Probabilistic Nonlinear Spatial Filtering for Multi-Talker Speech Enhancement

Multi-Source Speech Presence Probability Estimation and DNN-Guided MVDR Beamforming in Multichannel Reverberant Environments

MSc Thesis — Electrical Engineering, Signals and Systems Track

Yuning Liu

Delft University of Technology

TU Delft

Probabilistic Nonlinear Spatial Filtering for Multi-Talker Speech Enhancement

Multi-Source Speech Presence Probability Estimation and DNN-Guided MVDR Beamforming in Multichannel Reverberant Environments

by

Yuning Liu

Instructor:	Richard.C.Hendriks
Teaching Assistant:	Giovanni Bologni
Project Duration:	11,2025 - 9,2026
Department:	Electrical Engineering, Mathematics and Computer Science, Delft
Faculty:	SPS,Microelectronics

TU Delft

Use of Generative AI

Generative AI tools were used during this project, and their usage was restricted to the tasks listed here. In the literature and background work, they were used to locate relevant publications and source material identification; each of these sources was then read and verified in the original. In the writing, they were used for language editing of text that had already been drafted by the author, for assistance in typesetting equations in LaTeX, and for checking the internal consistency of citations and cross-references. No derivation, experimental result or figure in this thesis was produced by such tools, and all technical content is the author's own and has been verified by the author.

Summary

A linear MVDR beamformer followed by a single-channel post-filter is MMSE-optimal only when the interference is Gaussian. That assumption fails when competing talkers are present, and the optimal estimator is then nonlinear. This thesis derives that estimator from a composite hypothesis structure which splits the per-bin signal-activity state into two questions: whether the target is present, and which interferer dominates. In closed form the estimator is a set of target-directed MVDR–Wiener filters, one per interference hypothesis, each weighted by an interferer posterior, with the weighted sum multiplied by a target posterior. These two posteriors are the multi-source speech presence probabilities, and the chain rule supplies the product structure exactly, without assuming that they factorise.

The estimator is shown to be structurally equivalent to the nonlinear spatial filter of Tesch and Gerkmann. Two sub-networks with no shared parameters estimate only these posteriors; all filtering is performed by the model-based filters. On a simulated five-talker reverberant corpus captured with a three-microphone array, the system improves SI-SDR by 11.26 dB and PESQ by 0.53, exceeding an end-to-end complex-mask baseline of comparable cost and more than doubling the gain of an oracle MVDR. Because the network outputs posteriors rather than a mask, the estimated quantities can be inspected directly and are shown to recover the intended signal-activity decomposition.

Contents

Nomenclature

Abbreviations

Abbreviation	Definition
ASR	Automatic Speech Recognition
ATF	Acoustic Transfer Function
BLSTM	Bidirectional Long Short-Term Memory
CASA	Computational Auditory Scene Analysis
cIRM	Complex Ideal Ratio Mask
CSS	Continuous Speech Separation
DCCRN	Deep Complex Convolution Recurrent Network
DFT	Discrete Fourier Transform
DNN	Deep Neural Network
DoA	Direction of Arrival
DPRNN	Dual-Path Recurrent Neural Network
ESTOI	Extended Short-Time Objective Intelligibility
FT-JNF	Frequency–Time Joint Nonlinear Filter
GMACs	Giga Multiply–Accumulate operations
GSVD	Generalised Singular Value Decomposition
IBM	Ideal Binary Mask
IMCRA	Improved Minima-Controlled Recursive Averaging
IRM	Ideal Ratio Mask
JNF	Joint Nonlinear Filter
KL	Kullback–Leibler (divergence)
LCMV	Linearly Constrained Minimum Variance
LSA	Log-Spectral Amplitude
LSTM	Long Short-Term Memory
MAC	Multiply–Accumulate operation
MCRA	Minima-Controlled Recursive Averaging
MC-SPP	Multichannel Speech Presence Probability
MMSE	Minimum Mean-Square Error
MMSE-LSA	Minimum Mean-Square-Error Log-Spectral Amplitude estimator
MSE	Mean Squared Error
MVDR	Minimum Variance Distortionless Response
MWF	Multichannel Wiener Filter
PESQ	Perceptual Evaluation of Speech Quality
PSD	Power Spectral Density
PSM	Phase-Sensitive Mask
RIR	Room Impulse Response
RNN	Recurrent Neural Network
RTF	Relative Transfer Function
SAP	Speech Absence Probability
SGMSE	Score-based Generative Model for Speech Enhancement
SI-SDR	Scale-Invariant Signal-to-Distortion Ratio
SNR	Signal-to-Noise Ratio
SPP	Speech Presence Probability
SSF	Spatially Selective Filter
STFT	Short-Time Fourier Transform
STOI	Short-Time Objective Intelligibility

Abbreviation	Definition
StoRM	Stochastic Regeneration Model
STSA	Short-Time Spectral Amplitude
TCN	Temporal Convolutional Network
TF	Time–Frequency
TSP	TSP speech corpus
UCA	Uniform Circular Array
uPIT	Utterance-level Permutation Invariant Training
WDO	W-Disjoint Orthogonality
WPE	Weighted Prediction Error
WSJ0	Wall Street Journal (speech corpus)

Symbols

Throughout, a hat $(\hat{\cdot})$ denotes a quantity estimated from the observation, and a superscript $(\cdot)^{\mathrm{o}}$ denotes an oracle quantity computed from ground-truth signals. Boldface lowercase denotes a vector, boldface uppercase a matrix. The frequency-bin and time-frame indices (k, l) are suppressed where unambiguous.

Symbol	Definition	Unit
Latin symbols		
A	Short-time spectral amplitude of the target, $S = A\,e^{j\varphi}$	[–]
$A_m(k)$	Acoustic transfer function from the source to microphone m	[–]
$\mathbf{a}(k)$	Acoustic transfer function (steering) vector, $\mathbf{a} = [A_1, \ldots, A_M]^T$	[–]
$\mathbf{b}_i$	Relative transfer function (steering vector) of source i	[–]
c_m	Weight of Gaussian-mixture component m (Tesch–Gerkmann model)	[–]
$\mathbf{d},\ \hat{\mathbf{d}}$	True / estimated relative transfer function of the target, $\mathbf{d} \triangleq \mathbf{b}_j$	[–]
f_s	Sampling rate	[Hz]
g_i	Constant per-source gain applied to source i in the mixture	[–]
$\tilde{g}_i$	Interferer detection statistic of hypothesis $\mathcal{H}_i$, Eq. (5.38)	[–]
G_{Wiener}	Single-channel Wiener gain, $\xi/(1+\xi)$	[–]
$G_{\mathrm{W},i},\ \hat{G}_{\mathrm{W},i}$	True / estimated Wiener post-filter gain of branch i	[–]
$G_{\min}$	Lower bound (floor) imposed on the Wiener gain	[–]
$h(S)$	Arbitrary functional of the target coefficient in the MMSE expansion	[–]
$\mathcal{H}_i$	Interferer hypothesis i ($\mathcal{H}_0$: noise only; $\mathcal{H}_i$, $i \geq 1$: interferer i dominant)	[–]
I	Total number of speech sources (the target and $I-1$ interferers); also the number of interferer hypotheses	[–]
$\mathbf{I}_M$	$M \times M$ identity matrix	[–]
k	Frequency-bin index	[–]
K	Number of frequency bins	[–]
l	Time-frame index	[–]
L	Number of time frames	[–]
$\mathcal{L}$	Total training loss (calligraphic; distinct from the frame count L), Eq. (6.13)	[–]
$\mathcal{L}_{\mathrm{enh}}$	Enhancement (dual-path ℓ_1) loss, Eq. (6.14)	[–]
$\mathcal{L}_{\mathcal{H}}$	Interferer-posterior supervision loss (mean squared error), Eq. (6.15)	[–]
$\mathcal{L}_{\mathcal{H}}^{\mathrm{KL}}$	Kullback–Leibler variant of the posterior-supervision loss, Appendix B	[–]
M	Number of microphones	[–]

Symbol	Definition	Unit
$M(a,b,x)$	Confluent hypergeometric (Kummer) function; distinct from the microphone count M	[–]
$\mathcal{M}_1,\ \mathcal{M}_0$	Target-present / target-absent hypothesis	[–]
m_{ref}	Reference-microphone index	[–]
$\mathbf{n}$	Multichannel additive-noise vector	[–]
$N(k,l)$	Single-channel noise STFT coefficient	[–]
N_{c}	Number of components in the Gaussian-mixture noise model of Section 5.6; equal to I under the correspondence (5.51)	[–]
$p_A(a)$	Generalised-Gamma prior density of the target amplitude	[–]
$P(\mathcal{H}_i)$	A priori probability of interferer hypothesis $\mathcal{H}_i$	[–]
$P(\mathcal{M}_m)$	A priori probability of target presence ($m=1$) or absence ($m=0$)	[–]
$p(\mathcal{H}_i \mid \mathbf{y}, \mathcal{M}_1)$	*Interferer posterior*: probability that interferer i dominates the bin, given that the target is present, Eq. (5.39)	[–]
$p(\mathcal{H}_i \mid \mathbf{y})$	Marginal interferer posterior, Eq. (5.47)	[–]
$p(\mathcal{M}_1 \mid \mathbf{y})$	*Target posterior*: probability that the target is present, Eq. (5.49)	[–]
$\hat{p}(\cdot)$	Network-estimated posterior	[–]
$p^{\mathrm{o}}(\cdot)$	Oracle posterior	[–]
P_m	Argument of the confluent hypergeometric function under mixture component m	[–]
q	A priori speech absence probability (single-source, Ch. 3), Eq. (3.23)	[–]
q_i	Target-absence prior term under $\mathcal{H}_i$, $q_i = 1 - \frac{\mathrm{tr}[\mathbf{\Phi}_{xx}]}{\mathrm{tr}[\mathbf{\Phi}_{xx}]+\mathrm{tr}[\mathbf{\Sigma}_i]}$	[–]
Q_m	Reparameterisation factor $(\nu + \mathbf{d}^H \mathbf{\Phi}_m^{-1} \mathbf{d}\, \sigma_S^2)^{-\nu}$, Eq. (5.50)	[–]
$R,\ \hat{R}$	Reference / estimated residual STFT, $R \triangleq Y_{m_{\mathrm{ref}}} - S$	[–]
$r,\ \hat{r}$	Reference / estimated residual waveform	[–]
$S,\ \hat{S}$	True / estimated target STFT coefficient, $S \triangleq V_j$	[–]
S_{src}	STFT coefficient of the target at the source (before propagation)	[–]
$s,\ \hat{s}$	Reference / estimated target waveform	[–]
T_{60}	Reverberation time	[s]
T_{MVDR}	Output of the linear MVDR beamformer (Fig. 1.1)	[–]
$T_{\mathrm{MVDR\text{-}MMSE}}$	Output of the beamformer–postfilter cascade (Fig. 1.1)	[–]
T_{MMSE}	Output of the MMSE-optimal nonlinear spatial filter, Eq. (5.50)	[–]
$T^{(m)}_{\mathrm{MVDR}}$	MVDR output under mixture component m; counterpart of z_i	[–]
$\mathbf{u}$	Aggregate of undesired components (interferers and noise)	[–]
$\mathbf{u}_1$	Reference-channel selection vector, $\mathbf{u}_1 = [1, 0, \ldots, 0]^T$	[–]
$u_{\max}(\cdot)$	Eigenvector associated with the largest eigenvalue	[–]
V_i	STFT coefficient of source i at m_{ref} (V_j for the target)	[–]
$\mathbf{v}_i$	Interference-plus-noise vector under $\mathcal{H}_i$, $\mathbf{v}_i \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Sigma}_i)$	[–]
$\mathbf{w}_i,\ \hat{\mathbf{w}}_i$	True / estimated MVDR weight vector of hypothesis i	[–]
$\mathbf{w}_{\mathrm{MVDR}}$	Single-source MVDR weight vector (Ch. 3), Eq. (3.27)	[–]
$\mathbf{y}$	Multichannel mixture (observation) STFT vector	[–]
$\mathbf{y}_i^{\mathrm{rev}}$	Gain-scaled reverberant image of source i at the array	[–]
$Y,\ Y_{m_{\mathrm{ref}}}$	Single-channel / reference-channel mixture STFT	[–]
z_i	Target-directed MVDR output of hypothesis i, $z_i = \mathbf{w}_i^H \mathbf{y}$	[–]
$\tilde{z}_i$	Interferer-directed MVDR output within the common base $\mathbf{\Sigma}_{S,0}$, Eq. (5.40)	[–]
Greek symbols		
α	Nominal smoothing constant of the covariance recursions (Ch. 6)	[–]
α_s	Recursive smoothing factor, single-channel noise PSD, Eq. (3.11)	[–]
$\alpha_v,\ \alpha_x$	Recursive smoothing factors, noise / speech covariance (Ch. 3)	[–]
α_G	Smoothing factor of the beamformer noise-floor recursion, Eq. (6.10)	[–]

Symbol	Definition	Unit
α_ℓ	Weight of the time-domain group in the enhancement loss, Eq. (6.14)	[–]
$\alpha_{\text{eff},i}$	Posterior-controlled effective smoothing factor of $\hat{\mathbf{\Sigma}}_i$, Eq. (6.1)	[–]
$\alpha_{\text{eff},S}$	Posterior-controlled effective smoothing factor of $\hat{\mathbf{\Sigma}}_{S,0}$, Eq. (6.4)	[–]
$\beta(k,l)$	Target-presence detection statistic (single-source, Ch. 3), Eq. (3.21)	[–]
β_{sc}	Prior ratio $P(\mathcal{H}_0)/P(\mathcal{H}_1)$ (single-channel SPP)	[–]
$\beta_{S,i}$	Target-presence detection statistic under $\mathcal{H}_i$, $\beta_{S,i} = \mathbf{y}^H\mathbf{\Sigma}_i^{-1}\mathbf{\Phi}_{xx}\mathbf{\Sigma}_i^{-1}\mathbf{y}$	[–]
γ	A posteriori SNR (single-channel), $\lvert Y\rvert^2/\Phi_{nn}$	[–]
γ_i	*Normalised* target-presence detection statistic under $\mathcal{H}_i$, $\gamma_i = \beta_{S,i}/(1+\xi_{S,i})$; *not* an SNR	[–]
$\Gamma(\cdot)$	Gamma function	[–]
δ	Dimensionless diagonal-loading factor, Eq. (6.3)	[–]
$\delta(\cdot)$	Dirac delta (target-absent component of the spike-and-slab prior)	[–]
$\mathbf{\Delta}(k)$	Utterance-averaged covariance difference, Eq. (6.6)	[–]
θ	Source azimuth angle	[deg]
λ_i	Largest-magnitude entry of $\hat{\mathbf{\Sigma}}_i$, used to scale the loading	[–]
$\lambda_{\mathcal{H}}$	Weight of the posterior-supervision loss, Eq. (6.13)	[–]
μ	Trade-off (over-subtraction) parameter of the branch Wiener gain, Eq. (6.11)	[–]
$\mu_{S\lvert y,i}$	Posterior mean of S under $(\mathcal{M}_1,\mathcal{H}_i)$	[–]
ν	Shape parameter of the generalised-Gamma speech prior	[–]
ξ	A priori SNR (single-channel), Φ_{ss}/Φ_{nn}	[–]
$\xi_{\mathcal{H}_1}$	Fixed a priori SNR under $\mathcal{H}_1$ (single-channel SPP)	[–]
$\xi_{S,i},\ \hat{\xi}_{S,i}$	True / estimated target a priori SNR under $\mathcal{H}_i$, $\xi_{S,i} = \mathbf{d}^H\mathbf{\Sigma}_i^{-1}\mathbf{d}\,\sigma_S^2 = \text{tr}(\mathbf{\Sigma}_i^{-1}\mathbf{\Phi}_{xx})$	[–]
$\tilde{\xi}_i$	Interferer a priori SNR of hypothesis $\mathcal{H}_i$ relative to $\mathbf{\Sigma}_{S,0}$	[–]
σ_S^2	Target-speech power spectral density	[–]
$\hat{\sigma}_{S,i}^2$	Estimated target PSD at the output of branch i, Eq. (6.11)	[–]
$\sigma_{V_i}^2$	Power spectral density of source i	[–]
$\sigma_{S\lvert y,i}^2$	Posterior variance of S under $(\mathcal{M}_1,\mathcal{H}_i)$	[–]
$\sigma_{n,\text{post},i}^2$	Analytical noise floor at the output of branch i, $(\mathbf{d}^H\mathbf{\Sigma}_i^{-1}\mathbf{d})^{-1}$	[–]
$\mathbf{\Sigma}_i,\ \hat{\mathbf{\Sigma}}_i$	True / estimated interference-plus-noise spatial covariance under $\mathcal{H}_i$	[–]
$\hat{\mathbf{\Sigma}}_i^\dagger$	Inverse of the diagonally loaded covariance, $(\hat{\mathbf{\Sigma}}_i + \delta\lambda_i\mathbf{I}_M)^{-1}$	[–]
$\mathbf{\Sigma}_N$	Noise-only spatial covariance matrix	[–]
$\mathbf{\Sigma}_{S,i}$	Target-inclusive covariance under $(\mathcal{M}_1,\mathcal{H}_i)$, $\mathbf{\Sigma}_i + \mathbf{d}\mathbf{d}^H\sigma_S^2$	[–]
$\mathbf{\Sigma}_{S,0},\ \hat{\mathbf{\Sigma}}_{S,0}$	True / estimated common-base covariance shared by all interferer hypotheses, Eq. (5.34)	[–]
φ	Phase of the target STFT coefficient, uniform on $[0,2\pi)$	[rad]
ϕ	Microphone-array orientation azimuth	[rad]
$\Phi_{ss},\ \Phi_{nn},\ \Phi_{yy}$	Speech / noise / observation PSD (single-channel, scalar)	[–]
$\mathbf{\Phi}_{xx},\ \hat{\mathbf{\Phi}}_{xx}$	True / estimated target-speech PSD matrix, $\mathbf{\Phi}_{xx} = \mathbf{d}\mathbf{d}^H\sigma_S^2$	[–]
$\hat{\Phi}_{z_i}$	Smoothed output power of branch i, Eq. (6.9)	[–]
$\hat{\Phi}_{n,i}$	Smoothed noise floor of branch i, Eq. (6.10)	[–]
$\mathbf{\Phi}_m$	Covariance of Gaussian-mixture component m (counterpart of $\mathbf{\Sigma}_i$)	[–]
ω_k	Frequency weight at bin k in the posterior-supervision loss	[–]
Operators and notation		
$\mathcal{CN}(\cdot,\cdot)$	Circularly-symmetric complex Gaussian distribution	[–]
$\mathbb{E}[\cdot]$	Expectation	[–]
$\text{tr}(\cdot)$	Matrix trace	[–]

Symbol	Definition	Unit
$(\cdot)^T$	Transpose	[–]
$(\cdot)^H$	Conjugate (Hermitian) transpose	[–]
$\lvert \cdot \rvert$	Matrix determinant (matrix argument) / modulus (scalar argument)	[–]
$\lVert \cdot \rVert_1$	ℓ_1 norm	[–]
$\hat{(\cdot)}$	Quantity estimated from the observation	[–]
$(\cdot)^{\mathrm{o}}$	Oracle quantity computed from ground truth	[–]
$\propto$	Proportional to	[–]
$\triangleq$	Defined as	[–]

1

Introduction

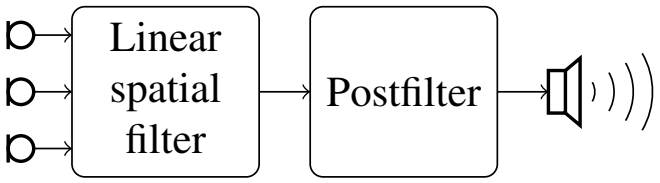
1.1. Background and Motivation

Speech enhancement is a foundational perceptual task in audio signal processing that seeks to recover a target speech signal from observations corrupted by additive noise, competing speakers, and room reverberation. The significance of this task extends to a broad range of real-world applications, from hearing prostheses and mobile telecommunications to robust automatic speech recognition (ASR) front-ends [1].

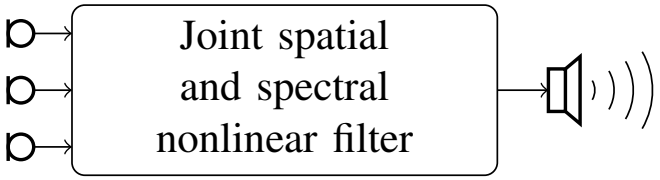
Conventional approaches to this problem are grounded in statistical frameworks for signal processing, wherein the short-time spectral characteristics of speech and noise are modelled under simplified distributional assumptions—typically complex Gaussian priors on the discrete Fourier transform (DFT) coefficients—and closed-form estimators are derived accordingly [2]. Representative single-channel methods include the minimum mean-square error (MMSE) short-time spectral amplitude estimator [3], the log-spectral amplitude (LSA) estimator [4], and the Wiener filter. The functionality of these methods relies on accurate estimates of the noise power spectral density (PSD), obtained through recursive tracking schemes such as minimum controlled recursive averaging (MCRA) [5] and its improved variant IMCRA [6].

While single-channel methods exploit mainly temporal–spectral statistics to achieve target speech retrieval, signal array processing extends the scope to the multichannel domain, where spatial information can additionally be leveraged to improve the estimation. Spatial filters, or beamformers, provide a complementary mechanism by combining the microphone signals with complex beamforming weights so as to shape the spatial response of the array. In this case, a prescribed response is maintained towards the target while components with a different spatial signature are attenuated [7]. A prominent example is the minimum variance distortionless response (MVDR) beamformer, whose weights are determined by the noise spatial covariance matrix and a steering vector characterising the acoustic transfer function from the target source to the array [8].

Despite their analytical elegance and computational efficiency, classical frameworks are subject to well-documented limitations. The Gaussian assumption on the noise distribution, while convenient for deriving closed-form solutions, is frequently violated in practical situations when the background interference comprises non-stationary sources such as competing speakers [9]. Hendriks et al. [10] demonstrated that, under Gaussian noise, the multichannel MMSE-optimal estimator is separable into a linear MVDR beamformer followed by a single-channel post-filter; however, for more general noise distributions, the optimal estimator is inherently a joint spatial–spectral nonlinear filter that cannot be decomposed into such a two-step cascade. Tesch and Gerkmann [9] provided extensive empirical evidence for this theoretical result, showing that a nonlinear spatial filter can suppress more than $D-1$ directional interferers with a D-element array without spatial adaptation—a capability fundamentally beyond the reach of any linear beamformer (as shown in Fig. 1.1 and Fig. 1.2). Moreover, the noise PSD tracking algorithms that underpin both single-channel post-filtering and beamformer weight estimation (e.g., MCRA, IMCRA, decision-directed methods) are predicated on the assumption of slowly varying noise statistics,



(a)



(b)

Figure 1.1: (a) Conventional two-step multichannel processing: a linear spatial filter (beamformer) followed by a single-channel postfilter. Under Gaussian noise, this cascade is MMSE-optimal [10]. (b) The MMSE-optimal nonlinear spatial filter for non-Gaussian noise: spatial and spectral processing are joined into a single non-separable nonlinear operation. Reproduced from [9].

rendering them inadequate for scenarios involving rapidly fluctuating speech-like interference [11].

The advent of deep learning shifted the field from classical beamforming-based estimators to learned ones—through mask estimation, direct spectral or waveform mapping, and more recently generative modelling—with substantial gains in perceptual quality under reverberant, multi-source conditions; Chapter 2 reviews these in detail. The gains come at a cost. The learned representations are not interpretable in terms of beamforming, noise statistics or source activity, so failure modes are hard to diagnose and the trade-off between suppression and distortion is hard to control [12]. Architectures that perform spatial and spectral processing jointly and implicitly also need substantially more capacity and training data than designs that carry the structure of the signal model [13]. This has motivated DNN-guided hybrid systems, in which the network estimates the statistical parameters that a classical beamformer or post-filter then uses [14, 15]. One such parameter is the a posteriori speech presence probability, which assigns each time–frequency bin a soft probability of speech activity. It provides a well-defined interface between the network and the classical stage, both for noise PSD tracking and for setting the beamformer parameters [16, 11, 17, 18].

The gap this thesis addresses lies between these two lines of work. The MMSE-optimal estimator for multi-talker interference is already known in closed form, and its posterior weights are source-activity probabilities; what is missing is a method that estimates those posteriors and uses them *as such*. End-to-end nonlinear filters attain the performance while discarding the structure, so neither a beamformer, a post-filter, nor a per-bin activity probability can be established from the network output. DNN-guided MC-SPP pipelines retain the structure, but are formulated for a single target against a stationary noise field, so the decision that makes the multi-talker estimator nonlinear—which competing talker dominates a given time–frequency bin—is never represented. Section 1.2 makes this concrete and states the resulting research questions.

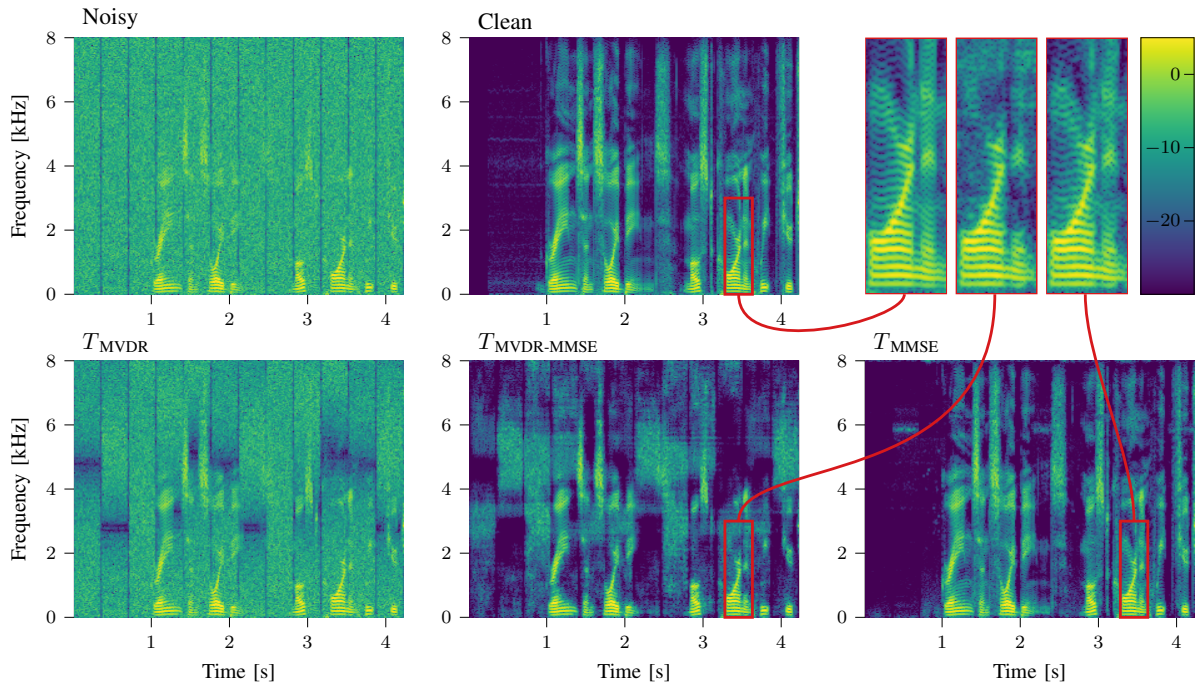


Figure 1.2: Spectrograms comparing the linear MVDR beamformer (T_{MVDR}), the conventional beamformer–postfilter cascade ($T_{\mathrm{MVDR\text{-}MMSE}}$), and the MMSE-optimal nonlinear spatial filter (T_{MMSE}) in a five-interferer scenario. The zoomed-in regions (top right) show that the nonlinear filter preserves the harmonic structure of voiced speech while achieving substantially greater noise suppression than the linear cascade. Reproduced from [9].

1.2. Problem Analysis and Research Questions

The theoretical result of Hendriks et al. [10] shows that the Bayesian MMSE-optimal multichannel estimator under Gaussian mixture noise takes the form of a weighted sum over source hypotheses, where each summand is an MVDR beamformer matched to a specific interferer, followed by a spectral gain, and the weights show that each interferer is active at the current time–frequency bin. In chapter 4 and chapter 5 we prove from the Bayesian framework that These weights are precisely the *multi-source speech presence probabilities*. Establishing this connection between posterior weights and multi-source speech presence probabilities is one of the contributions of the present thesis: Chapter 4 develops the Bayesian MMSE framework, and Chapter 5 derives the corresponding posteriors in closed form and shows their structural equivalence to the analytic nonlinear spatial filter of Tesch and Gerkmann's work. Overall, the estimator is nonlinear because the weights depend on the observation through the likelihood ratios of the competing hypotheses.

Two recent lines of work have attempted to realise this nonlinear estimator in practice, each with limitations that motivate the present thesis.

Tesch and Gerkmann [9, 19] proposed a family of *deep nonlinear filters* in which a neural network directly predicts a complex ideal ratio mask, effectively implementing the nonlinear spatial filter end-to-end. Their approach achieves strong empirical performance, but precisely because the mask is learned without explicit beamformer structure, the estimator lacks interpretability: it is not possible to decompose the output into spatial filtering, spectral post-filtering, and hypothesis weighting, nor to extract a measure of source-presence uncertainty at each time–frequency bin.

Tao et al. [18] took the complementary approach: they trained a DNN to predict the MC-SPP and used it to drive a classical MVDR beamformer. However, their MC-SPP formulation addresses only the single-source case, where the noise is modelled as stationary and the SPP serves primarily as a soft mask for PSD tracking. The connection to the multi-source Bayesian estimator is not established, and the overall pipeline is an empirical design rather than one derived from a Bayesian formulation.

This thesis bridges the two approaches by grounding the DNN-guided beamforming pipeline in the

Bayesian MMSE theory of nonlinear spatial filtering. The central question is:

> *Can a multichannel speech enhancement framework for multi-talker scenarios be derived, directly from the Bayesian MMSE-optimal nonlinear spatial filter, such that it retains the modular structure of classical beamformer, while a neural network supplies only the posterior source-presence probabilities?*

This question decomposes into the following sub-questions:

1. **Multi-source MC-SPP derivation.** Can the single-source MC-SPP be extended to a multi-source setting under the W-disjoint orthogonality assumption, and if so, does the resulting posterior for the activity of each source admit a closed form?
2. **Connection to the MMSE-optimal nonlinear filter.** What is the precise relationship between the multi-source MC-SPP and the Bayesian MMSE estimator, and under what conditions, if any, does the optimal estimator factorize into an MVDR beamformer, a spectral post-filter, with probably posterior weight activation?
3. **DNN-guided estimation and probabilistic output.** Can a neural network be trained to predict the multi-source MC-SPP from multichannel observations, and can the predicted probabilities drive a classical MVDR and Wiener pipeline to achieve multi-talker enhancement with interpretable, probabilistic output—providing not only enhanced signals but also a per-bin uncertainty measure for source activity?
4. **Sensitivity to parameter estimation.** How sensitive is the enhancement performance to the accuracy of the per-source PSD estimates used in the MC-SPP computation, and what is the performance gap between oracle and estimated quantities?

Together, these sub-questions lead to a single controlled comparison, which the experiments of Chapter 7 are designed to answer. Given an identical multichannel observation of the same acoustic scene, what is gained by replacing an end-to-end learned mask with the composite-hypothesis estimator developed here, in which the network estimates only the source-activity posteriors and the spatial–spectral filtering is left to a filter prescribed by the signal model?

1.3. Main Contributions and Structure of this Thesis

In this thesis, we propose a cascaded learning framework for multichannel speech enhancement grounded on the a posteriori MC-SPP.

Ahead of the experiments, we first revisit the MC-SPP theory within the classical statistical paradigm and examine its integration with multichannel Wiener filtering (MWF) for the conventional speech denoising task. Building on this theory, three signal model configurations are considered, describing target signal retrieval in reverberant non-stationary noisy environments, competing speaker scenarios, and composite conditions with interferers and noise.

Second, we propose a DNN-guided framework. Classical noise trackers such as MCRA and IMCRA fail when the interference consists of competing speakers, because their spectro-temporal behaviour is not distinguishable from that of the target by stationarity alone. A network instead learns the MC-SPP from the multichannel observation, and the resulting probabilities drive the estimation of the noise covariance and the acoustic transfer function that the beamformer needs. This keeps the interpretability of the statistical model while handling interference that varies over time.

Third, we replace the single speech/noise dichotomy by a per-source hypothesis set: instead of one presence probability for the target against a lumped interference covariance, the model carries one posterior per interferer, so that the interference is *resolved* into individual competing talkers rather than absorbed into a single noise term. The target remains a single designated source throughout—*multi-source* here refers to the resolution of the interference, not to multiple simultaneous targets. Throughout this progression, the MC-SPP serves as a unifying probabilistic interface between the data-driven front-end and the model-based spatial–spectral filtering back-end, ensuring that domain knowledge embodied in the beamformer design is preserved while the critical parameter estimation step benefits from the representational capacity of modern neural networks.

