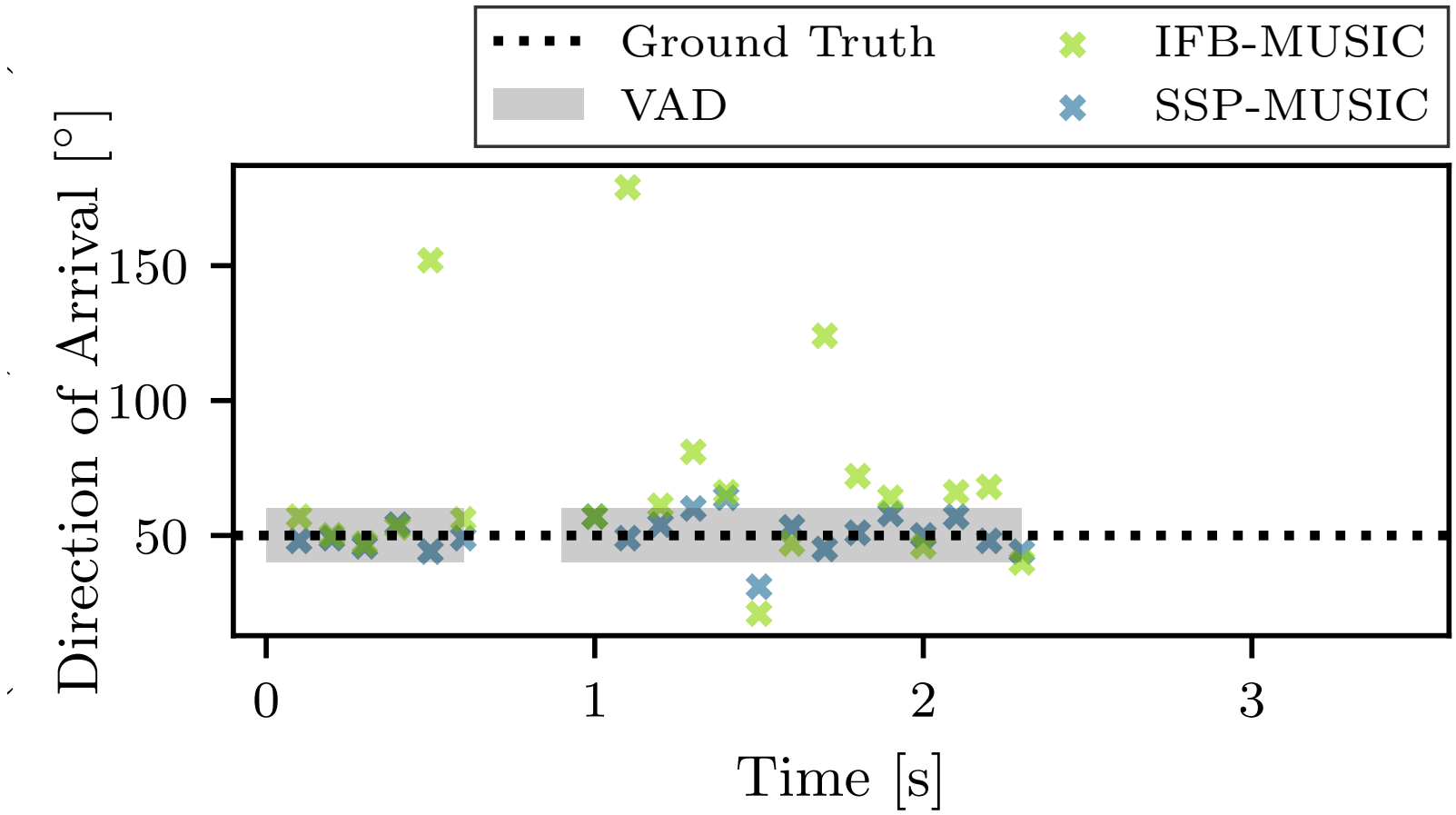
SNR: 0 dB, T60: 0.25 s



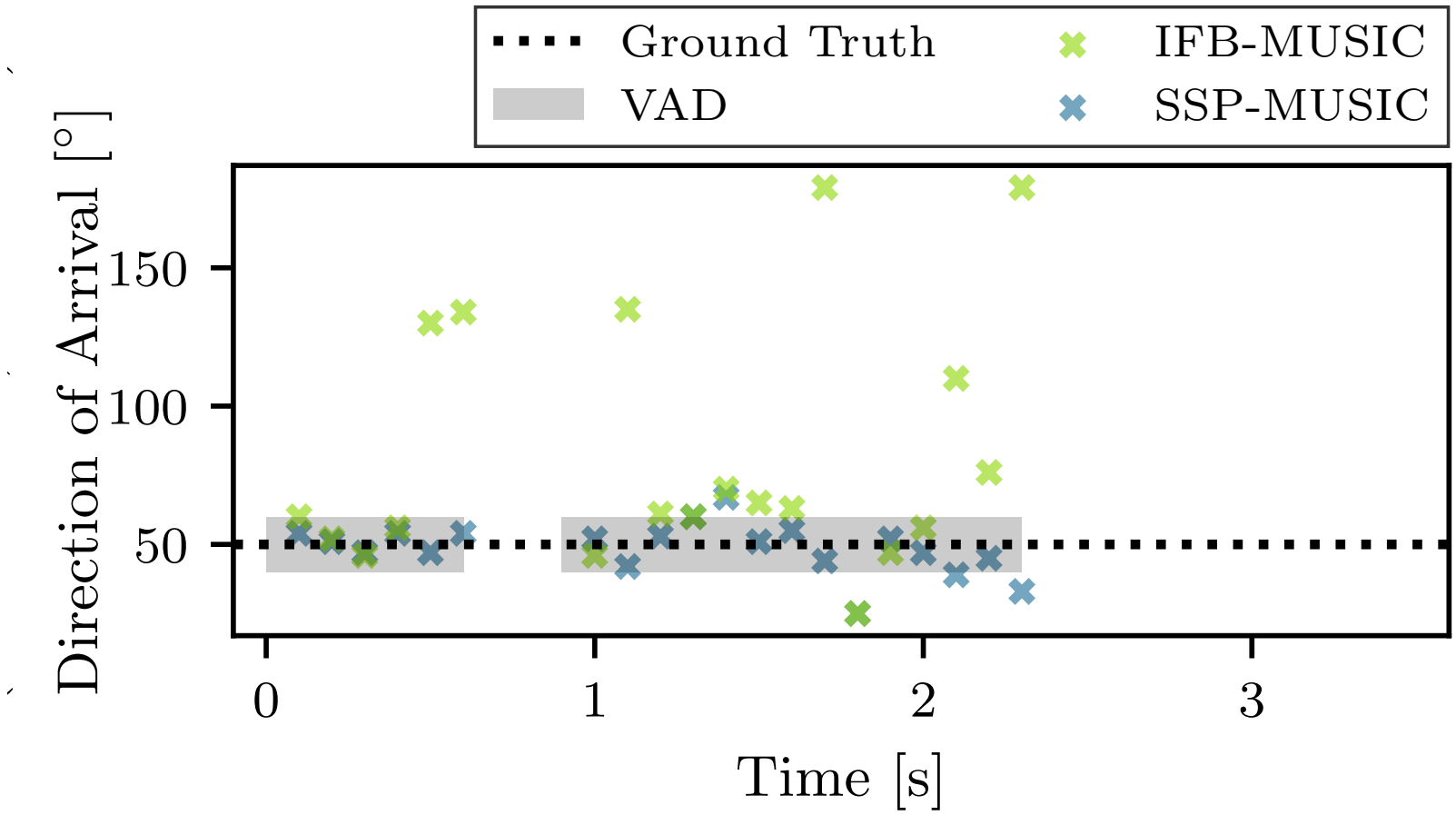
SNR: -5 dB, T60: 0.25 s



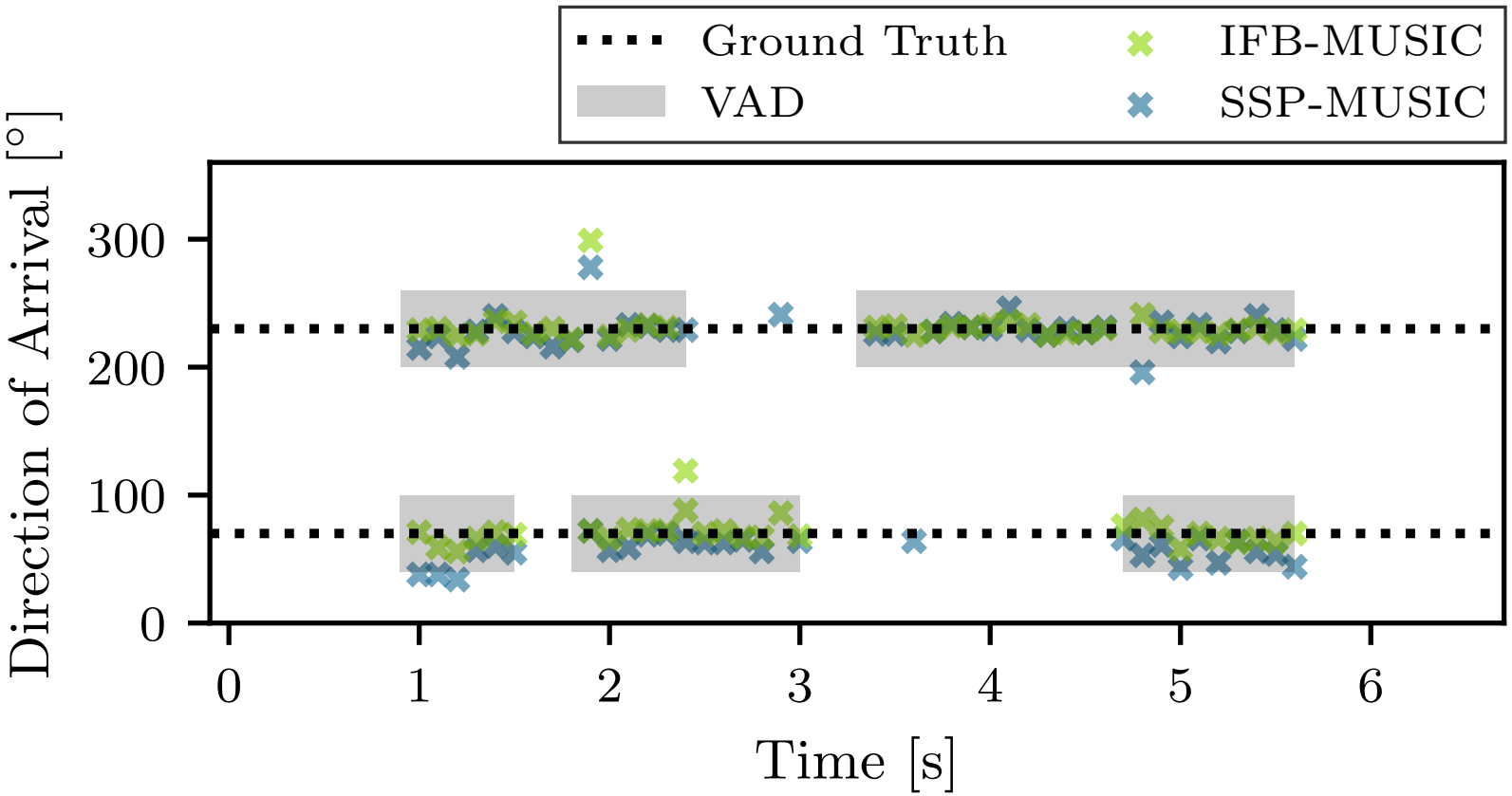
SNR: -10 dB, T60: 0.25 s



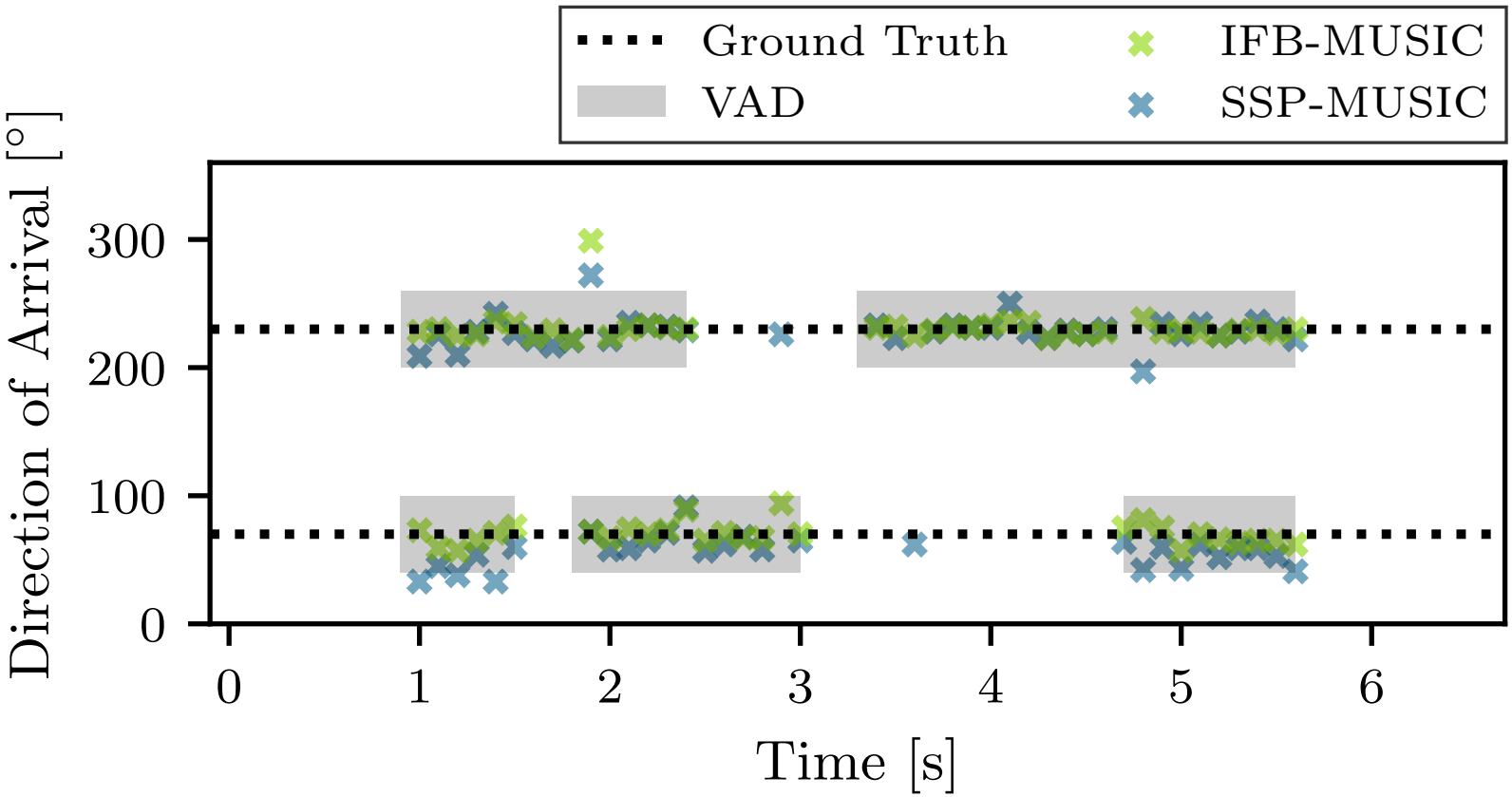
