Multi-Microphone Speech Enhancement

A Signal Subspace Approach

Richard Heusdens

May 25, 2020

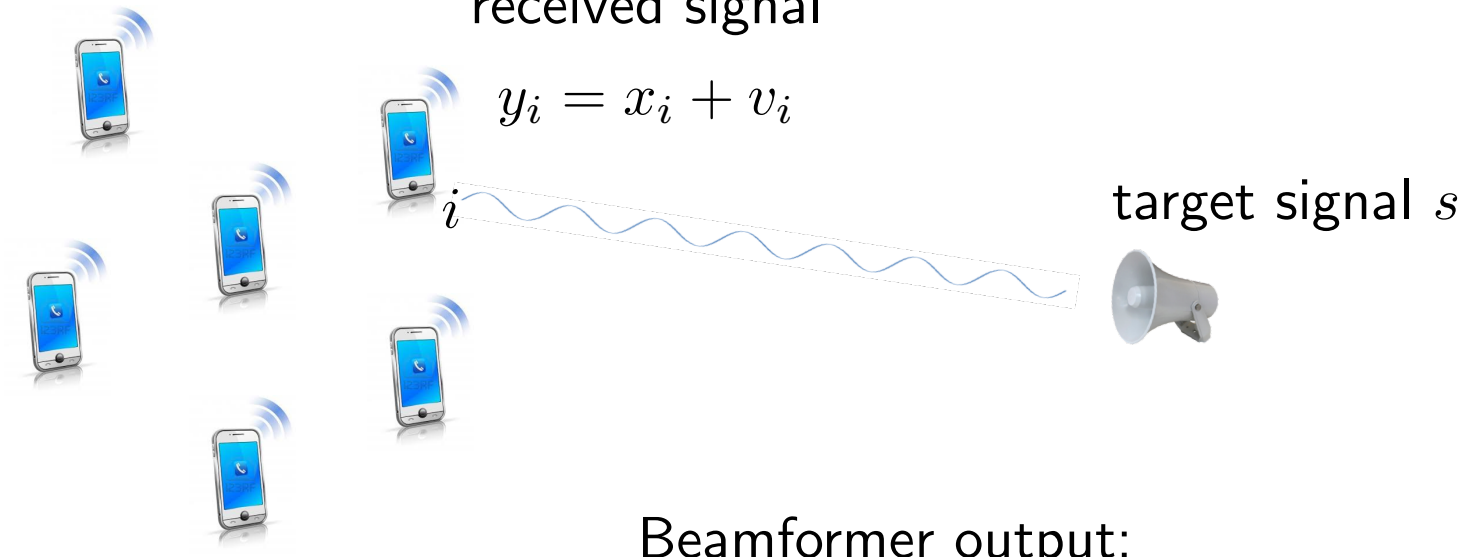
TU Delft

Delft University of Technology

Content

- Introduction
- Room impulse response
- Eigenvalue decomposition
  - Estimation of R_X for (spatially) uncorrelated noise process
- Generalised eigenvalue decomposition
  - Estimation of R_X for arbitrary noise processes
- Beamforming

Introduction



Beamformer output:

$$w^H y = w^H x + w^H v$$

where $y = (y_1, \ldots, y_M)^T$

Room impulse response

Given a source of sound and a receiving position in a room a mathematical description of all possible sound paths from the source to the receiver, which includes the reflections due to the walls, floor, ceiling and other obstacles, is given by what is called the *room impulse response (RIR)*.