The remainder of this thesis is organised as follows. Chapter 2 reviews the relevant literature on statistical and DNN-based speech enhancement, multichannel beamforming, nonlinear spatial filtering, and speech presence probability estimation. Chapter 3 formulates the single- and multichannel signal models and derives the a posteriori speech presence probability in both settings. Chapter 4 develops the Bayesian MMSE framework for multi-talker target extraction. It introduces the multi-source signal model, states the distributional priors on the target speech and the interference, and casts the signal-activity space as a composite hypothesis structure. It also identifies the Gaussian-mixture nature of the interference as the origin of the nonlinearity in the MMSE-optimal estimator. Chapter 5 derives the composite-hypothesis estimator in closed form. The estimator is a set of MVDR–Wiener filters, one per interference hypothesis, each weighted by the corresponding multi-source MC-SPP posterior. The chapter also establishes its structural equivalence to the analytic nonlinear spatial filter of Tesch and Gerkmann. Chapter 6 presents the DNN-guided realisation: the two networks that estimate the posteriors, the model-based filtering stage they control, and the oracle-guided training and loss design. Chapter 7 reports the experimental evaluation: the simulated multi-talker corpus, the baselines, the metrics, and the results. Chapter 8 concludes the thesis and outlines directions for future work.

2

Related Works

This chapter surveys the literature that forms the context for the present thesis. We begin with single-channel statistical estimators and multichannel beamforming (§2.1–2.2), then review DNN-based approaches to speech enhancement and separation (§2.3), including the emerging class of generative diffusion models (§2.4). We subsequently discuss the MMSE-optimal nonlinear spatial filter theory that underpins this work (§2.5), and conclude with the speech presence probability framework (§2.6) and multi-source speech processing (§2.7).

2.1. Statistical Single-Channel Speech Enhancement

The statistical approach to single-channel speech enhancement models the DFT coefficients of clean speech and noise as independent random variables and derives estimators that minimise a given distortion criterion under these distributional assumptions.

The foundational work of Ephraim and Malah [3] derived the MMSE estimator of the short-time spectral amplitude (STSA) under a complex Gaussian prior for both speech and noise. Their subsequent log-spectral amplitude (LSA) estimator [4] minimises the expected log-spectral distortion, yielding reduced musical-noise artefacts. Both estimators require two quantities: the a priori SNR ξ and the a posteriori SNR γ, estimated recursively using the decision-directed approach [3].

Erkelens et al. [20] generalised these results by replacing the Gaussian (Rayleigh magnitude) speech prior with a generalised Gamma distribution parametrised by a shape factor ν. For $\nu < 1$, the resulting super-Gaussian prior better captures the heavy-tailed statistics of speech DFT coefficients [20]. The Gaussian case is recovered for $\nu = 1$, and the resulting MMSE estimator involves confluent hypergeometric functions——a mathematical structure that reappears in the multichannel nonlinear filter of Chapter 5.

Accurate noise PSD estimation is critical for all MMSE estimators. Martin [21] proposed the minimum statistics approach, which tracks the noise floor by following the minima of a smoothed periodogram within a sliding window. Cohen and Berdugo [6] improved on this with MCRA/IMCRA, which combines recursive averaging with a soft speech-presence indicator. Gerkmann and Hendriks [11] showed that an unbiased MMSE-optimal noise PSD estimator is obtained by weighting the recursive update with the speech presence probability, so that each time–frequency bin contributes to the noise estimate in proportion to its speech-absence probability; the estimator is discussed in Section 2.6.

For a comprehensive treatment of single-channel methods, we refer the reader to Loizou [1] and the survey of Hendriks et al. [2].

2.2. Multichannel Beamforming and Spatial Filtering

With regard to multichannel speech enhancement acoustic task, more information can be leveraged, in addition to temporal and spectral features. Beamforming exploits the spatial diversity offered by a

microphone array to separate sources based on their direction of arrival. Early adaptive designs include the linearly constrained minimum variance (LCMV) beamformer of Frost [22] and the generalised sidelobe canceller. Van Veen and Buckley [7] provided a widely cited tutorial that unified these designs within a constrained optimisation framework.

The MVDR beamformer [8] minimises the output noise power subject to a distortionless constraint on the target steering vector. Its weights are determined by the noise spatial covariance matrix $\mathbf{\Sigma}_N$ and the relative transfer function (RTF) $\boldsymbol{d}$ of the target source. Doclo and Moonen [23] proposed a GSVD-based approach that jointly diagonalises the speech and noise covariance matrices, providing a rank-reduced optimal filter. Souden et al. [8] showed that the MVDR beamformer, when followed by a single-channel Wiener post-filter, achieves the minimum output noise power among all distortionless filters; the multichannel Wiener filter (MWF) follows as a special case.

A central challenge in practical beamforming is the estimation of the required second-order statistics—the noise PSD matrix $\mathbf{\Sigma}_N$ and the speech PSD matrix $\mathbf{\Phi}_{xx}$—from the noisy observation. In the single-source case, Souden et al. [17] proposed an integrated framework in which a multichannel SPP drives recursive PSD tracking; the RTF is then estimated from the speech PSD matrix by covariance subtraction. This approach avoids explicit voice activity detection and instead provides a continuous, probabilistically motivated estimate of the spatial statistics required for beamformer computation.

For an overview of multichannel speech enhancement and source separation, we refer to the survey of Gannot et al. [14].

2.3. Discriminative Models for Speech Enhancement

The introduction of DNNs has transformed multichannel speech enhancement. Three broad families of approaches have emerged, which we review in turn.

2.3.1. Mask-Based Learning

Mask-based methods train a DNN to estimate a time–frequency mask that, when applied to the mixture spectrogram or used to compute spatial statistics, yields the enhanced speech. Wang and Chen [12] provided a comprehensive overview distinguishing ideal binary masks (IBM), ideal ratio masks (IRM), phase-sensitive masks (PSM) [24], and complex ideal ratio masks (cIRM). The multichannel use of such masks—estimating the spatial covariance matrices that parameterise a beamformer—is treated separately in Section 2.3.3.

The paradigm applies equally to single-channel enhancement. Hu et al. [25] proposed the deep complex convolution recurrent network (DCCRN), which operates on both the real and imaginary parts of the STFT to estimate a complex ratio mask, preserving phase information.

A related line replaces the mask by a direct regression: the network maps the corrupted spectrum or waveform to the clean one, so that magnitude and phase are produced jointly and no masking constraint is imposed [26]. Mapping avoids the implicit bound that a bounded mask places on the achievable gain, at the cost of a harder regression target.

2.3.2. End-to-End Time-Domain Models

A parallel line of research bypasses the STFT entirely and operates directly on time-domain waveforms. Luo and Mesgarani [27] introduced Conv-TasNet (Fig. 2.1), which uses a learned encoder, a temporal convolutional network for mask estimation, and a learned decoder, achieving state-of-the-art separation on the WSJ0-2mix benchmark.

Luo et al. [28] subsequently proposed the dual-path RNN (DPRNN), which models both local and global temporal dependencies through interleaved intra-chunk and inter-chunk recurrent layers. Subakan et al. [29] replaced the recurrent layers with self-attention, yielding the SepFormer architecture.

These end-to-end models achieve impressive objective scores but do not naturally incorporate spatial information from multiple microphones, and their internal representations are difficult to interpret in terms of classical signal processing concepts such as beamforming, noise statistics, or source presence.

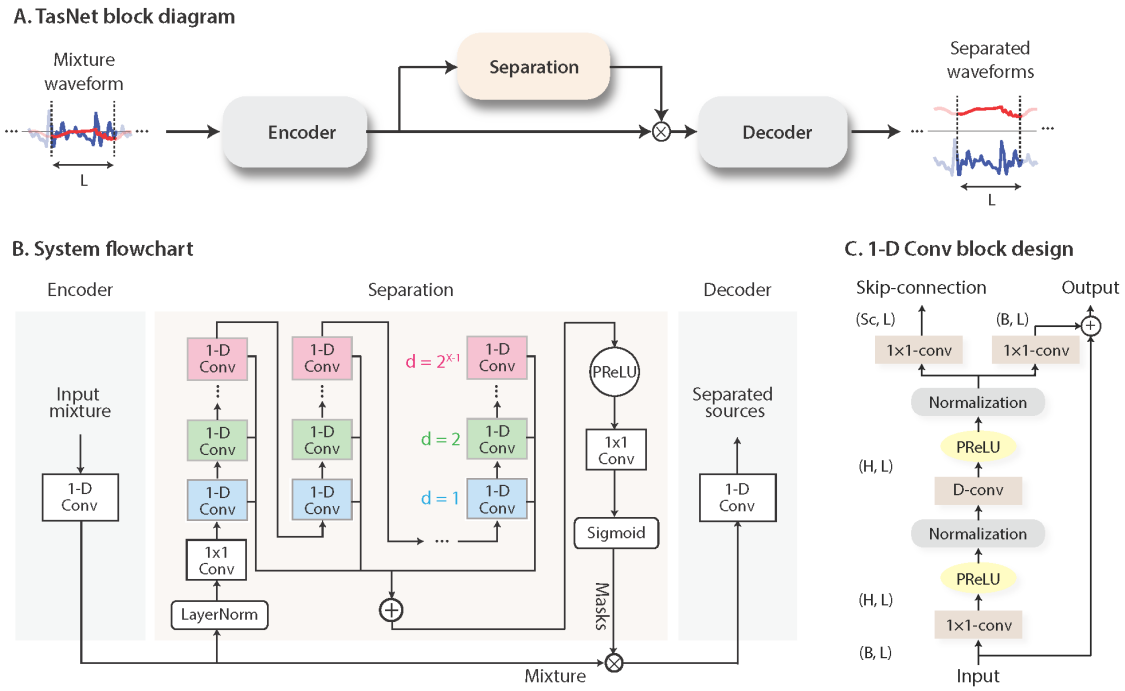


Figure 2.1: Architecture of Conv-TasNet [27]. (A) High-level block diagram: a learned encoder maps mixture waveform segments to a latent representation, a separation module estimates per-source masks, and a learned decoder reconstructs the individual waveforms. (B) Detailed system flowchart with the temporal convolutional network (TCN) separator. (C) Design of the 1-D convolutional block with dilated depthwise convolution, skip connections, and residual paths. The architecture operates on a single channel and contains no explicit beamformer structure or spatial model. Reproduced from [27]

2.3.3. DNN-Guided Beamforming

DNN-guided (or hybrid) methods retain the model-based beamformer structure but use a neural network for parameter estimation. Beyond the mask-based PSD estimation of Heymann et al. [15], several variants have been explored: direct RTF estimation by the network, joint mask-and-beamformer training with differentiable MVDR layers, and neural architectures that combine spatial and spectral features [14].

Tao et al. [18] proposed a two-stage DNN-guided pipeline in which a first network estimates the MC-SPP for MVDR beamforming, and a second network estimates a single-channel SPP for MMSE-LSA post-filtering. Their work demonstrated that the MC-SPP provides a probabilistically motivated interface between the DNN and the classical beamformer, yielding better noise tracking than conventional mask-based PSD estimation.

However, their formulation is restricted to a single target source with stationary noise, and the MC-SPP is used solely as a soft weighting factor for PSD updates rather than as a posterior probability within a Bayesian estimator.

The present thesis extends this line of work by deriving the multi-source MC-SPP from the Bayesian MMSE framework, establishing its role as the posterior weight in the MMSE-optimal nonlinear spatial filter, and using a DNN to predict it for all sources simultaneously.

2.4. Generative Models for Speech Enhancement

In contrast to the discriminative (predictive) approaches described above, generative models learn the distribution of clean speech and perform enhancement by sampling from the posterior distribution conditioned on the noisy observation. This paradigm offers two key advantages: the ability to generate multiple plausible estimates (capturing uncertainty), and the natural integration of Bayesian inference tools for posterior sampling.

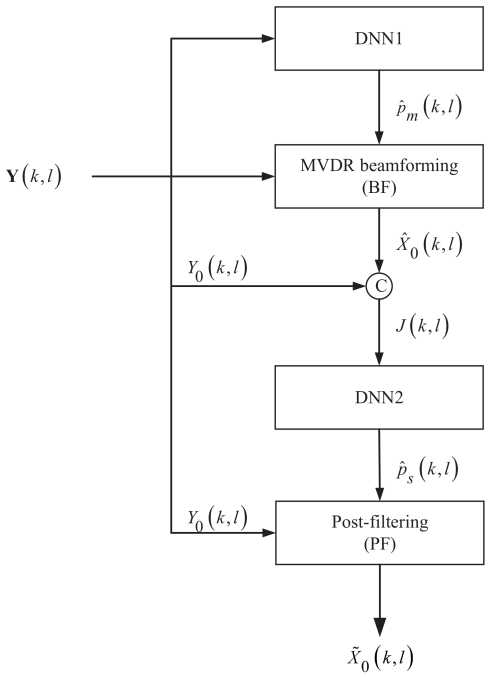


Figure 2.2: Two-stage DNN-guided multichannel speech enhancement pipeline of Tao et al. [30]. Stage 1: a DNN estimates the MC-SPP $\hat{p}_m(k,l)$, which drives recursive PSD updates and MVDR beamforming. Stage 2: the beamformer output is concatenated with the reference microphone signal and fed to a second DNN that estimates a single-channel SPP $\hat{p}_s(k,l)$ for MMSE-LSA post-filtering. The present thesis extends this single-source framework to the multi-source setting by replacing the scalar MC-SPP with a per-source posterior vector derived from the Bayesian MMSE expansion. Reproduced from [30].

Diffusion models [31, 32] have emerged as a particularly powerful class of generative models for audio. They define a forward process that gradually corrupts data with Gaussian noise, and a reverse process—parametrised by a learned score model—that iteratively denoises the signal. Welker et al. [33] first applied score-based diffusion to speech enhancement in the complex STFT domain; Richter et al. [34] extended this to the SGMSE+ framework for joint enhancement and dereverberation. Lemercier et al. [35] proposed StoRM, a stochastic regeneration model that initialises the diffusion process with the output of a predictive network, significantly reducing the number of sampling steps. A comprehensive review of diffusion models for audio restoration is provided by Lemercier et al. [36].

From the perspective of the present thesis, generative models are relevant for two reasons. First, diffusion-based enhancement operates within a probabilistic (Bayesian) framework, using Bayes' rule to combine a learned prior with a measurement likelihood; this parallels the Bayesian MMSE derivation that underpins our multi-source MC-SPP. Second, the output of a diffusion model is inherently stochastic, providing a natural measure of uncertainty—analogous to the per-bin posterior probabilities produced by our MC-SPP estimator, which quantify the confidence of source-activity decisions at each time–frequency point. Unlike diffusion models, however, our framework preserves the deterministic, analytically grounded structure of classical beamforming and post-filtering, and the probabilistic output (the MC-SPP) has a direct physical interpretation as a source-presence probability rather than an implicit latent variable.

2.5. MMSE-Optimal Nonlinear Spatial Filtering

The theoretical foundation for nonlinear spatial filtering was laid by Hendriks et al. [10], who proved that the multichannel MMSE-optimal estimator decomposes into a linear MVDR beamformer followed by a single-channel MMSE post-filter *if and only if* the noise DFT coefficients are Gaussian distributed. For non-Gaussian noise—in particular, for the Gaussian mixture that arises when multiple interferers are present under a sparsity assumption—the optimal estimator is a nonlinear function of the multichannel observation. This result was derived under a generalised Gamma prior for the speech magnitude,

which includes the Gaussian (Rayleigh) case as a special instance [20].

Balan and Rosca [37] had earlier shown, in a simpler setting, that the multichannel Bayesian estimator factors into an MVDR beamformer and a post-filter when the noise is Gaussian. The contribution of Hendriks et al. [10] was to prove this as a necessary and sufficient condition and to provide the general (nonlinear) estimator for arbitrary noise distributions.

Tesch and Gerkmann [9] provided the first comprehensive empirical investigation of the nonlinear spatial filter. Using an oracle implementation with known signal statistics, they demonstrated that the analytic MMSE estimator with a super-Gaussian speech prior ($\nu = 0.25$) can suppress more than $M-1$ directional interferers with an M-element array, outperforming the linear MVDR by a significant margin. The physical mechanism is that the estimator implicitly selects, at each time–frequency bin, the MVDR beamformer whose noise covariance matrix places a null towards the interferer that is actually active; the "selection" is performed probabilistically through the posterior hypothesis weights.

In subsequent work, Tesch and Gerkmann proposed to learn the nonlinear filter directly with a DNN. Their *joint nonlinear filter* (JNF) [13](Fig. 2.3) predicts a complex mask that implicitly captures the nonlinear spatial–spectral processing.

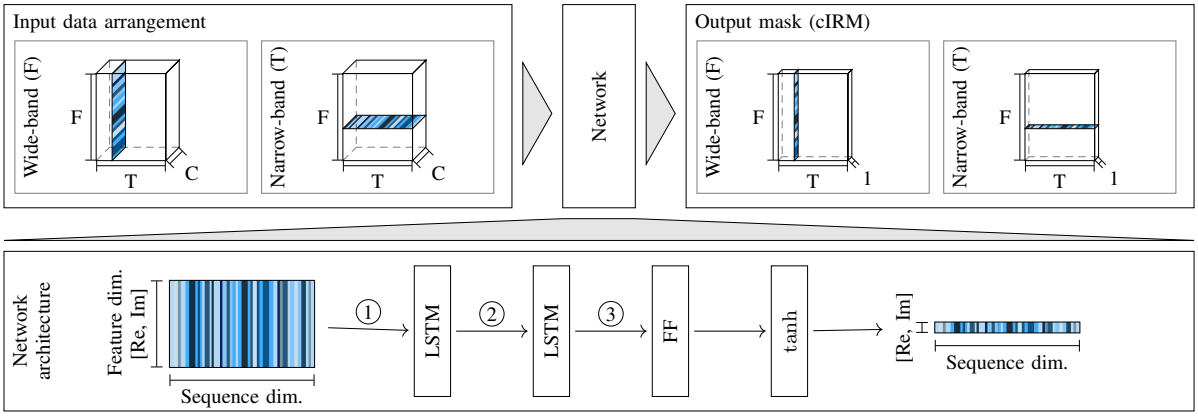


Figure 2.3: Base network architecture of the joint nonlinear filter (JNF) [13], whose backbone topology is also used by the posterior network of Chapter 6. The multichannel STFT input is arranged either in wide-band (F) or narrow-band (T) format and processed by two LSTM layers followed by a feed-forward layer with $\tanh$ activation, producing a complex ideal ratio mask (cIRM). Reproduced from [13].

In their successive work, they establish the *spatially selective filter* (SSF) [38, 19] that conditions the mask on a target direction, enabling speaker extraction. The structure is shown in Fig. 2.4 .

These methods achieve excellent separation quality and confirmed the practical relevance of nonlinear spatial filtering, but they do not preserve the decomposition into beamforming, post-filtering, and hypothesis weighting that the analytic MMSE estimator reveals. The present thesis can be seen as occupying the middle ground between the oracle analytic estimator of [9] and the end-to-end learned filter of [19]: it retains the beamformer structure of the former while employing a DNN for the hypothesis weights, as in the latter.

2.6. Speech Presence Probability Estimation

The concept of speech presence probability was introduced by McAulay and Malpass [39] in the context of noise spectral subtraction and formalised by Malah et al. [16] for tracking speech-presence uncertainty in non-stationary noise environments. In the STFT domain, the SPP at each time–frequency bin is the posterior probability $p(\mathcal{H}_1 \mid Y)$ that speech is present, computed via Bayes' rule from the likelihoods under the speech-present and speech-absent hypotheses.

Gerkmann and Hendriks [11] showed that using a *fixed* a priori SNR for the SPP computation—rather than an adaptively estimated value—yields an unbiased noise power estimator with low complexity. The fixed-prior approach decouples the noise estimator from the post-filter, avoiding the positive feedback

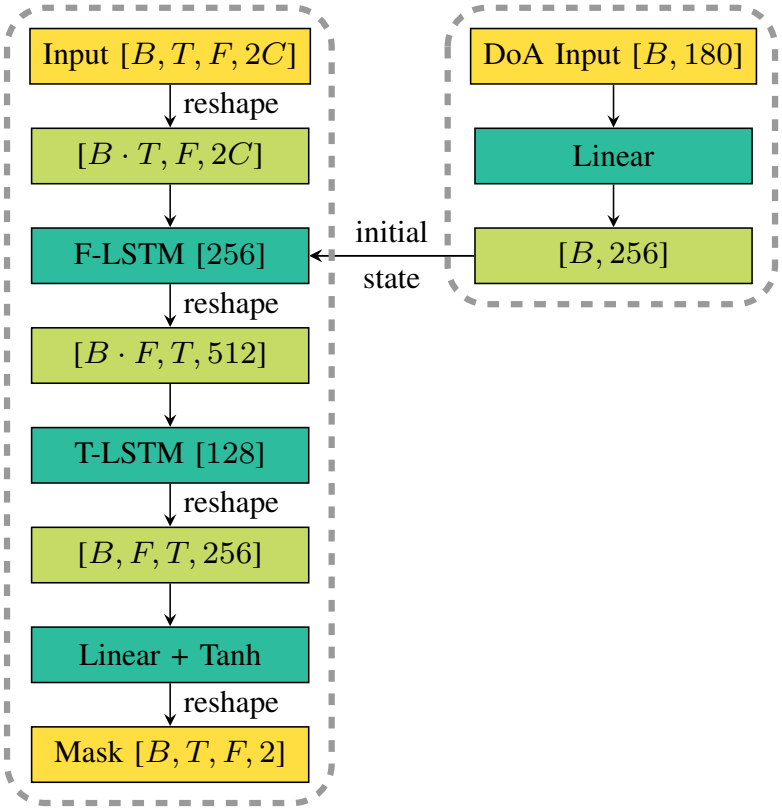


Figure 2.4: Spatially selective filter (SSF) architecture from [19]. The target direction-of-arrival (DoA) is encoded via a linear layer and injected as the initial hidden state of the F-LSTM, enabling the same network to extract different speakers by conditioning on their spatial location. The present thesis replaces the implicit nonlinear mask output with an explicit MC-SPP posterior that drives a classical MVDR–Wiener cascade. Reproduced from [19].

loop that plagues adaptive methods and providing a theoretically clean probabilistic interface for noise PSD tracking.

Souden et al. [17] generalised the SPP to the multichannel setting (MC-SPP) by formulating the likelihood functions under a multivariate complex Gaussian signal model. The resulting MC-SPP incorporates spatial information through the noise spatial covariance matrix and the target RTF, providing a detection statistic that is proportional to the squared output of an MVDR beamformer steered towards the target. Their integrated framework uses the MC-SPP to recursively update both the noise and speech PSD matrices, from which the steering vector and beamformer weights are subsequently computed. The MC-SPP thus serves as a unified probabilistic interface between spatial statistics estimation and beamformer design.

The MC-SPP formulation of [17] considers only two hypotheses (speech present or absent) for a single target source. When multiple sources are present, a natural extension is to define I mutually exclusive hypotheses—one for each source plus a noise-only state—and to derive the posterior probability for each. Such an extension is developed in Chapter 5.

2.7. Multi-Source Speech Separation and the W-Disjoint Orthogonality Assumption

Multi-talker speech separation aims to recover each individual speaker's signal from a mixture. Early approaches relied on computational auditory scene analysis (CASA) principles, grouping time–frequency regions by pitch and onset cues. The introduction of deep learning methods has dramatically advanced the field.

Hershey et al. [40] proposed deep clustering, which embeds each time–frequency bin into a high-dimensional space where bins belonging to the same speaker are close together; spectral masks are

then obtained by clustering the embeddings. Kolbæk et al. [41] addressed the permutation ambiguity inherent in multi-speaker separation by introducing utterance-level permutation invariant training (uPIT), which considers all possible label assignments and selects the one with the lowest loss. These techniques enabled the training of single-channel separation models such as Conv-TasNet [27] and DPRNN [28].

In the multichannel setting, spatial information provides a powerful cue for separation. Yoshioka et al. [42] proposed a multi-microphone neural separation system that combines spectral and spatial features for far-field multi-talker ASR. Chen et al. [43] introduced continuous speech separation (CSS), addressing the more realistic scenario of overlapping utterances of variable duration. These multichannel approaches typically operate end-to-end, mapping the multi-microphone input directly to separated waveforms or masks without explicit beamformer structure.

W-disjoint orthogonality (WDO), introduced by Yilmaz and Rickard [44], posits that the STFT representations of distinct speech sources have approximately disjoint supports: at any given time–frequency bin, at most one source contributes significant energy. This sparsity property is well supported empirically for speech signals and underpins ideal binary mask estimation, sparse demixing algorithms, and the single-dominant-source hypothesis structure used in the multi-source MC-SPP derivation of Chapter 5.

The WDO assumption converts the multi-source problem into a hypothesis testing problem: at each time–frequency bin, the system must decide which of the I sources (or none) is active. The posterior probabilities for these hypotheses—the multi-source MC-SPP—then drive source-specific beamformers that place spatial nulls towards the identified interferers. This probabilistic formulation stands in contrast to end-to-end separation methods, which learn the assignment implicitly, and to binary masking approaches, which make hard decisions. The soft, posterior-probability-based assignment preserves the full uncertainty of the source-activity decision and, as shown in Chapter 5, coincides with the weights prescribed by the Bayesian MMSE-optimal nonlinear spatial filter.

2.8. Positioning of This Thesis

The literature reviewed above leaves the multi-talker case between two incomplete answers. The MMSE-optimal nonlinear spatial filter of Section 2.5 is known in closed form, and its posterior weights are source-activity probabilities; but the deep realisations of that filter predict a mask directly and expose neither the beamformer nor the posteriors. The MC-SPP literature of Section 2.6 provides exactly such posteriors—and Tao et al. [18] showed they can be learned—but only for a single target against a covariance that lumps all interference together, so the decision that makes the multi-talker estimator nonlinear, namely which competing talker dominates a given time–frequency bin, is never represented.

This thesis joins the two. It extends the MC-SPP of [17] to a per-source hypothesis set, under the WDO assumption of Section 2.7. It then shows that the resulting posteriors are precisely the weights of the nonlinear filter of [10, 9], so that the network is trained on a quantity that has both a statistical meaning and a defined role in the optimal estimator. Finally, it carries those posteriors through to the recursive PSD tracking that parameterises the filters, generalising the single-source scheme of [17]. Unlike the generative approaches of Section 2.4, the resulting estimator remains deterministic and analytically grounded, and its probabilistic output is a source-presence probability with a direct physical reading rather than an implicit latent variable. The derivation occupies Chapters 4 and 5; the realisation and its evaluation, Chapters 6 and 7.

3

Signal Models and Speech Presence Probability

This chapter establishes the mathematical signal models and derives the a posteriori speech presence probability that underpins the remainder of the thesis. We proceed from the single-channel case to the multichannel setting, and in each case first formulate the signal model and then derive the corresponding SPP.

3.1. Single-Channel Signal Model

Let $s(t)$ denote a clean speech signal observed through a single microphone in the presence of additive noise $n(t)$. The noisy observation is $y(t) = s(t)+n(t)$. Applying the short-time Fourier transform (STFT) with frame index l and frequency bin index k, the narrow-band additive model becomes

$$Y(k,l) = S(k,l) + N(k,l). \tag{3.1}$$

Under the standard assumption that the DFT coefficients of speech and noise are mutually independent, zero-mean, complex Gaussian random variables, the power spectral density (PSD) of the observation decomposes as

$$\Phi_{yy}(k,l) = \Phi_{ss}(k,l) + \Phi_{nn}(k,l), \tag{3.2}$$

where Φ_{ss} and Φ_{nn} denote the speech and noise PSDs, respectively. Defining the *a priori* and *a posteriori* signal-to-noise ratios as

$$\xi(k,l) \triangleq \frac{\Phi_{ss}(k,l)}{\Phi_{nn}(k,l)}, \qquad \gamma(k,l) \triangleq \frac{|Y(k,l)|^2}{\Phi_{nn}(k,l)}, \tag{3.3}$$

the Wiener filter takes the familiar form

$$G_{\mathrm{Wiener}}(k,l) = \frac{\xi(k,l)}{1+\xi(k,l)}. \tag{3.4}$$

3.2. Single-Channel Speech Presence Probability

Both the Wiener filter and the more advanced MMSE-LSA estimator [4] require an accurate estimate of the noise PSD Φ_{nn}, which must be continuously tracked from the noisy observation. The a posteriori speech presence probability provides an elegant probabilistic mechanism for this purpose.

3.2.1. Two-State Hypothesis Model

At each time–frequency point (k,l), we define two mutually exclusive hypotheses:

$$\begin{cases} \mathcal{H}_0: & Y(k,l) = N(k,l), \\ \mathcal{H}_1: & Y(k,l) = S(k,l) + N(k,l). \end{cases} \tag{3.5}$$

Under the complex Gaussian assumption, the likelihood functions under the two hypotheses are

$$p(Y \mid \mathcal{H}_0) = \frac{1}{\pi \Phi_{nn}} \exp\!\left(-\frac{|Y|^2}{\Phi_{nn}} \right), \tag{3.6}$$

$$p(Y \mid \mathcal{H}_1) = \frac{1}{\pi (\Phi_{ss} + \Phi_{nn})} \exp\!\left(-\frac{|Y|^2}{\Phi_{ss} + \Phi_{nn}} \right). \tag{3.7}$$

3.2.2. Derivation of the Single-Channel SPP

Applying Bayes' rule, the a posteriori speech presence probability is

$$p(\mathcal{H}_1 \mid Y) = \frac{P(\mathcal{H}_1)\, p(Y \mid \mathcal{H}_1)}{P(\mathcal{H}_0)\, p(Y \mid \mathcal{H}_0) + P(\mathcal{H}_1)\, p(Y \mid \mathcal{H}_1)}. \tag{3.8}$$

Substituting (3.6) and (3.7) and defining $\beta_{\mathrm{sc}} = P(\mathcal{H}_0)/P(\mathcal{H}_1)$ yields

$$p(\mathcal{H}_1 \mid Y) = \left[1 + \beta_{\mathrm{sc}}\,(1 + \xi_{\mathcal{H}_1})\, \exp\!\left(-\frac{|Y|^2}{\Phi_{nn}} \cdot \frac{\xi_{\mathcal{H}_1}}{1 + \xi_{\mathcal{H}_1}} \right) \right]^{-1}, \tag{3.9}$$

where $\xi_{\mathcal{H}_1}$ denotes the a priori SNR under hypothesis $\mathcal{H}_1$.

Gerkmann and Hendriks [11] showed that using a fixed a priori SNR $\xi_{\mathcal{H}_1}$ representing the SNR typical of speech presence, rather than an adaptively estimated value, yields an unbiased noise power estimator with low complexity. The fixed-prior approach decouples the noise power estimator from subsequent processing stages and avoids the pathological overlap of the two likelihood functions in speech-absent regions that plagues adaptive methods [11].

3.2.3. SPP-Based Noise Power Estimation

With the SPP estimate, the noise PSD can be updated without imposing a hard voice activity decision. The MMSE-optimal noise periodogram estimator under speech presence uncertainty is:

$$\mathbb{E}\big[|N|^2 \mid Y \big] = \big(1 - p(\mathcal{H}_1 \mid Y) \big)\, |Y|^2 + p(\mathcal{H}_1 \mid Y)\, \hat{\Phi}_{nn}, \tag{3.10}$$

which replaces the hard decision of a voice activity detector by a soft weighting between the current noisy periodogram and the previous noise PSD estimate. The noise PSD is then obtained by recursive temporal smoothing:

$$\hat{\Phi}_{nn}(k,l) = \alpha_s\, \hat{\Phi}_{nn}(k, l{-}1) + (1 - \alpha_s)\, \mathbb{E}\big[|N(k,l)|^2 \mid Y(k,l) \big], \tag{3.11}$$

where $0 < \alpha_s < 1$ is a smoothing factor.

3.3. Multichannel Signal Model

Consider an M-element microphone array receiving a single target source in a reverberant enclosure. Denoting the acoustic transfer function (ATF) from the source to the m-th microphone as $A_m(k)$, the multichannel observation vector in the STFT domain is

$$\boldsymbol{y}(k,l) = \boldsymbol{a}(k)\, S(k,l) + \boldsymbol{n}(k,l), \tag{3.12}$$

where $\boldsymbol{y}(k,l) = [Y_1(k,l), \ldots, Y_M(k,l)]^T \in \mathbb{C}^M$, the relative transfer function is $\boldsymbol{a}(k) = [A_1(k), \ldots, A_M(k)]^T$, and $\boldsymbol{n}(k,l)$ represents the multichannel noise.

The ATF carries no frame index, which encodes the assumption that the source positions and the room impulse responses are time-invariant over the observation interval. This static-source assumption is maintained throughout the thesis and is reflected in the fixed-position simulated scenes of Chapter 7.