SNR: -15 dB, T60: 0.25 s



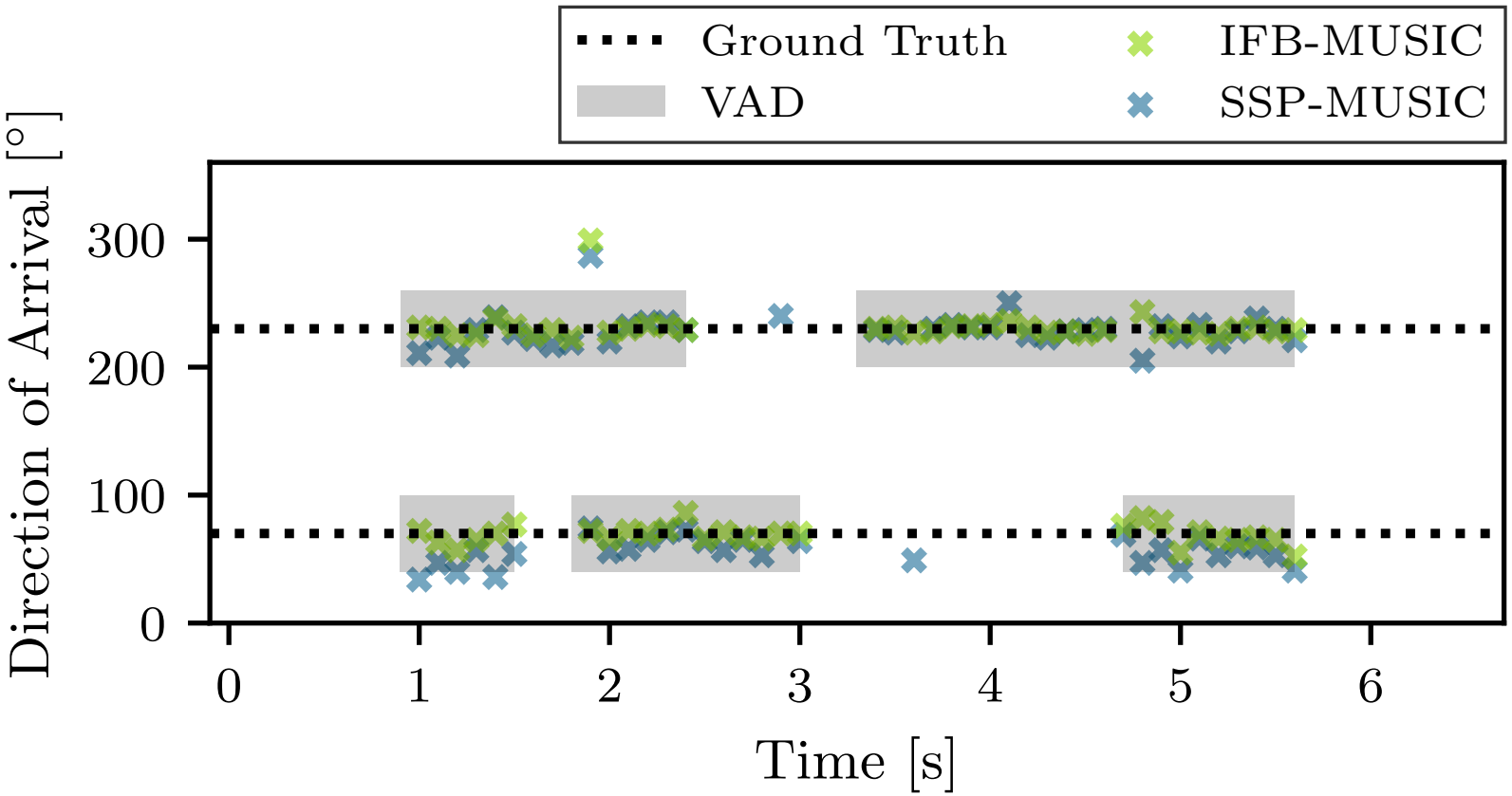
SNR: 25 dB, T60: 0.25 s



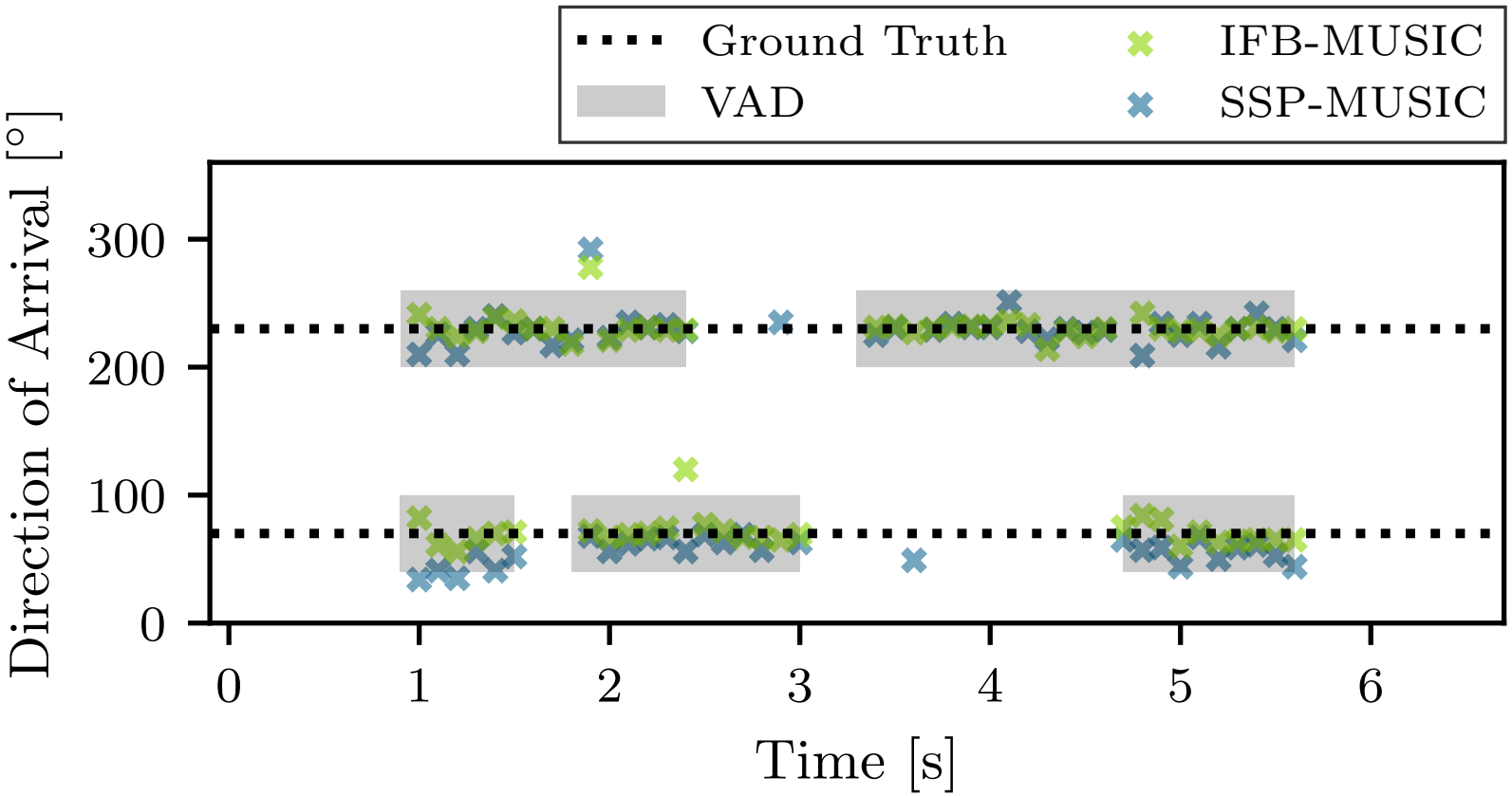
SNR: 20 dB, T60: 0.25 s



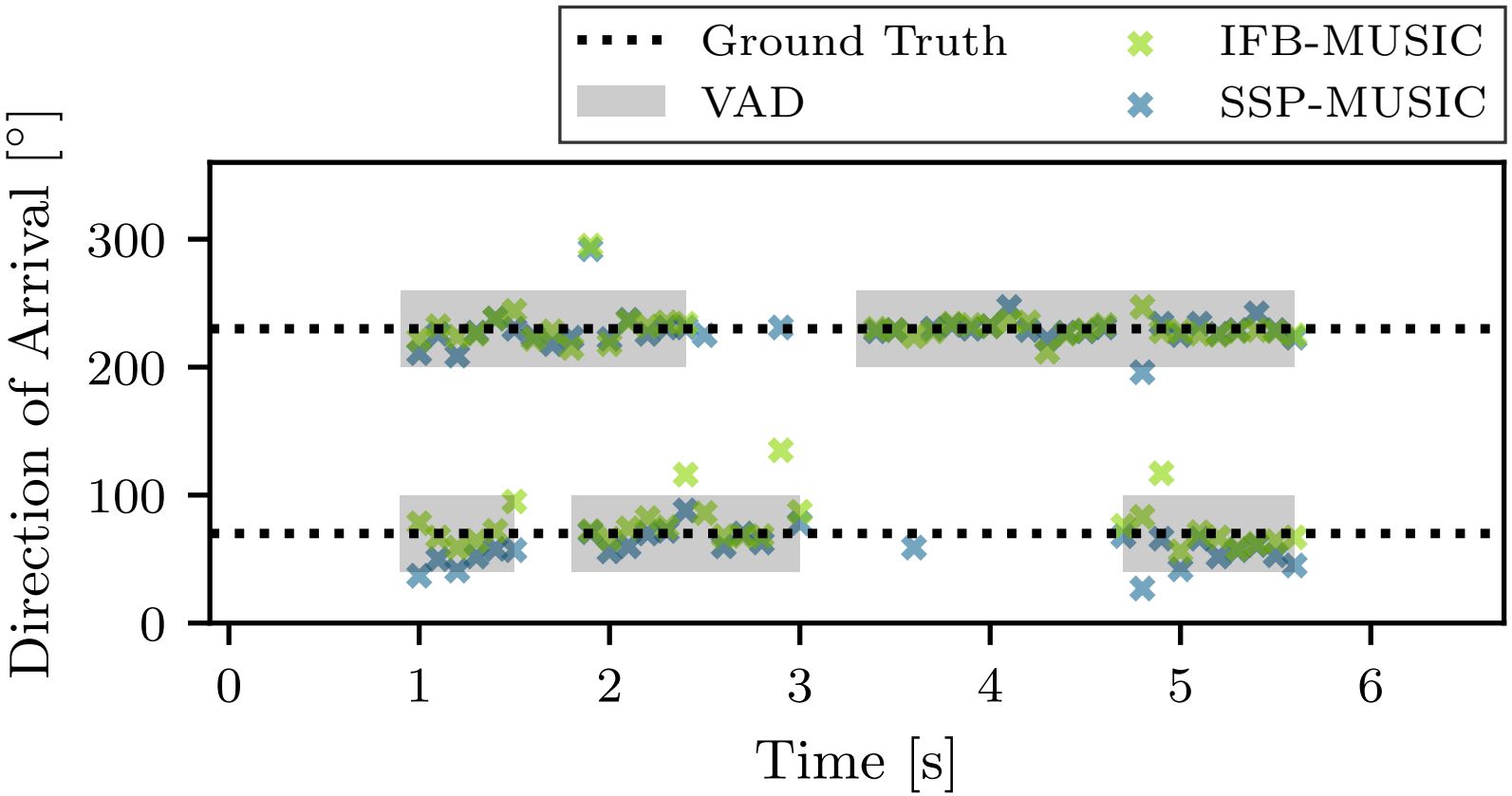
SNR: 15 dB, T60: 0.25 s



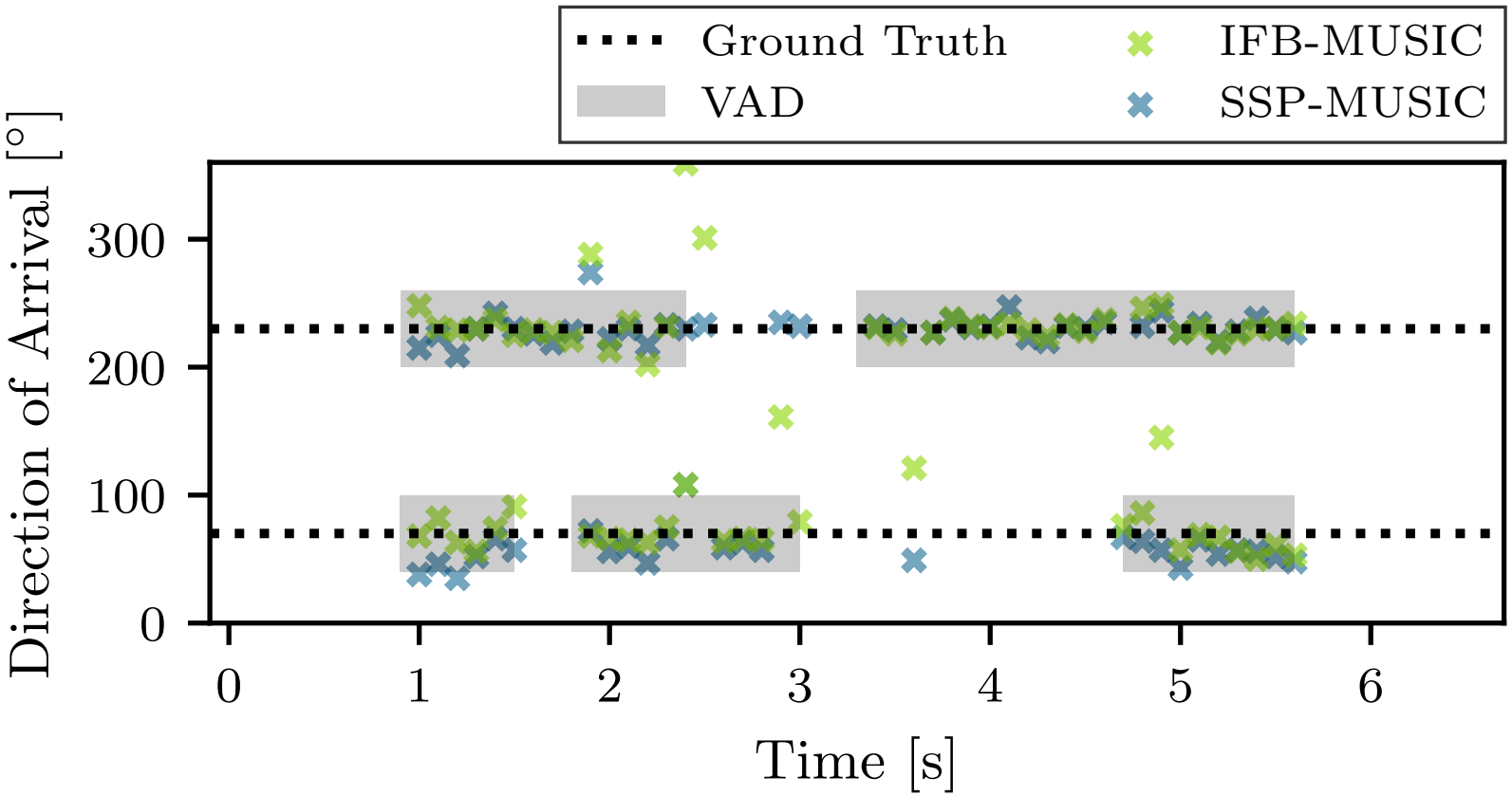