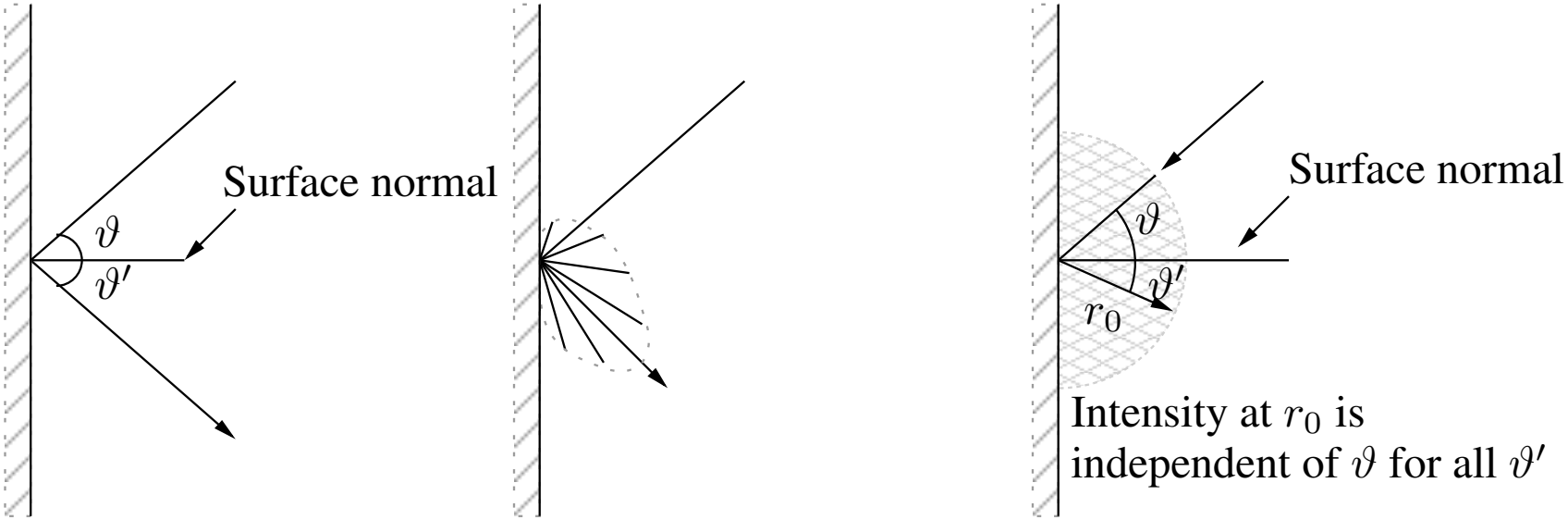
If the source is modeled as a point in space and emits an impulse, i.e. a mathematical idealisation of an explosive, very short in time and loud sound, then what is measured at the receiver (e.g. a microphone) is the RIR.

Room impulse response

The acoustical phenomena (of interest in room acoustics) that is implicitly carried by the RIR can be classified into the following categories:

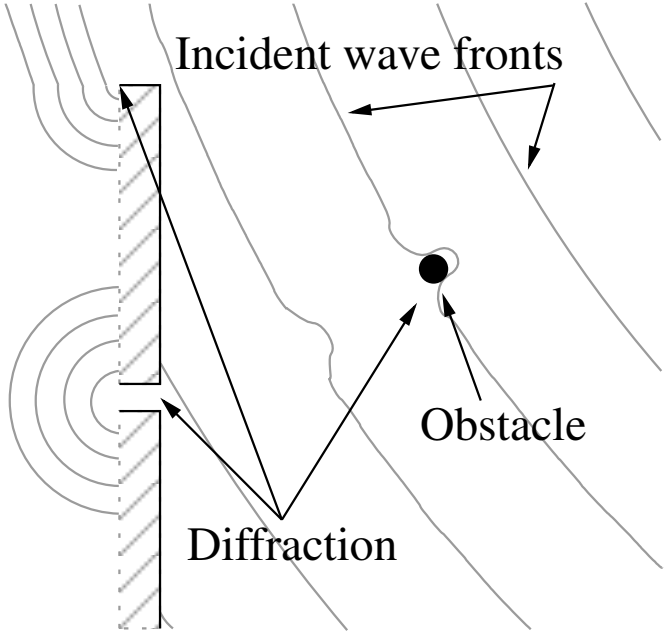
- Specular reflections. Here the sound reflects in a specular way, following a simple geometric law.
- Diffuse reflections. The sound is reflected in a scattered, not necessarily homogeneous, way.
- Absorption and refraction. When a sound wave encounters a boundary part, energy will be absorbed, and/or the angle of incidence at the boundaries changes (refraction, Snell's law).
- Diffraction. It is caused by scattering effects that occur in some situations when sound waves encounter an obstacle.