Working with the relative transfer function (RTF) with respect to a reference microphone m_{ref},

$$\boldsymbol{d}(k) \triangleq \frac{\boldsymbol{a}(k)}{A_{m_{\mathrm{ref}}}(k)}, \tag{3.13}$$

so that $d_{m_{\mathrm{ref}}}(k) = 1$, the model becomes

$$\boldsymbol{y}(k,l) = \boldsymbol{d}(k)\, S(k,l) + \boldsymbol{n}(k,l), \tag{3.14}$$

where $S(k,l) \triangleq A_{m_{\mathrm{ref}}}(k)\, S_{\mathrm{src}}(k,l)$ is the reverberant speech image at the reference microphone, with instantaneous PSD $\sigma_S^2(k,l) = \mathbb{E}[|S(k,l)|^2]$. The noise cross-PSD matrix is $\boldsymbol{\Sigma}_N(k,l) = \mathbb{E}\big[\boldsymbol{n}(k,l)\,\boldsymbol{n}^H(k,l)\big]$, and the observation cross-PSD matrix decomposes as

$$\boldsymbol{\Phi}_{yy}(k,l) = \sigma_S^2(k,l)\,\boldsymbol{d}\,\boldsymbol{d}^H + \boldsymbol{\Sigma}_N(k,l). \tag{3.15}$$

3.4. Multichannel Speech Presence Probability

The single-channel SPP exploits only temporal–spectral information to distinguish speech from noise. When multiple microphones are available, the spatial information encoded in the inter-microphone phase and level differences can additionally be leveraged. Souden et al. [17] generalised the SPP to the multichannel setting by formulating the likelihood functions under the multivariate Gaussian model.

3.4.1. Multichannel Likelihood Functions

Under the two-state model for the single-source multichannel case, the observation vector is distributed as

$$\mathcal{H}_0: \quad \boldsymbol{y} \sim \mathcal{CN}(\boldsymbol{0},\, \boldsymbol{\Sigma}_N), \tag{3.16}$$

$$\mathcal{H}_1: \quad \boldsymbol{y} \sim \mathcal{CN}\big(\boldsymbol{0},\, \sigma_S^2\,\boldsymbol{d}\,\boldsymbol{d}^H + \boldsymbol{\Sigma}_N\big). \tag{3.17}$$

The corresponding multivariate likelihood functions are

$$p(\boldsymbol{y} \mid \mathcal{H}_0) = \frac{1}{\pi^M |\boldsymbol{\Sigma}_N|} \exp\!\big(-\boldsymbol{y}^H \boldsymbol{\Sigma}_N^{-1}\, \boldsymbol{y}\big), \tag{3.18}$$

$$p(\boldsymbol{y} \mid \mathcal{H}_1) = \frac{1}{\pi^M |\boldsymbol{\Sigma}_N + \sigma_S^2\,\boldsymbol{d}\,\boldsymbol{d}^H|} \exp\!\big(-\boldsymbol{y}^H (\boldsymbol{\Sigma}_N + \sigma_S^2\,\boldsymbol{d}\,\boldsymbol{d}^H)^{-1}\, \boldsymbol{y}\big). \tag{3.19}$$

3.4.2. Derivation of the MC-SPP

Define the multichannel a priori SNR and the detection statistic:

$$\xi(k,l) \triangleq \mathrm{tr}\big(\boldsymbol{\Sigma}_N^{-1}\,\sigma_S^2\,\boldsymbol{d}\,\boldsymbol{d}^H\big) = \boldsymbol{d}^H \boldsymbol{\Sigma}_N^{-1} \boldsymbol{d}\; \sigma_S^2(k,l), \tag{3.20}$$

$$\beta(k,l) \triangleq \boldsymbol{y}^H \boldsymbol{\Sigma}_N^{-1}\, (\sigma_S^2\,\boldsymbol{d}\,\boldsymbol{d}^H)\, \boldsymbol{\Sigma}_N^{-1}\, \boldsymbol{y}. \tag{3.21}$$

The multichannel a priori SNR ξ is also the theoretical output SNR of the parameterised multichannel Wiener filter [8]. Applying Bayes' rule and using the matrix determinant lemma to simplify the likelihood ratio yields the a posteriori MC-SPP [17]:

$$\boxed{p(\mathcal{H}_1 \mid \boldsymbol{y}) = \left[1 + \frac{q(k,l)}{1 - q(k,l)}\,\big(1 + \xi(k,l)\big)\,\exp\!\left(-\frac{\beta(k,l)}{1 + \xi(k,l)}\right)\right]^{-1}} \tag{3.22}$$

where $q(k,l)$ is the a priori speech absence probability (SAP). It is not a fixed model prior. Following [18], it is evaluated at each bin as an energy ratio between the target and the interference-plus-noise spatial covariances,

$$q(k,l) = 1 - \frac{\mathrm{tr}\big[\boldsymbol{\Phi}_{xx}(k,l)\big]}{\mathrm{tr}\big[\boldsymbol{\Phi}_{xx}(k,l)\big] + \mathrm{tr}\big[\boldsymbol{\Phi}_{vv}(k,l)\big]}. \tag{3.23}$$

so that q is close to one where the interference dominates and close to zero where the target does. Both covariances are themselves estimated from the observation, so (3.23) makes the prior data-dependent; this is a departure from a strict Bayesian prior, adopted here to follow the convention of [17, 18].

The detection statistic β can equivalently be written as $\beta = |\boldsymbol{d}^H \boldsymbol{\Sigma}_N^{-1} \boldsymbol{y}|^2\, \sigma_S^2$, revealing that it is proportional to the squared output of an MVDR beamformer steered towards the target, scaled by the source PSD—a quantity that is inherently spatial in nature.

Souden et al. [17] showed that when the noise emanates from a point source, perfect speech detection is theoretically possible, and when the noise is spatially white, a coherent summation of the speech components across microphones enhances the detection accuracy.

3.4.3. MC-SPP-Guided Parameter Estimation

With the MC-SPP $\hat{p}(k,l) \triangleq p(\mathcal{H}_1 \mid \boldsymbol{y}(k,l))$ evaluated from (3.22), the noise and speech cross-PSD matrices are updated recursively [17]:

$$\hat{\boldsymbol{\Sigma}}_N(k,l) = \alpha_v \, \hat{\boldsymbol{\Sigma}}_N(k,l-1) + (1-\alpha_v)\big(1-\hat{p}(k,l)\big)\, \boldsymbol{y}\boldsymbol{y}^H, \tag{3.24}$$

$$\hat{\boldsymbol{\Phi}}_{xx}(k,l) = \alpha_x \, \hat{\boldsymbol{\Phi}}_{xx}(k,l-1) + (1-\alpha_x)\, \hat{p}(k,l)\, \boldsymbol{y}\boldsymbol{y}^H, \tag{3.25}$$

where $\alpha_v, \alpha_x \in (0,1)$ are temporal smoothing factors. The relative transfer function is estimated via the covariance subtraction method (The alternative covariance-whitening method is more accurate, but it needs a generalised eigenvalue decomposition at every bin [45], here we adopt the subtraction method for simplicity) :

$$\hat{\boldsymbol{d}}(k,l) = \frac{\hat{\boldsymbol{\Phi}}_{xx}(k,l)\, \boldsymbol{u}_1}{\boldsymbol{u}_1^H \, \hat{\boldsymbol{\Phi}}_{xx}(k,l)\, \boldsymbol{u}_1}, \tag{3.26}$$

where $\boldsymbol{u}_1 = [1, 0, \ldots, 0]^T$ selects the reference channel. The MVDR beamformer weights are then

$$\boldsymbol{w}_{\mathrm{MVDR}}(k,l) = \frac{\hat{\boldsymbol{\Sigma}}_N^{-1}\, \hat{\boldsymbol{d}}}{\hat{\boldsymbol{d}}^H \, \hat{\boldsymbol{\Sigma}}_N^{-1}\, \hat{\boldsymbol{d}}}. \tag{3.27}$$

The MC-SPP thus serves as a unified probabilistic interface that drives the downstream parameter estimation required by the classical MVDR beamforming and post-filtering pipeline.

4

A Bayesian MMSE Framework for Target Speech Extraction

In realistic multi-talker acoustic scenes, the single-source speech presence probability developed in Chapter 3 is insufficient: competing speakers introduce highly non-stationary spectro-temporal energy patterns that cannot be absorbed into a single noise covariance matrix. The enhancement objective in such scenarios is *target speech extraction*—isolating a single designated speaker from a mixture of interferers and background noise. This task demands a signal-activity model that accounts for both the presence of the target and the activity pattern of every competing source.

This chapter constructs the Bayesian MMSE framework that underpins the composite-hypothesis estimator derived in Chapter 5. The starting point is the total-probability expansion of the conditional MMSE estimator over a set of mutually exclusive hypotheses indexing the signal-activity configuration at each time–frequency bin. We develop the multi-source signal model, state the distributional priors on speech and interference, and identify the Gaussian-mixture nature of the resulting interference distribution as the fundamental source of nonlinearity in the MMSE-optimal multichannel estimator. A composite hypothesis structure is then introduced that factorises the signal-activity space into two independent axes—target presence and interferer dominance—yielding the Cartesian-product expansion from which the closed-form estimator of Chapter 5 follows.

Throughout this chapter, the frequency-bin and time-frame indices (k, l) are suppressed unless required for disambiguation.

4.1. Multi-Source Signal Model

Consider an M-element microphone array receiving I point speech sources in a reverberant enclosure with additive noise. Source $i \in \{1, \dots, I\}$ is characterised by its relative transfer function (RTF) $\boldsymbol{b}_i \in \mathbb{C}^M$ with respect to a reference microphone m_{ref}, and by the STFT coefficient $V_i \in \mathbb{C}$ of its image at the reference microphone, with instantaneous power spectral density (PSD) $\sigma_{V_i}^2 \triangleq \mathbb{E}[|V_i|^2]$. The multichannel observation is

$$\boldsymbol{y} \;=\; \sum_{i=1}^{I} \boldsymbol{b}_i \, V_i \;+\; \boldsymbol{n}, \tag{4.1}$$

with additive noise $\boldsymbol{n}(k,l) \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{\Sigma}_N(k))$ assumed zero-mean, uncorrelated across STFT frames, wide-sense stationary over the frames used to estimate its statistics and statistically independent of all speech sources. The noise cross-PSD matrix $\boldsymbol{\Sigma}_N$ is generally full-rank and captures both diffuse ambient noise and uncorrelated sensor noise.

When a specific source—say source j—is designated as the target, it is convenient to partition (4.1) as

$$\boldsymbol{y} \;=\; \underbrace{\boldsymbol{b}_j\, V_j}_{\text{target}} \;+\; \underbrace{\sum_{\substack{i=1\\ i\neq j}}^{I} \boldsymbol{b}_i\, V_i \;+\; \boldsymbol{n}}_{\triangleq\, \boldsymbol{u}}, \tag{4.2}$$

with $\boldsymbol{u}$ collecting all undesired components (the index j is fixed once the target is designated and is therefore suppressed). Writing $\boldsymbol{d} \triangleq \boldsymbol{b}_j$ for the target RTF and $S \triangleq V_j$ for the target STFT coefficient, the model takes the compact target-centric form

$$\boldsymbol{y} \;=\; \boldsymbol{d}\, S \;+\; \boldsymbol{u}, \tag{4.3}$$

where $\boldsymbol{u}$ collects the $I-1$ interferers and noise. This target-centric formulation is the basis for the composite-hypothesis estimator developed in Chapter 5.

4.2. The Single-Dominant-Interferer Approximation

The hypothesis structure adopted in this chapter rests on the W-disjoint orthogonality (WDO) of speech reviewed in Section 2.7. The physical basis is the harmonic structure of voiced speech: energy concentrates on the harmonics of a speaker-specific fundamental, and simultaneous talkers differ in fundamental frequency, phonemic content and temporal envelope, so the TF regions in which each talker contributes appreciable energy are largely non-overlapping (Fig. 4.1). At a given bin, at most one interferer is therefore taken to be dominant. The approximation has identifiable failure modes: low frequencies, where the harmonic spacing is dense; consonant–vowel transitions; and bins where two talkers share a fundamental frequency or a formant. At exactly those bins the Bayesian formulation degrades gracefully, in the sense that the posteriors derived in Chapter 5 are continuous in $[0,1]$ and spread probability across hypotheses instead of committing to one.

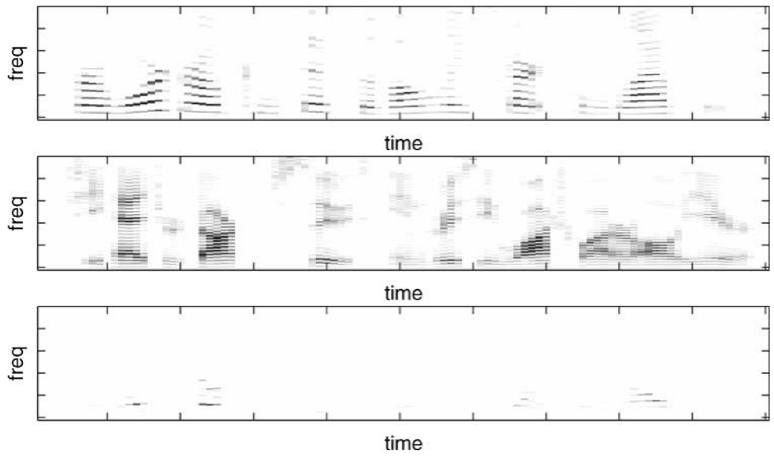


Figure 4.1: Illustration of approximate W-disjoint orthogonality for two speech signals. Top: spectrogram magnitude $|\hat{s}_1(\tau,\omega)|$ of speaker 1. Middle: $|\hat{s}_2(\tau,\omega)|$ of speaker 2. Bottom: pointwise product $|\hat{s}_1(\tau,\omega)\hat{s}_2(\tau,\omega)|$, which is nearly zero everywhere—the time–frequency regions carrying significant energy for each speaker are approximately disjoint. This property motivates the single-dominant-interferer hypothesis adopted in Section 4.5. Reproduced from [44].

4.3. The Bayesian MMSE Expansion

Let $h(S)$ denote an arbitrary functional of the target speech coefficient—for instance, $h(S)=S$ for the complex DFT coefficient or $h(S)=|S|$ for the spectral magnitude. For any mutually exclusive and jointly exhaustive hypothesis set $\{\mathcal{H}_0,\mathcal{H}_1,\dots,\mathcal{H}_P\}$ indexing the signal-activity configuration at the current TF bin, the conditional expectation admits the marginalisation

$$\mathbb{E}[\,h(S)\mid \boldsymbol{y}\,] \;=\; \sum_{i=0}^{P} \mathbb{E}[\,h(S)\mid \boldsymbol{y},\,\mathcal{H}_i\,]\; p(\mathcal{H}_i\mid \boldsymbol{y}), \tag{4.4}$$

with posterior weights given by Bayes' rule,

$$p(\mathcal{H}_i \mid \boldsymbol{y}) \;=\; \frac{p(\boldsymbol{y} \mid \mathcal{H}_i)\, P(\mathcal{H}_i)}{\sum_{m=0}^{P} p(\boldsymbol{y} \mid \mathcal{H}_m)\, P(\mathcal{H}_m)}. \tag{4.5}$$

Equation (4.4) is the common analytical starting point for the estimator developed in the remainder of this thesis. It admits a compact interpretation: the MMSE estimator is a *posterior-weighted combination of conditional estimators*, one per hypothesis, in which the conditional estimators act as hypothesis-specific spatial–spectral filters and the posterior probabilities act as soft weights. The concrete closed form of (4.4) is determined by two modelling choices:

1. the *distributional priors* on the target speech, the interferers, and the noise, which dictate the likelihoods $p(\boldsymbol{y} \mid \mathcal{H}_i)$ and the conditional estimators $\mathbb{E}[\, h(S) \mid \boldsymbol{y}, \mathcal{H}_i \,]$; and
2. the *hypothesis set* $\{\mathcal{H}_i\}$ itself, which enumerates which signal-activity configurations are indexed by i.

The first choice is treated in Section 4.4; the second, in Section 4.5.

4.4. Distributional Assumptions

The distributional priors adopted in this work are consistent with the statistical signal models employed in the foundational analyses of Hendriks et al. [10, 20] and Tesch and Gerkmann [9, 13]. This consistency serves two purposes: it ensures that the closed-form estimators derived from these priors remain directly comparable with the existing body of multichannel MMSE results, and—as we show in Chapter 5—it enables us to establish a formal structural equivalence between the composite-hypothesis estimator and the analytic nonlinear spatial filter of [9].

4.4.1. Target Speech Prior

Let $S = A\, e^{j\varphi}$ with the phase φ distributed uniformly on $[0, 2\pi)$ and statistically independent of the magnitude A. The magnitude is modelled by a generalised Gamma distribution with shape parameter $\nu > 0$ [20]:

$$p_A(a) \;=\; \frac{2\,(\nu/\sigma_S^2)^{\nu}}{\Gamma(\nu)}\, a^{2\nu-1}\, \exp\!\left(-\frac{\nu}{\sigma_S^2}\, a^2\right), \qquad a \geq 0, \tag{4.6}$$

where $\sigma_S^2 = \mathbb{E}[|S|^2]$ denotes the target PSD and $\Gamma(\cdot)$ is the Gamma function. This distributional family was adopted in [10] and [9] precisely to model the super-Gaussian, heavy-tailed nature of speech DFT coefficients: empirical measurements of short-time spectral amplitudes consistently reveal distributions with higher kurtosis than a Gaussian, reflecting the peaky, intermittent energy structure of voiced speech. Values of $\nu < 1$ capture this excess kurtosis, producing sharper concentration near zero and heavier tails relative to the Gaussian case (Figure 4.2). In the analytic framework of [10, 9], this general prior leads to MMSE estimators involving confluent hypergeometric functions $M(\cdot, \cdot, \cdot)$ [46].

To achieve computational tractability while maintaining structural consistency of the speech signal model across all components of the estimator, we specialise to the case $\nu = 1$. Under this choice, the generalised Gamma distribution (4.6) reduces to the Rayleigh distribution for the magnitude,

$$p_A(a)\big|_{\nu=1} \;=\; \frac{2}{\sigma_S^2}\, a\, \exp\!\left(-\frac{a^2}{\sigma_S^2}\right), \qquad a \geq 0, \tag{4.7}$$

and, equivalently, the complex DFT coefficient S follows a zero-mean complex Gaussian distribution:

$$S \;\sim\; \mathcal{CN}(0,\, \sigma_S^2). \tag{4.8}$$

The Gaussian specialisation (4.8) yields closed-form MVDR–Wiener cascades for each conditional MMSE branch and enables the likelihoods $p(\boldsymbol{y} \mid \mathcal{H}_i)$ to be expressed entirely in terms of matrix determinants and quadratic forms, avoiding the hypergeometric functions required by the general prior. We adopt (4.8) as the working distributional assumption for all subsequent derivations and experiments.

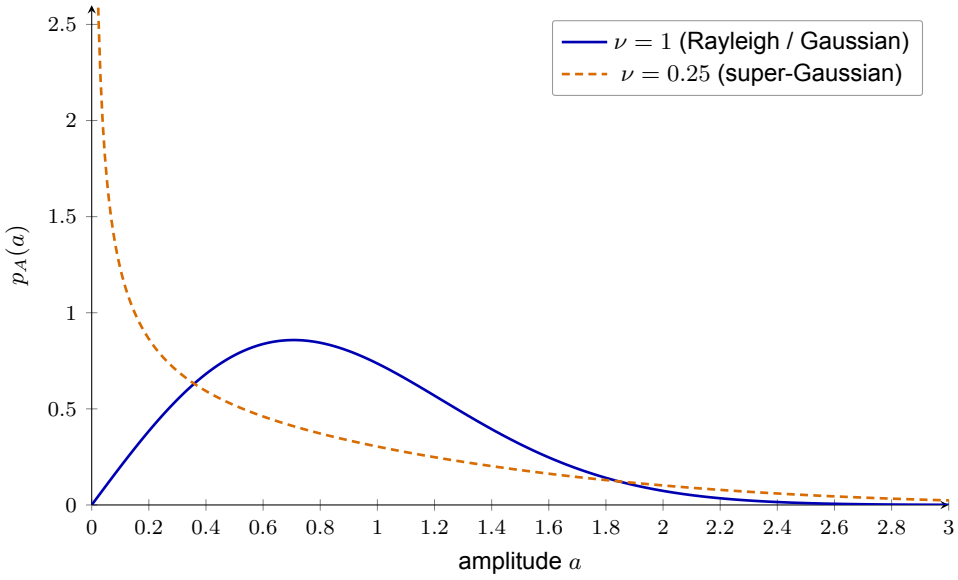


Figure 4.2: Generalised-Gamma amplitude prior (4.6) for two shape parameters, at fixed target PSD $\sigma_S^2 = 1$. For $\nu = 1$ the prior reduces to a Rayleigh density (equivalently, a zero-mean complex Gaussian for the DFT coefficient S), with a finite mode away from the origin. For $\nu = 0.25 < 1$ the density is strongly super-Gaussian: mass concentrates sharply near zero and the tail decays more slowly, matching the peaky, intermittent energy of voiced speech DFT coefficients. The Gaussian specialisation $\nu = 1$ adopted in this work trades this heavy-tailed modelling capacity for the closed-form MVDR–Wiener tractability developed in Chapter 5.

4.4.2. Interferers and Noise: Complex Gaussian

Each interferer V_i (for $i \neq j$) is modelled as zero-mean complex Gaussian when active, so that its rank-one multichannel contribution is distributed as

$$\boldsymbol{b}_i V_i \;\sim\; \mathcal{CN}(\mathbf{0},\, \boldsymbol{\Sigma}_{V_i}), \qquad \boldsymbol{\Sigma}_{V_i} \;=\; \boldsymbol{b}_i \boldsymbol{b}_i^H \sigma_{V_i}^2. \tag{4.9}$$

The noise is likewise assumed zero-mean complex Gaussian and always active, $\boldsymbol{n} \sim \mathcal{CN}(\mathbf{0}, \boldsymbol{\Sigma}_N)$. Although each individual interferer is Gaussian and the noise is Gaussian, the marginal distribution of the aggregate interference term $\boldsymbol{u}$ in (4.3) is *not* Gaussian once a hypothesis structure is imposed: it is a Gaussian mixture with components indexed by the active-interferer hypothesis, as we show next.

4.4.3. Gaussian-Mixture Interference and the Nonlinearity of the Optimal Estimator

Under the composite hypothesis structure introduced in Section 4.5, the marginal distribution of the interference term is obtained by marginalising over the discrete hypothesis index,

$$p(\boldsymbol{u}) \;=\; \sum_i P(\mathcal{H}_i)\, p(\boldsymbol{u} \mid \mathcal{H}_i) \;=\; \sum_i P(\mathcal{H}_i)\, \mathcal{CN}(\boldsymbol{u};\, \mathbf{0},\, \boldsymbol{\Sigma}_i)\,, \tag{4.10}$$

with $\boldsymbol{\Sigma}_i$ the interference-plus-noise covariance under $\mathcal{H}_i$. A convex combination of distinct Gaussians is not itself Gaussian, and (4.10) is therefore a *Gaussian mixture*: a heavy-tailed, generally non-Gaussian distribution whose empirical presence in multi-talker mixtures has been verified by direct measurement [9].

This observation is what makes the estimator nonlinear. Balan and Rosca [37] proved—and Tesch and Gerkmann [9] gave a simplified proof—that the multichannel MMSE-optimal estimator separates into an MVDR beamformer followed by a single-channel post-filter *if and only if* the interference is Gaussian (Section 2.5). Under the mixture (4.10) separability therefore fails, and the optimal estimator is a jointly spatial–spectral nonlinear functional of $\boldsymbol{y}$.

The nonlinearity enters through the posterior weights of the expansion (4.4). The posterior weights $p(\mathcal{H}_i \mid \boldsymbol{y})$ depend on $\boldsymbol{y}$ through exponential and determinant factors in (4.5), and the summation over hypotheses therefore produces an estimator that is nonlinear in $\boldsymbol{y}$ as a whole, even under the Gaussian

speech assumption (4.8) where each conditional estimator $\mathbb{E}[\, S \mid \boldsymbol{y}, \mathcal{H}_i]$ is itself linear. The optimal multi-source estimator thus takes the form of a probabilistically weighted combination of linear spatial–spectral filters, and this architecture is concretely realised in Chapter 5.

4.5. Composite Hypothesis Structure

Based on the distributional assumptions established above, we now specify the hypothesis set over which the Bayesian MMSE expansion (4.4) is evaluated. In target speech extraction, one source is designated as the target and the remaining $I-1$ sources are treated as interferers. Two distinct signal-activity questions are physically relevant at each TF bin:

(i) whether the target is present at all, and

(ii) which of the $I-1$ interferers—if any—contributes dominantly to the interference.

These questions concern *different speakers*, and we take their answers to be governed by statistically independent processes. The appropriate hypothesis structure is therefore a *Cartesian product* of two hypothesis sets, one over the target and one over the interferers:

$$\begin{aligned} \text{Target hypotheses:} \quad & \begin{cases} \mathcal{M}_0 : & \text{target is absent,} \\ \mathcal{M}_1 : & \text{target is present;} \end{cases} \\ \text{Interferer hypotheses:} \quad & \begin{cases} \mathcal{H}_0 : & \text{noise only (no interferer active),} \\ \mathcal{H}_i : & \text{interferer } i \text{ dominant,} \quad i = 1, \ldots, I-1. \end{cases} \end{aligned} \tag{4.11}$$

Under the single-dominant-interferer approximation motivated by WDO (Section 4.2), at most one interferer is active per TF bin, so the I interferer hypotheses $\{\mathcal{H}_i\}_{i=0}^{I-1}$ are mutually exclusive; the two target hypotheses $\{\mathcal{M}_m\}$ are likewise mutually exclusive.

The product structure of (4.11) spans $2 \times I$ composite hypotheses $(\mathcal{M}_m, \mathcal{H}_i)$. The Bayesian MMSE expansion is then applied over all $2I$ composite states. The target-absent terms vanish, and the chain rule splits the remaining joint posterior into two factors: the target posterior $p(\mathcal{M}_1 \mid \boldsymbol{y})$, and the interferer posterior $p(\mathcal{H}_i \mid \boldsymbol{y}, \mathcal{M}_1)$, which is conditioned on target presence.

In Chapter 5, we derive the closed-form MMSE estimator that results from this expansion under the Gaussian signal assumptions (4.8)–(4.9), including the chain-rule factorisation that writes the joint posterior as a product of two factors that can be estimated separately. We further show that the resulting set of posterior-weighted filters is structurally equivalent to the analytic MMSE-optimal nonlinear spatial filter derived independently by Hendriks et al. [10] and Tesch and Gerkmann [9] from a Gaussian-mixture noise model—establishing that the composite-hypothesis expansion and the mixture-component summation are two representations of the same optimal nonlinear filter, differing only in their probabilistic interpretation of the summation index.

5

Closed-Form Composite-Hypothesis Estimator

Chapter 4 identified a composite hypothesis structure that factorises the signal-activity space at each time–frequency bin into two axes: target presence and interferer dominance. This chapter derives the complete closed-form MMSE estimator induced by this structure.

The derivation proceeds in the following order. Sections 5.1 and 5.2 establish the hypothesis set and the per-hypothesis signal model with their associated likelihoods. Section 5.3 then presents the core result: the full MMSE expansion over the composite hypothesis space and the chain-rule factorisation that yields a set of posterior-weighted MVDR–Wiener filters. Section 5.4 derives the conditional MMSE estimator under each hypothesis—the MVDR–Wiener filter that constitutes each element of that set. Section 5.5 computes the two sets of closed-form posteriors that activate the bank. Finally, Section 5.6 demonstrates that the resulting estimator is structurally equivalent to the analytic nonlinear spatial filter of Tesch and Gerkmann [9], establishing that the probabilistic composite-hypothesis expansion and the Gaussian-mixture MMSE derivation are two perspectives on the same optimal nonlinear filter.

Throughout this chapter the indices (k, l) are suppressed unless required.

5.1. Composite Hypothesis Structure

We consider the extraction of a single designated target source from a mixture of $I - 1$ interferers and noise, captured by an M-element microphone array. The target is characterised by RTF $\boldsymbol{d} \in \mathbb{C}^M$ and STFT coefficient S with PSD σ_S^2; interferer i by RTF $\boldsymbol{b}_i \in \mathbb{C}^M$ and STFT coefficient V_i with PSD $\sigma_{V_i}^2$.

As established in Section 4.5, the signal-activity state at each TF bin is modelled by a Cartesian product of two mutually exclusive hypothesis sets:

$$\begin{aligned} &\text{Target:} \quad \mathcal{M}_0 \ (\text{absent}), \quad \mathcal{M}_1 \ (\text{present}); \\ &\text{Interferer:} \quad \mathcal{H}_0 \ (\text{noise only}), \quad \mathcal{H}_i \ (\text{interferer } i \text{ dominant}), \quad i = 1, \ldots, I-1. \end{aligned} \tag{5.1}$$

The composite hypothesis space thus comprises $2I$ states $\{(\mathcal{M}_m, \mathcal{H}_i)\}$ with $m \in \{0, 1\}$ and $i \in \{0, 1, \ldots, I-1\}$.

Two properties of (5.1) are used in the sequel and should be distinguished carefully. First, the $2I$ composite states are *mutually exclusive and jointly exhaustive*, so that the two posterior sets each normalise to unity; this is a property of the hypothesis set itself and requires no probabilistic assumption. Second, we take the two axes to be independent *a priori*,

$$P(\mathcal{M}_m, \mathcal{H}_i) \ = \ P(\mathcal{M}_m)\, P(\mathcal{H}_i), \tag{5.2}$$

which reflects the physical separation between the target speaker's voice-activity process and the interferers' activity processes: the target speaks or is silent according to its own phonetic content, independently of which competitor happens to dominate a given TF bin.

It is important to note that a priori independence (5.2) does *not* imply independence of the corresponding posteriors: the observation $\boldsymbol{y}$ couples the two axes, because the evidence for target presence depends on which interference model is assumed and vice versa. The derivations of Sections 5.3 and 5.5 therefore rely on the chain rule, which is an identity, rather than on any posterior factorisation. Assumption (5.2) is invoked exactly once, in Section 5.5.1, to identify the prior carried by the target-conditional interferer posterior.

5.2. Per-Hypothesis Signal Model and Likelihoods

5.2.1. Signal Model Under Each Composite Hypothesis

Under the composite hypothesis $(\mathcal{M}_1, \mathcal{H}_i)$—target present and interferer i dominant—the multichannel observation is

$$\boldsymbol{y} \;=\; \boldsymbol{d}\, S \;+\; \boldsymbol{v}_i, \qquad \boldsymbol{v}_i \;\sim\; \mathcal{CN}(\boldsymbol{0},\, \boldsymbol{\Sigma}_i), \tag{5.3}$$

with per-hypothesis interference-plus-noise covariance

$$\boldsymbol{\Sigma}_i \;=\; \begin{cases} \boldsymbol{\Sigma}_N, & i = 0, \\ \boldsymbol{\Sigma}_N \;+\; \boldsymbol{b}_i\, \boldsymbol{b}_i^H\, \sigma_{V_i}^2, & i = 1, \ldots, I{-}1. \end{cases} \tag{5.4}$$

Under $(\mathcal{M}_0, \mathcal{H}_i)$—target absent—the target contribution vanishes and $\boldsymbol{y} = \boldsymbol{v}_i$.

A structural feature of (5.3) is critical: the target RTF $\boldsymbol{d}$ is *invariant* across all interferer hypotheses $\mathcal{H}_i$. Only the interference-plus-noise covariance $\boldsymbol{\Sigma}_i$ changes from one hypothesis to the next. This invariance produces a bank of MVDR beamformers that all steer in the same target direction but differ in null placement—each branch places its spatial null towards a different interferer.