SNR: 10 dB, T60: 0.25 s



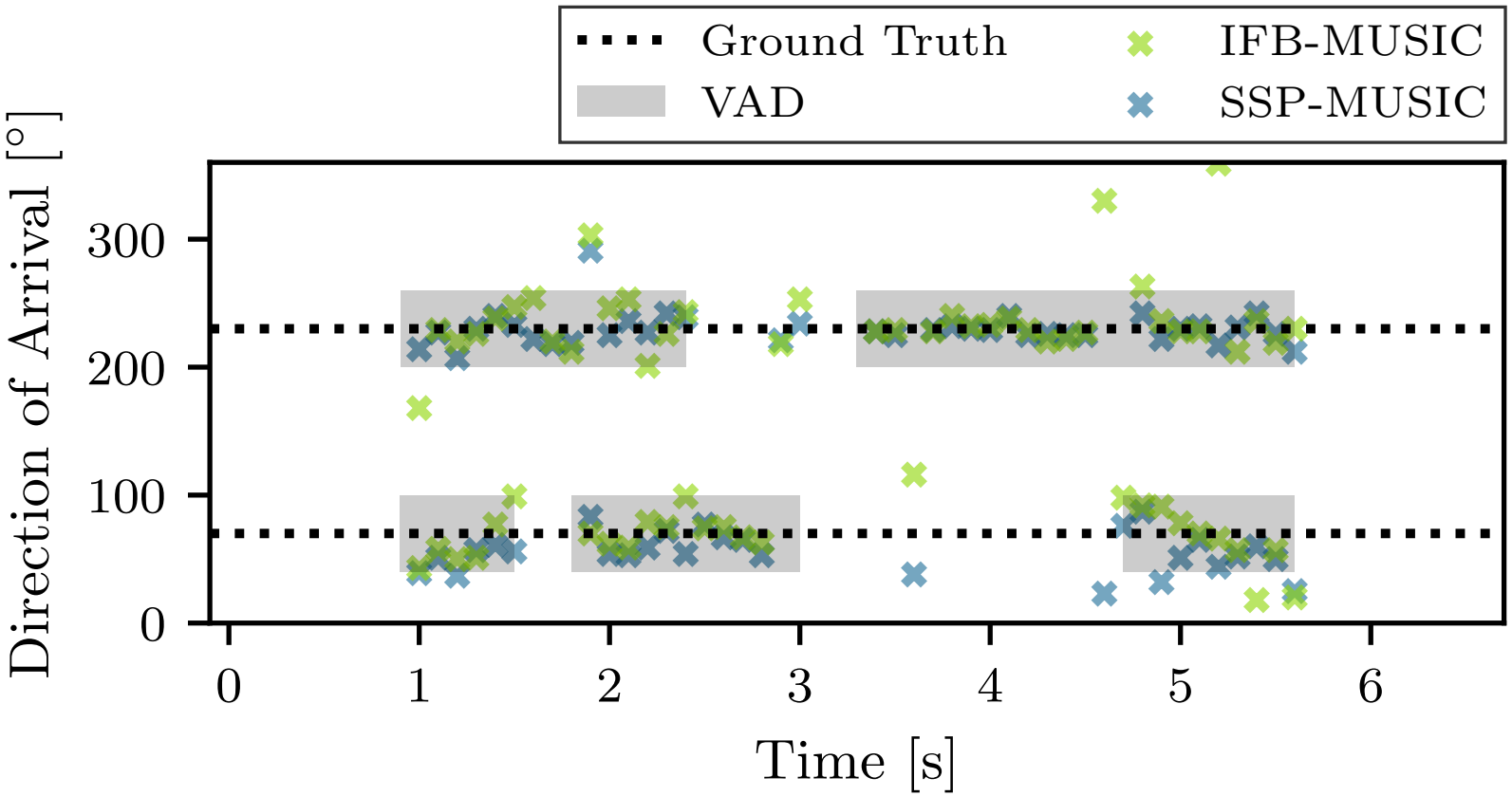
SNR: 5 dB, T60: 0.25 s



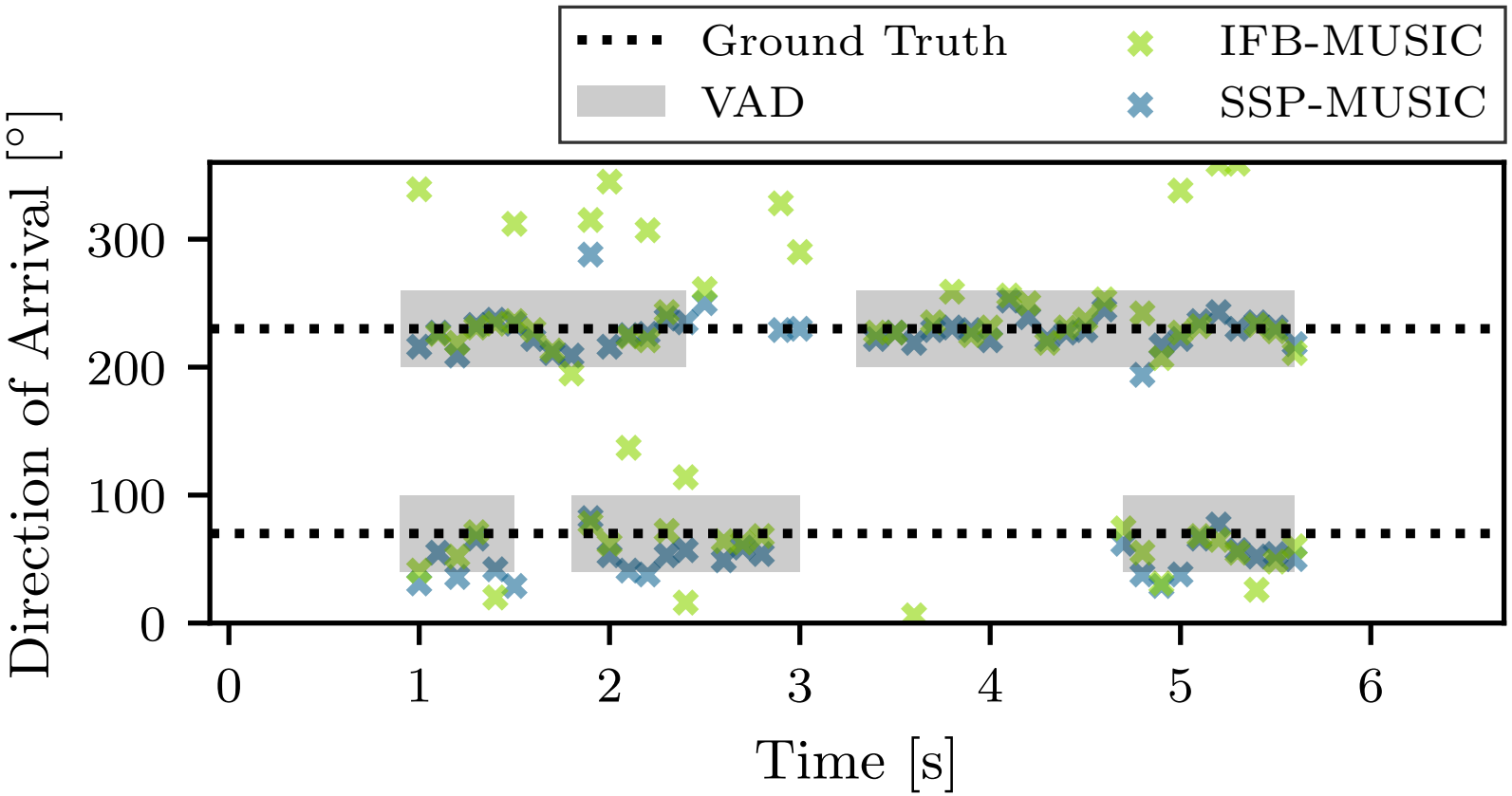
SNR: 0 dB, T60: 0.25 s



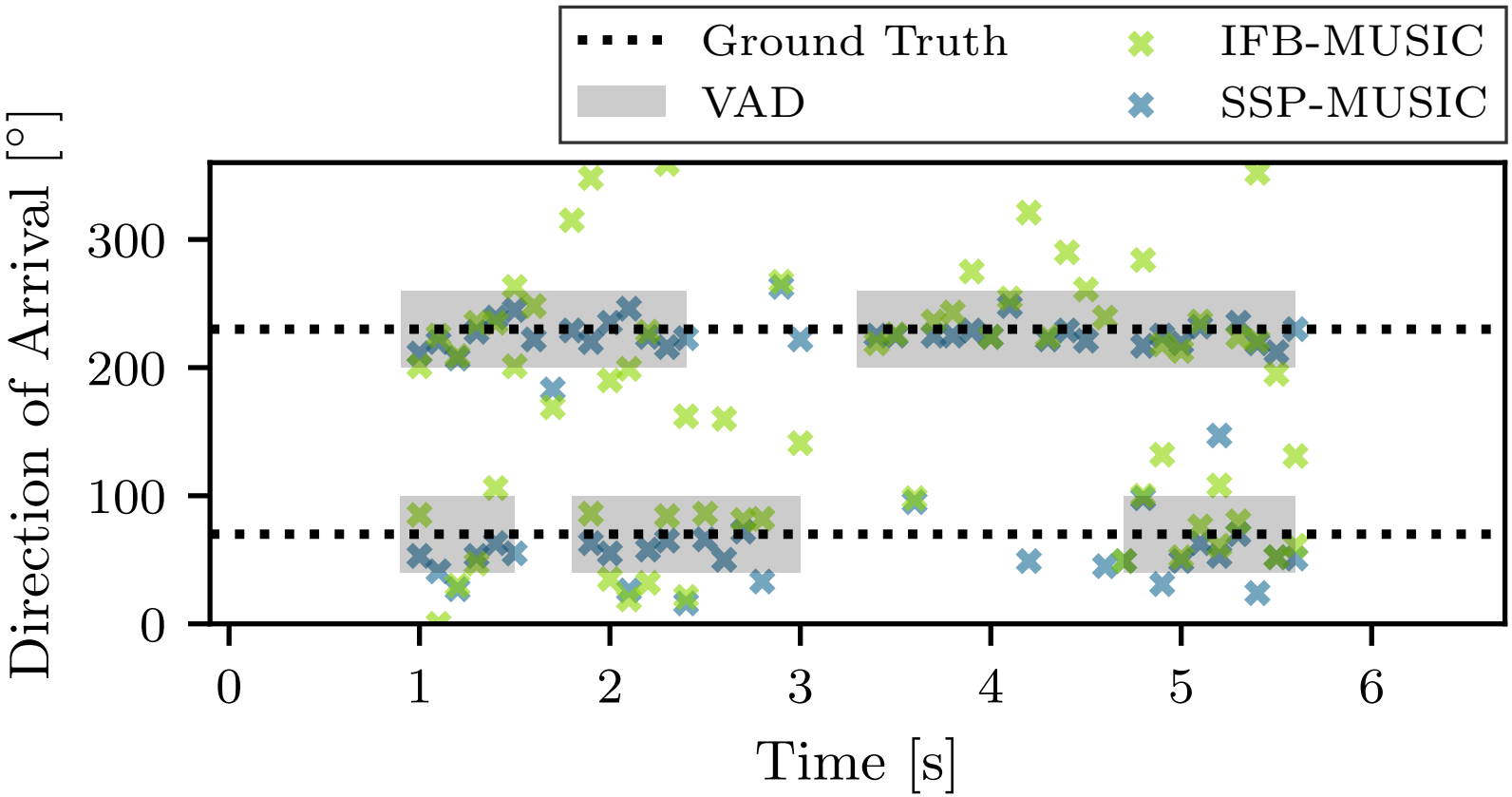
SNR: -5 dB, T60: 0.25 s



SNR: -10 dB, T60: 0.25 s