Room impulse response

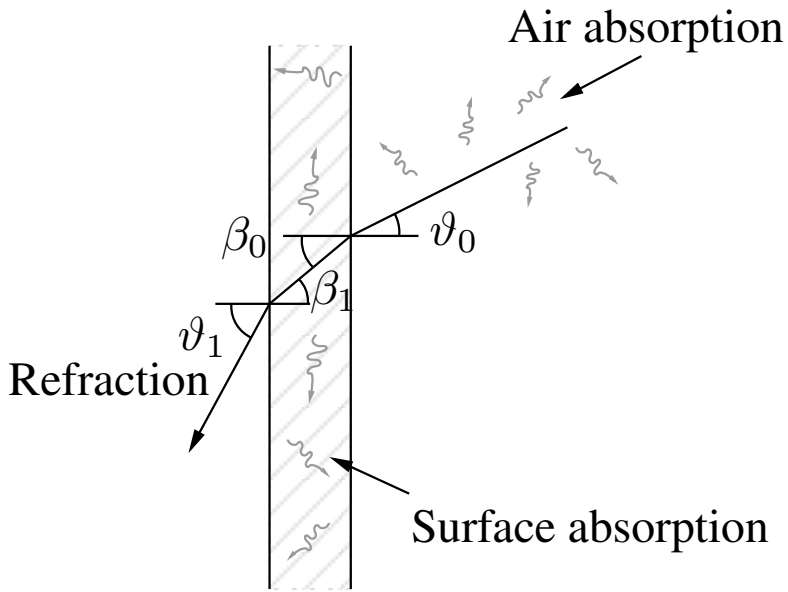


(a) *Specular reflection* (b) *Semi-diffuse reflection* (c) *Totally diffuse reflection*