5.2.2. Conditional Likelihood Under Target Presence

Since the noise and interferers are modelled as complex Gaussian (Section 4.4.2) and the target prior under $\nu = 1$ is likewise complex Gaussian $S \sim \mathcal{CN}(0, \sigma_S^2)$, the observation under $(\mathcal{M}_1, \mathcal{H}_i)$ is the sum of two independent Gaussian components and is itself Gaussian, with the target-inclusive covariance

$$\boldsymbol{\Sigma}_{S,i} \;\triangleq\; \boldsymbol{\Sigma}_i \;+\; \boldsymbol{d}\, \boldsymbol{d}^H\, \sigma_S^2. \tag{5.5}$$

The corresponding likelihood is

$$p(\boldsymbol{y} \mid \mathcal{M}_1,\, \mathcal{H}_i) \;=\; \frac{1}{\pi^M\, |\boldsymbol{\Sigma}_{S,i}|}\; \exp\!\big(-\boldsymbol{y}^H\, \boldsymbol{\Sigma}_{S,i}^{-1}\, \boldsymbol{y}\big). \tag{5.6}$$

Conditional Likelihood Under Target Absence. Under $\mathcal{M}_0$ the target coefficient is identically zero, $S \equiv 0$; this is the *definition* of the target-absent hypothesis. The observation is then governed by the interference plus noise alone:

$$p(\boldsymbol{y} \mid \mathcal{M}_0,\, \mathcal{H}_i) \;=\; \frac{1}{\pi^M\, |\boldsymbol{\Sigma}_i|}\; \exp\!\big(-\boldsymbol{y}^H\, \boldsymbol{\Sigma}_i^{-1}\, \boldsymbol{y}\big). \tag{5.7}$$

5.2.3. Simplified Likelihood via Matrix Identities

Applying the matrix determinant lemma to (5.5),

$$|\boldsymbol{\Sigma}_{S,i}| \;=\; |\boldsymbol{\Sigma}_i|\, \big(1 + \xi_{S,i}\big), \tag{5.8}$$

with the per-hypothesis target *a priori* SNR

$$\xi_{S,i} \;\triangleq\; \boldsymbol{d}^H\, \boldsymbol{\Sigma}_i^{-1}\, \boldsymbol{d}\;\; \sigma_S^2. \tag{5.9}$$

Applying the Sherman–Morrison formula to $\boldsymbol{\Sigma}_{S,i}^{-1}$,

$$\boldsymbol{\Sigma}_{S,i}^{-1} \;=\; \boldsymbol{\Sigma}_i^{-1} \;-\; \frac{\boldsymbol{\Sigma}_i^{-1}\, \boldsymbol{d}\, \boldsymbol{d}^H\, \boldsymbol{\Sigma}_i^{-1}\, \sigma_S^2}{1 + \xi_{S,i}}, \tag{5.10}$$

the quadratic form in (5.6) simplifies:

$$\boldsymbol{y}^H \boldsymbol{\Sigma}_{S,i}^{-1} \boldsymbol{y} \;=\; \boldsymbol{y}^H \boldsymbol{\Sigma}_i^{-1} \boldsymbol{y} \;-\; \frac{|\boldsymbol{d}^H \boldsymbol{\Sigma}_i^{-1} \boldsymbol{y}|^2 \, \sigma_S^2}{1 + \xi_{S,i}}. \tag{5.11}$$

Substituting (5.8) and (5.11) into (5.6):

$$p(\boldsymbol{y} \mid \mathcal{M}_1, \mathcal{H}_i) \;=\; \frac{1}{\pi^M \, |\boldsymbol{\Sigma}_i| \, (1 + \xi_{S,i})} \; \exp\!\Big(-\boldsymbol{y}^H \boldsymbol{\Sigma}_i^{-1} \boldsymbol{y} \;+\; \frac{|\boldsymbol{d}^H \boldsymbol{\Sigma}_i^{-1} \boldsymbol{y}|^2 \, \sigma_S^2}{1 + \xi_{S,i}} \Big). \tag{5.12}$$

The second term in the exponent is recognised as the per-hypothesis *detection statistic*:

$$\gamma_i \;\triangleq\; \frac{|\boldsymbol{d}^H \boldsymbol{\Sigma}_i^{-1} \boldsymbol{y}|^2 \, \sigma_S^2}{1 + \xi_{S,i}}. \tag{5.13}$$

This statistic measures the energy of the observation along the target direction, whitened by the per-hypothesis interference covariance and normalised by the per-hypothesis SNR.

Likelihood Ratio. The ratio of the target-present to target-absent likelihoods under the same interferer hypothesis is

$$\frac{p(\boldsymbol{y} \mid \mathcal{M}_1, \mathcal{H}_i)}{p(\boldsymbol{y} \mid \mathcal{M}_0, \mathcal{H}_i)} \;=\; \frac{1}{1 + \xi_{S,i}} \; \exp\!\big(\gamma_i\big). \tag{5.14}$$

This ratio drives the per-hypothesis target-presence decision: large γ_i indicates strong target energy under interference model i, pushing the posterior towards $\mathcal{M}_1$.

5.3. Core Result: Probability-Activated Beamformer Bank

We now derive the central result of this chapter: the closed-form MMSE estimator over the composite hypothesis space and its factorisation into a probability-activated bank of spatial–spectral filters.

5.3.1. Full MMSE Expansion

Applying the total-probability MMSE expansion (4.4) with $h(S) = S$ over all $2I$ composite hypotheses:

$$\hat{S}(\boldsymbol{y}) \;=\; \sum_{m \in \{0,1\}} \sum_{i=0}^{I-1} \mathbb{E}[\, S \mid \boldsymbol{y}, \mathcal{M}_m, \mathcal{H}_i \,] \; p(\mathcal{M}_m, \mathcal{H}_i \mid \boldsymbol{y}) \,. \tag{5.15}$$

Elimination of the Target-Absent Terms. Under $\mathcal{M}_0$ we have $S \equiv 0$ by the definition of the hypothesis, so the observation carries no information about S. Therefore

$$\mathbb{E}[\, S \mid \boldsymbol{y}, \mathcal{M}_0, \mathcal{H}_i \,] \;=\; 0, \qquad \forall \, i \in \{0, 1, \ldots, I{-}1\}. \tag{5.16}$$

All $m = 0$ terms in (5.15) vanish, reducing the double sum to

$$\hat{S}(\boldsymbol{y}) \;=\; \sum_{i=0}^{I-1} \mathbb{E}[\, S \mid \boldsymbol{y}, \mathcal{M}_1, \mathcal{H}_i \,] \; p(\mathcal{M}_1, \mathcal{H}_i \mid \boldsymbol{y}) \,. \tag{5.17}$$

At this stage, the estimator is a weighted sum of I conditional MMSE estimators (one per interferer hypothesis), each weighted by the corresponding joint posterior $p(\mathcal{M}_1, \mathcal{H}_i \mid \boldsymbol{y})$.

5.3.2. Chain-Rule Factorisation

The joint posterior in (5.17) couples the target-presence and interferer-activity decisions. It is tempting to decouple them by assuming that the two posteriors factorise, $p(\mathcal{M}_1, \mathcal{H}_i \mid \boldsymbol{y}) = p(\mathcal{M}_1 \mid \boldsymbol{y})\, p(\mathcal{H}_i \mid \boldsymbol{y})$. No such assumption is required, and none is made here: the chain rule of conditional probability supplies the same product structure *exactly*,

$$\boxed{\; p(\mathcal{M}_1, \mathcal{H}_i \mid \boldsymbol{y}) \;=\; p(\mathcal{M}_1 \mid \boldsymbol{y}) \;\cdot\; p(\mathcal{H}_i \mid \boldsymbol{y}, \mathcal{M}_1) \;} \tag{5.18}$$

Equation (5.18) is an identity valid for any joint law on $\{\mathcal{M}_m\} \times \{\mathcal{H}_i\}$. The price of exactness is a change of meaning in the second factor: it is the interferer posterior *conditioned on target presence*, $p(\mathcal{H}_i \mid \boldsymbol{y}, \mathcal{M}_1)$, not the marginal $p(\mathcal{H}_i \mid \boldsymbol{y})$. This is not a restriction. What (5.17) requires is the joint posterior, and the chain rule delivers it exactly, as the product of a target-presence factor and this target-conditional interferer factor; Section 5.5.1 derives the latter in closed form.

Substituting (5.18) into (5.17), the target-presence posterior factors out of the summation over the interferer axis, yielding the *probability-activated beamformer bank*:

$$\hat{S}(\boldsymbol{y}) \;=\; \underbrace{p(\mathcal{M}_1 \mid \boldsymbol{y})}_{\text{target posterior}} \;\cdot\; \sum_{i=0}^{I-1} \underbrace{p(\mathcal{H}_i \mid \boldsymbol{y},\, \mathcal{M}_1)}_{\text{interferer posterior}} \;\cdot\; \underbrace{\mathbb{E}[\,S \mid \boldsymbol{y},\, \mathcal{M}_1,\, \mathcal{H}_i\,]}_{\text{conditional estimator}}. \tag{5.19}$$

Equation (5.19) is the central equation of this thesis. It is an exact rewriting of the MMSE estimator (5.17), not a factorisation obtained by assumption. The three factors have the following roles.

- The **conditional estimator** $\mathbb{E}[S \mid \boldsymbol{y}, \mathcal{M}_1, \mathcal{H}_i]$ estimates the target under the assumption that interferer i dominates. Section 5.4 shows that it is an MVDR beamformer steered at $\boldsymbol{d}$ with interference covariance $\boldsymbol{\Sigma}_i$, followed by a Wiener post-filter. There is one such filter per hypothesis; we call the set of them the *filter bank*.
- The **interferer posterior** $p(\mathcal{H}_i \mid \boldsymbol{y}, \mathcal{M}_1)$ is the probability that interferer i dominates the current bin, given that the target is present. It weights the corresponding filter. A bin dominated by interferer i therefore takes most of its output from the filter that nulls interferer i.
- The **target posterior** $p(\mathcal{M}_1 \mid \boldsymbol{y})$ is the probability that the target is present in the current bin. It multiplies the weighted sum, so the output is suppressed where the target is absent.

5.3.3. Explicit Four-Factor Form

With the branch decomposition (5.29) now in hand, we write each conditional estimator as $\mathbb{E}[S \mid \boldsymbol{y}, \mathcal{M}_1, \mathcal{H}_i] = G_{\mathrm{W},i}\; z_i(\boldsymbol{y})$, where $z_i(\boldsymbol{y})$ is the per-hypothesis MVDR output and $G_{\mathrm{W},i}$ the per-hypothesis Wiener gain. Substituting into (5.19):

$$\hat{S}(\boldsymbol{y}) \;=\; \underbrace{p(\mathcal{M}_1 \mid \boldsymbol{y})}_{\text{target posterior}} \;\cdot\; \sum_{i=0}^{I-1} \underbrace{p(\mathcal{H}_i \mid \boldsymbol{y},\, \mathcal{M}_1)}_{\text{interferer posterior}} \;\cdot\; \underbrace{G_{\mathrm{W},i}}_{\text{Wiener}} \;\cdot\; \underbrace{z_i(\boldsymbol{y})}_{\text{target MVDR}}. \tag{5.20}$$

The estimator is a bank of I MVDR–Wiener filters, one per interference model. Each filter is weighted by the interferer posterior, and the weighted sum is multiplied by the target posterior. Only two quantities in (5.20) are probabilistic: $p(\mathcal{M}_1 \mid \boldsymbol{y})$ and $\{p(\mathcal{H}_i \mid \boldsymbol{y}, \mathcal{M}_1)\}_{i=0}^{I-1}$. Everything else is a deterministic function of the signal-model parameters. These two are what the network of Chapter 6 is trained to produce.

5.4. Conditional MMSE Estimator: The MVDR--Wiener Branch

We now derive the conditional estimator $\mathbb{E}[S \mid \boldsymbol{y}, \mathcal{M}_1, \mathcal{H}_i]$ that constitutes each branch of the bank in (5.19).

5.4.1. Derivation Under the Gaussian Prior

Under $(\mathcal{M}_1, \mathcal{H}_i)$ with the Gaussian target prior ($\nu = 1$, i.e., $S \sim \mathcal{CN}(0, \sigma_S^2)$), the observation model (5.3) is a standard linear-Gaussian model. The conditional posterior of S given $\boldsymbol{y}$ is therefore Gaussian:

$$S \mid \boldsymbol{y},\, \mathcal{M}_1,\, \mathcal{H}_i \;\sim\; \mathcal{CN}\Big(\mu_{S|y,i},\; \sigma^2_{S|y,i}\Big). \tag{5.21}$$

The posterior mean and variance are obtained by standard Gaussian conditioning (cf. [47]):

$$\mu_{S|y,i} \;=\; \sigma_S^2\, \boldsymbol{d}^H\, \boldsymbol{\Sigma}_{S,i}^{-1}\, \boldsymbol{y}, \tag{5.22}$$

$$\sigma^2_{S|y,i} \;=\; \sigma_S^2 \;-\; \sigma_S^4\, \boldsymbol{d}^H\, \boldsymbol{\Sigma}_{S,i}^{-1}\, \boldsymbol{d}. \tag{5.23}$$

The conditional MMSE estimator is the posterior mean $\mathbb{E}[S \mid \boldsymbol{y}, \mathcal{M}_1, \mathcal{H}_i] = \mu_{S|y,i}$.

5.4.2. Reduction to the MVDR--Wiener Cascade

Substituting the Sherman–Morrison expansion (5.10) into (5.22):

$$\mu_{S|y,i} \;=\; \sigma_S^2\, \boldsymbol{d}^H \left(\boldsymbol{\Sigma}_i^{-1} - \frac{\boldsymbol{\Sigma}_i^{-1}\, \boldsymbol{d}\, \boldsymbol{d}^H\, \boldsymbol{\Sigma}_i^{-1}\, \sigma_S^2}{1+\xi_{S,i}} \right) \boldsymbol{y}. \tag{5.24}$$

Expanding and collecting terms,

$$\mu_{S|y,i} \;=\; \sigma_S^2\, \boldsymbol{d}^H\, \boldsymbol{\Sigma}_i^{-1}\, \boldsymbol{y} \;-\; \frac{\sigma_S^4\, (\boldsymbol{d}^H\, \boldsymbol{\Sigma}_i^{-1}\, \boldsymbol{d})\, (\boldsymbol{d}^H\, \boldsymbol{\Sigma}_i^{-1}\, \boldsymbol{y})}{1+\xi_{S,i}}. \tag{5.25}$$

Factoring out $\boldsymbol{d}^H\, \boldsymbol{\Sigma}_i^{-1}\, \boldsymbol{y}$:

$$\mu_{S|y,i} \;=\; (\boldsymbol{d}^H\, \boldsymbol{\Sigma}_i^{-1}\, \boldsymbol{y})\; \sigma_S^2 \left(1 - \frac{\sigma_S^2\, \boldsymbol{d}^H\, \boldsymbol{\Sigma}_i^{-1}\, \boldsymbol{d}}{1+\xi_{S,i}} \right). \tag{5.26}$$

Recalling $\xi_{S,i} = \boldsymbol{d}^H\, \boldsymbol{\Sigma}_i^{-1}\, \boldsymbol{d}\, \sigma_S^2$, the parenthetical simplifies:

$$1 - \frac{\xi_{S,i}}{1+\xi_{S,i}} \;=\; \frac{1}{1+\xi_{S,i}}, \tag{5.27}$$

yielding

$$\mu_{S|y,i} \;=\; \frac{\boldsymbol{d}^H\, \boldsymbol{\Sigma}_i^{-1}\, \boldsymbol{y}}{\boldsymbol{d}^H\, \boldsymbol{\Sigma}_i^{-1}\, \boldsymbol{d}} \;\cdot\; \frac{\xi_{S,i}}{1+\xi_{S,i}}. \tag{5.28}$$

This is recognised as the cascade of two classical components:

$$\boxed{\mathbb{E}[\, S \mid \boldsymbol{y},\, \mathcal{M}_1,\, \mathcal{H}_i \,] \;=\; \underbrace{\frac{\boldsymbol{d}^H\, \boldsymbol{\Sigma}_i^{-1}\, \boldsymbol{y}}{\boldsymbol{d}^H\, \boldsymbol{\Sigma}_i^{-1}\, \boldsymbol{d}}}_{z_i(\boldsymbol{y})\ \text{[MVDR]}} \;\cdot\; \underbrace{\frac{\xi_{S,i}}{1+\xi_{S,i}}}_{G_{\mathrm{W},i}\ \text{[Wiener]}},} \tag{5.29}$$

where:

- $z_i(\boldsymbol{y})$ is the MVDR beamformer steered at the target direction $\boldsymbol{d}$ with interference covariance $\boldsymbol{\Sigma}_i$, satisfying the distortionless constraint $\mathbb{E}[z_i \mid S] = S$;
- $G_{\mathrm{W},i} = \xi_{S,i}/(1+\xi_{S,i})$ is the single-channel Wiener gain.

5.4.3. Per-Hypothesis MVDR Beamformer Properties

Each MVDR beamformer $z_i(\boldsymbol{y})$ solves the constrained optimisation

$$\boldsymbol{w}_i \;=\; \arg\min_{\boldsymbol{w}}\; \boldsymbol{w}^H\, \boldsymbol{\Sigma}_i\, \boldsymbol{w} \qquad \text{s.t.} \quad \boldsymbol{w}^H\, \boldsymbol{d} = 1, \tag{5.30}$$

with the closed-form solution

$$\boldsymbol{w}_i \;=\; \frac{\boldsymbol{\Sigma}_i^{-1}\, \boldsymbol{d}}{\boldsymbol{d}^H\, \boldsymbol{\Sigma}_i^{-1}\, \boldsymbol{d}}, \qquad z_i(\boldsymbol{y}) \;=\; \boldsymbol{w}_i^H\, \boldsymbol{y}. \tag{5.31}$$

For $i \geq 1$, the covariance $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}_N + \boldsymbol{b}_i\, \boldsymbol{b}_i^H\, \sigma_{V_i}^2$ explicitly encodes the rank-one spatial signature of interferer i. The beamformer therefore places a spatial null towards $\boldsymbol{b}_i$ while preserving the target along $\boldsymbol{d}$. Under $\mathcal{H}_0$ ($i = 0$), $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_N$ and no interferer null is placed—only diffuse noise is suppressed.

The analytical noise floor at the output of each beamformer is

$$\sigma_{n,\mathrm{post},i}^2 \;=\; \boldsymbol{w}_i^H\, \boldsymbol{\Sigma}_i\, \boldsymbol{w}_i \;=\; \frac{1}{\boldsymbol{d}^H\, \boldsymbol{\Sigma}_i^{-1}\, \boldsymbol{d}}, \tag{5.32}$$

and the per-hypothesis a priori SNR (5.9) can be equivalently expressed as

$$\xi_{S,i} \;=\; \frac{\sigma_S^2}{\sigma_{n,\mathrm{post},i}^2}. \tag{5.33}$$

5.5. Closed-Form Posteriors

The estimator (5.20) needs two posteriors: the interferer posterior and the target posterior. Both are derived here in closed form under the Gaussian target prior $\nu = 1$, which is in force throughout this section. The likelihoods (5.6) and (5.7) are Gaussian only under that prior, and every closed form below follows from them.

The chain rule (5.18) organises the derivation, with two consequences. First, the interferer posterior required by (5.20) is $p(\mathcal{H}_i \mid \boldsymbol{y}, \mathcal{M}_1)$, conditioned on target presence, and not the marginal $p(\mathcal{H}_i \mid \boldsymbol{y})$. Section 5.5.1 derives it. Second, because target presence is carried by the other factor, the likelihood used for the interferer posterior is evaluated under $\mathcal{M}_1$ throughout. It is therefore a single Gaussian, not a mixture over the target axis.

5.5.1. Interferer Posterior

Under the composite hypothesis $(\mathcal{M}_1, \mathcal{H}_i)$ the observation is, by the per-hypothesis signal model (5.3) and the target-inclusive covariance (5.5), the sum of independent complex Gaussian components and is therefore itself complex Gaussian, $\boldsymbol{y} \mid \mathcal{M}_1, \mathcal{H}_i \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{\Sigma}_{S,i})$. Substituting $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}_N + \boldsymbol{b}_i \boldsymbol{b}_i^H \sigma_{V_i}^2$ into $\boldsymbol{\Sigma}_{S,i} = \boldsymbol{\Sigma}_i + \boldsymbol{d}\boldsymbol{d}^H \sigma_S^2$ and grouping the interferer-independent terms defines the *common base* covariance,

$$\boldsymbol{\Sigma}_{S,0} \triangleq \boldsymbol{\Sigma}_{S,i}\big|_{i=0} = \boldsymbol{\Sigma}_N + \boldsymbol{d}\boldsymbol{d}^H \sigma_S^2, \tag{5.34}$$

comprising the noise and the target with *no* interferer active (the $\mathcal{H}_0$ configuration), so that

$$\boldsymbol{y} \mid \mathcal{M}_1, \mathcal{H}_i \sim \mathcal{CN}\big(\boldsymbol{0},\ \boldsymbol{\Sigma}_{S,0} + \boldsymbol{b}_i \boldsymbol{b}_i^H \sigma_{V_i}^2\big). \tag{5.35}$$

No approximation is involved: (5.35) is the same law rewritten about the base (5.34). The common base $\boldsymbol{\Sigma}_{S,0}$ is shared across all interferer hypotheses and only the rank-one term $\boldsymbol{b}_i \boldsymbol{b}_i^H \sigma_{V_i}^2$ distinguishes one hypothesis from the next. From the perspective of interferer detection the target simply joins the always-active background: $\boldsymbol{\Sigma}_{S,0}$ plays here exactly the role that the noise covariance $\boldsymbol{\Sigma}_N$ plays in the single-source multichannel SPP of Chapter 3. The intermittency of the target requires no separate weighting factor, because $\sigma_S^2(k,l)$ denotes the *instantaneous* target PSD at bin (k,l) and vanishes where the target is silent, at which point $\boldsymbol{\Sigma}_{S,0} \to \boldsymbol{\Sigma}_N$ and the detector reduces to its noise-only form.

Applying the matrix determinant lemma and the Sherman–Morrison formula to the covariance in (5.35) *relative to the common base* $\boldsymbol{\Sigma}_{S,0}$ gives

$$\big|\boldsymbol{\Sigma}_{S,0} + \boldsymbol{b}_i \boldsymbol{b}_i^H \sigma_{V_i}^2\big| = \big|\boldsymbol{\Sigma}_{S,0}\big|\,(1 + \tilde{\xi}_i), \qquad \boldsymbol{y}^H \big(\boldsymbol{\Sigma}_{S,0} + \boldsymbol{b}_i \boldsymbol{b}_i^H \sigma_{V_i}^2\big)^{-1} \boldsymbol{y} = \boldsymbol{y}^H \boldsymbol{\Sigma}_{S,0}^{-1} \boldsymbol{y} - \tilde{g}_i, \tag{5.36}$$

where the interferer-detection statistics are defined relative to the common base:

$$\tilde{\xi}_i \triangleq \boldsymbol{b}_i^H \boldsymbol{\Sigma}_{S,0}^{-1} \boldsymbol{b}_i \ \sigma_{V_i}^2, \qquad \tilde{\xi}_0 = 0, \tag{5.37}$$

$$\tilde{g}_i \triangleq \frac{|\boldsymbol{b}_i^H \boldsymbol{\Sigma}_{S,0}^{-1} \boldsymbol{y}|^2 \ \sigma_{V_i}^2}{1 + \tilde{\xi}_i}, \qquad \tilde{g}_0 = 0. \tag{5.38}$$

Substituting (5.36) into the Gaussian likelihood $p(\boldsymbol{y} \mid \mathcal{M}_1, \mathcal{H}_i) = \pi^{-M} |\boldsymbol{\Sigma}_{S,0} + \boldsymbol{b}_i \boldsymbol{b}_i^H \sigma_{V_i}^2|^{-1} \exp(-\boldsymbol{y}^H (\boldsymbol{\Sigma}_{S,0} + \boldsymbol{b}_i \boldsymbol{b}_i^H \sigma_{V_i}^2)^{-1} \boldsymbol{y})$ and forming the posterior by Bayes' rule, the factor $\pi^{-M} |\boldsymbol{\Sigma}_{S,0}|^{-1} \exp(-\boldsymbol{y}^H \boldsymbol{\Sigma}_{S,0}^{-1} \boldsymbol{y})$ cancels between numerator and denominator, leaving

$$\boxed{p(\mathcal{H}_i \mid \boldsymbol{y}, \mathcal{M}_1) = \left[\sum_{j=0}^{I-1} \frac{P(\mathcal{H}_j)}{P(\mathcal{H}_i)} \cdot \frac{1 + \tilde{\xi}_i}{1 + \tilde{\xi}_j} \cdot \exp\big(\tilde{g}_j - \tilde{g}_i\big) \right]^{-1}}. \tag{5.39}$$

The detection statistic $\tilde{g}_i$ admits the MVDR-output form

$$\tilde{g}_i = \frac{|\tilde{z}_i(\boldsymbol{y})|^2 \ \tilde{\xi}_i \ \boldsymbol{b}_i^H \boldsymbol{\Sigma}_{S,0}^{-1} \boldsymbol{b}_i}{1 + \tilde{\xi}_i}, \qquad \tilde{z}_i(\boldsymbol{y}) \triangleq \frac{\boldsymbol{b}_i^H \boldsymbol{\Sigma}_{S,0}^{-1} \boldsymbol{y}}{\boldsymbol{b}_i^H \boldsymbol{\Sigma}_{S,0}^{-1} \boldsymbol{b}_i}, \tag{5.40}$$

where $\tilde{z}_i$ is an MVDR beamformer steered at interferer i within the common base field $\boldsymbol{\Sigma}_{S,0}$ (noise plus target). The functional form of (5.39) is identical to the multi-source MC-SPP formula established in Chapter 3, with the substitution $\boldsymbol{\Sigma}_N \to \boldsymbol{\Sigma}_{S,0}$. This substitution reflects the fact that, from the perspective of interferer detection, the target appears as additional noise.

5.5.2. Target-Presence Posterior

Per-hypothesis presence probability. Conditioned on a single interferer hypothesis $\mathcal{H}_i$, the target-presence decision is an ordinary binary test between the composite hypotheses $(\mathcal{M}_0, \mathcal{H}_i)$ and $(\mathcal{M}_1, \mathcal{H}_i)$. Its posterior odds are the prior odds times the reciprocal of the likelihood ratio (5.14), which defines

$$A_i \triangleq \frac{p(\mathcal{M}_0 \mid \boldsymbol{y}, \mathcal{H}_i)}{p(\mathcal{M}_1 \mid \boldsymbol{y}, \mathcal{H}_i)} = \frac{q_i}{1-q_i}\,(1+\xi_{S,i})\,\exp\!\left(-\frac{\beta_{S,i}}{1+\xi_{S,i}}\right), \tag{5.41}$$

and therefore yields the standard single-source MC-SPP form, evaluated with the per-hypothesis interference covariance $\boldsymbol{\Sigma}_i$:

$$p(\mathcal{M}_1 \mid \boldsymbol{y}, \mathcal{H}_i) = \big[\,1 + A_i\,\big]^{-1} = \left[1 + \frac{q_i}{1-q_i}\,(1+\xi_{S,i})\,\exp\!\left(-\frac{\beta_{S,i}}{1+\xi_{S,i}}\right)\right]^{-1}, \tag{5.42}$$

with per-hypothesis parameters:

$$\xi_{S,i} = \mathrm{tr}\big(\boldsymbol{\Sigma}_i^{-1}\,\boldsymbol{\Phi}_{xx}\big), \qquad \boldsymbol{\Phi}_{xx} \triangleq \boldsymbol{d}\,\boldsymbol{d}^H\,\sigma_S^2, \tag{5.43}$$

$$\beta_{S,i} \triangleq \boldsymbol{y}^H\,\boldsymbol{\Sigma}_i^{-1}\,\boldsymbol{\Phi}_{xx}\,\boldsymbol{\Sigma}_i^{-1}\,\boldsymbol{y}, \tag{5.44}$$

$$q_i \triangleq 1 - \frac{\mathrm{tr}[\boldsymbol{\Phi}_{xx}]}{\mathrm{tr}[\boldsymbol{\Phi}_{xx}] + \mathrm{tr}[\boldsymbol{\Phi}_{vv,i}]}, \tag{5.45}$$

where $\boldsymbol{\Phi}_{vv,i} = \boldsymbol{\Sigma}_i$ is the interference-plus-noise spatial covariance of hypothesis i. Comparing (5.44) with the detection statistic (5.13) confirms $\beta_{S,i}/(1+\xi_{S,i}) = \gamma_i$, so that (5.41) is indeed (5.14) inverted and weighted by the prior odds. The model prior underlying q_i is the target-absence probability $P(\mathcal{M}_0)$, which the a priori independence of Section 5.1 makes common to all interferer configurations; (5.45) is its per-bin plug-in estimate, evaluated separately under each interference model because the reference against which target dominance is judged changes with the hypothesis, following the single-source convention of Chapter 3.

The gate required by (5.20) is the unconditional $p(\mathcal{M}_1 \mid \boldsymbol{y})$, which follows from (5.42) in three exact steps. Writing the joint posterior in its two chain-rule expansions,

$$p(\mathcal{H}_i \mid \boldsymbol{y})\,p(\mathcal{M}_1 \mid \boldsymbol{y}, \mathcal{H}_i) = p(\mathcal{M}_1, \mathcal{H}_i \mid \boldsymbol{y}) = p(\mathcal{M}_1 \mid \boldsymbol{y})\,p(\mathcal{H}_i \mid \boldsymbol{y}, \mathcal{M}_1), \tag{5.46}$$

and dividing by $p(\mathcal{M}_1 \mid \boldsymbol{y}, \mathcal{H}_i)$, which by (5.42) is strictly positive for every finite A_i, isolates the marginal interferer posterior $p(\mathcal{H}_i \mid \boldsymbol{y})$,

$$p(\mathcal{H}_i \mid \boldsymbol{y}) = p(\mathcal{M}_1 \mid \boldsymbol{y})\,p(\mathcal{H}_i \mid \boldsymbol{y}, \mathcal{M}_1)\,\big(1 + A_i\big). \tag{5.47}$$

Summing (5.47) over i and taking the i-independent factor $p(\mathcal{M}_1 \mid \boldsymbol{y})$ outside the sum,

$$\underbrace{\sum_{i=0}^{I-1} p(\mathcal{H}_i \mid \boldsymbol{y})}_{=\,1} = p(\mathcal{M}_1 \mid \boldsymbol{y}) \left[\underbrace{\sum_{i=0}^{I-1} p(\mathcal{H}_i \mid \boldsymbol{y}, \mathcal{M}_1)}_{=\,1} + \sum_{i=0}^{I-1} p(\mathcal{H}_i \mid \boldsymbol{y}, \mathcal{M}_1)\,A_i\right], \tag{5.48}$$

where both underbraced sums equal unity because the interferer hypotheses $\{\mathcal{H}_i\}_{i=0}^{I-1}$ are mutually exclusive and jointly exhaustive (5.1), a property preserved under conditioning on $\mathcal{M}_1$. Rearranging (5.48) gives the composite target-presence posterior in closed form,

$$\boxed{p(\mathcal{M}_1 \mid \boldsymbol{y}) = \left[1 + \sum_{i=0}^{I-1} p(\mathcal{H}_i \mid \boldsymbol{y}, \mathcal{M}_1)\,A_i\right]^{-1}, \qquad A_i = \frac{q_i}{1-q_i}\,(1+\xi_{S,i})\,\exp\!\left(-\frac{\beta_{S,i}}{1+\xi_{S,i}}\right).} \tag{5.49}$$

The marginal interferer posterior $p(\mathcal{H}_i \mid \boldsymbol{y})$ appears only as an intermediate quantity in (5.47) and is eliminated by its own normalisation; the result is driven entirely by the target-conditional posterior (5.39) and the per-hypothesis statistics (5.43)–(5.45). Every step from (5.46) to (5.49) is an identity: (5.49) invokes no independence between the two hypothesis axes, no approximation to any distribution, and it remains exact even if the target-absence prior varies across interferer configurations.

Single-source reduction. For $I = 1$ only $\mathcal{H}_0$ exists, so $p(\mathcal{H}_0 \mid \boldsymbol{y}, \mathcal{M}_1) = 1$ and $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_N$. Equation (5.49) then collapses to $p(\mathcal{M}_1 \mid \boldsymbol{y}) = [1 + A_0]^{-1}$, which is exactly the single-source multichannel SPP of Chapter 3, consistent with property (iv) of Section 5.7.

Summary. The estimator (5.20) needs only two posteriors: the interferer posterior $\{p(\mathcal{H}_i \mid \boldsymbol{y}, \mathcal{M}_1)\}_{i=0}^{I-1}$ of (5.39), and the target posterior $p(\mathcal{M}_1 \mid \boldsymbol{y})$ of (5.49). Their product is the joint posterior $p(\mathcal{M}_1, \mathcal{H}_i \mid \boldsymbol{y})$ used in (5.20). This product follows from the chain rule (5.18); no factorisation is assumed. Both posteriors can be computed from the signal parameters $\{\boldsymbol{b}_i, \sigma_{V_i}^2, \boldsymbol{d}, \sigma_S^2, \boldsymbol{\Sigma}_N\}$, or estimated by a neural network from the observation. In either case the posterior estimation stays separate from the filtering in (5.20).

5.6. Structural Equivalence with the Tesch--Gerkmann Nonlinear Spatial Filter

Tesch and Gerkmann [9] derived the MMSE-optimal estimator for a multichannel observation corrupted by Gaussian-mixture noise with N_{c} mixture components, each with covariance $\boldsymbol{\Phi}_m$ and mixture weight c_m. The component count is written N_{c} rather than M, which denotes the number of microphones throughout this thesis. Under a generalised-Gamma speech prior with shape parameter ν, their estimator (cf. [9, Eq. (17)]) reads

$$T_{\mathrm{MMSE}}(\boldsymbol{y}) \;=\; \frac{\nu \displaystyle\sum_{m=1}^{N_{\mathrm{c}}} \frac{c_m\, Q_m}{|\boldsymbol{\Phi}_m|}\, e^{-\boldsymbol{y}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{y}}\, \frac{\sigma_S^2\, T_{\mathrm{MVDR}}^{(m)}(\boldsymbol{y})}{\nu\, (\boldsymbol{d}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{d})^{-1} + \sigma_S^2}\, M(\nu{+}1,\, 2,\, P_m)}{\displaystyle\sum_{m=1}^{N_{\mathrm{c}}} \frac{c_m\, Q_m}{|\boldsymbol{\Phi}_m|}\, e^{-\boldsymbol{y}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{y}}\, M(\nu,\, 1,\, P_m)}, \tag{5.50}$$

with:

- $Q_m = (\nu + \boldsymbol{d}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{d} \sigma_S^2)^{-\nu}$, as part of $T_{\mathrm{MMSE}}(\boldsymbol{y})$ reparameterisation.
- $T_{\mathrm{MVDR}}^{(m)}(\boldsymbol{y}) = \frac{\boldsymbol{d}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{y}}{\boldsymbol{d}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{d}}$, which is exactly the standard architecture of the MVDR beamformer.
- $P_m = \frac{\sigma_S^2 \boldsymbol{d}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{d} |T_{\mathrm{MVDR}}^{(m)}|^2}{\nu (\boldsymbol{d}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{d})^{-1} + \sigma_S^2}$, as part of $T_{\mathrm{MMSE}}(\boldsymbol{y})$ reparameterisation.