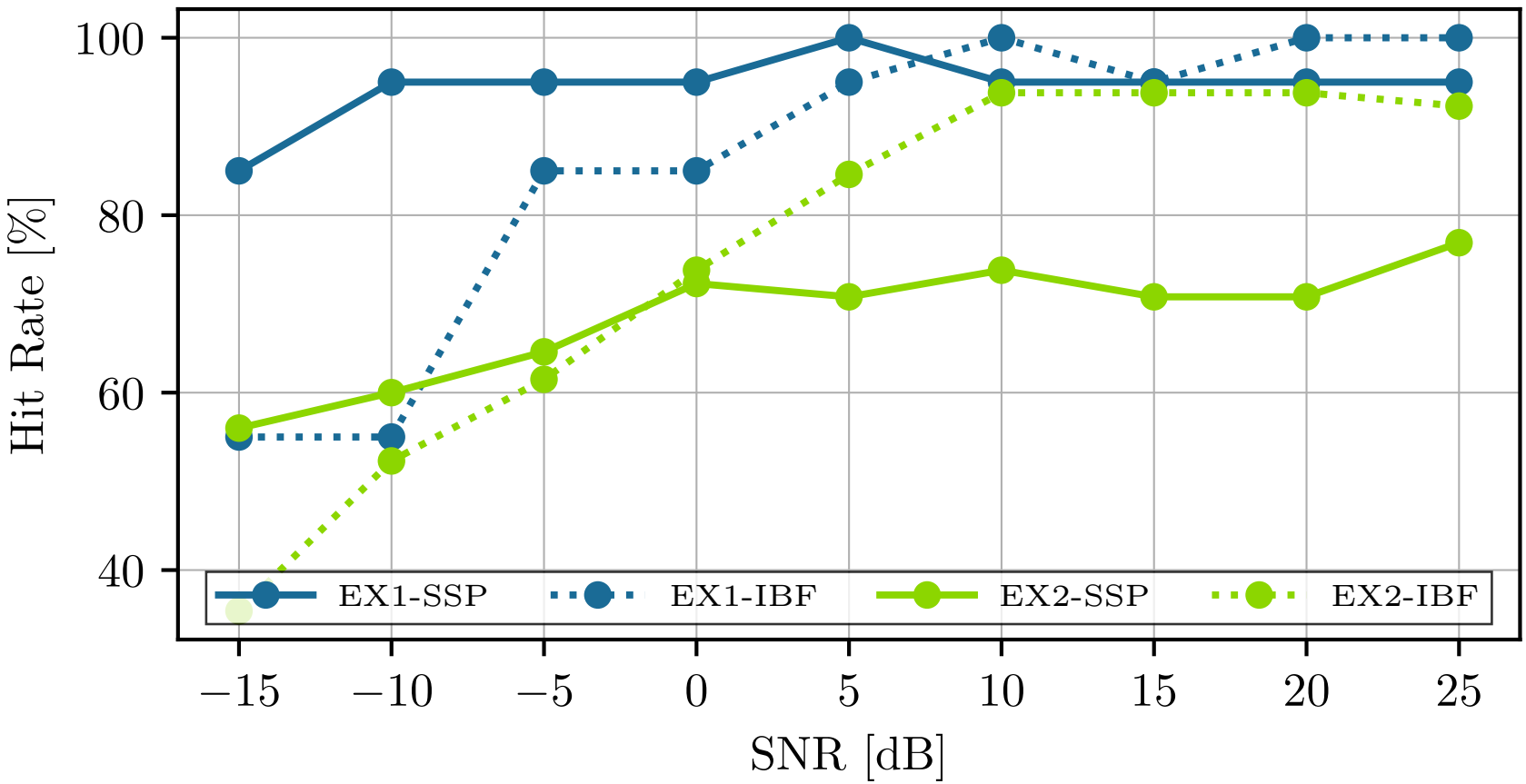


SNR: -15 dB, T60: 0.25 s

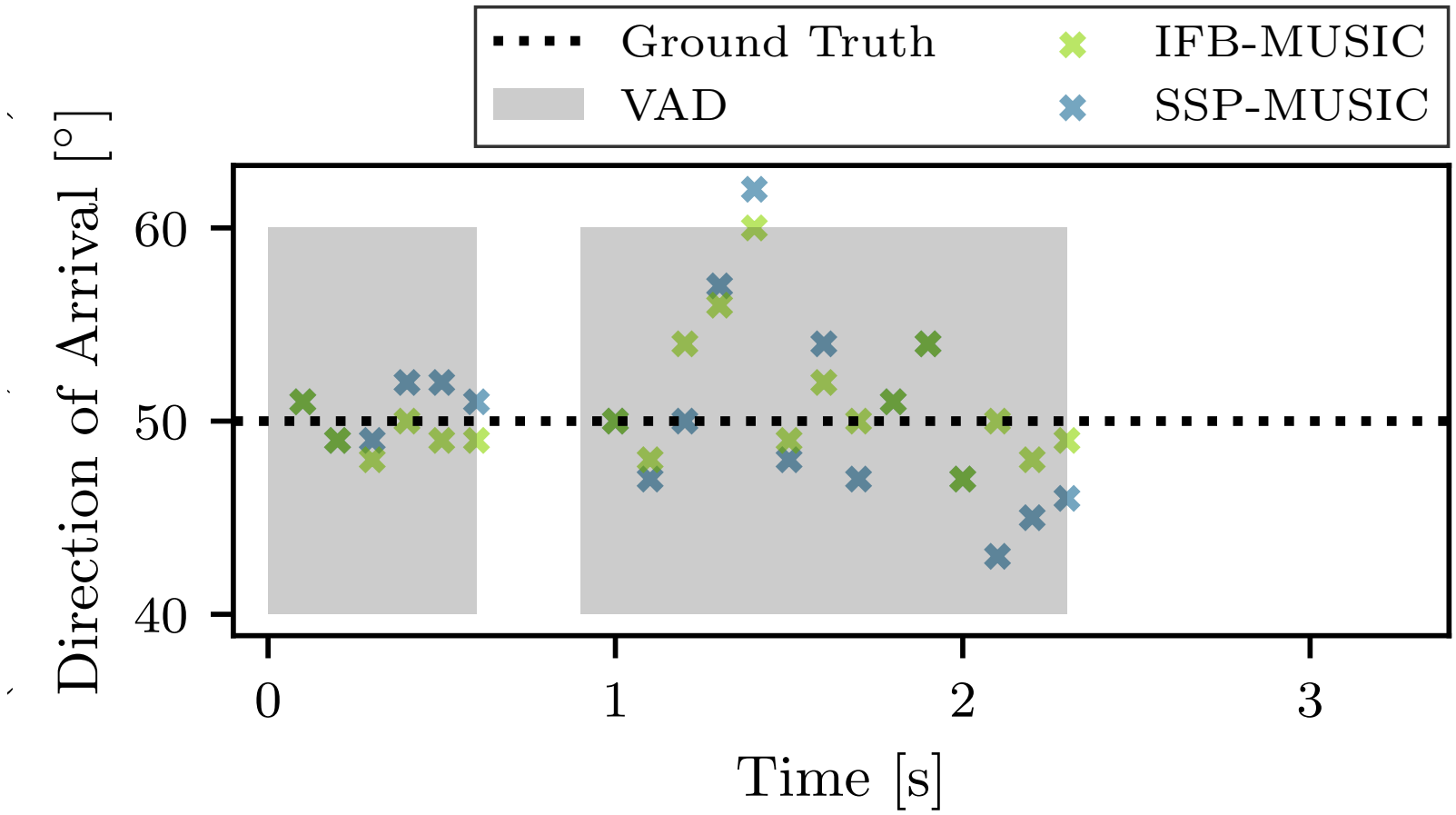


EX1: SINGLE SOUND SOURCE IN A SIMULATED ROOM

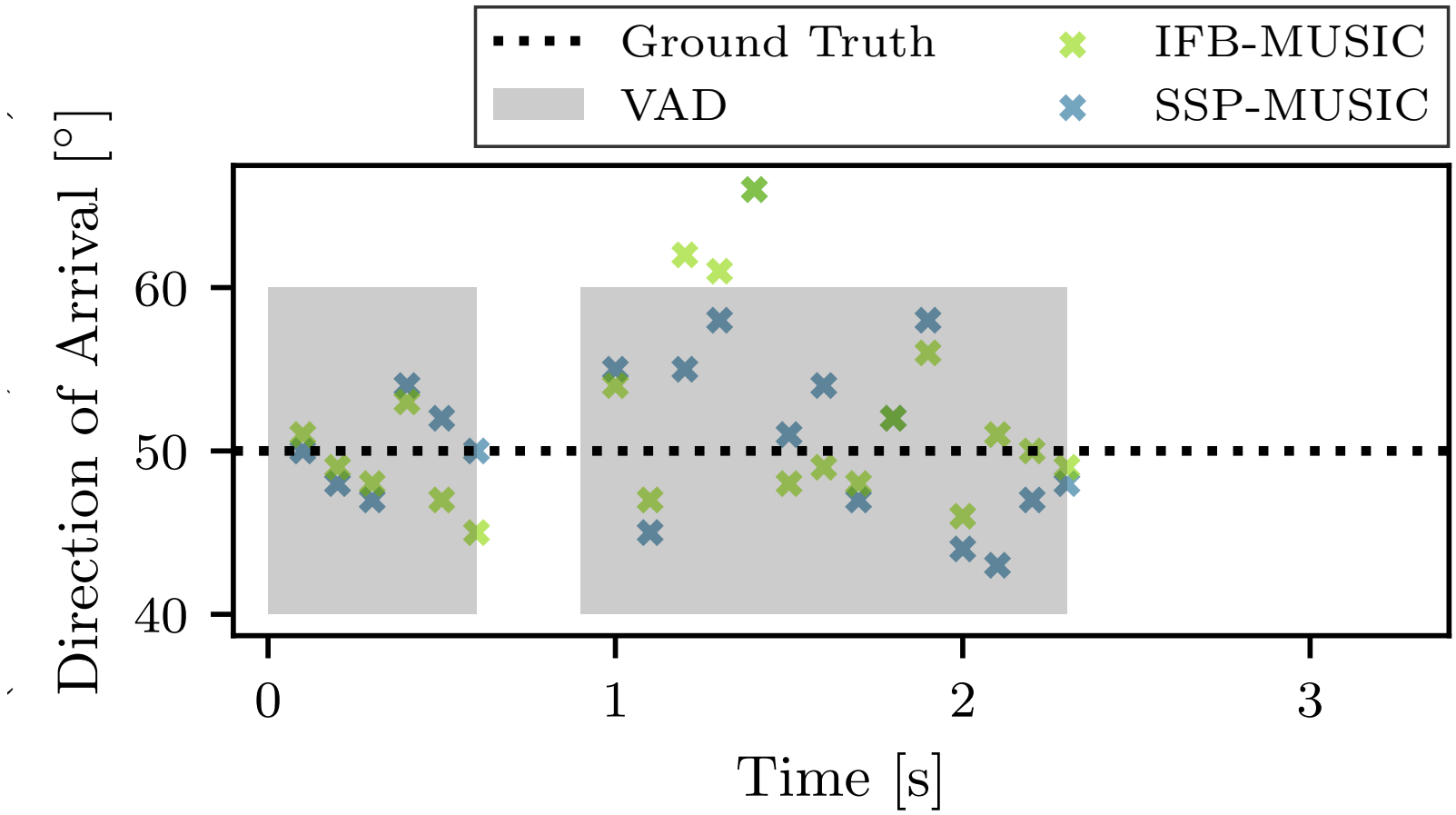
EX2: TWO SOUND SOURCES IN A SIMULATED ROOM