Note that capital P_m is part of the parameter definition from $T_{\mathrm{MMSE}}(\boldsymbol{y})$ formulation, rather than probability. We now show that (5.50) and (5.19) have identical algebraic structure.

5.6.1. Notation Correspondence

The mapping between the probabilistic composite-hypothesis framework and the Gaussian-mixture framework of Tesch and Gerkmann is:

$$\begin{aligned} \text{Gaussian-mixture component } m &\;\longleftrightarrow\; \text{Interferer hypothesis } \mathcal{H}_i, \\ \text{Mixture weight } c_m &\;\longleftrightarrow\; \text{Hypothesis prior } P(\mathcal{H}_i), \\ \text{Component covariance } \boldsymbol{\Phi}_m &\;\longleftrightarrow\; \text{Per-hypothesis covariance } \boldsymbol{\Sigma}_i, \\ T_{\mathrm{MVDR}}^{(m)}(\boldsymbol{y}) &\;\longleftrightarrow\; z_i(\boldsymbol{y}). \end{aligned} \tag{5.51}$$

The key conceptual difference is one of interpretation: the Tesch–Gerkmann framework treats the N_{c} mixture components as statistical parameters of a noise distribution model, fitted via the expectation-maximisation algorithm from observed noise data; the composite-hypothesis framework treats the I hypotheses as physical signal-activity states, each representing a specific interferer's spatial signature. Despite this difference in interpretation, the mathematical operations are identical.

We rewrite (5.50) as a posterior-weighted sum by defining the posterior weight of component m:

$$p_m(\boldsymbol{y}) \;\triangleq\; \frac{c_m\, Q_m\, |\boldsymbol{\Phi}_m|^{-1}\, e^{-\boldsymbol{y}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{y}}\, M(\nu,\, 1,\, P_m)}{\displaystyle\sum_{m'} c_{m'}\, Q_{m'}\, |\boldsymbol{\Phi}_{m'}|^{-1}\, e^{-\boldsymbol{y}^H \boldsymbol{\Phi}_{m'}^{-1} \boldsymbol{y}}\, M(\nu,\, 1,\, P_{m'})}. \tag{5.52}$$

The denominator in (5.50) is precisely $\sum_m(\cdots) = p(\boldsymbol{y})$, the marginal likelihood of the Gaussian-mixture observation model. Then (5.50) becomes

$$T_{\mathrm{MMSE}}(\boldsymbol{y}) \;=\; \sum_{m=1}^{N_{\mathrm{c}}} p_m(\boldsymbol{y}) \;\cdot\; \underbrace{\frac{\nu\,\sigma_S^2\; M(\nu{+}1,\,2,\,P_m)}{(\nu\,(\boldsymbol{d}^H\boldsymbol{\Phi}_m^{-1}\boldsymbol{d})^{-1}+\sigma_S^2)\; M(\nu,\,1,\,P_m)}}_{\text{generalised spectral gain under component } m} \cdot\; T_{\mathrm{MVDR}}^{(m)}(\boldsymbol{y}). \tag{5.53}$$

5.6.2. Specialisation to $\nu = 1$

For the Gaussian speech prior ($\nu = 1$), the confluent hypergeometric functions satisfy $M(1,1,x) = M(2,2,x) = e^x$ (verified from the Pochhammer-symbol series), so the hypergeometric ratio equals unity. The spectral gain then reduces to

$$\frac{1\cdot\sigma_S^2}{1\cdot(\boldsymbol{d}^H\boldsymbol{\Phi}_m^{-1}\boldsymbol{d})^{-1}+\sigma_S^2}\;\cdot\;1 \;=\; \frac{\boldsymbol{d}^H\boldsymbol{\Phi}_m^{-1}\boldsymbol{d}\;\sigma_S^2}{1+\boldsymbol{d}^H\boldsymbol{\Phi}_m^{-1}\boldsymbol{d}\;\sigma_S^2} \;=\; \frac{\xi_{S,m}}{1+\xi_{S,m}}, \tag{5.54}$$

which is the Wiener gain $G_{\mathrm{W},m}$. The posterior weight simplifies correspondingly. Using the Gaussian likelihood $p(\boldsymbol{y}\mid\mathcal{H}_m) = c_m/(\pi^M|\boldsymbol{\Phi}_m|(1+\xi_{S,m}))\cdot\exp(-\boldsymbol{y}^H\boldsymbol{\Phi}_m^{-1}\boldsymbol{y}+\gamma_m)$ and recognising $p_m(\boldsymbol{y}) = p(\mathcal{H}_m\mid\boldsymbol{y})$, the Tesch–Gerkmann estimator at $\nu = 1$ becomes

$$T_{\mathrm{MMSE}}(\boldsymbol{y})\big|_{\nu=1} \;=\; \sum_m p(\mathcal{H}_m\mid\boldsymbol{y}) \;\cdot\; G_{\mathrm{W},m} \;\cdot\; T_{\mathrm{MVDR}}^{(m)}(\boldsymbol{y}). \tag{5.55}$$

Applying the notation correspondence (5.51):

$$T_{\mathrm{MMSE}}(\boldsymbol{y})\big|_{\nu=1} \;\equiv\; \sum_{i=0}^{I-1} p(\mathcal{H}_i\mid\boldsymbol{y}) \;\cdot\; G_{\mathrm{W},i} \;\cdot\; z_i(\boldsymbol{y}). \tag{5.56}$$

5.6.3. Structural Identity: From Single-Axis to Two-Axis Hypothesis

The Tesch–Gerkmann estimator (5.56) is the MMSE-optimal nonlinear spatial filter under the assumption that the target is *always present*. Formally, the Tesch–Gerkmann signal model employs the prior

$$p(S) \;=\; \mathcal{CN}(0,\,\sigma_S^2), \tag{5.57}$$

and the MMSE expansion is performed over a *single-axis* hypothesis set $\{\mathcal{H}_i\}_{i=0}^{I-1}$ indexing the interferer activity alone:

$$T_{\mathrm{MMSE}}(\boldsymbol{y}) \;=\; \sum_{i=0}^{I-1} p(\mathcal{H}_i\mid\boldsymbol{y}) \;\cdot\; \mathbb{E}[\,S\mid\mathcal{H}_i,\boldsymbol{y}\,]. \tag{5.58}$$

In the composite-hypothesis framework, the target prior is generalised to allow for target absence:

$$p(S) \;=\; P(\mathcal{M}_1)\,\mathcal{CN}(0,\,\sigma_S^2) \;+\; P(\mathcal{M}_0)\,\delta(S). \tag{5.59}$$

The MMSE expansion is now performed over the *two-axis* hypothesis space $\{\mathcal{M}_m\}\times\{\mathcal{H}_i\}$. After the target-absent terms vanish ($\mathbb{E}[S\mid\mathcal{M}_0,\mathcal{H}_i,\boldsymbol{y}] = 0$, cf. (5.16)), the estimator reduces to

$$\hat{S}(\boldsymbol{y}) \;=\; \sum_{i=0}^{I-1} p(\mathcal{M}_1,\mathcal{H}_i\mid\boldsymbol{y}) \;\cdot\; \mathbb{E}[\,S\mid\mathcal{M}_1,\mathcal{H}_i,\boldsymbol{y}\,]. \tag{5.60}$$

Two observations establish that (5.60) inherits the same nonlinear spatial filter structure as (5.58).

Joint posterior as a two-axis extension. The single-axis posterior $p(\mathcal{H}_i\mid\boldsymbol{y})$ extended from the Tesch framework weights each branch by the probability that interferer i dominates, under the implicit assumption that the target is present. The two-axis joint posterior $p(\mathcal{M}_1,\mathcal{H}_i\mid\boldsymbol{y})$ plays the same role for the interferers, and in addition carries the target-presence probability. Both decisions are thus encoded in a single posterior over the product hypothesis space.

Substituting (5.29) into (5.60):

$$\hat{S}(\boldsymbol{y}) = \sum_{i=0}^{I-1} \underbrace{p(\mathcal{M}_1, \mathcal{H}_i \mid \boldsymbol{y})}_{\text{joint posterior}} \cdot \underbrace{G_{\mathrm{W},i} \cdot z_i(\boldsymbol{y})}_{\text{MVDR–Wiener branch}} . \tag{5.61}$$

Equations (5.58) and (5.61) share the same nonlinear spatial filter: a probability-weighted bank of MVDR–Wiener branches, each placing its null towards a different interferer. The transition from the Tesch framework to the composite-hypothesis framework does not alter this filter architecture; it extends the single-axis posterior $p(\mathcal{H}_i \mid \boldsymbol{y})$ to the two-axis joint posterior $p(\mathcal{M}_1, \mathcal{H}_i \mid \boldsymbol{y})$, thereby embedding target-presence detection into the same posterior. Under $P(\mathcal{M}_1) = 1$, the prior (5.59) reduces to (5.57), and the joint posterior collapses to the Tesch posterior, recovering the single-axis estimator exactly.

The superior spatial selectivity demonstrated by Tesch and Gerkmann [9]—in particular, the ability of the nonlinear filter to suppress more than $M - 1$ directional interferers with an M-element array—is therefore shared with the composite-hypothesis estimator. Because the interferer posterior changes from bin to bin, a different filter dominates in different bins, so the bank is not limited by the degrees of freedom of a single linear beamformer.

This structural identity motivates the implementation of Chapter 6: a neural network estimates the interferer posterior and the target posterior, while the model-based MVDR–Wiener bank (5.20) performs the filtering.

5.7. Properties of the Composite-Hypothesis Estimator

(i) Common target-directed MVDR. All beamformers $\{z_i(\boldsymbol{y})\}_{i=0}^{I-1}$ share the target RTF $\boldsymbol{d}$, preserving target energy at every TF bin regardless of which interferer dominates.

(ii) The two posteriors can be estimated separately. The chain rule (5.18) writes the joint posterior as a product of two factors. They are different quantities, so they can be produced by two separate networks or estimation pipelines. This does not assume that they are independent: both depend on the same observation. It only means that neither has to be computed inside the other. In the analytic evaluation the two are computed in sequence, because the target posterior (5.49) uses the interferer posterior (5.39).

(iii) Posterior normalisation. Because the interferer hypotheses are mutually exclusive and jointly exhaustive, the interferer posteriors satisfy $\sum_{i=0}^{I-1} p(\mathcal{H}_i \mid \boldsymbol{y}, \mathcal{M}_1) = 1$, ensuring that the bank output is a convex combination of the branch outputs. Normalisation is preserved under conditioning on $\mathcal{M}_1$ and is what allows the marginal interferer posterior to be eliminated in (5.48).

(iv) Single-source reduction. When $I = 1$ (no interferer), only $\mathcal{H}_0$ exists, $p(\mathcal{H}_0 \mid \boldsymbol{y}, \mathcal{M}_1) = 1$, and the estimator reduces to $\hat{S} = p(\mathcal{M}_1 \mid \boldsymbol{y}) \cdot G_{\mathrm{W},0} \cdot z_0(\boldsymbol{y})$, the standard MC-SPP-weighted MVDR–Wiener cascade of Chapter 3.

(v) Nonlinearity through the posterior weights. The estimator (5.20) is nonlinear in $\boldsymbol{y}$ through the posterior weights, even though each branch is individually linear under the Gaussian prior $\nu = 1$. A second, intra-branch nonlinearity appears for $\nu < 1$, where the spectral gain itself depends on the observation; the branch estimator for that case is given in Appendix A.

(vi) Graceful WDO degradation. The continuous interferer posteriors distribute mass between hypotheses at TF bins where multiple interferers overlap, interpolating between ideal binary masking and ideal ratio masking behaviour.

6

Algorithm Design: DNN-Guided Realisation of the Estimator

Chapter 5 established *what* the optimal estimator is. Its result, the four-factor bank (5.20), is an exact statement—but a conditional one: it presumes that the two posteriors of Chapter 5 and the signal-model quantities they are evaluated from are available. In a deployed system none of them is. This chapter specifies *how* the estimator is realised: which quantities have to be produced from the observation, by what architecture, and under what supervision.

This chapter fixes the algorithm: the input it consumes, the architecture that maps that input to the two posteriors, the model-based bank those posteriors activate, the recursions that supply the spatial statistics, and the objective under which the whole is trained. The corpus, the reference systems, the metrics and the results are discussed in Chapter 7.

Section 6.1 states the estimation problem that Chapter 5 leaves open, and separates carefully the quantities that are available at inference from those that exist only during training. Section 6.2 gives the resulting signal flow. Section 6.3 specifies the learned posterior-estimation stage, and Section 6.4 the model-based filtering stage together with the posterior-controlled recursions that supply its parameters. Section 6.5 closes the chapter with the supervision and the training objective.

6.1. What Is Known and What Must Be Estimated

The estimator (5.20) consumes five quantities. Two of them are probabilistic—the interferer posterior $\{p(\mathcal{H}_i \mid \boldsymbol{y}, \mathcal{M}_1)\}_{i=0}^{I-1}$ and the target posterior $p(\mathcal{M}_1 \mid \boldsymbol{y})$—and three are signal-model parameters: the target relative transfer function $\boldsymbol{d}$, the per-hypothesis interference-plus-noise covariances $\{\boldsymbol{\Sigma}_i\}_{i=0}^{I-1}$ of (5.4), and the target power spectral density σ_S^2. Everything else in (5.20)—the beamformer weights, the Wiener gains, the branch outputs—is a deterministic function of these five.

At inference, the only input is the mixture. The system observes the multichannel STFT $\boldsymbol{y}(k,l)$ and nothing else. It is given no direction of arrival, no voice-activity information, no separated source, no noise-only segment and no room impulse response. All five quantities above must therefore be produced from $\boldsymbol{y}$ alone. They are obtained in two different ways, and the distinction organises the rest of the chapter:

- the **two posteriors** are *learned*: a neural front end maps the stacked STFT directly to them (Section 6.3). This is the only learned component of the system;
- the **three signal-model parameters** are *not* learned. They are formed by recursive averaging of the observation, arbitrated by the very posteriors the front end emits (Section 6.4): each covariance is updated only at those time–frequency bins that the posteriors attribute to the corresponding hypothesis, and the relative transfer function and the target PSD follow from those covariances by covariance subtraction and spectral subtraction respectively.

No oracle quantity enters this path at any point, and no quantity computed from ground truth is used to filter the observation.

At training, ground truth is available—but only as a loss target. Because the corpus is simulated (Chapter 7), three further quantities exist during training that do not exist at inference: the dry per-source images, the isolated noise field, and the direct-path room impulse responses. From these the true $\boldsymbol{\Sigma}_N$, $\boldsymbol{d}$, $\{\boldsymbol{b}_i\}$, σ_S^2 and $\{\sigma_{V_i}^2\}$ can be formed exactly, and substituting them into the closed forms of Chapter 5 yields *oracle* posteriors: the values the network would emit if the signal statistics were known. They are used as regression targets for the network (Section 6.5). They are treated as constants when the loss is differentiated, so no gradient is propagated back through the oracle computation; this is what is meant below by a *stop-gradient*.

Ground-truth quantities appear in exactly one place, the loss; they appear in no equation of the filtering path; and at inference they are unavailable and unused. Table 6.1 in Section 6.4 tabulates the provenance of every quantity for the proposed system and for the two reference systems of Chapter 7, and confirms that the proposed system is the only one of the three that receives no oracle information at test time.

Why the parameters are estimated *through* the posteriors. A conventional DNN-guided beamformer would have the network predict the spatial statistics themselves—a mask from which covariances are accumulated, or a relative transfer function directly. The composite-hypothesis structure suggests a different division of labour. The network is asked only to resolve the *discrete* signal-activity ambiguity—which interferer dominates a bin, and whether the target is present—which is a classification problem with a well-defined probabilistic answer and, because the corpus is simulated, a computable oracle target. Every operation that touches the observation itself is then a deterministic function prescribed by the signal model. This is the same principle as the classical MC-SPP-guided recursion of Chapter 3, applied on the two-axis hypothesis space of Chapter 5: a soft attribution of each bin to a hypothesis, used to arbitrate which statistics that bin is allowed to update.

6.2. Signal-Processing Pipeline

The proposed system instantiates the four-factor estimator (5.20) as a two-stage pipeline that cleanly separates a data-driven *posterior-estimation* stage from a model-based *spatial–spectral filtering* stage. Figure 6.1 summarises the end-to-end signal flow. The multichannel time-domain mixture is first transformed to the STFT domain and its real and imaginary parts are stacked across channels to form the network input. A neural front end then produces the two posteriors required by the estimator: the interferer posterior $\{p(\mathcal{H}_i \mid \boldsymbol{y}, \mathcal{M}_1)\}$ and the target posterior $p(\mathcal{M}_1 \mid \boldsymbol{y})$. These posteriors do not filter the signal themselves. They weight a bank of MVDR–Wiener filters that operate directly on the observed STFT, as detailed in Section 6.4 and shown in Figure 6.3. The bank output is the probability-weighted combination of branch estimates, which is finally returned to the time domain by the inverse STFT to yield the extracted target.

A defining feature of this architecture is that the learned component never produces the enhanced signal directly. This division of labour is what allows the estimator to inherit the analytic optimality properties established in Chapter 5, and it is the principal structural difference from the end-to-end baseline.

Note that our system does not, however, remove the parameter-estimation problem: the bank still requires a target RTF, a per-hypothesis covariance and a target PSD, none of which is available at inference. These are obtained from the observation under the control of the same posteriors that route the bank, as specified in Section 6.4; no oracle quantity enters the filtering path.

6.3. Network Architecture: Two Decoupled Posterior Estimators

The chain rule (5.18) writes the joint posterior as an exact product of the target posterior and the interferer posterior. The two are different quantities, so they can be produced by two networks that share no parameters, and the split costs nothing in optimality. The front end is therefore implemented as two parallel sub-networks—an *interferer network* and a *target network*—each mapping the same stacked-STFT input to one of the two posteriors. Figure 6.2 shows this arrangement.

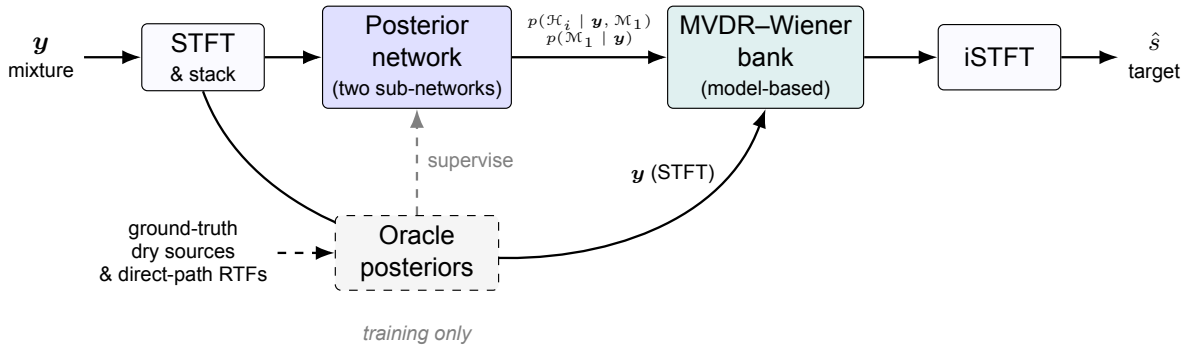


Figure 6.1: End-to-end signal-processing pipeline of the proposed composite-hypothesis system. The multichannel mixture $\boldsymbol{y}$ is transformed to the STFT domain and mapped by the posterior network to the interferer posterior $\{p(\mathcal{H}_i \mid \boldsymbol{y}, \mathcal{M}_1)\}$ and the target posterior $p(\mathcal{M}_1 \mid \boldsymbol{y})$. These posteriors weight a model-based MVDR–Wiener bank that filters the observed STFT directly (Figure 6.3); the inverse STFT returns the extracted target $\hat{s}$. During training only (dashed), oracle posteriors computed from the ground-truth dry sources and direct-path relative transfer functions supervise the network; at inference no oracle information is used.

The two posteriors are driven by different cues, which is the reason for keeping the networks separate. Deciding which competing talker dominates a bin is a spatial question: it requires the inter-microphone phase and level differences. Deciding whether the target is present is mainly a spectro-temporal energy question about one source whose position does not change. A single shared network would have to trade these two feature types against each other in its hidden representation; two networks let each specialise. This is the practical use of property (ii) in Section 5.7.

A lighter alternative is a shared backbone with two separate output projections, which would reduce both parameters and computation. It is not used here because it reintroduces the shared representation that the separation is meant to avoid, but it is a natural extension if model size becomes a constraint.

Both sub-networks share the same backbone topology, chosen to match the joint narrow-band/wide-band structure that has proven effective for multichannel mask estimation [13]. Each processes the input with two stacked recurrent layers operating on complementary axes of the time–frequency plane: the first traverses the frequency axis, allowing the network to exploit cross-band spatial structure at a fixed frame, and the second traverses the time axis, capturing temporal context at a fixed band. A final linear projection maps the recurrent features to the required number of outputs. The two sub-networks differ only in their output layer and activation:

- the **interferer network** projects to I logits—one per interferer hypothesis $\mathcal{H}_0, \ldots, \mathcal{H}_{I-1}$—followed by a softmax across the hypothesis axis, which enforces the normalisation $\sum_i p(\mathcal{H}_i \mid \boldsymbol{y}, \mathcal{M}_1) = 1$ required by property (iii) of Section 5.7;
- the **target network** projects to a single logit followed by a sigmoid, yielding a probability $p(\mathcal{M}_1 \mid \boldsymbol{y}) \in [0, 1]$ per TF bin.

The softmax and the sigmoid are the probabilistic counterparts of the two hypothesis sets: a categorical distribution over the mutually exclusive interferer hypotheses, and a Bernoulli distribution over target presence. The network therefore outputs exactly the two posteriors that (5.20) requires. It emits no power spectral densities and no filter coefficients.

6.4. Model-Based Filtering Stage

The two posteriors weight the MVDR–Wiener bank (5.20), shown in Figure 6.3. Branch i is the target-directed MVDR beamformer (5.31) with interference covariance $\boldsymbol{\Sigma}_i$, followed by the Wiener gain (5.29). All branches share the same target RTF $\boldsymbol{d}$ and differ only through $\boldsymbol{\Sigma}_i$, so each places its null towards a different interferer. Nothing in this stage is learned: given $\boldsymbol{d}$, $\{\boldsymbol{\Sigma}_i\}$ and σ_S^2, every quantity in (5.20) follows from those two equations. The deployed form is (6.12) at the end of this section.

The bank is nonlinear in $\boldsymbol{y}$ even though each branch is linear, because the interferer posterior depends on the observation through the exponential and determinant factors of (5.39). This is property (v) of

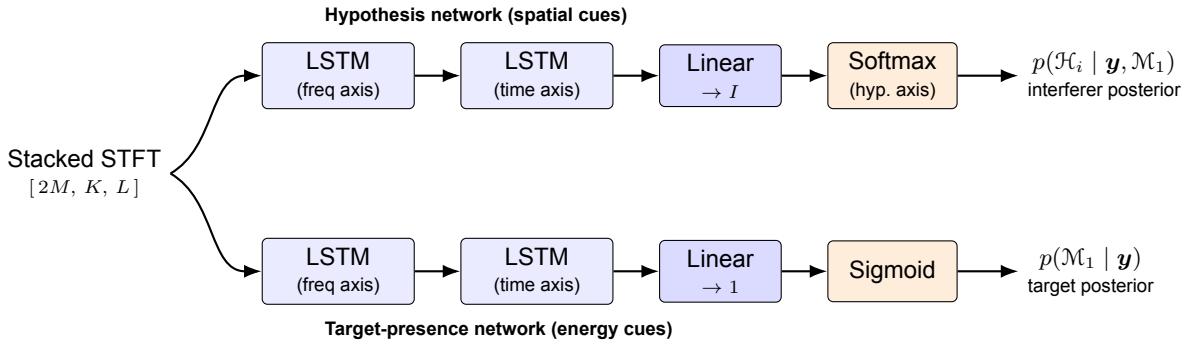


Figure 6.2: The two posterior networks. Two sub-networks with no shared parameters map the same stacked-STFT observation to the two posteriors. Each applies a frequency-axis and a time-axis recurrent layer followed by a linear projection, and they differ only in their output size and activation. The interferer network emits a softmax-normalised categorical posterior over the I interferer hypotheses; the target network emits a per-bin Bernoulli posterior. Keeping them separate means neither has to represent both the spatial cues that identify the dominant interferer and the spectro-temporal cues that indicate target presence.

Section 5.7.

The three signal-model parameters—the target RTF $\boldsymbol{d}$, the per-hypothesis covariances $\{\boldsymbol{\Sigma}_i\}_{i=0}^{I-1}$ of (5.4), and the target PSD σ_S^2—are not available at inference. Chapter 5 treats them as known, as is standard in the analytic MMSE literature [10, 9], so the optimality of (5.20) is conditional on them. The rest of this section specifies how they are estimated. The same two posteriors that weight the bank also control the recursive averaging from which $\{\hat{\boldsymbol{\Sigma}}_i\}$ and $\hat{\boldsymbol{d}}$ are formed, so no oracle quantity enters the filtering path.

The scheme is the multi-source counterpart of the MC-SPP-guided recursion of Chapter 3, applied on the two-axis hypothesis space: the covariance of hypothesis $\mathcal{H}_i$ is updated only at those time–frequency bins that the posteriors attribute to that hypothesis.

6.4.1. Posterior-controlled Interference Covariances.

The interference covariance of hypothesis $\mathcal{H}_i$ describes the observation when the target is *absent* and interferer i is dominant. Both conditions are read off the two posteriors, and their product forms an adaptive smoothing factor

$$\alpha_{\mathrm{eff},i}(k,l) \;=\; 1 \;-\; \big(1-\alpha\big)\,\hat{p}(\mathcal{H}_i \mid \boldsymbol{y})\,\big(1-\hat{p}(\mathcal{M}_1 \mid \boldsymbol{y})\big), \tag{6.1}$$

which drives the recursive update

$$\hat{\boldsymbol{\Sigma}}_i(k,l) \;=\; \alpha_{\mathrm{eff},i}(k,l)\,\hat{\boldsymbol{\Sigma}}_i(k,l-1) \;+\; \big(1-\alpha_{\mathrm{eff},i}(k,l)\big)\,\boldsymbol{y}(k,l)\,\boldsymbol{y}^H(k,l), \qquad \hat{\boldsymbol{\Sigma}}_i(k,0) = \boldsymbol{y}(k,0)\,\boldsymbol{y}^H(k,0), \tag{6.2}$$

with $\alpha \in (0,1)$ the nominal smoothing constant. The product of the two posteriors is what gives (6.2) its meaning. When both posteriors agree on the interference-only condition ($\hat{p}(\mathcal{H}_i \mid \boldsymbol{y}) \to 1$ and $\hat{p}(\mathcal{M}_1 \mid \boldsymbol{y}) \to 0$) the smoothing factor collapses to $\alpha_{\mathrm{eff},i} \to \alpha$ and the estimate tracks the observation rapidly; when either condition fails, $\alpha_{\mathrm{eff},i} \to 1$ and the recursion *freezes*, so that frames contaminated by the target, or attributed to a different interferer, contribute nothing to $\hat{\boldsymbol{\Sigma}}_i$. The weight is written inside the smoothing factor rather than applied to the instantaneous outer product, as it is in the single-source form of Chapter 3. This keeps each update a convex combination, so the scale of the estimate is preserved; a weight applied to $\boldsymbol{y}\boldsymbol{y}^H$ alone would shrink the covariance towards zero in the very regions where the recursion is meant to hold it constant.

One point requires care. The network emits the target-conditional interferer posterior $p(\mathcal{H}_i \mid \boldsymbol{y}, \mathcal{M}_1)$, whereas (6.1) needs the marginal $\hat{p}(\mathcal{H}_i \mid \boldsymbol{y})$, since the update is performed on bins where the target is *absent*. The marginal follows from the two estimated posteriors through (5.47).

6.4.2. Diagonal Loading

Chapter 5 presumes $\boldsymbol{d}$ and $\boldsymbol{\Sigma}_i$ to be known exactly, whereas both are estimated here. Accumulated error in the recursions, together with imperfect posterior weighting, can leave $\hat{\boldsymbol{d}}$ inaccurate. In a real

deployment the estimated RTF is also perturbed by deviations of the microphone positions from their nominal geometry, by reverberation, and by small movements of the target speaker. We therefore apply diagonal loading, which improves the numerical conditioning of the matrix inverse and penalises the norm of the beamformer weights, making the branch more robust to RTF error:

$$\hat{\boldsymbol{\Sigma}}_i^{\dagger} \;\triangleq\; \Big(\hat{\boldsymbol{\Sigma}}_i \;+\; \delta\,\lambda_i\,\boldsymbol{I}_M\Big)^{-1}, \qquad \lambda_i \;=\; \max_{m,m'}\big|[\hat{\boldsymbol{\Sigma}}_i]_{mm'}\big|, \tag{6.3}$$

where scaling the loading by the magnitude λ_i of the matrix itself makes δ dimensionless and level-independent. The loading used inside the bank is deliberately larger than the purely numerical regularisation applied to the oracle covariances of Section 6.5.1, trading a small amount of null depth for robustness against leakage-induced cancellation.

6.4.3. Target-present Common-base Covariance and the RTF.

The target RTF is obtained by *covariance subtraction* followed by a rank-one projection [48], which requires a second recursion accumulating the observation when the target *is* present. Using the complementary condition—target present and no interferer dominant—gives

$$\alpha_{\mathrm{eff},S}(k,l) \;=\; 1 \;-\; \big(1-\alpha\big)\,\hat{p}(\mathcal{H}_0 \mid \boldsymbol{y})\,\hat{p}(\mathcal{M}_1 \mid \boldsymbol{y}), \tag{6.4}$$

$$\hat{\boldsymbol{\Sigma}}_{S,0}(k,l) \;=\; \alpha_{\mathrm{eff},S}(k,l)\,\hat{\boldsymbol{\Sigma}}_{S,0}(k,l-1) \;+\; \big(1-\alpha_{\mathrm{eff},S}(k,l)\big)\,\boldsymbol{y}(k,l)\,\boldsymbol{y}^{H}(k,l). \tag{6.5}$$

The notation is deliberate: the configuration selected by (6.4) is target present with the $\mathcal{H}_0$ (noise-only) interference model, which is exactly the common-base covariance $\boldsymbol{\Sigma}_{S,0} = \boldsymbol{\Sigma}_N + \boldsymbol{d}\,\boldsymbol{d}^H\sigma_S^2$ of (5.34). The weight is moreover exact: because the network emits the target-conditional interferer posterior, the product $\hat{p}(\mathcal{H}_0 \mid \boldsymbol{y}, \mathcal{M}_1)\,\hat{p}(\mathcal{M}_1 \mid \boldsymbol{y})$ in (6.4) is exactly $\hat{p}(\mathcal{M}_1, \mathcal{H}_0 \mid \boldsymbol{y})$ by the chain rule (5.18), which is the configuration that the $\boldsymbol{\Sigma}_{S,0}$ update requires.

Subtracting the two recursions removes the noise contribution and leaves the rank-one target term. Because the source azimuths and the array are fixed within an utterance (Section 7.1.3), the RTF is frequency-dependent but time-invariant, and the subtraction is performed on the utterance-averaged matrices to suppress the variance of the per-frame estimates:

$$\hat{\boldsymbol{\Phi}}_{xx}(k) \;=\; \tfrac{1}{2}\Big(\boldsymbol{\Delta}(k) + \boldsymbol{\Delta}^{H}(k)\Big), \qquad \boldsymbol{\Delta}(k) \;=\; \frac{1}{L}\sum_{l=1}^{L}\hat{\boldsymbol{\Sigma}}_{S,0}(k,l) \;-\; \frac{1}{L}\sum_{l=1}^{L}\hat{\boldsymbol{\Sigma}}_{0}(k,l) \;\approx\; \boldsymbol{d}(k)\,\boldsymbol{d}^{H}(k)\,\overline{\sigma_S^2}(k), \tag{6.6}$$

the Hermitian symmetrisation removing the small anti-Hermitian part left by finite-precision arithmetic. Since $\hat{\boldsymbol{\Phi}}_{xx}$ is theoretically rank one, the RTF is recovered as its principal eigenvector, normalised so that the reference-microphone entry is unity as the definition of an RTF requires:

$$\hat{\boldsymbol{d}}(k) \;=\; \frac{\boldsymbol{u}_{\max}\big(\hat{\boldsymbol{\Phi}}_{xx}(k)\big)}{\big[\boldsymbol{u}_{\max}\big(\hat{\boldsymbol{\Phi}}_{xx}(k)\big)\big]_{m_{\mathrm{ref}}}}, \qquad \boldsymbol{u}_{\max}(\cdot) \;\triangleq\; \text{eigenvector of the largest eigenvalue.} \tag{6.7}$$

The normalisation in (6.7) simultaneously resolves the arbitrary complex scaling of the eigenvector and enforces $[\hat{\boldsymbol{d}}]_{m_{\mathrm{ref}}} = 1$. Taking the principal eigenvector rather than a single column of $\hat{\boldsymbol{\Phi}}_{xx}$ makes the estimate a best rank-one approximation in the least-squares sense and is markedly more robust when the subtraction in (6.6) leaves a residual full-rank component. Note that this is the covariance-*subtraction* family of RTF estimators, in which the target covariance is formed by subtracting the interference covariance and then projected onto its dominant subspace [48]. It is distinct from covariance *whitening*, which instead takes the principal generalised eigenvector of the pencil $(\hat{\boldsymbol{\Phi}}_{yy}, \hat{\boldsymbol{\Sigma}}_N)$ [45]. Note that a *single* RTF is estimated and shared by all I branches, in accordance with property (i) of Section 5.7: the branches differ only through $\hat{\boldsymbol{\Sigma}}_i$. Note also that the utterance-level averaging in (6.6) makes the RTF estimate non-causal, whereas the covariance recursions (6.2) and (6.5) are strictly causal; a block-online variant, in which the averages are taken over a sliding window, is a direct extension.

6.4.4. Branch a priori SNR and Wiener Gain.

The branch gain $G_{\mathrm{W},i} = \xi_{S,i}/(1+\xi_{S,i})$ of (5.29) requires the per-branch a priori SNR, and hence σ_S^2. Rather than estimating the target PSD separately, we exploit the fact that the MVDR output noise floor

Table 6.1: Provenance of the parameters required by each system at inference. Only the oracle MVDR receives ground-truth statistics; the proposed system estimates every quantity from the observation under posterior control, and the end-to-end baseline forms no explicit statistics at all.

Quantity	Proposed (MC-SPP bank)	Baseline (FT-JNF)	Oracle MVDR
Target RTF $\boldsymbol{d}$	estimated, Eqs. (6.6)–(6.7)	not formed	oracle (direct-path RIR)
Interference cov. $\boldsymbol{\Sigma}_i$	estimated, Eq. (6.2)	not formed	oracle $\boldsymbol{\Sigma}_N$ (isolated noise)
Target PSD σ_S^2	estimated, Eq. (6.11)	not formed	not used (no post-filter)
Posteriors $\hat{p}(\cdot \mid \boldsymbol{y})$	network, from $\boldsymbol{y}$ only	not formed	not used
Oracle information at inference	none	none	$\boldsymbol{\Sigma}_N$, RTF
Number of spatial nulls	I, soft-selected per bin	implicit	$M-1$, fixed

is available in closed form from quantities already computed. With the estimated statistics substituted into (5.31) and (5.32), each branch produces

$$\hat{\boldsymbol{w}}_i(k,l) \;=\; \frac{\hat{\boldsymbol{\Sigma}}_i^{\dagger}\,\hat{\boldsymbol{d}}}{\hat{\boldsymbol{d}}^H\,\hat{\boldsymbol{\Sigma}}_i^{\dagger}\,\hat{\boldsymbol{d}}}, \qquad z_i(k,l) \;=\; \hat{\boldsymbol{w}}_i^H\,\boldsymbol{y}, \qquad \hat{\sigma}^2_{n,\mathrm{post},i}(k,l) \;=\; \frac{1}{\hat{\boldsymbol{d}}^H\,\hat{\boldsymbol{\Sigma}}_i^{\dagger}\,\hat{\boldsymbol{d}}}. \tag{6.8}$$

The output power and the noise floor are smoothed recursively,

$$\hat{\Phi}_{z_i}(k,l) \;=\; \alpha\,\hat{\Phi}_{z_i}(k,l-1) \;+\; (1-\alpha)\,\big|z_i(k,l)\big|^2, \tag{6.9}$$

$$\hat{\Phi}_{n,i}(k,l) \;=\; \alpha_G\,\hat{\Phi}_{n,i}(k,l-1) \;+\; (1-\alpha_G)\,\hat{\sigma}^2_{n,\mathrm{post},i}(k,l), \tag{6.10}$$

and the target PSD at the beamformer output follows by spectral subtraction, giving the a priori SNR and the gain

$$\hat{\sigma}^2_{S,i} \;=\; \max\big(\hat{\Phi}_{z_i} - \hat{\Phi}_{n,i},\; 0\big), \qquad \hat{\xi}_{S,i} \;=\; \frac{\hat{\sigma}^2_{S,i}}{\hat{\Phi}_{n,i}}, \qquad \hat{G}_{\mathrm{W},i} \;=\; \max\!\left(\frac{\hat{\xi}_{S,i}}{\hat{\xi}_{S,i}+\beta},\; G_{\min}\right). \tag{6.11}$$

For $\beta = 1$ the gain in (6.11) is exactly the analytic Wiener gain of (5.29); $\beta > 1$ trades speech distortion for additional residual suppression, in the manner of the parameterised multichannel Wiener filter, and the floor $G_{\min}$ bounds the attenuation applied to any single bin and thereby limits the musical-noise artefacts to which spectral-subtraction SNR estimates are prone. The distortionless constraint $\hat{\boldsymbol{w}}_i^H\hat{\boldsymbol{d}} = 1$ is what makes (6.11) legitimate: because the target passes each branch unit-gain, the excess of the output power over the analytic noise floor is an estimate of the target PSD itself, so no separate PSD estimator is required.

Substituting (6.8)–(6.11) and the two estimated posteriors into (5.20) gives the deployed estimator in full,

$$\hat{S}(k,l) \;=\; \hat{p}(\mathcal{M}_1 \mid \boldsymbol{y}) \sum_{i=0}^{I-1} \hat{p}(\mathcal{H}_i \mid \boldsymbol{y}, \mathcal{M}_1)\;\hat{G}_{\mathrm{W},i}(k,l)\;z_i(k,l), \tag{6.12}$$

in which every factor is a function of the observation alone.

Table 6.1 summarises the provenance of each quantity for the three systems compared in Chapter 7, and answers directly the question of whether that comparison is fair with respect to parameter knowledge. The proposed system receives no oracle information at inference. The end-to-end baseline of Section 7.2 requires no spatial statistics at all—an RTF and a covariance model are never formed, since the complex ratio mask absorbs the entire spatial–spectral operation into a learned function—so there is no corresponding estimation problem to solve, and none to grant it an advantage in. The oracle MVDR of Section 7.2.1, by contrast, is deliberately given both its noise covariance and its relative transfer function from ground truth, precisely so that it measures the ceiling of *linear* spatial filtering rather than the loss incurred by parameter estimation. The comparison of Section 7.5.1 is therefore conservative with respect to the proposed system on two counts: it is scored against an anechoic reference that a spatial filter cannot attain, and it is compared against a linear reference whose statistics are exact while its own are estimated from the mixture.

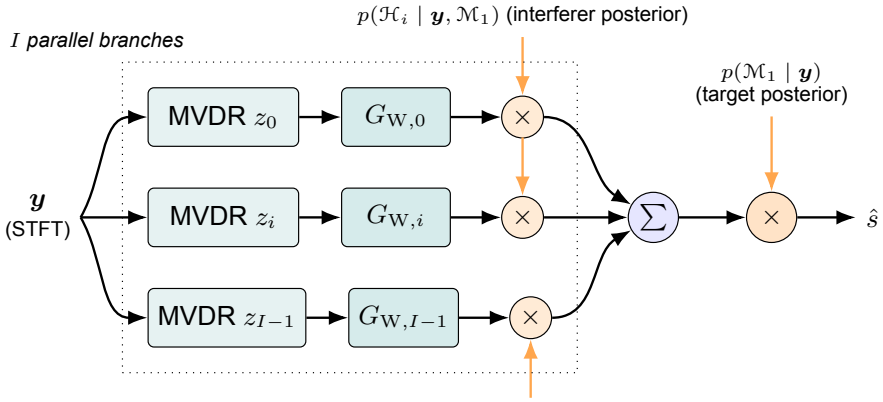


Figure 6.3: Probability-activated MVDR–Wiener bank realising the four-factor estimator (5.20). Each of the I branches applies a target-directed MVDR beamformer z_i with a per-hypothesis interference covariance, followed by a Wiener gain $G_{\mathrm{W},i}$. Branch outputs are weighted by the interferer posterior $p(\mathcal{H}_i \mid \boldsymbol{y}, \mathcal{M}_1)$, summed, and the sum is multiplied by the target posterior $p(\mathcal{M}_1 \mid \boldsymbol{y})$. The spatial statistics parameterising the bank are estimated recursively from the observation under the control of the same two posteriors (Section 6.4), so no oracle information enters at inference.

A final remark closes the loop with the training procedure. The oracle posteriors of Section 6.5.1 are computed from ground-truth quantities, but they enter the system only as regression targets for the network, under the stop-gradient of Section 6.1; they never appear in (6.2)–(6.12). The dashed branch of Figure 6.1 is therefore severed at inference, and the statistics that parameterise the bank at test time are formed exclusively by the recursions of this section.

6.5. Oracle-Guided Training and Loss Design

Training the proposed system poses a supervision problem that does not arise for end-to-end mask estimators: the network must learn to emit posteriors that carry a specific probabilistic meaning, yet the only externally available target is the clean speech waveform. We resolve this with a two-part objective that combines a signal-level enhancement loss with a dense posterior-supervision loss derived from oracle quantities.

6.5.1. Oracle Posterior Supervision

Because the corpus is fully simulated, the ground-truth separated sources, the isolated noise field, and the direct-path relative transfer functions are all available at training time (Section 7.1 of Chapter 7). From these we compute the *oracle* interferer posteriors and target-presence posterior—evaluations of the closed-form expressions (5.39) and (5.49) with the true signal statistics substituted for their estimates. These oracle posteriors provide dense, per-TF-bin references for the network's posterior outputs. As detailed in Section 6.5.2, the oracle interferer posterior enters the training objective directly, anchoring the interferer network to its intended probabilistic meaning, while the oracle target posterior serves as an analysis reference against which the learned one is compared (Section 7.5.3). The oracle computation is performed under a stop-gradient and only during training. At inference the network relies solely on the observation, and the model-based bank is parameterised entirely by observation-driven recursive estimates, so no oracle information is present in the deployed system.

The target-presence oracle is itself a posterior-weighted average of per-hypothesis presence probabilities (5.49), evaluated per TF bin from the true target and interference power spectra; the interferer oracle is the categorical posterior over the interferer hypotheses induced by the true per-hypothesis covariances. The interferer oracle supplies the training signal for the interferer network of Section 6.3; the target oracle provides the reference used to assess the learned target posterior in Section 7.5.3.

6.5.2. Loss Design

The training objective combines a signal-level enhancement loss on the extracted target with a posterior-supervision loss on the interferer network, weighted by a fixed coefficient throughout training:

$$\mathcal{L} \;=\; \mathcal{L}_{\mathrm{enh}} \;+\; \lambda_{\mathcal{H}}\, \mathcal{L}_{\mathcal{H}}. \tag{6.13}$$

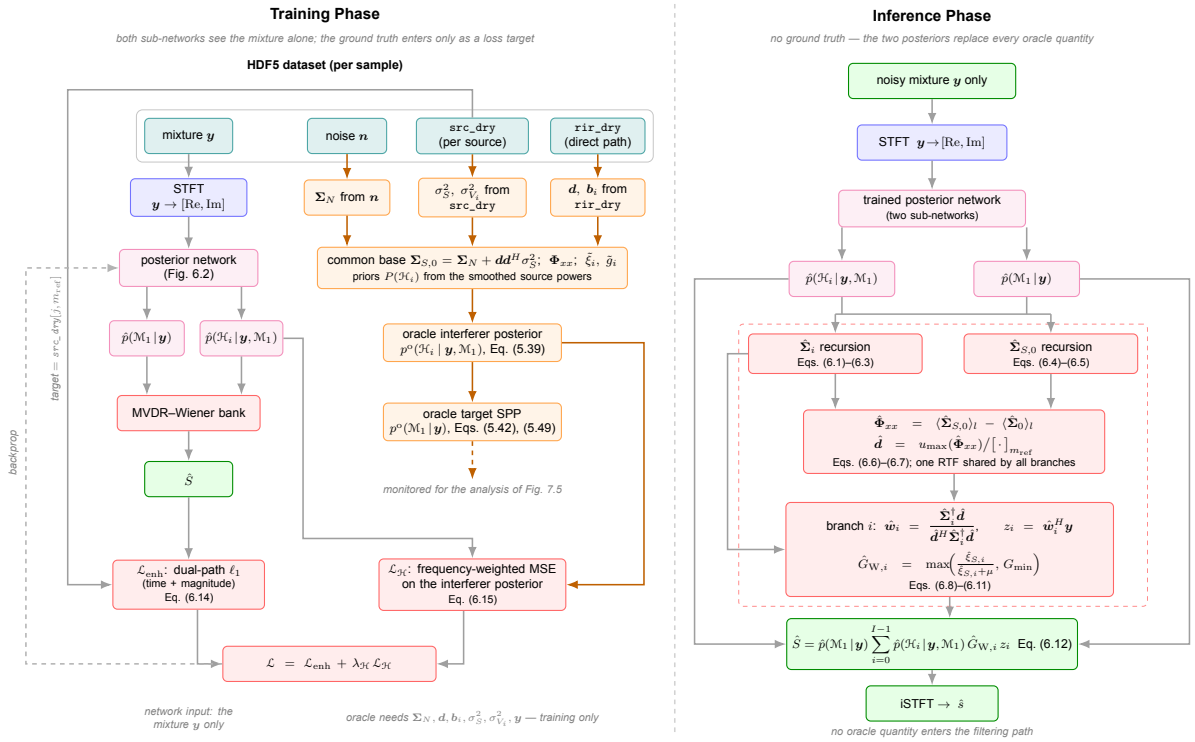


Figure 6.4: Training and inference phases of the proposed composite-hypothesis system. During training (left) the network sees only the multichannel mixture and produces the interferer posterior $\hat{p}(\mathcal{H}_i \mid \boldsymbol{y}, \mathcal{M}_1)$ and the target posterior $\hat{p}(\mathcal{M}_1 \mid \boldsymbol{y})$. These weight the same model-based MVDR–Wiener bank that is used at inference, so the enhanced estimate $\hat{S}$ is produced inside the training loop. The ground-truth quantities enter only through the loss: the oracle interferer posterior of Eq. (5.39), computed from the isolated noise, the per-source dry images and the direct-path RIRs, supervises the interferer network, while the dry target supplies the reference of the dual-path ℓ_1 enhancement loss. The oracle target posterior is evaluated for monitoring only and is not a term of $\mathcal{L}$; the target network is trained through $\mathcal{L}_{\mathrm{enh}}$ alone. At inference (right) the two posteriors replace every oracle quantity: they control the recursive estimation of the per-hypothesis covariances $\hat{\boldsymbol{\Sigma}}_i$ and of the common base $\hat{\boldsymbol{\Sigma}}_{S,0}$, from which a single relative transfer function $\hat{\boldsymbol{d}}$—shared by all branches—is recovered by covariance subtraction, and they then weight the branch outputs of the bank. The dashed box marks the model-based stage of Section 6.4; no oracle information enters the filtering path at inference. In the smoothing-factor expressions $\hat{p}(\mathcal{H}_i)$ abbreviates $\hat{p}(\mathcal{H}_i \mid \boldsymbol{y}, \mathcal{M}_1)$ and $\hat{p}(\mathcal{M}_1)$ abbreviates $\hat{p}(\mathcal{M}_1 \mid \boldsymbol{y})$.

Both terms are active from the first epoch, with a constant weight $\lambda_{\mathcal{H}}$; the interferer posterior and the enhancement mapping are therefore learned jointly rather than in separate stages.

Enhancement loss. The enhancement term is a dual-path ℓ_1 loss evaluated jointly in the time and spectral-magnitude domains, penalising both the retained target and the removed residual. Let $\hat{S}$ denote the estimated target STFT produced by the bank (5.20) and S the anechoic target reference of Section 7.1.3 of Chapter 7, with corresponding waveforms $\hat{s}$ and s. Let $Y_{m_{\mathrm{ref}}}$ be the reference-channel mixture STFT, and define the residual signals $R \triangleq Y_{m_{\mathrm{ref}}} - S$ and $\hat{R} \triangleq Y_{m_{\mathrm{ref}}} - \hat{S}$, with waveforms r and $\hat{r}$. The enhancement loss is

$$\mathcal{L}_{\mathrm{enh}} = \frac{\alpha_\ell}{2}\Big(\underbrace{\big\| s - \hat{s} \big\|_1}_{\text{target, time}} + \underbrace{\big\| r - \hat{r} \big\|_1}_{\text{residual, time}}\Big) + \underbrace{\big\| |S| - |\hat{S}| \big\|_1}_{\text{target, magnitude}} + \underbrace{\big\| |R| - |\hat{R}| \big\|_1}_{\text{residual, magnitude}}, \tag{6.14}$$

where α_ℓ weights the time-domain group relative to the magnitude group. The residual is defined on the shared reference channel, so that the two time-domain terms are linked through the identity $r - \hat{r} = -(s - \hat{s})$; the residual path therefore contributes new information principally through its magnitude term, and the factor $\frac{1}{2}$ on the time-domain group keeps its aggregate weight comparable to a single-path time-domain loss. Penalising the residual in addition to the target discourages the estimator from suppressing target energy in order to reduce interference leakage, since any target component removed into $\hat{R}$ is penalised by the residual terms. This dual-path form is identical to the criterion used by the baseline of Section 7.2, so that both systems are optimised towards the same signal-fidelity objec-

tive and any performance difference is attributable to the estimator structure rather than to a mismatch in training loss.

Interferer-posterior supervision. The interferer network is anchored to its intended probabilistic meaning by penalising the discrepancy between the estimated and oracle interferer posteriors of Section 6.5.1. Write $\hat{p}(\mathcal{H}_i \mid \boldsymbol{y}, \mathcal{M}_1)$ for the estimate, $p^{\mathrm{o}}(\mathcal{H}_i \mid \boldsymbol{y}, \mathcal{M}_1)$ for its oracle counterpart (5.39), and ω_k for the frequency weight at bin k. The posteriors are treated as bounded regression targets, and the discrepancy is measured by a frequency-weighted mean squared error:

$$\mathcal{L}_{\mathcal{H}} \;=\; \frac{1}{I\,K\,L} \sum_{i=0}^{I-1} \sum_{k=1}^{K} \sum_{l=1}^{L} \omega_k \left(\hat{p}(\mathcal{H}_i \mid \boldsymbol{y}, \mathcal{M}_1) - p^{\mathrm{o}}(\mathcal{H}_i \mid \boldsymbol{y}, \mathcal{M}_1) \right)^2. \tag{6.15}$$

At each bin both quantities are categorical distributions, so a Kullback–Leibler objective is also natural; that variant was implemented as well and is reported in Appendix B. The oracle posteriors are computed under a stop-gradient and only during training (Section 6.5.1); at inference the interferer network operates on the observation alone.

Target posterior. The target posterior $p(\mathcal{M}_1 \mid \boldsymbol{y})$ is *not* supervised by a dedicated posterior loss. Because it multiplies the entire bank output in (5.20), it receives gradients directly from the enhancement loss (6.14) and is learned end-to-end. This ties the target-presence estimate to the signal-fidelity objective rather than to an oracle that is itself only an approximation at the frame level, and it avoids imposing a target whose optimal value need not coincide with the level of suppression that minimises the enhancement loss. The learned posterior is compared with its oracle in Section 7.5.3.

7

Experimental Evaluation

Chapter 5 derived the estimator from the mathematical perspective and Chapter 6 specified the algorithm that realises it. This chapter evaluates that algorithm. Section 7.1 describes the simulated multi-talker corpus and the acoustic conditions under which every system is measured; Section 7.2 defines the two reference systems—an end-to-end complex-mask network and a non-neural oracle MVDR beamformer that fixes the ceiling of linear spatial filtering; Section 7.3 states the error metrics and the scoring convention; Section 7.4 lists the configuration of each system together with the controls that reduce the comparison to a single axis; and Section 7.5 reports and analyses the results.

7.1. Dataset Construction and Acoustic Environment

This section describes the simulated multi-talker corpus on which all subsequent experiments are conducted. The dataset is generated by an image-source room-acoustics simulation that places a designated target speaker together with several competing talkers and a directional noise source in a reverberant enclosure, and renders the resulting multichannel mixture at a compact microphone array. The construction is designed to instantiate exactly the composite-hypothesis signal model of Chapter 5: a single target source of interest, a fixed number of spatially separated interferers, and an always-active background-noise field.

7.1.1. Source Material

Speakers clean speech is drawn from the TSP speech corpus [49], resampled to a working sampling rate of $f_s = 16\,\mathrm{kHz}$. Speakers are partitioned at the *speaker* level into disjoint training, validation, and test sets, so that no speaker identity appearing in training recurs at validation or test time. Of the available speakers, three are held out for validation and three for test, with the remainder assigned to training; the split is drawn once from a fixed random seed and held constant across all experiments to guarantee reproducibility. Within each generated mixture, all I speech sources—the target and the $I - 1$ competing talkers—are drawn from this same split-specific pool: each of the target utterances within every group is taken in turn from the pool, and the $I - 1$ interferer utterances are sampled from the corpus independently. Role assignment is by spatial position: the source at 0° is the target and the remaining sources are interferers. Because interferers are sampled with replacement from the full pool, the same speaker may in rare cases contribute more than one source within a mixture; To avoid this, individual target speakers are manually separated with chosen interferers. Also, the fixed per-source azimuths and independent random time offsets nonetheless render such sources spatially and temporally distinct.

The directional interfering noise is drawn from the NoiseX-92 database [50], resampled to $16\,\mathrm{kHz}$. A single noise recording is selected per mixture and rendered as a point source in the room (Section 7.1.4), providing a spatially coherent, non-speech interferer in addition to the competing talkers.

7.1.2. Room and Array Geometry

Each mixture is simulated in an independently randomised shoebox room using the image-source method [51]. The room dimensions are sampled uniformly per mixture, with width in $[2.5, 5.0]\,\mathrm{m}$, length in $[3.0, 9.0]\,\mathrm{m}$, and height in $[2.2, 3.5]\,\mathrm{m}$. Reverberation is controlled through the reverberation time T_{60}, sampled uniformly in $[0.2, 0.5]\,\mathrm{s}$ and converted to wall absorption via the inverse Sabine formula.

The array is a uniform circular array of $M = 3$ omnidirectional microphones with radius $5\,\mathrm{cm}$. The array centre is placed at a randomised position within the room interior, at a fixed height of $1.5\,\mathrm{m}$, with a randomised azimuthal orientation ϕ. All sources are positioned on the horizontal plane relative to this array-centred, rotated reference frame.

7.1.3. Source Configuration

Each mixture comprises $I = 5$ simultaneously active speech sources placed at *fixed* azimuths relative to the array orientation, namely

$$\theta \in \{0^\circ,\, 72^\circ,\, 144^\circ,\, 216^\circ,\, 288^\circ\}, \tag{7.1}$$

i.e. equi-spaced around the array at 72° separation. The source at 0° is designated as the target; the remaining four are treated as interfering talkers. Fixing the source azimuths across the corpus isolates the effect of the spectro-temporal activity patterns and the interference statistics from that of angular variability, consistent with the fixed-geometry assumption under which the closed-form estimator of Chapter 5 is derived.

Each source is placed at a randomised radial distance in $[1.0, 1.5]\,\mathrm{m}$ from the array centre, at a height sampled around $1.6\,\mathrm{m}$; positions are redrawn as needed to ensure they lie within the room interior. Every source signal is normalised and time-shifted by a random offset prior to rendering, so that the competing utterances are temporally misaligned. This misalignment emulates the asynchronous nature of real multi-talker speech scenario, in which speakers do not begin or pause altogether, while increasing the randomness of each mixture samples of the acoustic datasets. To introduce level variability across the interferers, a per-source gain is sampled uniformly in $[0.3, 0.7]$ and applied to each competing talker, while the target source gain is fixed to unity so that the target level is held constant across the corpus.

Each source is rendered through the simulated room to obtain its reverberant image at all M microphones. In parallel, an anechoic (direct-path) rendering of the target is retained as the clean reference signal for training and evaluation; the corresponding direct-path relative transfer function is likewise retained, providing the target steering information required by the oracle quantities of Chapter 5.

One aspect of the level convention governs the interference statistics seen by the estimator: each source, target and interferers alike, is scaled by a single scalar gain that is constant for the whole utterance, and no time-varying gain schedule is imposed.

Figure 7.1 illustrates a representative set of reference-microphone source envelopes under this convention: the target retains unit gain, the interferers are attenuated by their respective constant gains, and the block heights—depicting short-time energy—vary over time solely as a consequence of the underlying speech.

7.1.4. Noise Field

The background-noise field combines two components. A spatially diffuse noise field is synthesised from the microphone geometry to model incoherent ambient noise: per-frequency inter-microphone coherences are imposed through a spherically-isotropic (sinc) coherence model and applied to independent complex-Gaussian sequences, yielding a multichannel diffuse field consistent with the array geometry [52, 53]. The diffuse field is scaled to a target signal-to-noise ratio sampled uniformly in $[5, 10]\,\mathrm{dB}$ relative to the aggregate speech mixture.

Superimposed on the diffuse field is a single directional noise source, drawn from NoiseX-92 and rendered as a point source at a fixed azimuth of 60° and distance $1\,\mathrm{m}$ from the array, at the same T_{60} as the speech sources. This directional component is scaled to a signal-to-noise ratio sampled uniformly in $[5, 10]\,\mathrm{dB}$ relative to the speech mixture. The total noise field is the sum of the diffuse and directional

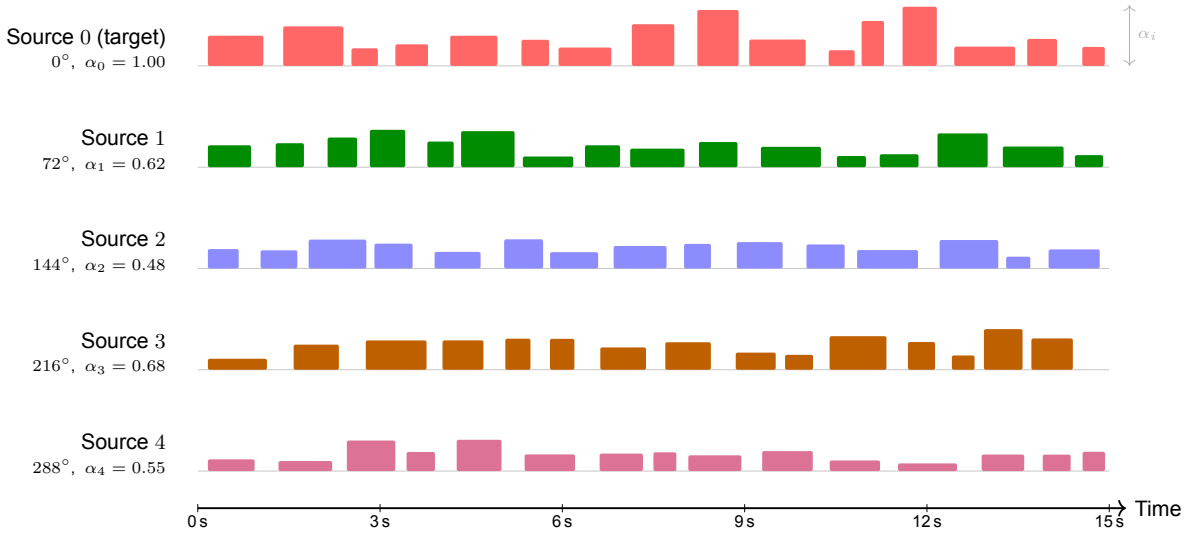


Figure 7.1: Representative reference-microphone energy envelopes of the five sources over one $15\,\mathrm{s}$ mixture. Block heights depict short-time speech energy, which varies over time as an intrinsic property of the speech waveforms; all five sources are active throughout the utterance. Each source is scaled by a single constant gain α_i that is fixed for the duration of the utterance—the target at unit gain ($\alpha_0 = 1$) and the interferers attenuated by $\alpha_i \sim \mathcal{U}[0.3, 0.7]$—so that the per-source level is time-invariant while the source energy is not. No time-varying gain schedule is applied to any source.

components; it is always active, in accordance with the always-active noise assumption of the signal model. Microphone self-noise, which is commonly added at 30–$50\,\mathrm{dB}$ SNR to model sensor noise, is not simulated here. Its omission makes the array output slightly cleaner at high frequencies than a physical array would be, and is noted as a limitation in Section 8.2.

7.1.5. Mixture Signal and Dataset Composition

The multichannel mixture presented to every system is the sum of all reverberant source images and the total noise field,

$$\boldsymbol{y} = \sum_{i=1}^{I} \boldsymbol{y}_i^{\mathrm{rev}} + \boldsymbol{n}, \tag{7.2}$$

where $\boldsymbol{y}_i^{\mathrm{rev}}$ is the gain-scaled reverberant image of source i and $\boldsymbol{n}$ the total (diffuse plus directional) noise field. Each rendered example is $15\,\mathrm{s}$ in duration at $16\,\mathrm{kHz}$.

The corpus comprises 1000 training, 100 validation, and 100 test mixtures (1200 in total), drawn from the speaker-disjoint partition of Section 7.1.1. At 15 s per mixture and a 16 kHz sampling rate, this corresponds to approximately 4.2 h of rendered training audio and 25 min each of validation and test material. This is a small training set by the standards of the field, where corpora of tens of hours are common; the consequences are discussed in Section 8.2. Each mixture is generated from a single random seed fixed by its index within the split, and every stochastic parameter of the acoustic scene—the room dimensions, the array placement and orientation, the source distances and temporal offsets, the per-source gains, the selected speech and noise recordings, and the diffuse- and directional-noise signal-to-noise ratios—is drawn in a fixed order from a random number generator initialised with that seed. The mapping from index to rendered multichannel signal is therefore deterministic: regenerating a given index reproduces the corresponding mixture sample-for-sample.

In these settings, two benefits are preserved: First, the dataset is fully specified by the generation code together with the index set and need not be archived as audio, as it can be regenerated on demand. Second, any two systems evaluated on the same indices are presented with identical acoustic scenes, which is what makes the controlled comparison of Section 6.6 exact: any difference observed between the proposed system and the baseline is attributable to the estimators themselves rather than to a discrepancy in the data on which they were trained and evaluated.

7.1.6. Short-Time Fourier Transform Configuration

All processing is carried out in the STFT domain. Signals are analysed with a 512-sample window ($32\,\mathrm{ms}$ at $16\,\mathrm{kHz}$) and a 256-sample hop ($16\,\mathrm{ms}$, 50% overlap), using a square-root Hann window for both analysis and synthesis to satisfy the constant-overlap-add reconstruction condition. This yields $K = 257$ non-redundant frequency bins per frame, matching the frequency-domain signal model on which the estimator of Chapter 5 operates.

A visualization of acoustic environment in this experiment is shown as follows:

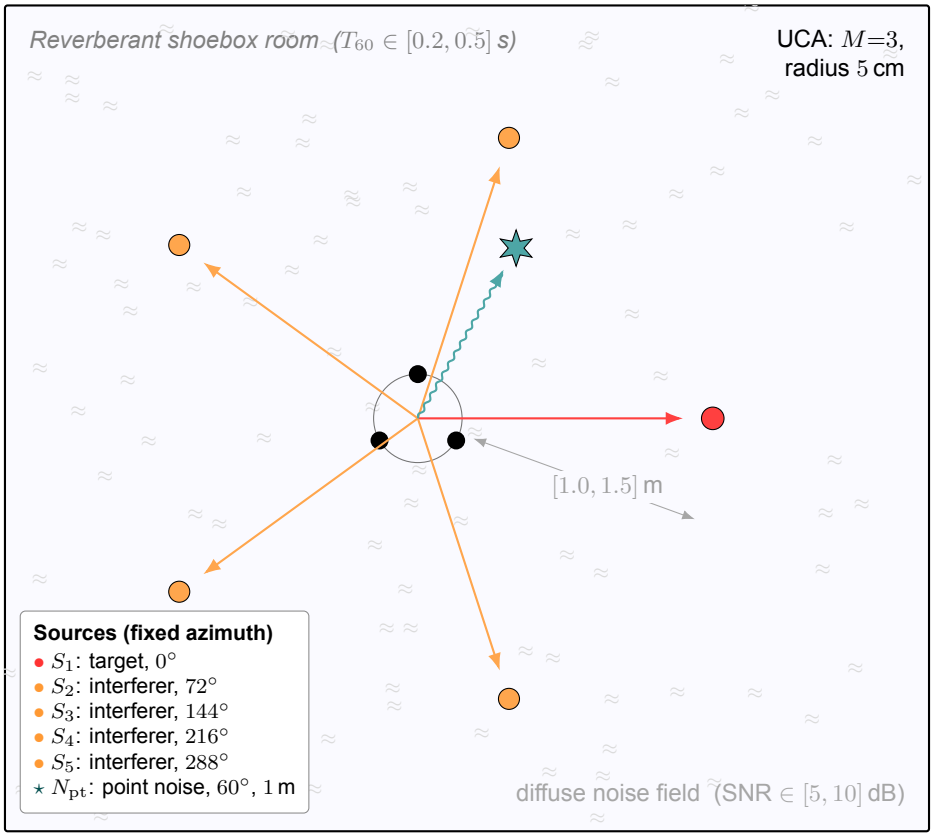


Figure 7.2: Top-view schematic of one representative realisation of the simulated acoustic environment. A uniform circular array of $M=3$ omnidirectional microphones (radius 5 cm, drawn enlarged for visibility) is placed at a randomised position and orientation within a reverberant shoebox room ($T_{60} \in [0.2, 0.5]$ s). Five speech sources S_1–S_5 are fixed at equi-spaced azimuths $\{0°, 72°, 144°, 216°, 288°\}$ relative to the array orientation (listed in the legend), at randomised radial distances in $[1.0, 1.5]$ m; S_1 is the designated target (red) and S_2–S_5 are interfering talkers (orange). A directional point-noise source N_{pt} (teal, drawn from NoiseX-92) is placed at a fixed azimuth of $60°$ and distance 1 m, superimposed on a spatially diffuse background noise field; both noise components are scaled to $\mathrm{SNR} \in [5, 10]$ dB relative to the speech mixture. Room dimensions, array placement and orientation, source distances and temporal offsets, per-interferer gains, and noise realisations are re-randomised for every mixture.