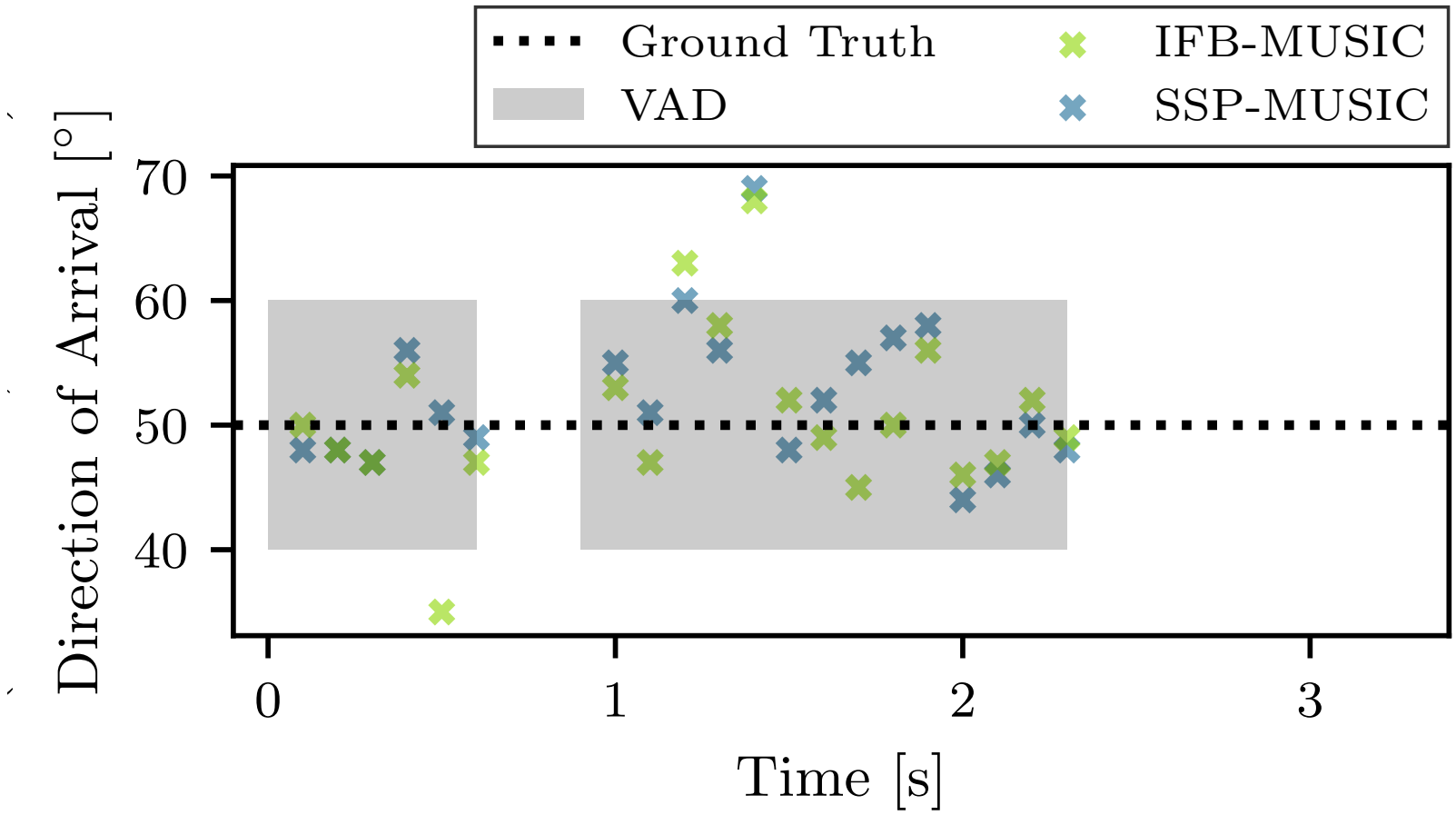
SNR: 15 dB, T60: 0.1 s



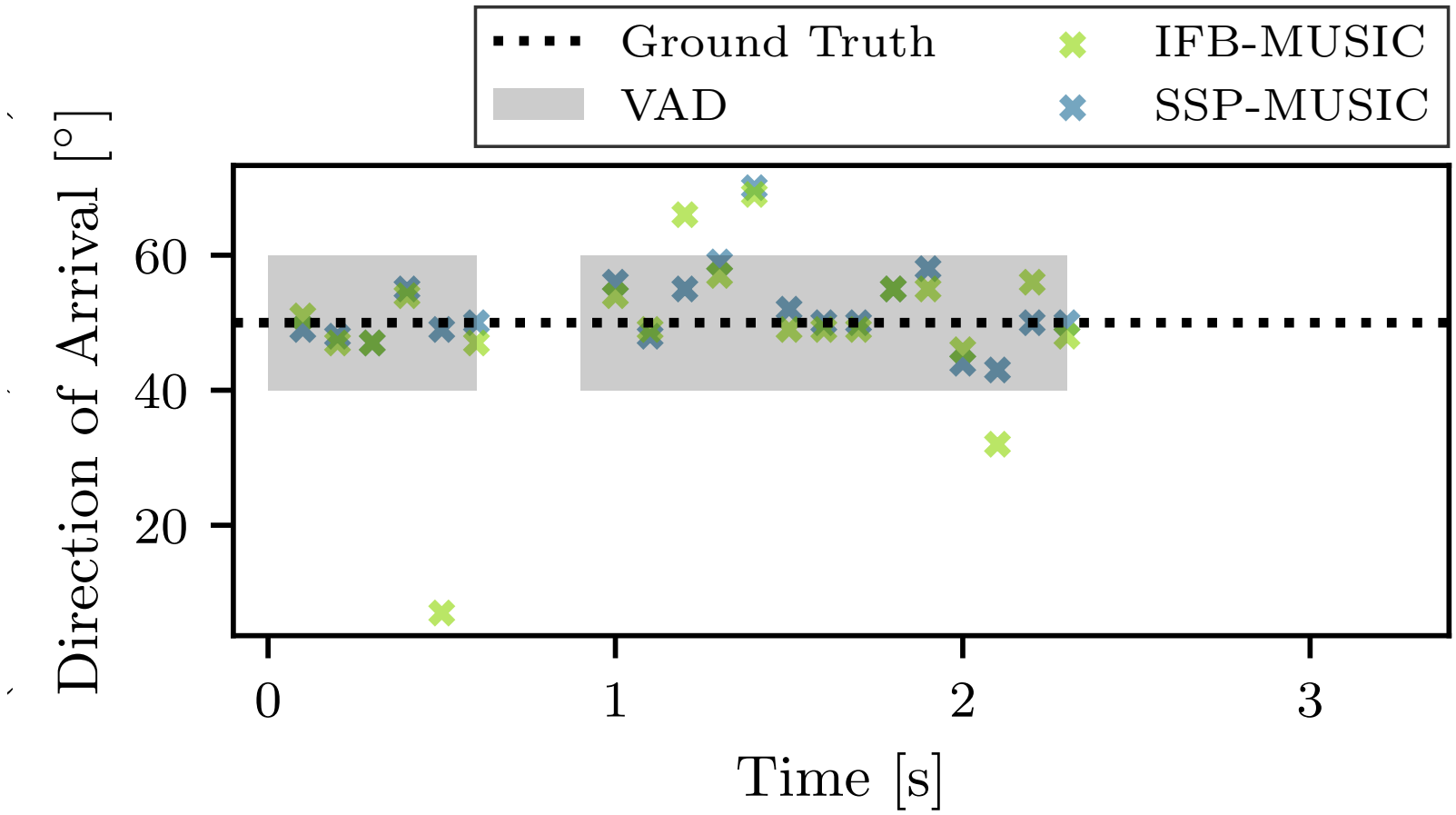
SNR: 15 dB, T60: 0.3 s



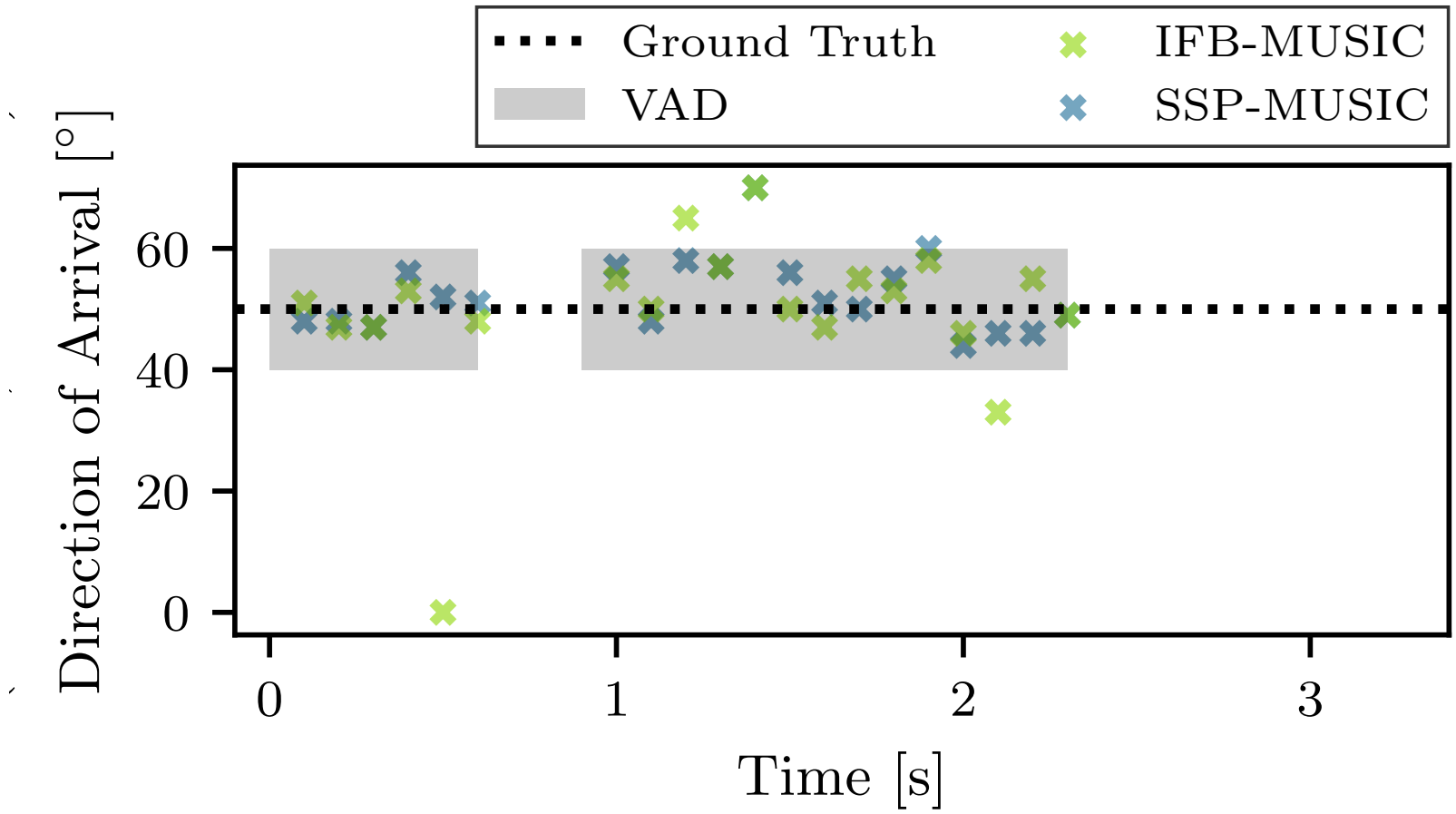
SNR: 15 dB, T60: 0.5 s



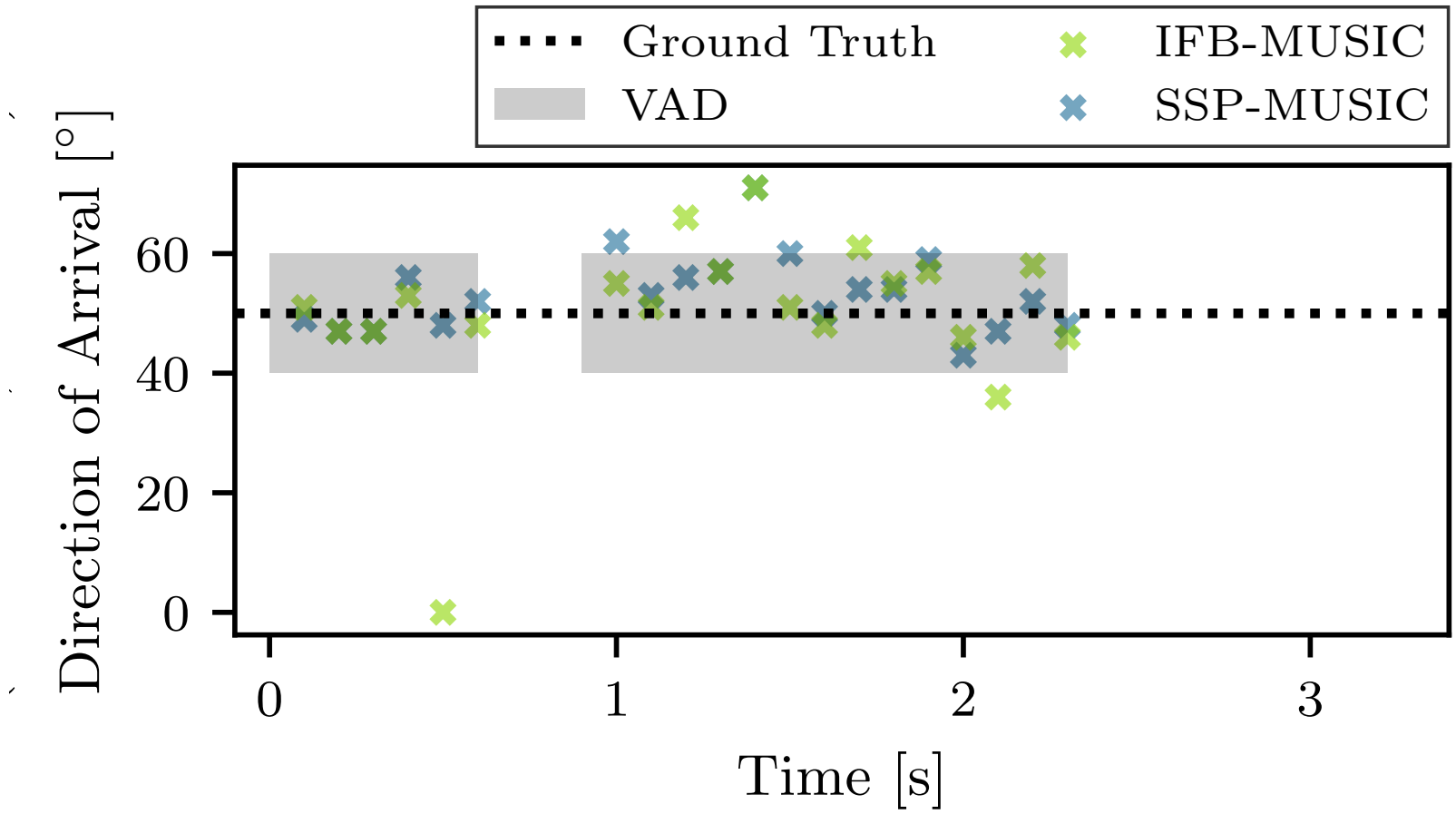
SNR: 15 dB, T60: 0.7 s



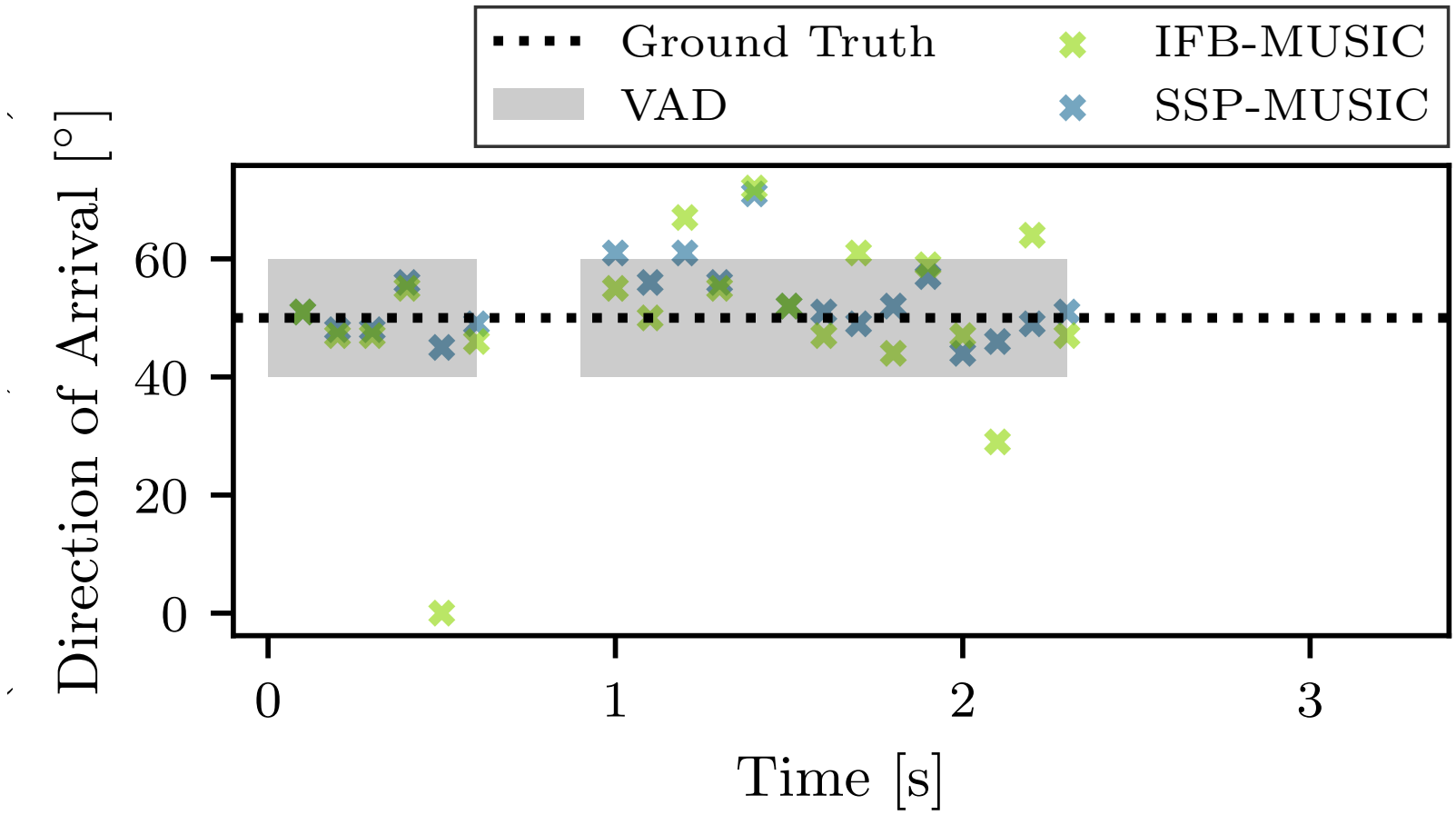
SNR: 15 dB, T60: 0.9 s



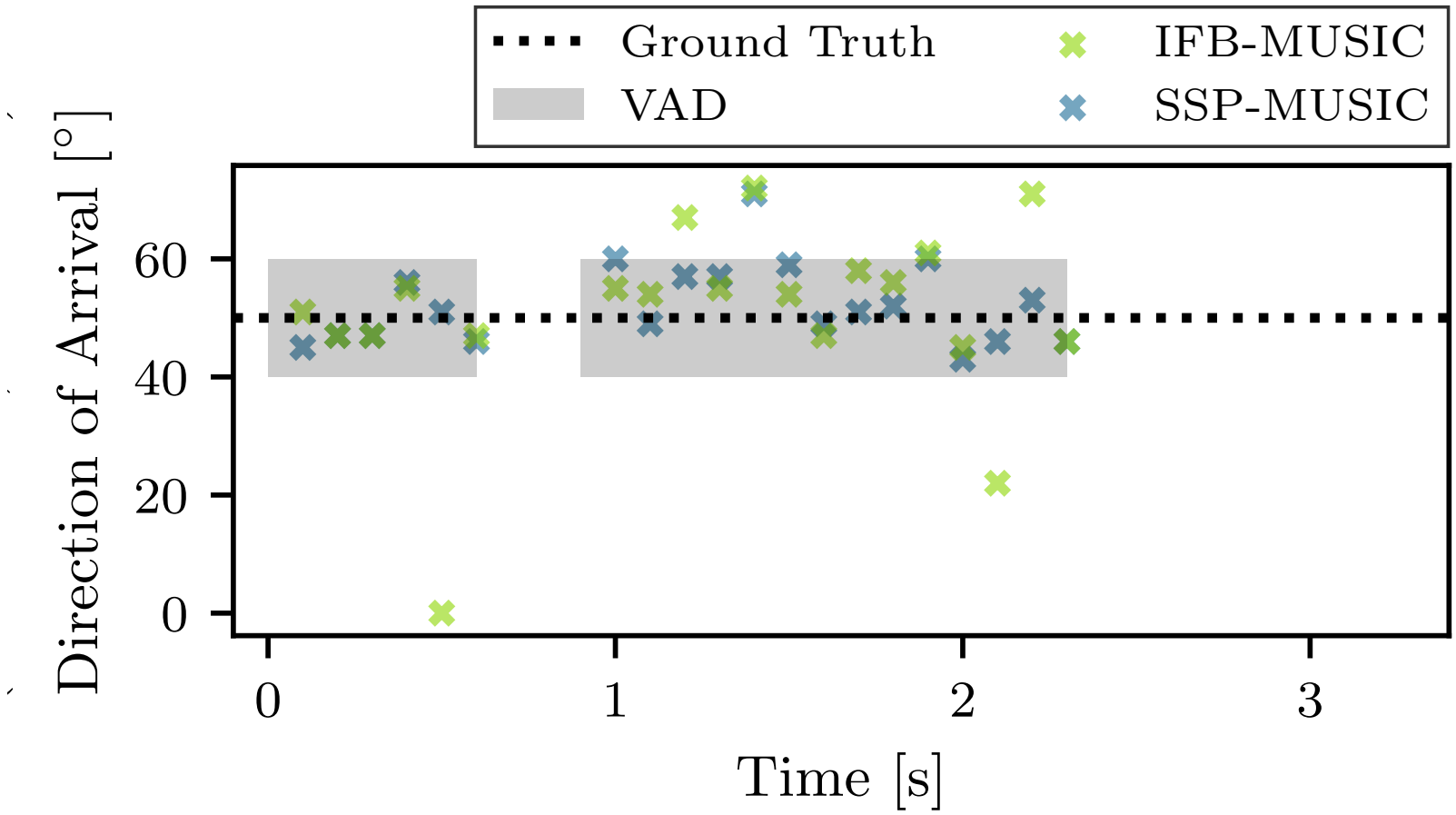
SNR: 15 dB, T60: 1.1 s



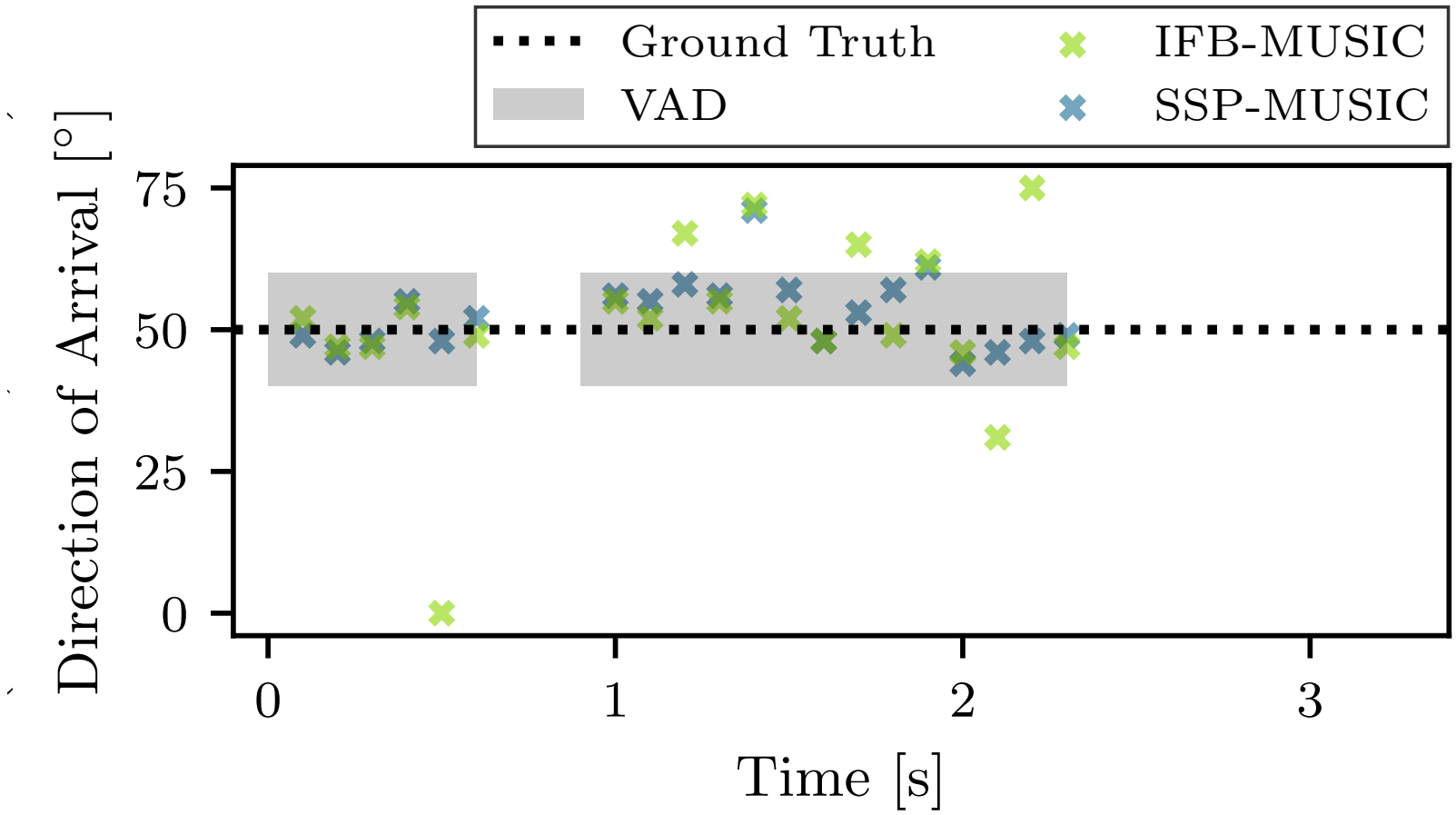
SNR: 15 dB, T60: 1.3 s



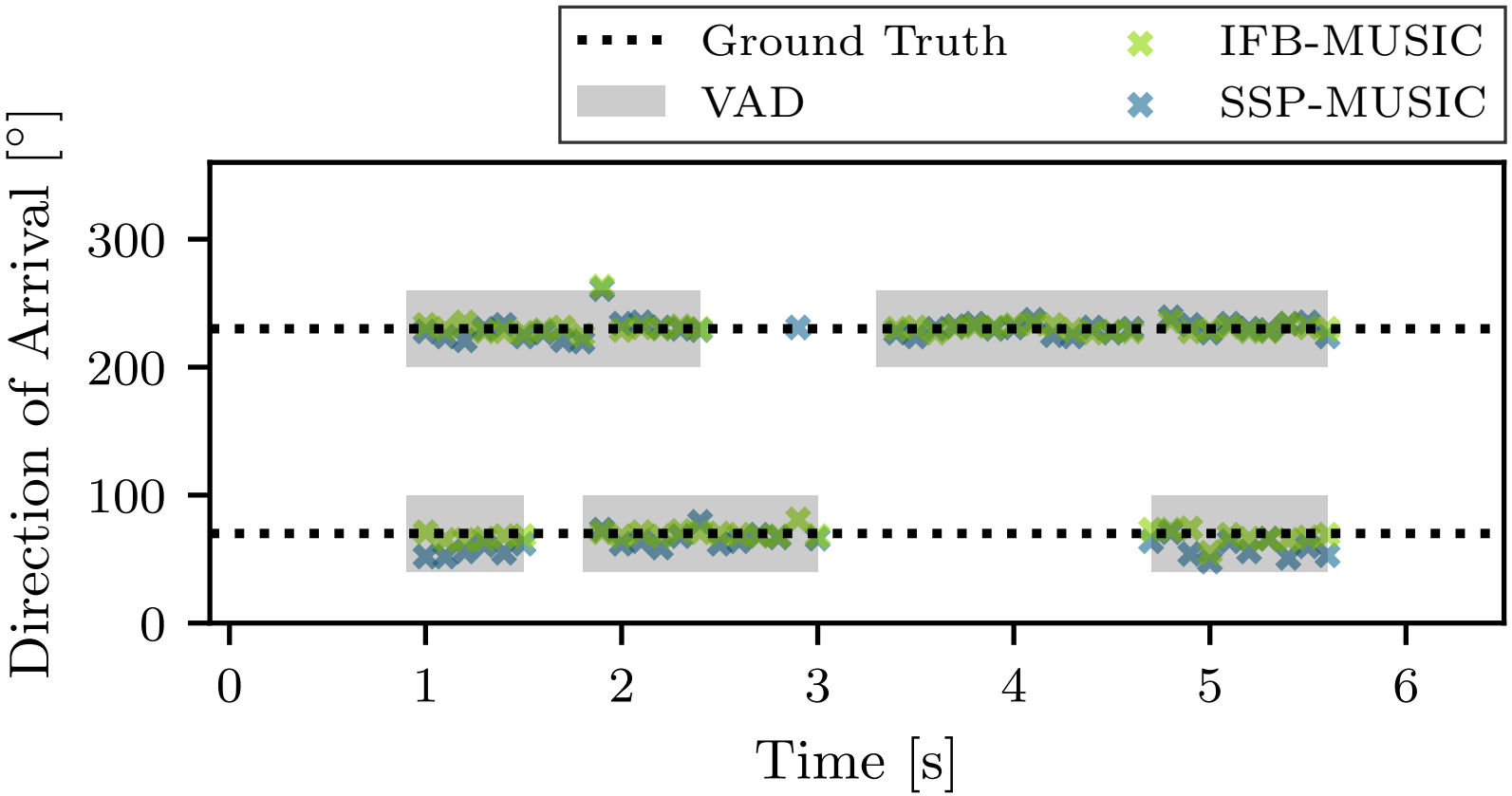