7.2. Baseline System Design

The baseline is a direct, end-to-end multichannel mask estimator representative of the deep non-linear spatial filters of Tesch and Gerkmann [9, 13], against which the composite-hypothesis estimator was shown to be structurally related in Section 5.6. It uses the same joint frequency–time recurrent backbone as the proposed front end, but is applied to the enhancement task in the conventional way: from the stacked-STFT observation it predicts a complex-valued ratio mask, which is applied multiplicatively to the reference-channel mixture to produce the enhanced target. The mask is trained with a joint time-domain and spectral-magnitude ℓ_1 objective evaluated on both the extracted target and the residual, matching the enhancement criterion of Section 6.5.2.

The baseline is deliberately constructed to differ from the proposed system along exactly one axis—the estimator structure—while holding every other experimental factor fixed. Concretely, the two systems

are aligned as follows. They consume the same multichannel mixture (7.2), regress towards the same anechoic target reference, are evaluated on the same held-out test scenes, and share the same STFT front end, reference microphone, and enhancement-loss form. Under these controls, the comparison isolates the contribution of the proposed architecture: an end-to-end network that produces the enhanced signal directly, versus a network that estimates only the activation posteriors and delegates the spatial–spectral filtering to a model-based bank.

Two structural distinctions between the systems are worth making explicit, as they frame the interpretation of the results. First, the baseline is a single network that jointly resolves spatial separation and spectral enhancement, whereas the proposed system decomposes the problem into a learned posterior-estimation stage and an analytic filtering stage. Second, the baseline mask acts end-to-end and can in principle learn a joint separation-and-dereverberation mapping, whereas the model-based bank of the proposed system is a spatial filter and performs no explicit dereverberation. A per-bin spatial filter preserves the target component that is aligned with the target direction—the direct path together with the reflections that fall inside one analysis window—at unit gain, and attenuates the later reverberation only to the extent that it behaves like the interference field being minimised. Some of the target's reverberation therefore survives at the beamformer output, and against an anechoic reference all of it counts as error. The end-to-end baseline is under no such constraint. This asymmetry is accounted for in the analysis of the results.

7.2.1. Non-Neural Classic Linear Beamforming Upper Bound: Oracle MVDR

To place the two neural systems against classical signal processing, we include a third reference point: a purely non-neural, model-based minimum-variance distortionless-response (MVDR) beamformer, granted *oracle* statistics so that it reflects the performance ceiling attainable by a *linear* spatial filter on this task rather than the loss incurred by parameter estimation. The pipeline is shown in Figure 7.3.

The beamformer is applied to the multichannel mixture in the Souden formulation, using two oracle quantities.[1] The noise spatial covariance $\mathbf{\Sigma}_N$ is estimated from the ground-truth isolated noise, and the the target relative transfer function is computed directly from the ground-truth direct-path room impulse response of the speech. The beamformer output is returned to the time domain. A spatial filter preserves the target's propagation delay, whereas the anechoic reference does not, so a delay compensation is applied before scoring: the enhanced signal is shifted to the lag of maximum cross-correlation with the reference, which removes the constant propagation offset without altering the waveform; the same compensation is applied to the unprocessed mixture so that the reported improvement is not an artefact of alignment. Crucially, only the direct path signals and early reflections of the target speech are predominantly retained and, when scored against the anechoic target, the latter count as residual error. The MVDR requires no training and uses the identical mixtures, reference, and test split as the two neural systems.

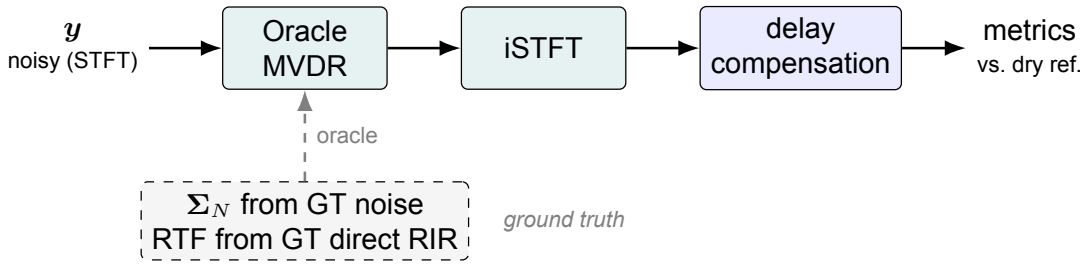


Figure 7.3: Non-neural upper bound: an oracle MVDR beamformer with delay-compensated evaluation. The noise covariance $\mathbf{\Sigma}_N$ is taken from the ground-truth isolated noise and the target relative transfer function from the ground-truth direct-path impulse response. No dereverberation is performed; the beamformer output is the target's reverberant image, delay-aligned to the anechoic reference before scoring. The cascade requires no training and uses the same mixtures, reference, and test split as the neural systems.

[1] The oracle beamformer follows the reference implementations in `github.com/Enny1991/beamformers`.

7.3. Evaluation Metrics

All systems are scored against the same reference—the anechoic (direct-path) image of the target at the reference microphone, retained during rendering as described in Section 7.1.3—using four standard objective measures that between them cover signal fidelity, perceptual quality and intelligibility. For every measure we report both the absolute value on the enhanced signal and the improvement Δ over the unprocessed reference-channel mixture, computed per utterance and then averaged over the test set. Reporting Δ alongside the absolute value removes the per-scene variability of the operating point, which differs from mixture to mixture through the randomised room, source distances and noise realisations.

Scale-invariant signal-to-distortion ratio (SI-SDR). SI-SDR measures waveform-level fidelity after projecting the estimate onto the reference and normalising out any overall gain, so that a system cannot improve its score by rescaling its output. It is the most direct measure of how much of the target waveform has been recovered and how much interference and distortion remain. Being a waveform measure it is sensitive to phase and to the reverberant tail, and therefore penalises a spatial filter that preserves the target's reverberant image when that image is scored against an anechoic reference.

Perceptual evaluation of speech quality (PESQ). PESQ models perceived quality on a mean-opinion-score scale and responds to artefacts—musical noise, spectral holes, residual interference—that a purely energetic measure need not penalise in proportion to their audibility. We report the wide-band variant at the 16 kHz working rate.

Short-time objective intelligibility (STOI) and its extended variant (ESTOI). STOI correlates short-time envelopes of reference and estimate within octave bands and predicts intelligibility. ESTOI additionally accounts for the correlation *between* spectral bands within a segment, which makes it markedly more reliable for modulated, speech-like interference—precisely the regime of the present corpus, in which the interference consists of competing talkers. Both are reported, since the two behave differently under multi-talker interference and their disagreement is itself informative.

7.4. Experimental Configuration

Both systems are trained and evaluated on the corpus of Section 7.1 ($I = 5$ sources, $M = 3$ microphones, 16 kHz), using the STFT front end of Section 7.1.6 and reference channel m_{ref}. Optimisation uses Adam with value-clipped gradients; each utterance is processed in randomly positioned fixed-length segments. The configurations of the two systems are summarised in Table 7.1; together they realise the controlled comparison of Section 7.2. We report the settings that bear on the comparison; the remaining constants—the recursive smoothing factors and the diagonal-loading terms that regularise the covariance inverses in the model-based bank, together with the fixed Wiener floor—follow the classical MC-SPP-guided pipeline of Chapter 3 and are held fixed across runs.

Posterior-supervision loss. The posterior-supervision term $\mathcal{L}_{\mathcal{H}}$ is the frequency-weighted mean squared error (6.15); the Kullback–Leibler variant of Appendix B was also implemented but is not used for the reported results. The reason is one of experimental control rather than of principle: the MSE form matches the regression-style (ℓ_1/ℓ_2) character of the enhancement loss and of the baseline objective of Section 7.2, so that the two neural systems are trained under a consistent class of criteria and the comparison isolates the estimator structure rather than the choice of posterior loss.

The recurrent width of each system and the time-domain loss weight α_ℓ are tuned independently for the two systems and reported as used.

7.5. Results and Analysis

7.5.1. Enhancement Performance

Table 7.2 reports full-utterance results on the held-out test set for the two neural systems and, as a non-neural reference point, the oracle MVDR beamformer of Section 7.2.1, measured by SI-SDR, PESQ, STOI, and extended STOI (ESTOI), together with the improvement Δ over the unprocessed mixture.

Table 7.1: Configuration of the proposed and baseline systems. Settings above the rule are held identical to realise the controlled comparison of Section 7.2; settings below are tuned independently.

Setting	Proposed	Baseline
Sampling rate	16 kHz	16 kHz
STFT window / hop	512 / 256	512 / 256
Analysis window	$\sqrt{\cdot}$-Hann	$\sqrt{\cdot}$-Hann
Reference channel	m_{ref}	m_{ref}
Segment length	2.5 s	2.5 s
Optimiser	Adam (10^{-3})	Adam (10^{-3})
Gradient clipping	value, 5.0	value, 5.0
Train / val / test	1000 / 100 / 100	1000 / 100 / 100
Enhancement loss	dual-path ℓ_1 (6.14)	dual-path ℓ_1
Estimator output	activation posteriors	complex ratio mask
Recurrent width	384 / 192	512 / 256
Time-domain weight α_ℓ	3	10
Posterior weight $\lambda_{\mathcal{H}}$	0.5	—
LR schedule	plateau ($\times 0.9$)	constant

The operating point is demanding: the input mixture has an SI-SDR of -8.3 dB, with five simultaneously active talkers, a directional interferer, and a diffuse noise field. All three systems improve over the noisy input, but by markedly different margins.

The oracle MVDR establishes the ceiling of *linear* spatial filtering on this task. Despite oracle noise statistics and an oracle relative transfer function—so that no estimation loss is incurred—it improves SI-SDR by only $+5.19$ dB and, tellingly, leaves the enhanced signal at a still-negative -3.11 dB; it barely moves perceptual quality ($+0.01$ PESQ) and yields only moderate intelligibility gains ($+0.09$ STOI, $+0.13$ ESTOI). Two structural limitations account for this ceiling. First, the spatial degrees of freedom: with $M = 3$ microphones the beamformer can place at most $M - 1 = 2$ independent spatial nulls, which is insufficient to suppress the four simultaneous interferers, so residual interference necessarily remains. Second, the absence of dereverberation: the beamformer is distortionless towards the direct-path of the target speech. However, as a spatial filter, MVDR itself does not contain any dereverberation capabilities, which means it is not capable of separating the early reflections from the direct path. Since the anechoic reference contains only the direct path signals, everything the beamformer retains beyond the direct path is scored as error. In modern signal processing operations the reverberant image usually requires special blocks (like weighted prediction error (WPE) [54]) to perform the dereverberation procedure. The MVDR thus quantifies precisely how far a classical linear front end can go here, and motivates the nonlinear approaches that follow.

Both neural systems improve SI-SDR by roughly twice the margin of the linear upper bound ($+10.28$ and $+11.26$ against $+5.19$ dB), and, unlike the MVDR, lift the enhanced signal to positive SI-SDR. This is the empirical counterpart of the argument in Section 5.6: a nonlinear filter is not bound by the $M - 1$ null budget of a single linear beamformer, and can suppress more interferers than the array has spatial degrees of freedom. Between the two neural systems, the proposed composite-hypothesis estimator attains the larger SI-SDR improvement, exceeding the end-to-end baseline by roughly one decibel ($+11.26$ against $+10.28$ dB), and the larger PESQ improvement ($+0.53$ against $+0.50$). On the two intelligibility measures the neural systems are statistically indistinguishable, the differences (0.01 in both STOI and ESTOI) lying well within a standard deviation. Both remove enough interference to restore intelligibility, so the intelligibility measures saturate to a common level; the advantage of the probability-activated bank appears in the signal-level fidelity of the recovered target—captured most directly by SI-SDR, and to a lesser extent by PESQ—which reflects the per-bin adaptation of the spatial null to the dominant interferer. The relatively small standard deviations of the improvements (e.g. ± 1.69 dB for the proposed SI-SDR gain) indicate that this behaviour is consistent across the test set rather than driven by a few favourable scenes.

It is worth noting the criterion against which these results are obtained. All systems are scored against the *anechoic* target. The end-to-end baseline can in principle attain this reference in full, since its mask

can learn a joint separation-and-dereverberation mapping; both the proposed bank and the oracle MVDR are spatial filters that preserve the target's reverberant image and perform no dereverberation, so both face the same anechoic-reference handicap. That the proposed system nonetheless matches or exceeds the baseline, and far exceeds the linear upper bound, indicates that the gain from the model-based spatial–spectral filtering more than compensates for the dereverberation handicap of the criterion.

Table 7.2: Full-utterance results on the test set (mean $\pm$ std). Δ is the improvement over the unprocessed mixture (mean $\pm$ std); all systems are scored against the same anechoic target reference on the same test scenes. The oracle MVDR is a non-neural, linear upper bound. Best Δ per metric in bold.

System		SI-SDR (dB)	PESQ	STOI	ESTOI
Unprocessed (noisy)		-8.30 ± 2.64	1.03 ± 0.01	0.58 ± 0.06	0.24 ± 0.06
Oracle MVDR	Enh.	-3.11 ± 2.66	1.04 ± 0.01	0.67 ± 0.07	0.37 ± 0.09
	Δ	$+5.19 \pm 1.08$	$+0.01 \pm 0.15$	$+0.09 \pm 0.02$	$+0.13 \pm 0.04$
Baseline (FT-JNF)	Enh.	1.97 ± 2.82	1.53 ± 0.13	0.84 ± 0.06	0.63 ± 0.12
	Δ	$+10.28 \pm 1.34$	$+0.50 \pm 0.13$	$\mathbf{+0.26 \pm 0.04}$	$\mathbf{+0.38 \pm 0.08}$
Proposed (MC-SPP bank)	Enh.	2.95 ± 1.98	1.56 ± 0.13	0.83 ± 0.06	0.61 ± 0.12
	Δ	$\mathbf{+11.26 \pm 1.69}$	$\mathbf{+0.53 \pm 0.13}$	$+0.25 \pm 0.03$	$+0.37 \pm 0.08$

7.5.2. Model Size and Computational Cost

Table 7.3 gives a per-layer breakdown of the parameter and MAC counts of the two neural systems, and Table 7.4 summarises the aggregate cost, with multiply–accumulate operations (MACs) reported both per 2.5 s input segment and normalised per second of audio so that the figure is independent of the segment length. The counts follow from the layer structure of each network: the baseline is a single bidirectional FT-JNF backbone with recurrent widths $512/256$, whereas the proposed system comprises two parameter-independent sub-networks (Section 6.3) with recurrent widths $384/192$ each, its per-layer total being the sum of the two sub-networks net_a and net_b in Table 7.3.

Table 7.3: Per-layer parameter and per-second MAC breakdown of the two neural systems. MACs/s is the multiply–accumulate rate in GMACs per second of audio; "$\times 2$" denotes the two parameter-independent sub-networks of the proposed system. BLSTM denotes a bidirectional LSTM layer.

Layer	Structure	Params	MACs/s (GMACs)
Proposed (MC-SPP bank), $384/192$, two sub-networks			
lstm1 $\times 2$	BLSTM($6 \to 384$)	$1{,}204{,}224 \times 2$	38.67
lstm2 $\times 2$	BLSTM($768 \to 192$)	$1{,}477{,}632 \times 2$	47.60
ff (net_a)	Linear($384 \to 5$)	$1{,}925$	0.03
ff (net_b)	Linear($384 \to 1$)	385	0.01
Total	net_a $2{,}683{,}781$ + net_b $2{,}682{,}241$	$\mathbf{5{,}366{,}022}$	$\mathbf{86.31}$
Baseline (FT-JNF), $512/256$, complex ratio mask			
lstm1	BLSTM($6 \to 512$)	$2{,}129{,}920$	34.24
lstm2	BLSTM($1024 \to 256$)	$2{,}625{,}536$	42.31
ff	Linear($512 \to 2$)	$1{,}026$	0.02
Total		$\mathbf{4{,}756{,}482}$	$\mathbf{76.57}$

Despite carrying two separate recurrent stacks, the proposed system is of comparable cost to the single-network baseline: at $384/192$ it uses about 12.8% more parameters and about 12.7% more compute than the baseline. The two effects nearly cancel—the doubling from the two sub-networks is offset by their smaller width—so that estimating the posteriors with two decoupled sub-networks incurs only a modest overhead over a single end-to-end mask network of the same family, while yielding the higher SI-SDR and PESQ of Table 7.2.

The decoupled architecture also degrades gracefully under width reduction. Halving the recurrent widths to $256/128$ brings the two sub-networks to almost exactly half the baseline's cost (roughly 50%

of its MACs), and the accuracy penalty is small: in that lighter configuration the system still attains a best PESQ improvement of $+0.45$ and STOI/ESTOI improvements of approximately $+0.22$ and $+0.32$, against $+0.53$, $+0.25$, and $+0.37$ for the full $384/192$ model. The method therefore offers a favourable accuracy–cost trade-off: a substantial reduction in computation costs only a small, controlled compromise in enhancement quality.

Table 7.4: Aggregate model size and computational cost. MACs are reported per $2.5\,\mathrm{s}$ input segment and per second of audio (GMACs/s); the last column is the cost relative to the baseline.

System	Params	MACs (2.5 s)	GMACs/s	rel. baseline
Baseline (FT-JNF, $512/256$)	$4.76\,\mathrm{M}$	$191.42\,\mathrm{G}$	76.57	100%
Proposed (MC-SPP, $384/192$)	$5.37\,\mathrm{M}$	$215.77\,\mathrm{G}$	86.31	112.7%

7.5.3. Interpretability of the Estimated Posteriors

A property that distinguishes the proposed system from an end-to-end mask estimator is that its learned front end produces quantities with an explicit probabilistic meaning, which can be inspected directly. This provides an interpretability handle that a complex ratio mask does not expose: one can ask whether the network has recovered the intended signal-activity decomposition, and not merely an opaque mapping that reduces the loss.

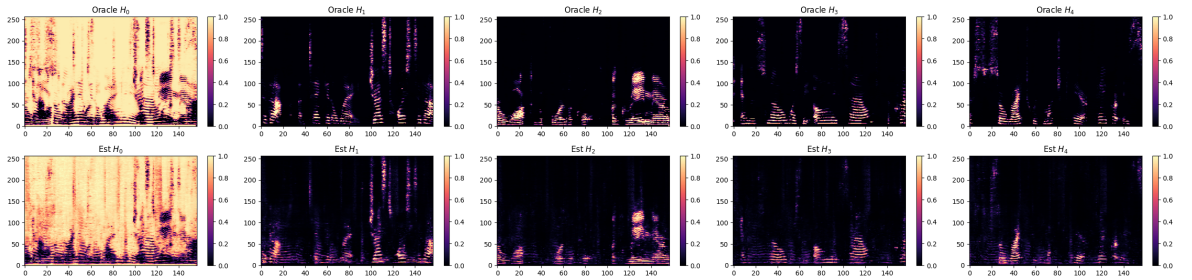


Figure 7.4: Estimated versus oracle interferer posteriors $\{p(\mathcal{H}_i \mid \boldsymbol{y}, \mathcal{M}_1)\}_{i=0}^{4}$ for a representative test utterance. Top row: oracle; bottom row: network estimate. $\mathcal{H}_0$ is the noise-only hypothesis; $\mathcal{H}_1$–$\mathcal{H}_4$ correspond to the four interfering talkers. Axes are frequency bin (vertical) against time frame (horizontal); colour encodes posterior probability in $[0, 1]$.

Figure 7.4 shows the interferer posteriors $\{p(\mathcal{H}_i \mid \boldsymbol{y}, \mathcal{M}_1)\}_{i=0}^{4}$, oracle above and network estimate below. The most conspicuous feature—the near-uniformly bright field of the noise-only panel $\mathcal{H}_0$ against the sparse, dark foregrounds of $\mathcal{H}_1$–$\mathcal{H}_4$—is a direct consequence of the posterior (5.39) and the sparsity of speech. The hypotheses are mutually exclusive and, through the softmax normalisation, satisfy $\sum_{i=0}^{I-1} p(\mathcal{H}_i \mid \boldsymbol{y}, \mathcal{M}_1) = 1$ at every TF bin; the five panels therefore partition a unit mass and must be read jointly. At a bin that carries no dominant interferer harmonic—that is, a bin excited only by the diffuse background—every interferer detection statistic in (5.39) collapses towards zero: the interferer SNR $\tilde{\xi}_i$ and the beamformer-output term $\tilde{g}_i$ vanish when no energy aligns with the corresponding steering direction $\boldsymbol{b}_i$, and the posterior mass defaults to $\mathcal{H}_0$. Because at most a few of the five talkers are simultaneously voiced at any single TF point, the overwhelming majority of bins are inactive for *any given* interferer, and $\mathcal{H}_0$ accordingly claims the entire sparse background—hence its bright, near-saturated field. Each interferer panel $\mathcal{H}_i$, by contrast, is illuminated only along the narrow harmonic ridges where talker i is momentarily the dominant source, and is dark elsewhere. The complementarity is exact: superimposing the four foreground panels reconstructs the support of all competing-talker activity, and its set complement is precisely the bright region of $\mathcal{H}_0$.

The estimated posteriors (lower row) reproduce this assignment faithfully. The differences from the oracle are systematic rather than random: the sharp oracle ridges broaden into gradients, and the $\mathcal{H}_0$ background is scattered with intermediate values. Both occur at bins where two interferers are active at once, so that the single-dominant-interferer assumption of Section 4.5 does not hold. At such a bin the network divides the probability between the two corresponding hypotheses instead of assigning it

to one, which is what the scattered intermediate values are. These are bounded departures rather than hard misclassifications: the assignment of harmonic regions to interferers is preserved.

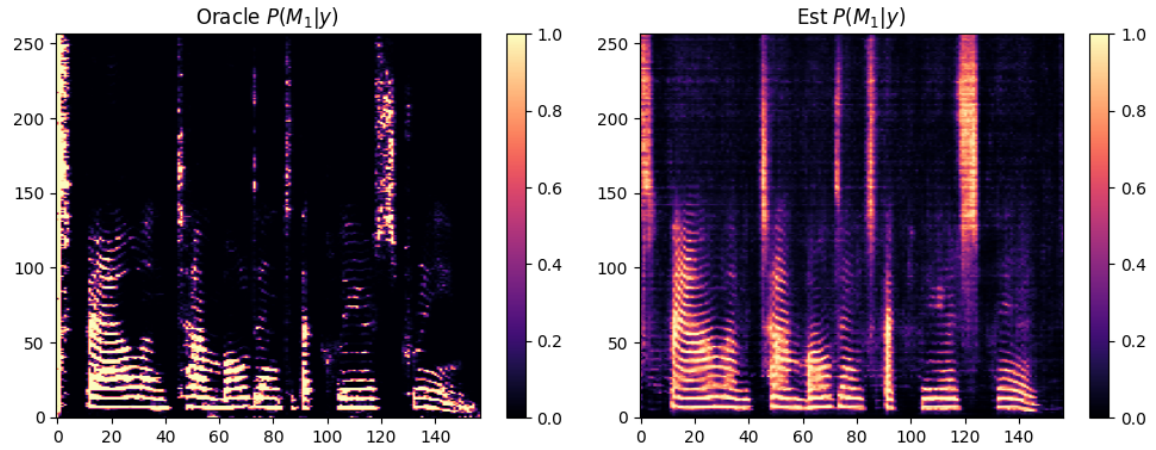


Figure 7.5: Estimated versus oracle target-presence posterior $p(\mathcal{M}_1 \mid \boldsymbol{y})$ for the same test utterance. The learned gate (right) tracks the oracle (left) and localises the target's harmonic activity, despite being trained only through the enhancement loss.

Figure 7.5 shows the target posterior $p(\mathcal{M}_1 \mid \boldsymbol{y})$, and its resemblance to a clean-speech spectrogram is not incidental but a property that the estimator is mathematically obliged to exhibit. From (5.49), $p(\mathcal{M}_1 \mid \boldsymbol{y})$ is a strictly decreasing function of the interferer-posterior-weighted average of the per-hypothesis odds A_i, each of which follows the multichannel MC-SPP form and is monotonically decreasing in a quadratic detection statistic $\beta_{S,i} = \boldsymbol{y}^H \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\Phi}_{xx} \boldsymbol{\Sigma}_i^{-1} \boldsymbol{y}$, in which the target power spectral density $\boldsymbol{\Phi}_{xx} = \boldsymbol{d}\,\boldsymbol{d}^H \sigma_S^2$ enters linearly. Since $\sigma_S^2(k,l) = |S(k,l)|^2$ is the instantaneous power of the clean target at the reference microphone, the statistic is large exactly where the target's own magnitude spectrum is large and small where it is not. The presence posterior is therefore, to first order, a soft-thresholded image of the anechoic target magnitude spectrogram, and must reproduce its defining structure: the harmonic comb of the target pitch, the formant envelope shaping the comb, and the syllabic pattern of onsets and offsets in time.

The two-dimensional posteriors of Figures 7.4 and 7.5 summarise the full time–frequency plane; to examine the temporal fidelity of the estimate more closely, Figure 7.6 plots the estimated and oracle posterior trajectories over time at a single frequency bin (1 kHz), for the target and the first two interferers. The estimated trajectory follows the oracle's on/off activity pattern frame by frame, rising and falling in step with each source's harmonic excitation at that frequency; the shaded intervals mark the frames in which the target is dominant. The residual discrepancies are the smooth, partial-activation excursions expected of a soft posterior at transitions and at frames where more than one source contributes—precisely the graceful degradation of the single-dominant-interferer approximation anticipated in Section 4.5, rather than hard estimation errors.

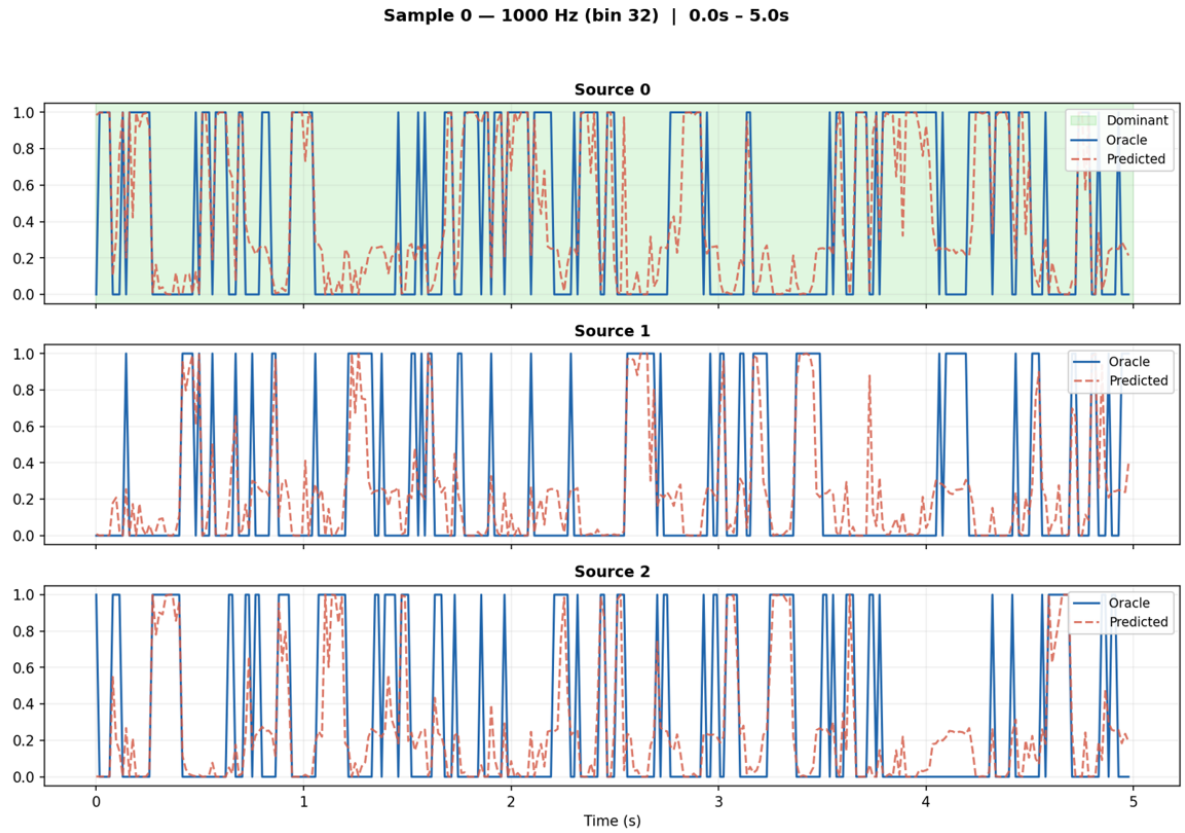


Figure 7.6: Tracking of the oracle posterior by the network estimate. Temporal posterior trajectories at a single frequency bin (1 kHz) over a 5 s excerpt, for the target (Source 0) and the first two interferers (Sources 1 and 2). Solid: oracle; dashed: network estimate. The shaded intervals in the top panel mark the frames in which the target is dominant. The estimate tracks the oracle's frame-by-frame activity, with soft transitions at bins where multiple sources overlap.

7.5.4. Spectrogram Comparison

We now examine the enhancement in the time–frequency domain. Figures 7.7 and 7.8 show reference-microphone spectrograms for the proposed system and the baseline respectively, on the same test utterance; in each, the three panels are, left to right, the noisy mixture, the anechoic target reference, and the enhanced output. The noisy and target panels are common to both figures and define the problem: the noisy mixture is a dense superposition of five talkers' harmonics filled in by broadband noise, in which the target is not visually separable, while the anechoic target is a sparse harmonic comb shaped by a formant envelope and punctuated by syllabic onsets and offsets—the structure each system must recover. We describe the two enhanced outputs in turn and then compare them directly. A single utterance is illustrative rather than conclusive; the quantitative ranking of the two systems is established over the full test set in Table 7.2, and the spectrograms serve only to characterise *how* the two enhanced outputs differ.

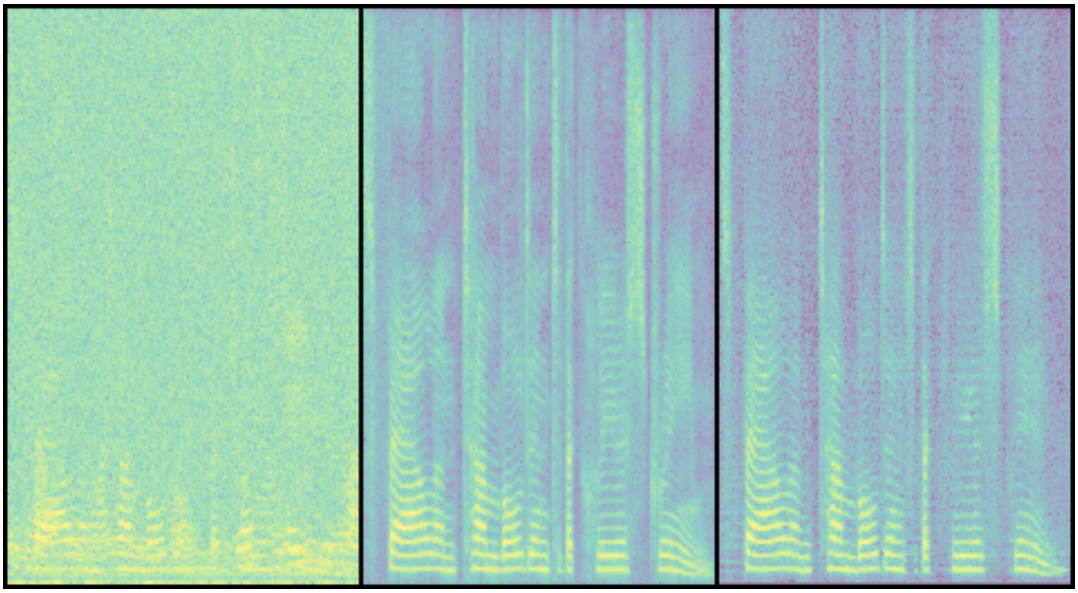


Figure 7.7: Proposed system (MC-SPP bank), common test utterance. Left to right: noisy mixture, anechoic target reference, enhanced output.