SNR: 15 dB, T60: 1.5 s



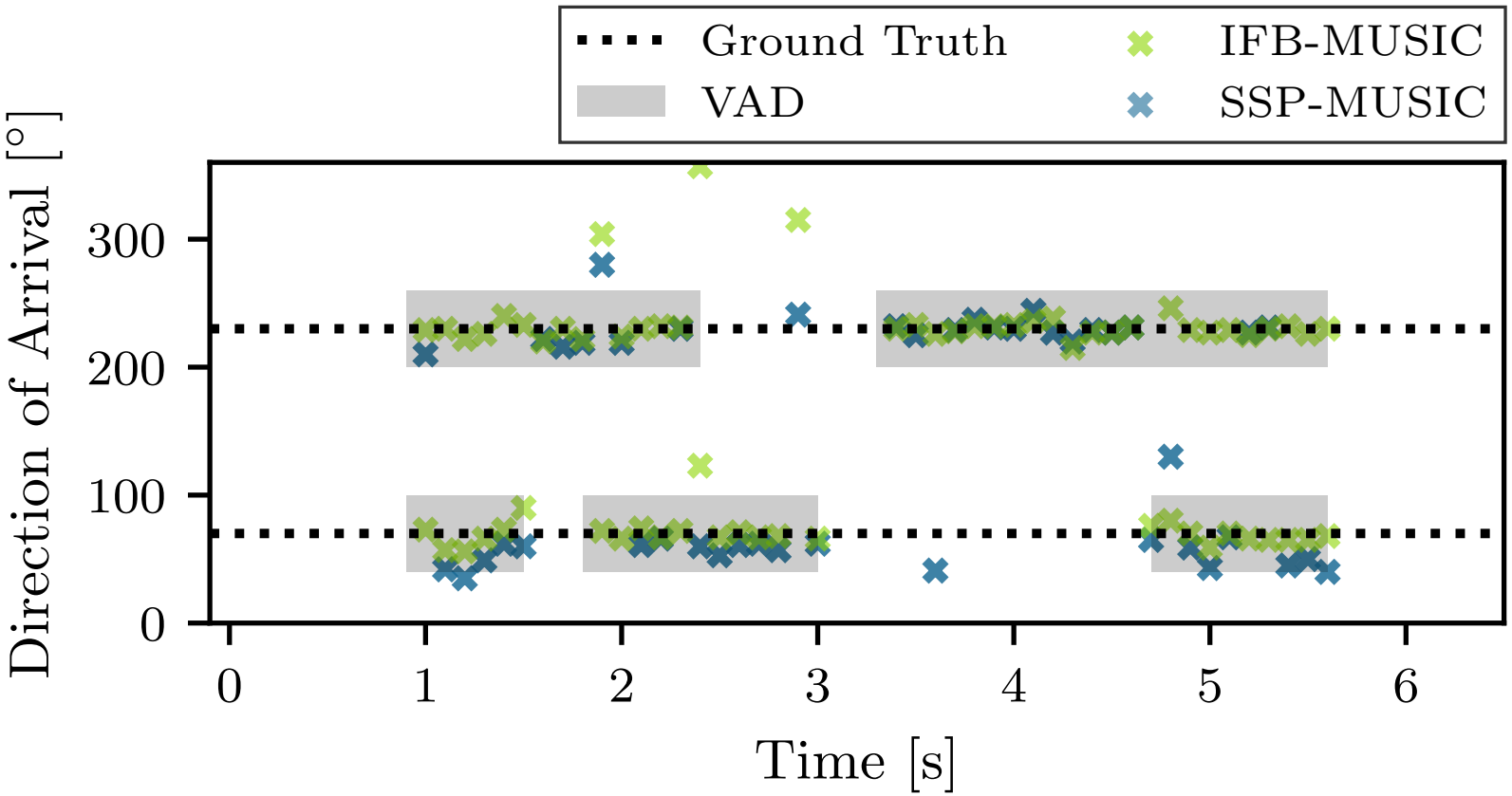
SNR: 15 dB, T60: 1.7 s



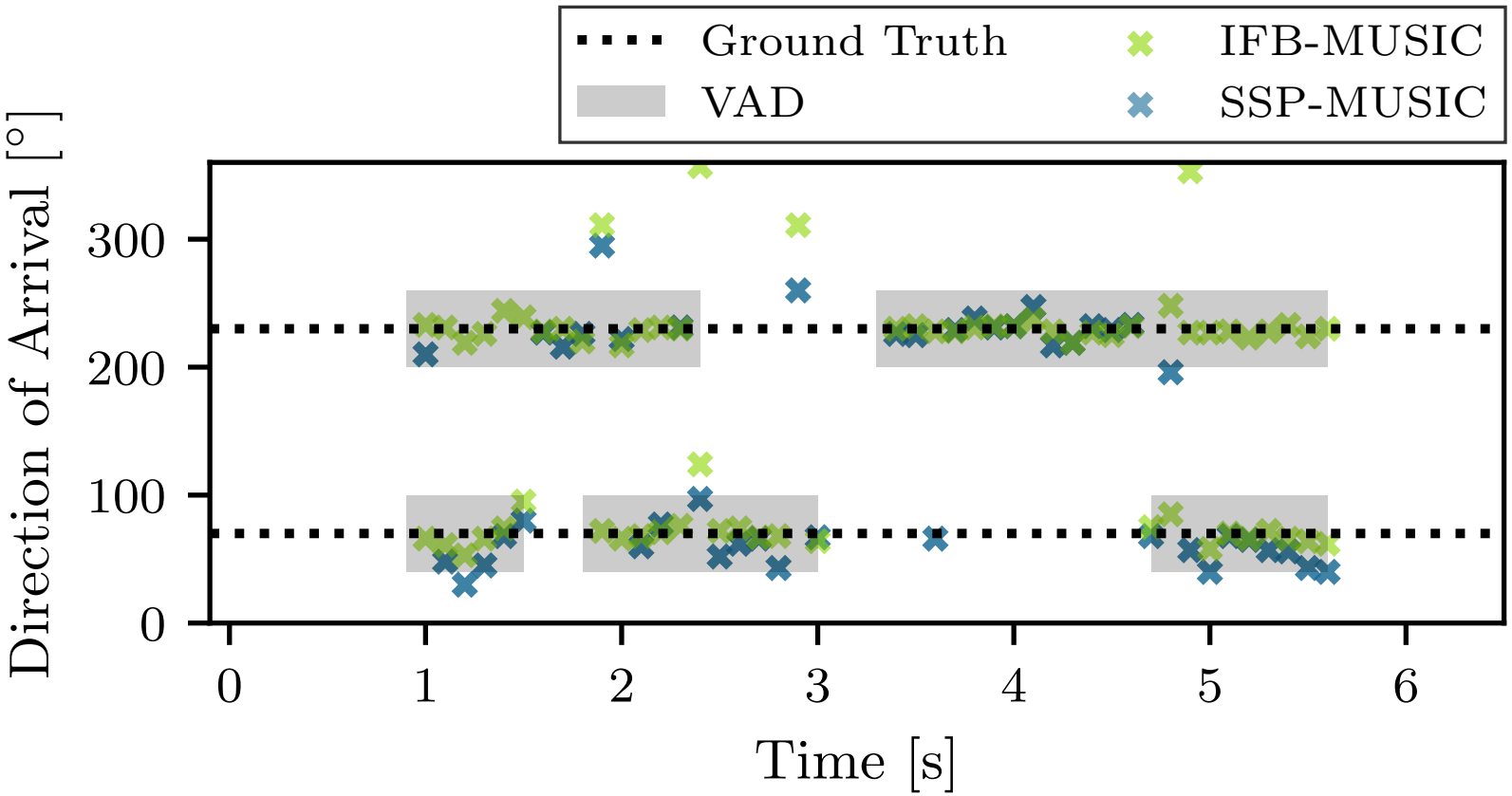
SNR: 15 dB, T60: 0.1 s



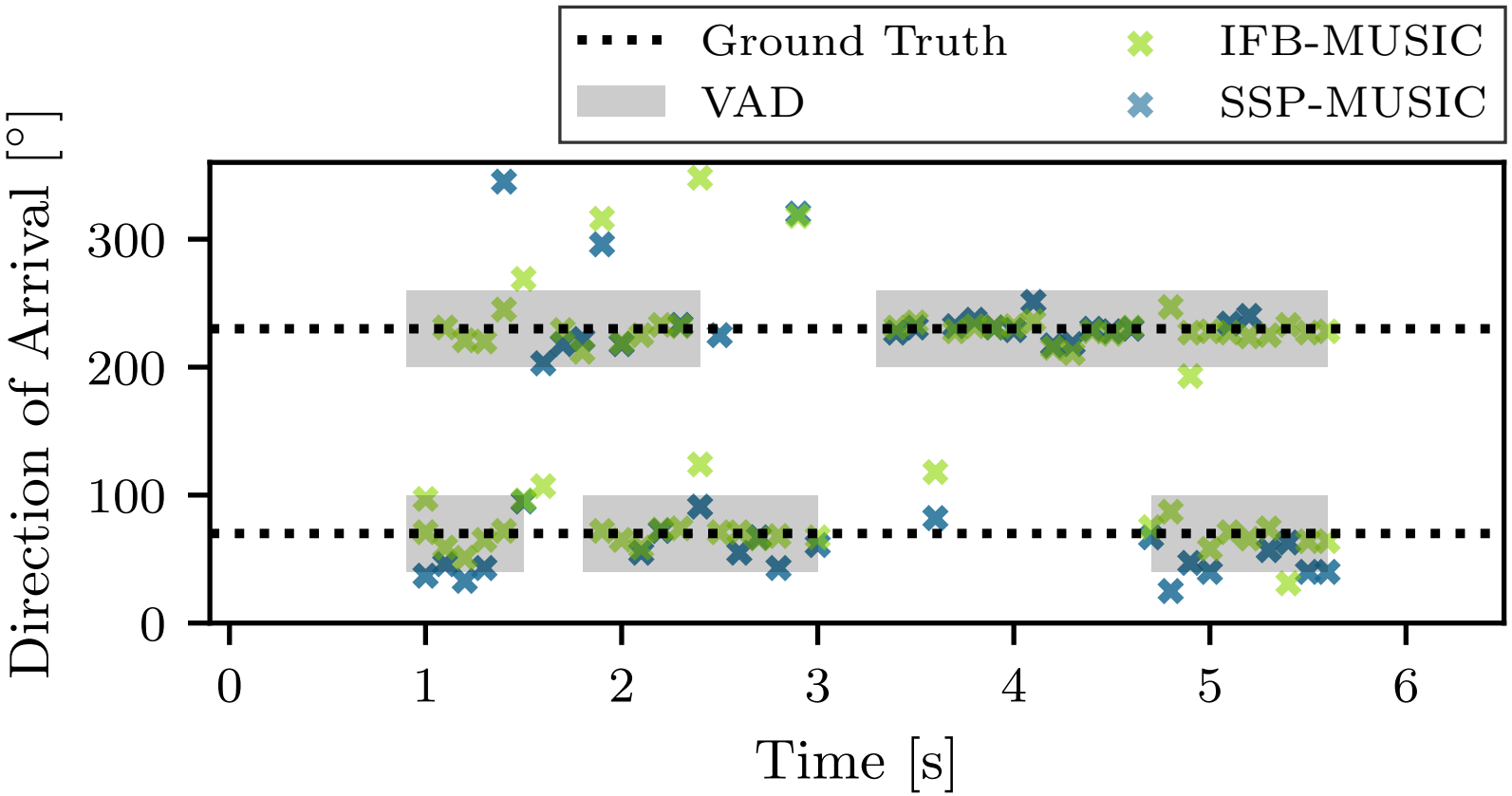
SNR: 15 dB, T60: 0.3 s



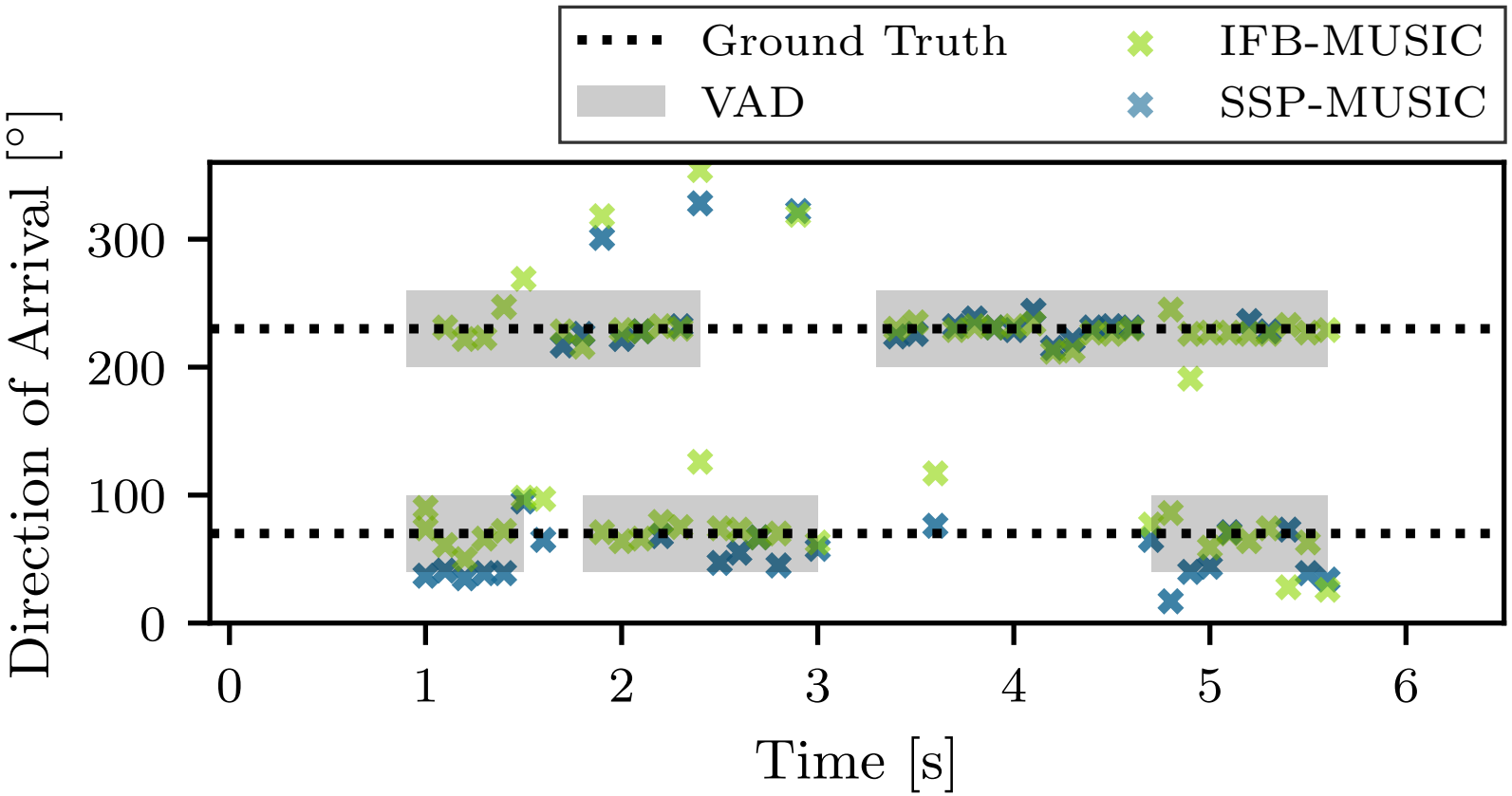
SNR: 15 dB, T60: 0.5 s



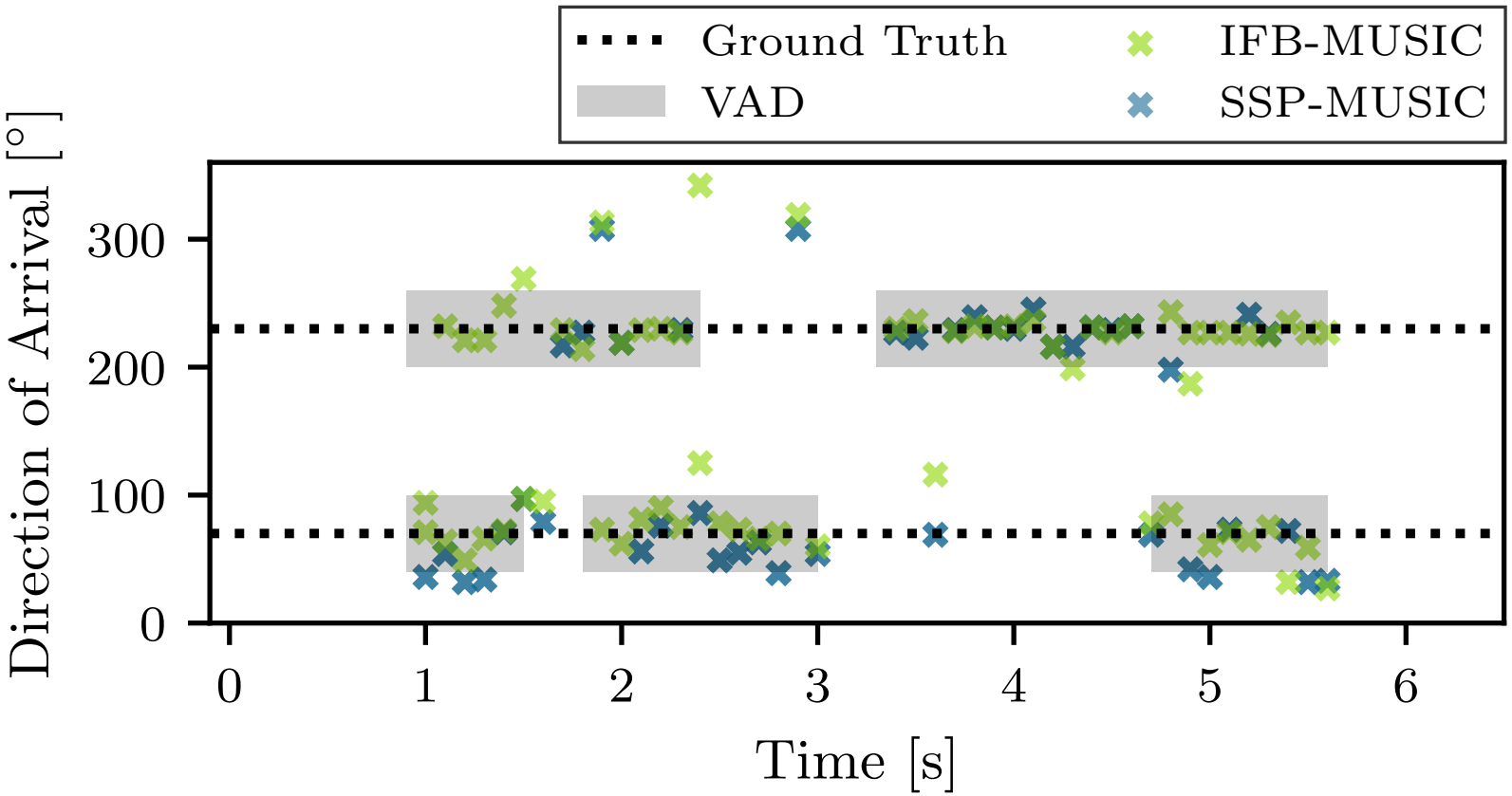
SNR: 15 dB, T60: 0.7 s



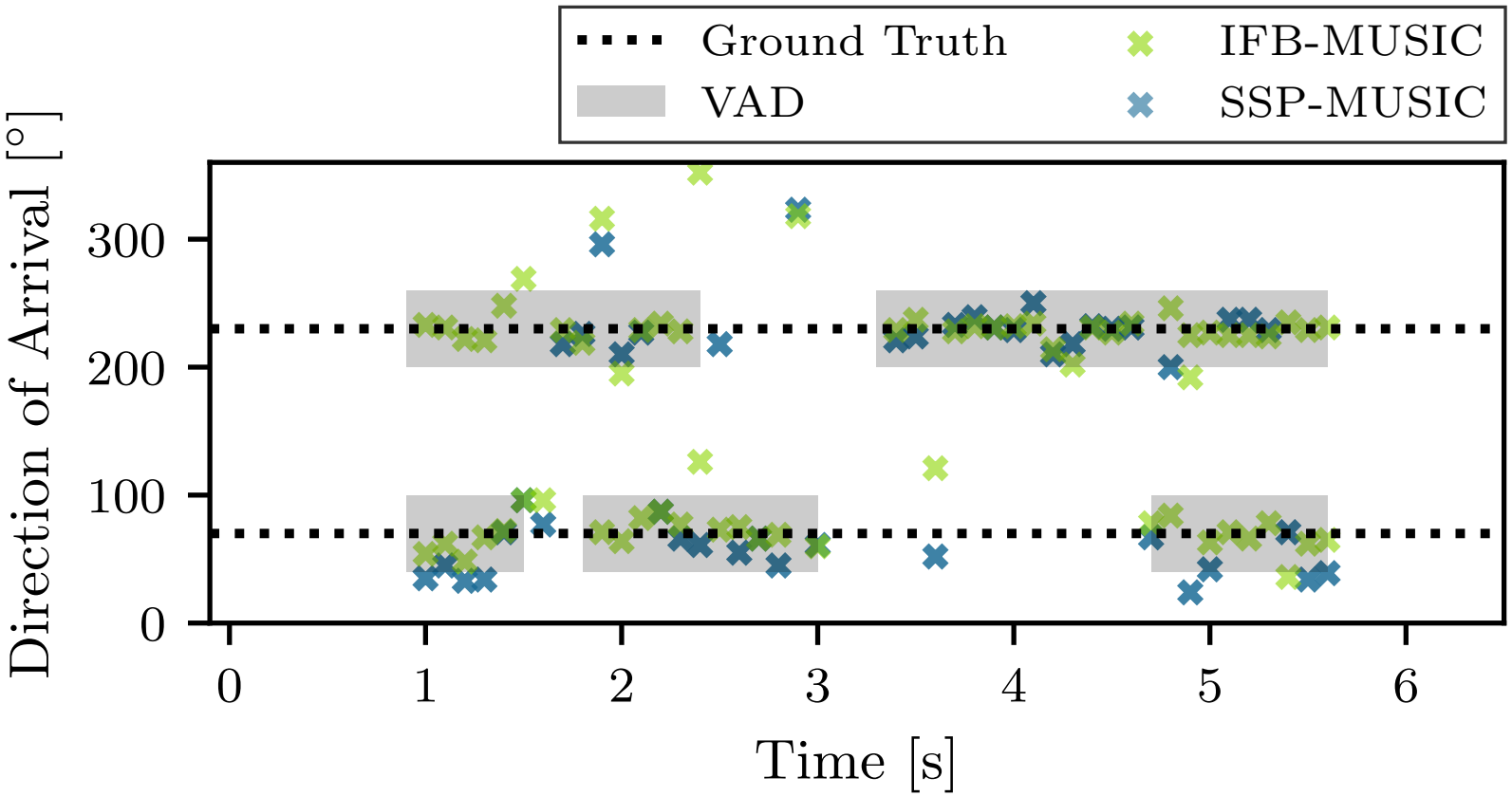
SNR: 15 dB, T60: 0.9 s



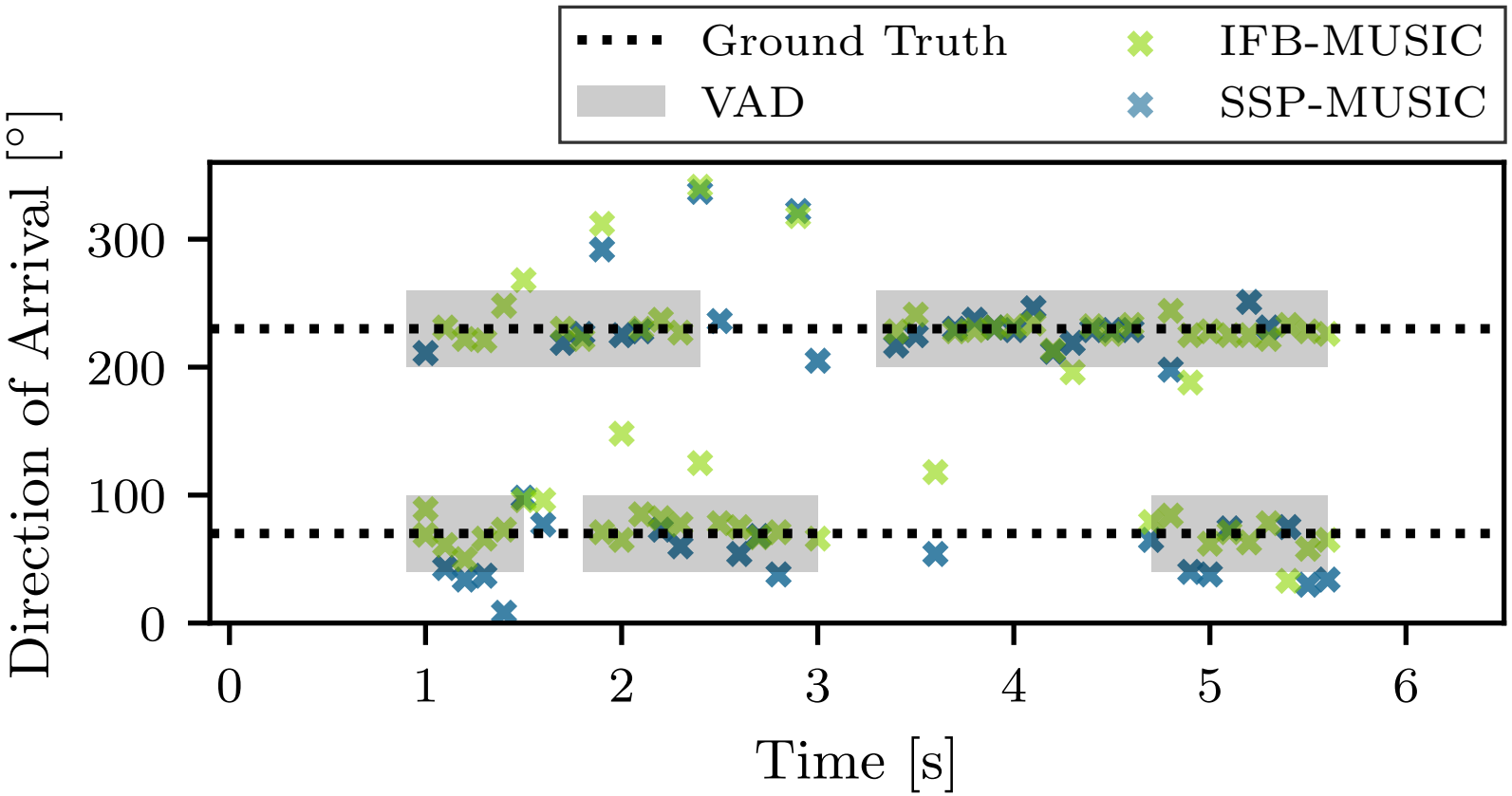