For the proposed system (Figure 7.7), the enhanced panel recovers the target's harmonic comb along the target-active frames and removes the dense competing-talker harmonics that dominate the noisy panel. Its residual, however, has a distinctly granular character: the target-inactive background is not a smooth floor but a finely fluctuating, noise-like field. This texture is the expected signature of the model-based bank. The per-branch spatial covariances and the Wiener post-filter gain are estimated recursively from the observation under posterior control, and in low-energy time–frequency bins these estimates fluctuate strongly, so that the applied gain varies rapidly from bin to bin, retaining a low-level granular residual across the background. Faint interferer traces additionally survive at isolated points where two talkers are simultaneously strong—the bins at which the single-dominant-interferer assumption of the interferer posterior is violated (Section 7.5.3).

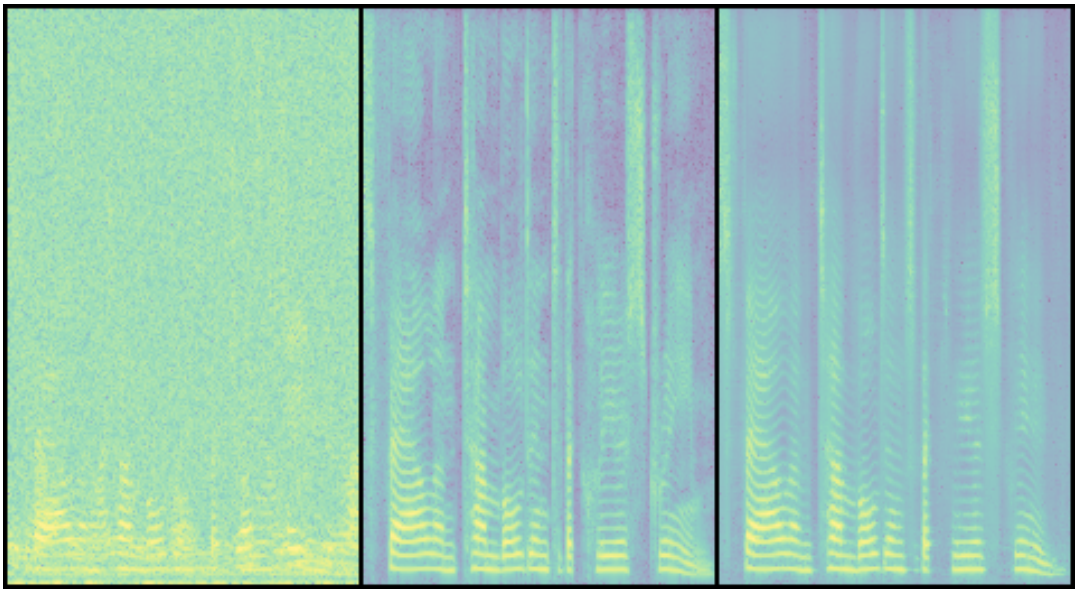


Figure 7.8: Baseline system (FT-JNF), same test utterance. Left to right: noisy mixture, anechoic target reference, enhanced output.

For the baseline (Figure 7.8), the enhanced panel likewise recovers the target harmonics and strongly

attenuates the competing talkers, but its residual is markedly smoother. The complex ratio mask is a single learned function that drives the target-inactive regions towards zero directly, without the per-bin covariance inversion and gain estimation that inject fluctuation in the model-based bank; the baseline background is consequently largely free of the granular, musical-noise texture of the proposed output. This smoothness is not without cost. Because the mask acts multiplicatively on the mixture, the output inherits the mixture's own time–frequency structure and can only attenuate, not synthesise, energy: where the mask is imperfect it tends to leave coherent, partially-suppressed structure rather than grain, and target harmonics buried below the interference tend to be recovered with reduced amplitude. These are structural rather than granular residuals, and are consistent with the slightly lower SI-SDR improvement of the baseline in Table 7.2.

Comparing the two enhanced outputs directly, both recover the target's harmonic comb and remove the bulk of the four competing talkers to a comparable degree, and their harmonic reconstructions are visually similar; this common suppression of the interference is why the two systems are statistically indistinguishable on the intelligibility measures (STOI and ESTOI) in Table 7.2. The salient difference lies in the character of the residual: the proposed bank leaves a low-level granular, noise-like background, whereas the mask-based baseline leaves a smoother but more structured residual. It must be emphasised that this granular appearance does *not* correspond to lower signal fidelity—the proposed system attains the higher SI-SDR and PESQ improvements of Table 7.2. The grain is a high-spatial-frequency, low-power residual that contributes little to the total error energy, whereas SI-SDR is dominated by the recovery of the target and the suppression of the high-energy interferers. Consequently, the proposed system trades a mild musical-noise residual for stronger interference suppression.

8

Conclusion

8.1. Summary

This thesis addressed the extraction of one target talker from a mixture of several competing talkers in reverberant, noisy conditions. The starting point is that a neural network need not produce the enhanced signal itself. It can instead estimate the multi-source speech presence probabilities, which appear exactly as the posterior weights of the MMSE-optimal nonlinear spatial filter. The filtering is then carried out by a deterministic filter prescribed by the signal model. This keeps the analytic structure of classical beamforming and uses the network only where model-based estimation is weakest, namely in estimating the signal statistics.

Chapter 3 established the statistical foundation. It derived the a posteriori speech presence probability from a two-state hypothesis test under complex Gaussian priors, together with its main use: a soft-weighted recursive noise PSD estimator that replaces a hard voice-activity decision. It then extended the model to the multichannel case and derived the MC-SPP. Two features of that derivation recur throughout the thesis. First, the MC-SPP detection statistic is proportional to the squared output of an MVDR beamformer steered at the target, so that detection and spatial filtering share the same quadratic form. Second, the MC-SPP acts as an interface for parameter estimation: it arbitrates the recursive updates of the noise and speech spatial covariances, from which the relative transfer function is recovered by covariance subtraction and the beamformer weights are then formed.

Chapter 4 built the Bayesian framework for the multi-talker case, where a single noise covariance is no longer adequate. It introduced the multi-source signal model and the single-dominant-interferer approximation, which follows from the W-disjoint orthogonality of speech. It then set out the Bayesian MMSE expansion and identified the two modelling choices that fix its closed form: the distributional priors, and the hypothesis set. The target prior was taken Gaussian, and the interferers and the noise were modelled as complex Gaussian. This gives the central observation of the chapter. Although every individual component is Gaussian, the aggregate interference is a Gaussian *mixture* once a hypothesis structure is imposed. The optimal estimator separates into an MVDR beamformer followed by a spectral post-filter only if the interference is Gaussian, so that separation fails here, and the MMSE-optimal estimator is a jointly spatial–spectral nonlinear function of the observation. The hypothesis set was then taken as the Cartesian product of target presence and interferer dominance, spanning $2I$ composite hypotheses.

Chapter 5 developed the main theoretical contribution. The optimal estimator was derived in closed form as a product of three factors: the target posterior, the interferer posterior, and a per-hypothesis MVDR–Wiener filter. Closed-form expressions were obtained for both posteriors, using the matrix determinant lemma and the Sherman–Morrison identity to express the detection statistics about a common base covariance shared by all interferer hypotheses. One methodological point is worth repeating. A natural route to the same architecture is to assume that the two posteriors factorise. That assumption is unnecessary, and in general it is false, because the observation couples the two hypothesis axes even when they are independent a priori. The chain rule gives the same product structure as an exact

identity, at the single cost of reading the second factor as the posterior *conditioned on target presence*. With that reading, the interferer posterior follows from a Gaussian conditional likelihood, and the target posterior follows from the per-hypothesis presence probabilities by an exact combination in the odds domain. The estimator was also shown to be structurally identical, under the Gaussian prior $\nu = 1$, to the Gaussian-mixture MMSE filter of Hendriks et al. [10] and Tesch and Gerkmann [9].

Chapter 6 turned that result into an algorithm. The closed-form estimator assumes that the target relative transfer function, the per-hypothesis interference covariances and the target power spectral density are known, and none of these is available at inference. The chapter therefore separated what the deployed system observes—the multichannel mixture, and nothing else—from what exists only during training. Two sub-networks with no shared parameters estimate the two posteriors from the observation alone. The three signal-model parameters are not learned. They are formed by recursive averaging of the observation, controlled by those same posteriors, so that each covariance is updated only at the time–frequency bins attributed to its hypothesis. Ground-truth quantities appear in one place only, as stop-gradient regression targets for the interferer sub-network; they appear in no equation of the filtering path. The target posterior receives no supervision of its own and is learned through the enhancement loss, which makes its agreement with the oracle an emergent property rather than a fitted one.

Chapter 7 evaluated the resulting system. A simulated corpus of five simultaneously active talkers, a directional noise source and a diffuse noise field in randomised reverberant rooms was constructed, with per-source levels held constant over each utterance. The proposed system was compared against two references under identical mixtures, target reference and test split: an end-to-end complex-mask network, and a non-neural oracle MVDR beamformer that fixes the ceiling of linear spatial filtering. The results support the thesis in three respects. In enhancement quality (Table 7.2), the proposed system gave the highest SI-SDR and PESQ improvements, matched the end-to-end baseline on intelligibility, and roughly doubled the SI-SDR improvement of the oracle linear MVDR. That last point is direct evidence that a nonlinear filter can suppress more interferers than the array has spatial degrees of freedom. In computational cost (Table 7.4), the two sub-networks together were comparable in size to the single-network baseline, and a half-width variant retained most of the enhancement quality at half the cost. In interpretability, the estimated posteriors recovered the intended signal-activity decomposition: the interferer posterior partitioned the time–frequency plane among the noise-only and the per-interferer hypotheses, and the target posterior reproduced the harmonic structure of the clean target spectrogram, as the MMSE construction requires.

Taken together, these chapters show that a network trained to estimate the multi-source speech presence probability can drive a model-based nonlinear spatial filter. That filter matches or exceeds an end-to-end mask network of the same family, exceeds the linear oracle by a wide margin, and keeps the analytic structure and the physical interpretation of classical beamforming.

8.2. Limitations and Future Work

The results above were obtained under a number of deliberate simplifications and assumptions. This section states them, and the further work to extend the system's robustness and the generalization ability of the experiment are also listed below.

Fixed source azimuths. The five sources occupy the same five azimuths in every mixture (Section 7.1.3). This is what makes the interferer hypotheses identifiable: hypothesis $\mathcal{H}_i$ always refers to the same direction, so the oracle labels of Section 6.5.1 are consistent across the corpus and the network outputs can be trained without a permutation-invariant criterion. If the azimuths were randomised per mixture, that binding would disappear. The network would have to infer from the mixture alone which output index corresponds to which direction, and the supervision would become ambiguous up to a permutation of the outputs. Two routes are available.

The first is to define the hypothesis set over a fixed angular grid instead of over source indices, so that $\mathcal{H}_i$ means "the dominant interferer lies in sector i" and stays well defined for any source placement. Indexing the outputs by direction rather than by source is established practice in multi-channel separation. Tesch and Gerkmann condition a non-linear filter on an arbitrary steering direction, and use the

same conditioning both to extract one speaker and to separate a variable number of them [38, 19]. Jenrungrot et al. train a network to isolate whatever falls inside an angular window $\theta \pm w/2$ and sweep the window to separate an unknown, possibly moving, set of sources [55]. Taherian et al. dispense with permutation-invariant training altogether by assigning each network output to a fixed azimuth range, which reduces the training complexity from factorial to linear in the number of speakers [56]. Earlier work conditioned extraction on a supplied direction of arrival [57]. The same device applies here, with the angular grid replacing the source index in the definition of $\mathcal{H}_i$.

The second route is to make the posterior loss permutation-invariant [41], at the cost of the direct probabilistic reading of each output. The first route preserves the interpretation on which Section 7.5.3 depends, and is the more natural extension of the present formulation. Neither has been evaluated here, so generalisation to unconstrained source positions remains untested and is the most immediate item of future work.

Size of the training corpus. The training split amounts to about 4.2 h of rendered audio (Section 7.1.5), which is small relative to the tens of hours commonly used for multichannel enhancement. Both neural systems are trained on the same data, so the comparison between them is unaffected, but the absolute scores of both are likely to be conservative. Because the corpus is generated procedurally from an index and a seed, enlarging it requires no new recordings and no additional storage. Retraining the model and re-running the comparison on a larger corpus or even in various datasets would test how much of the remaining gap to the oracle is due to data rather than to model structure, meanwhile enhancing the robustness of the pretrained system.

Anechoic scoring reference. All systems are scored against the anechoic image of the target (Section 7.3). As discussed in Section 7.2, this penalises the beamforming-based estimator, which passes the target reverberation falling inside the analysis window at unit gain and attenuates the rest only partially, whereas the end-to-end mask is free to learn a dereverberating mapping. A reference consisting of the direct path plus the early reflections would be the more informative choice for comparing a spatial filter against an end-to-end mask. Re-scoring against such a reference is a direct extension, since the required renderings are produced by the same simulation.

Sensor noise. Microphone self-noise is not simulated (Section 7.1.4). A physical array adds a spatially white floor, typically at 30–50 dB SNR, which limits how deep a spatial null can usefully be and raises the noise floor at frequencies where the diffuse field is weak. Adding a white sensor-noise floor to the simulation would give a more realistic estimate of how much diagonal loading the deployed system needs, and make the overall experiment more convincing as compared with the realistic acoustic scene.

Further extensions. Three directions follow from the formulation rather than from the experimental setup. The first is the generalised-Gamma target prior of Appendix A: the branch estimator for $\nu \neq 1$ is available in closed form, but the two posteriors would have to be re-derived, since the per-hypothesis likelihood is then no longer Gaussian. The second is causal, low-latency operation: the recursions of Section 6.4 are already causal, but the recurrent layers of Section 6.3 are not, and replacing them by unidirectional or masked layers would allow the system to be evaluated under a streaming constraint. The third is a moving target: the static-source assumption of Section 3.3 enters through the frame-independent relative transfer function, and relaxing it would require the covariance recursions to track a time-varying $\boldsymbol{d}$.

References

[1] Philipos C. Loizou. *Speech Enhancement: Theory and Practice*. 2nd ed. CRC Press, 2013. ISBN: 978-1-4665-0421-9.

[2] Richard C. Hendriks, Timo Gerkmann, and Jesper Jensen. *DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement: A Survey of the State of the Art*. Synthesis Lectures on Speech and Audio Processing. Morgan & Claypool, 2013. DOI: 10.2200/S00473ED1V01Y201301SAP011.

[3] Yariv Ephraim and David Malah. "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* ASSP-32.6 (1984), pp. 1109–1121. DOI: 10.1109/TASSP.1984.1164453.

[4] Yariv Ephraim and David Malah. "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* ASSP-33.2 (1985), pp. 443–445. DOI: 10.1109/TASSP.1985.1164550.

[5] Israel Cohen and Baruch Berdugo. "Noise estimation by minima controlled recursive averaging for robust speech enhancement". In: *IEEE Signal Processing Letters* 9.1 (2002), pp. 12–15. DOI: 10.1109/97.988717.

[6] Israel Cohen. "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging". In: *IEEE Transactions on Speech and Audio Processing* 11.5 (2003), pp. 466–475. DOI: 10.1109/TSA.2003.811544.

[7] Barry D. Van Veen and Kevin M. Buckley. "Beamforming: A versatile approach to spatial filtering". In: *IEEE ASSP Magazine* 5.2 (1988), pp. 4–24. DOI: 10.1109/53.665.

[8] Mehrez Souden, Jacob Benesty, and Sofiène Affes. "On optimal frequency-domain multichannel linear filtering for noise reduction". In: *IEEE Transactions on Audio, Speech, and Language Processing* 18.2 (2010), pp. 260–276. DOI: 10.1109/TASL.2009.2025790.

[9] Kristina Tesch and Timo Gerkmann. "Nonlinear spatial filtering in multichannel speech enhancement". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), pp. 1795–1805. DOI: 10.1109/TASLP.2021.3076372.

[10] Richard C. Hendriks et al. "On optimal multichannel mean-squared error estimators for speech enhancement". In: *IEEE Signal Processing Letters* 16.10 (2009), pp. 885–888. DOI: 10.1109/LSP.2009.2026781.

[11] Timo Gerkmann and Richard C. Hendriks. "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay". In: *IEEE Transactions on Audio, Speech, and Language Processing* 20.4 (2012), pp. 1383–1393. DOI: 10.1109/TASL.2011.2180896.

[12] DeLiang Wang and Jitong Chen. "Supervised speech separation based on deep learning: An overview". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26.10 (2018), pp. 1702–1726. DOI: 10.1109/TASLP.2018.2842159.

[13] Kristina Tesch and Timo Gerkmann. "Insights into deep non-linear filters for improved multichannel speech enhancement". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2023), pp. 563–575. DOI: 10.1109/TASLP.2022.3229555.

[14] Sharon Gannot et al. "A consolidated perspective on multimicrophone speech enhancement and source separation". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.4 (2017), pp. 692–730. DOI: 10.1109/TASLP.2016.2647702.

[15] Jahn Heymann, Lukas Drude, and Reinhold Haeb-Umbach. "Neural network based spectral mask estimation for acoustic beamforming". In: *Proc. IEEE ICASSP*. 2016, pp. 196–200. DOI: 10.1109/ICASSP.2016.7471664.

[16] David Malah, Richard V. Cox, and Anthony J. Accardi. “Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments”. In: *Proc. IEEE ICASSP*. Vol. 2. 1999, pp. 789–792. DOI: 10.1109/ICASSP.1999.759789.

[17] Mehrez Souden et al. “An integrated solution for online multichannel noise tracking and reduction”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.7 (2011), pp. 2159–2169. DOI: 10.1109/TASL.2010.2096216.

[18] Shuai Tao et al. “Learning-based multi-channel speech presence probability estimation using a low-parameter model and integration with MVDR beamforming for multi-channel speech enhancement”. In: *Proc. IWAENC*. 2024, pp. 100–104. DOI: 10.1109/IWAENC61483.2024.10694149.

[19] Kristina Tesch and Timo Gerkmann. “Multi-channel speech separation using spatially selective deep non-linear filters”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 32 (2024), pp. 542–553. DOI: 10.1109/TASLP.2023.3334101.

[20] Jan S. Erkelens et al. “Minimum mean-square error estimation of discrete Fourier coefficients with generalized gamma priors”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 15.6 (2007), pp. 1741–1752. DOI: 10.1109/TASL.2007.899233.

[21] Rainer Martin. “Noise power spectral density estimation based on optimal smoothing and minimum statistics”. In: *IEEE Transactions on Speech and Audio Processing* 9.5 (2001), pp. 504–512. DOI: 10.1109/89.928915.

[22] Otis Lamont Frost III. “An algorithm for linearly constrained adaptive array processing”. In: *Proceedings of the IEEE* 60.8 (1972), pp. 926–935. DOI: 10.1109/PROC.1972.8817.

[23] Simon Doclo and Marc Moonen. “GSVD-based optimal filtering for single and multimicrophone speech enhancement”. In: *IEEE Transactions on Signal Processing* 50.9 (2002), pp. 2230–2244. DOI: 10.1109/TSP.2002.801937.

[24] Hakan Erdogan et al. “Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks”. In: *Proc. IEEE ICASSP*. 2015, pp. 708–712. DOI: 10.1109/ICASSP.2015.7178061.

[25] Yanxin Hu et al. “DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement”. In: *Proc. Interspeech*. 2020, pp. 2472–2476. DOI: 10.21437/Interspeech.2020-2537.

[26] Ashutosh Pandey and DeLiang Wang. “TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain”. In: *Proc. IEEE ICASSP*. 2019, pp. 6875–6879. DOI: 10.1109/ICASSP.2019.8683634.

[27] Yi Luo and Nima Mesgarani. “Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.8 (2019), pp. 1256–1266. DOI: 10.1109/TASLP.2019.2915167.

[28] Yi Luo, Zhuo Chen, and Takuya Yoshioka. “Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation”. In: *Proc. IEEE ICASSP*. 2020, pp. 46–50. DOI: 10.1109/ICASSP40776.2020.9054266.

[29] Cem Subakan et al. “Attention is all you need in speech separation”. In: *Proc. IEEE ICASSP*. 2021, pp. 21–25. DOI: 10.1109/ICASSP39728.2021.9413901.

[30] Shuai Tao et al. “Multi-channel speech enhancement guided by learning-based *a posteriori* speech presence probability estimation”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2025). Early access.

[31] Jonathan Ho, Ajay Jain, and Pieter Abbeel. “Denoising diffusion probabilistic models”. In: *Proc. NeurIPS*. 2020, pp. 6840–6851.

[32] Yang Song et al. “Score-based generative modeling through stochastic differential equations”. In: *Proc. ICLR*. 2021.

[33] Simon Welker, Julius Richter, and Timo Gerkmann. “Speech enhancement with score-based generative models in the complex STFT domain”. In: *Proc. Interspeech*. 2022.

[34] Julius Richter et al. “Speech enhancement and dereverberation with diffusion-based generative models”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2023), pp. 2351–2364. DOI: 10.1109/TASLP.2023.3285241.

[35] Jean-Marie Lemercier et al. “StoRM: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2023), pp. 2724–2737. DOI: 10.1109/TASLP.2023.3294692.

[36] Jean-Marie Lemercier et al. “Diffusion models for audio restoration: A review”. In: *IEEE Signal Processing Magazine* 41.6 (2024), pp. 72–84. DOI: 10.1109/MSP.2024.3445871.

[37] Radu Balan and Justinian Rosca. “Microphone array speech enhancement by Bayesian estimation of spectral amplitude and phase”. In: *Proc. IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM)*. 2002, pp. 209–213. DOI: 10.1109/SAM.2002.1191030.

[38] Kristina Tesch and Timo Gerkmann. “Spatially selective deep non-linear filters for speaker extraction”. In: *Proc. IEEE ICASSP*. 2023.

[39] Robert J. McAulay and Michael L. Malpass. “Speech enhancement using a soft-decision noise suppression filter”. In: *Proc. IEEE ICASSP*. 1980, pp. 137–140. DOI: 10.1109/ICASSP.1980.1170788.

[40] John R. Hershey et al. “Deep clustering: Discriminative embeddings for segmentation and separation”. In: *Proc. IEEE ICASSP*. 2016, pp. 31–35. DOI: 10.1109/ICASSP.2016.7471631.

[41] Morten Kolbæk et al. “Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.10 (2017), pp. 1901–1913. DOI: 10.1109/TASLP.2017.2726762.

[42] Takuya Yoshioka et al. “Multi-microphone neural speech separation for far-field multi-talker speech recognition”. In: *Proc. IEEE ICASSP*. 2018, pp. 5739–5743.

[43] Zhuo Chen et al. “Continuous speech separation: Dataset and analysis”. In: *Proc. IEEE ICASSP*. 2020, pp. 7284–7288. DOI: 10.1109/ICASSP40776.2020.9053426.

[44] Özgür Yilmaz and Scott Rickard. “Blind separation of speech mixtures via time-frequency masking”. In: *IEEE Transactions on Signal Processing* 52.7 (2004), pp. 1830–1847. DOI: 10.1109/TSP.2004.828896.

[45] Shmulik Markovich-Golan and Sharon Gannot. “Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method”. In: *Proc. IEEE ICASSP*. Brisbane, Australia, 2015, pp. 544–548.

[46] Izrail S. Gradshteyn and Iosif M. Ryzhik. *Table of Integrals, Series, and Products*. 7th ed. Academic Press, 2007.

[47] Steven M. Kay. *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*. Upper Saddle River, NJ: Prentice Hall, 1993. ISBN: 978-0133457117.

[48] Reza Varzandeh, Maja Taseska, and Emanuël A. P. Habets. “An iterative multichannel subspace-based covariance subtraction method for relative transfer function estimation”. In: *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*. 2017, pp. 11–15. DOI: 10.1109/HSCMA.2017.7895552.

[49] Peter Kabal. “TSP speech database”. In: *McGill University, Database Version* 1.0 (2002), pp. 09–02.

[50] A Varga, HJM Steeneken, et al. “Noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems”. In: *Speech Commun* 12.3 (1993), pp. 247–253.

[51] Robin Scheibler, Eric Bezzam, and Ivan Dokmanić. “Pyroomacoustics: A python package for audio room simulation and array processing algorithms”. In: *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2018, pp. 351–355.

[52] Emanuël A. P. Habets and Sharon Gannot. “Generating sensor signals in isotropic noise fields”. In: *The Journal of the Acoustical Society of America* 122.6 (2007), pp. 3464–3470. DOI: 10.1121/1.2799929.

[53] Emanuël A. P. Habets, Israel Cohen, and Sharon Gannot. “Generating nonstationary multisensor signals under a spatial coherence constraint”. In: *The Journal of the Acoustical Society of America* 124.5 (2008), pp. 2911–2917. DOI: 10.1121/1.2987429.

[54] Tomohiro Nakatani et al. “Speech dereverberation based on variance-normalized delayed linear prediction”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 18.7 (2010), pp. 1717–1731.

[55] Teerapat Jenrungrot et al. “The cone of silence: Speech separation by localization”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 33. 2020, pp. 20925–20938.

[56] Hassan Taherian, Ke Tan, and DeLiang Wang. “Multi-channel talker-independent speaker separation through location-based training”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022), pp. 2791–2800. DOI: 10.1109/TASLP.2022.3202129.

[57] Rongzhi Gu et al. “Neural spatial filter: Target speaker speech separation assisted with directional information”. In: *Proc. Interspeech*. 2019, pp. 4290–4294.

A

Generalised-Gamma Branch Estimator

Section 4.4.1 adopted the generalised-Gamma amplitude prior (4.6) for consistency with the analytic multichannel MMSE literature [20, 10, 9], and then specialised to the Gaussian case $\nu = 1$, which is the prior in force throughout Chapters 4–7. This appendix records the conditional branch estimator for a general ν, and states which parts of Chapter 5 extend to that case and which do not.

A.1. The Branch Estimator for $\nu \neq 1$

Under the composite hypothesis $(\mathcal{M}_1, \mathcal{H}_i)$ with the signal model (5.3), replacing the Gaussian target prior by a generalised-Gamma prior of shape parameter ν gives the conditional MMSE estimator

$$\mathbb{E}[\, S \mid \boldsymbol{y},\, \mathcal{M}_1,\, \mathcal{H}_i \,] \;=\; \underbrace{\frac{\nu\, \sigma_S^2}{\nu\, (\boldsymbol{d}^H \boldsymbol{\Sigma}_i^{-1} \boldsymbol{d})^{-1} + \sigma_S^2} \cdot \frac{M(\nu{+}1,\, 2,\, \mathcal{P}_i[\boldsymbol{y}])}{M(\nu,\, 1,\, \mathcal{P}_i[\boldsymbol{y}])}}_{\text{generalised spectral gain}} \cdot\, z_i(\boldsymbol{y}), \tag{A.1}$$

where $z_i(\boldsymbol{y})$ is the per-hypothesis MVDR output (5.31), $M(\cdot,\cdot,\cdot)$ is the confluent hypergeometric function (Kummer's function) [46], and

$$\mathcal{P}_i[\boldsymbol{y}] \;\triangleq\; \frac{|\boldsymbol{d}^H \boldsymbol{\Sigma}_i^{-1} \boldsymbol{y}|^2\; \sigma_S^2}{\nu + \xi_{S,i}}. \tag{A.2}$$

The result is due to Hendriks et al. [10], who derived the multichannel MMSE estimator under a generalised-Gamma magnitude prior and showed that it separates into an MVDR beamformer followed by a single-channel post-filter if and only if the noise is Gaussian. Erkelens et al. [20] give the single-channel estimator from which the spectral gain in (A.1) inherits its hypergeometric form.

Only the scalar gain applied to the MVDR output changes with ν. The spatial part of the estimator does not: the observation still enters through $z_i(\boldsymbol{y})$ alone, so all branches remain steered at the same target RTF and differ only through $\boldsymbol{\Sigma}_i$.

Equation (A.1) is the same quantity that appears, in the notation of Tesch and Gerkmann, as the generalised spectral gain in (5.53); the correspondence $\boldsymbol{\Phi}_m \leftrightarrow \boldsymbol{\Sigma}_i$ is the notation map of (5.51). Section 5.6.2 carries out the reduction at $\nu = 1$, where $M(1,1,x) = M(2,2,x) = e^x$, the hypergeometric ratio equals unity, and the gain collapses to the Wiener form $\xi_{S,i}/(1+\xi_{S,i})$ of (5.29). That reduction is not repeated here.

For $\nu < 1$ the hypergeometric ratio depends on the observation through $\mathcal{P}_i[\boldsymbol{y}]$, so the branch itself becomes a nonlinear function of $\boldsymbol{y}$. This is a second source of nonlinearity, distinct from the posterior weighting of (5.20), which is present already at $\nu = 1$ (property (v) of Section 5.7). Tesch and Gerkmann [9] evaluate the estimator at $\nu = 0.25$ and report that the super-Gaussian gain accounts for part of the advantage of the nonlinear filter over a linear one.

A.2. Scope: What Does Not Extend

Equation (A.1) replaces the branch estimator only. The two posteriors of Section 5.5 do *not* extend unchanged to $\nu \neq 1$, and the reason is worth stating explicitly, because the two results are otherwise easy to combine.

The target posterior (5.49) is built from the per-hypothesis odds (5.41), which are the reciprocal of the likelihood ratio (5.14). That ratio is available in closed form because the target-present likelihood (5.6) is Gaussian, which holds only under the Gaussian prior. For $\nu \neq 1$ the observation $\boldsymbol{y} = \boldsymbol{d}\,S + \boldsymbol{v}_i$ is the sum of a generalised-Gamma component and a Gaussian one and is therefore not Gaussian. The same applies to the interferer posterior (5.39), whose derivation rests on the same Gaussian likelihood being rewritten about the common base (5.34).

The form the weights would take at general ν is visible in (5.52): the weight of component m carries the two additional factors $Q_m = (\nu + \boldsymbol{d}^H \boldsymbol{\Phi}_m^{-1} \boldsymbol{d}\,\sigma_S^2)^{-\nu}$ and $M(\nu, 1, P_m)$, both of which reduce to the Gaussian case only at $\nu = 1$. Combining (A.1) at $\nu \neq 1$ with the posteriors of Section 5.5 as derived would therefore be inconsistent. Both posteriors would have to be re-derived under the non-Gaussian likelihood before the estimator could be evaluated at $\nu \neq 1$. This extension is not pursued in this thesis.

B

Kullback--Leibler Posterior-Supervision Loss

Section 6.5.2 supervises the interferer network with the frequency-weighted mean squared error (6.15). At each time–frequency bin the estimated and oracle interferer posteriors are categorical distributions over the I mutually exclusive hypotheses, so a Kullback–Leibler (KL) divergence is an equally natural criterion. That variant was implemented and is recorded here.

Writing $\hat{p}(\mathcal{H}_i \mid \boldsymbol{y}, \mathcal{M}_1)$ for the estimate and $p^{\mathrm{o}}(\mathcal{H}_i \mid \boldsymbol{y}, \mathcal{M}_1)$ for the oracle of (5.39), and ω_k for the frequency weight at bin k, the KL form of the posterior-supervision term is

$$\mathcal{L}_{\mathcal{H}}^{\mathrm{KL}} \;=\; \frac{1}{K\,L} \sum_{k=1}^{K} \sum_{l=1}^{L} \omega_k \sum_{i=0}^{I-1} p^{\mathrm{o}}(\mathcal{H}_i \mid \boldsymbol{y}, \mathcal{M}_1)\, \log \frac{p^{\mathrm{o}}(\mathcal{H}_i \mid \boldsymbol{y}, \mathcal{M}_1)}{\hat{p}(\mathcal{H}_i \mid \boldsymbol{y}, \mathcal{M}_1)}. \tag{B.1}$$

It is minimised when the two distributions coincide, and it is the conventional objective for fitting a probability distribution. Either (6.15) or (B.1) can serve as $\mathcal{L}_{\mathcal{H}}$ in the total objective (6.13).

The two criteria differ in how they weight the small-probability region. The KL divergence is unbounded as $\hat{p}(\mathcal{H}_i \mid \boldsymbol{y}, \mathcal{M}_1) \to 0$ at a bin where the oracle assigns non-negligible mass, so it penalises confident errors much more heavily than the squared error does, and it requires the estimate to be floored away from zero for numerical stability. The squared error weights all bins equally and is bounded, which makes the scale of $\mathcal{L}_{\mathcal{H}}$ comparable to that of the enhancement loss (6.14) and hence makes the fixed weight $\lambda_{\mathcal{H}}$ of (6.13) easier to set. The MSE form is used for all results reported in Chapter 7; the reason is stated with the remaining experimental controls in Section 7.4.