SNR: 15 dB, T60: 1.1 s



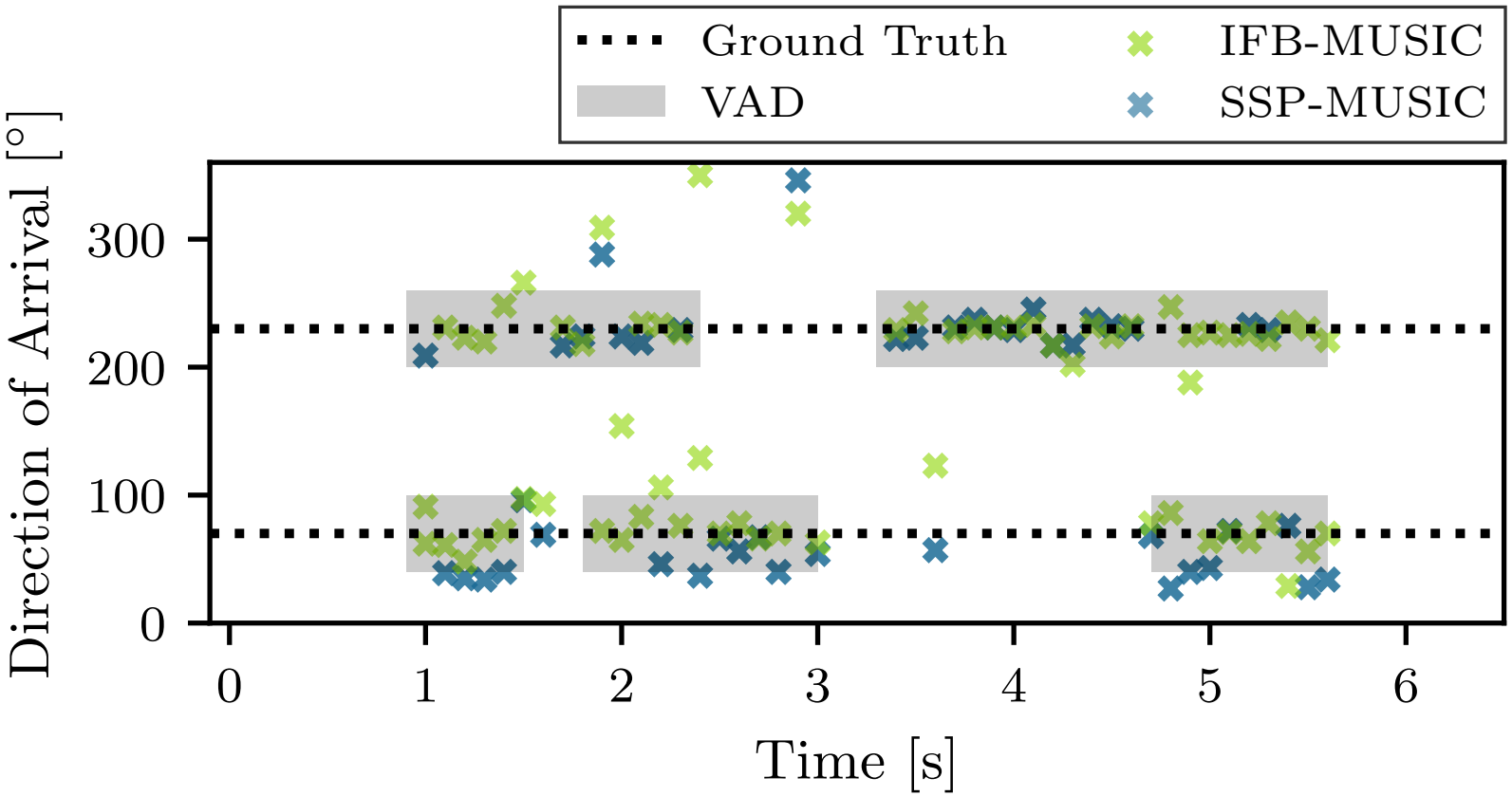
SNR: 15 dB, T60: 1.3 s



SNR: 15 dB, T60: 1.5 s

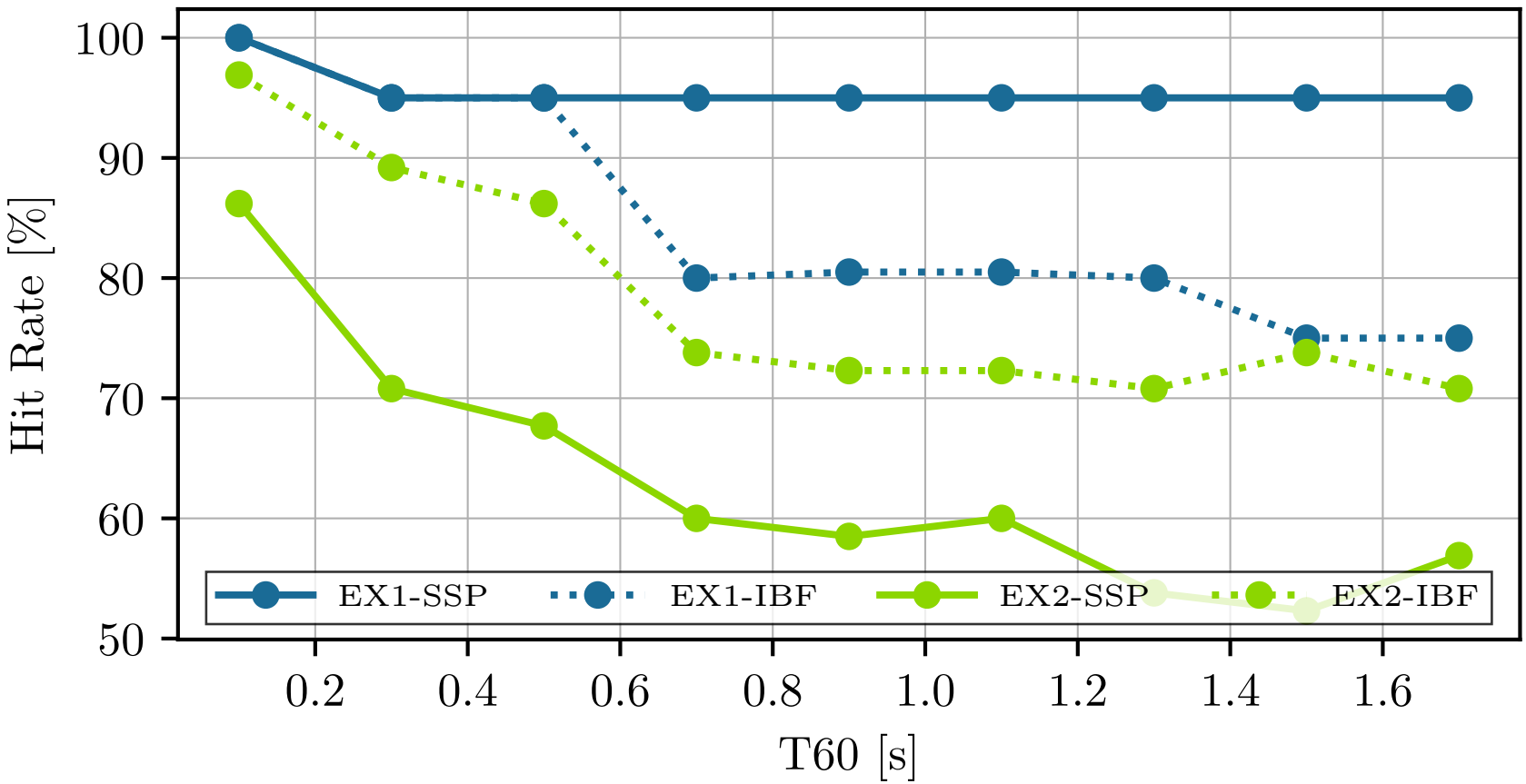


SNR: 15 dB, T60: 1.7 s



EX1: SINGLE SOUND SOURCE IN A SIMULATED ROOM

EX2: TWO SOUND SOURCES IN A SIMULATED ROOM



Method		SSP-MUSIC		IFB-MUSIC	
Metric		HR	FAR	HR	FAR
Recording	1	90.0	10.0	95.0	5.0
	2	64.7	35.3	67.6	32.4
	3	95.1	4.9	75.6	24.4

Comparison of SSP-MUSIC against IFB-MUSIC on the first 3 LOCATA recordings for Task 1.

CONCLUSION

In this talk, we have...

... developed and explored the potential of SSP-MUSIC, which is a polynomial extension of MUSIC.

... proposed some enhancements for sound source localization using SSP-MUSIC.

... demonstrated that SSP-MUSIC is more robust to noise and reverberation.

Questions?

Please email: aidan.hogg13@imperial.ac.uk

Imperial College London
University of Southampton
University of Strathclyde Glasgow
EPSRC Engineering and Physical Sciences Research Council