Room impulse response



(a) *Diffraction phenomena*

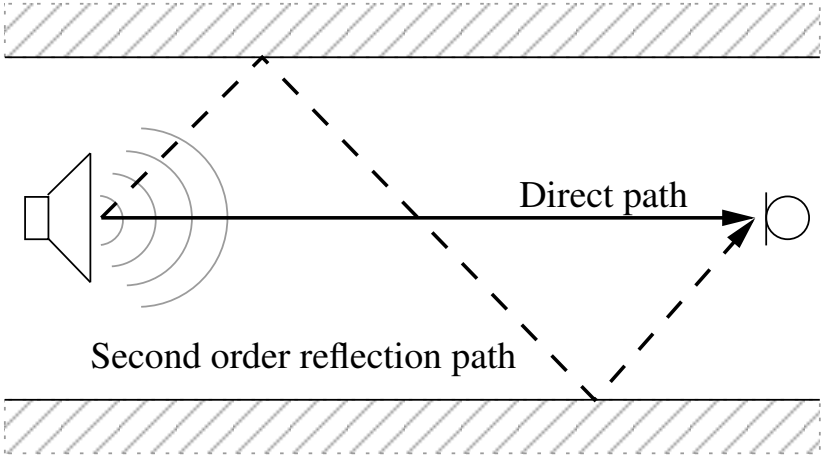


(b) *Transmitted wave. Absorption and refraction*

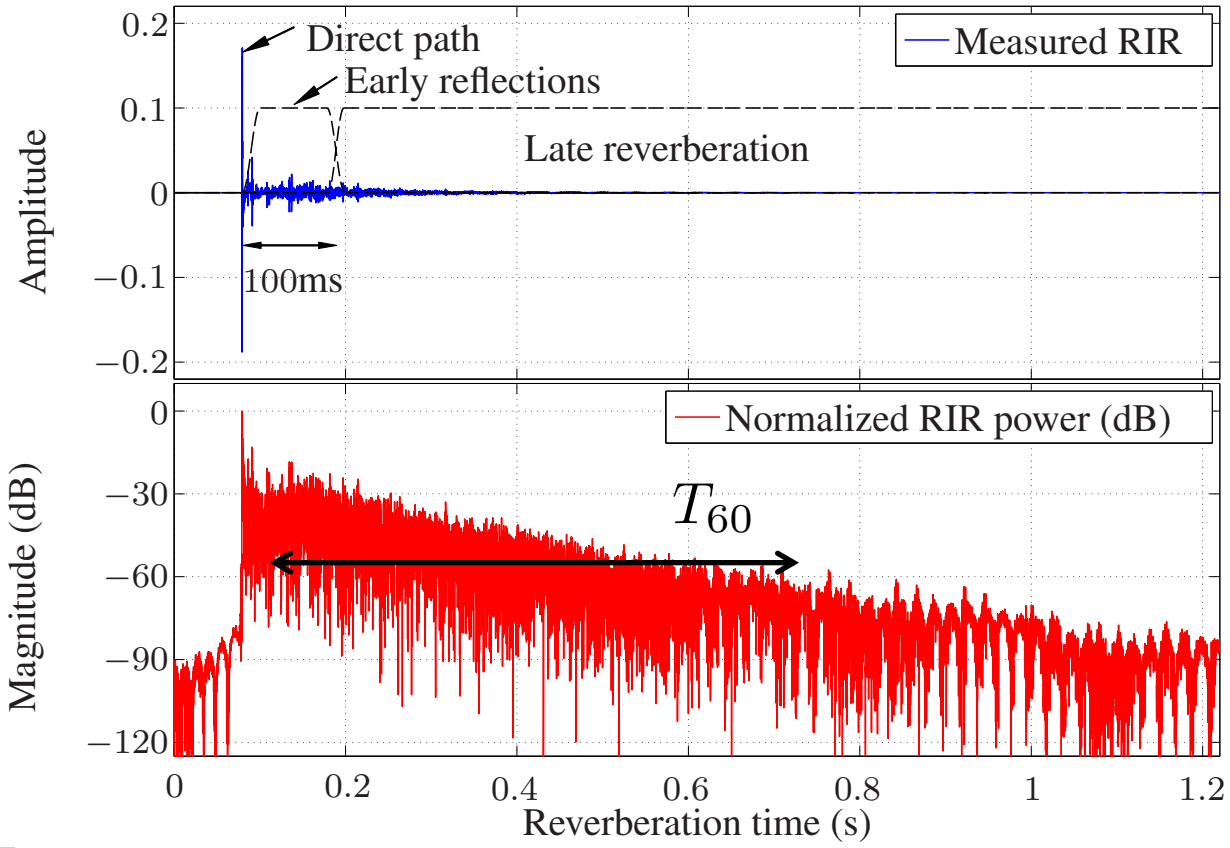
Room impulse response

The sound is modified by the different acoustic phenomena present along a given path until it reaches the receiver.

The RIR is thus a signal made up of all possible (modified) copies of the impulse that arrive at the receiver after traveling their corresponding paths



Room impulse response



Room impulse response

Early reflections arriving in a time window between 20 ms and 100 ms after the direct path can contribute to the intelligibility and definition of speech and the clarity of music.

Reflections arriving later than 100 ms after the direct path create a diffuse reverberant effect, the so-called *late reverberation zone*, which contributes, for example in concert halls, to the warmth and brilliance of music. However, in non-acoustically specialised enclosures, such as a swimming bath, this late reverberation field can be very detrimental for speech intelligibility.

Acoustic transfer function

The acoustic events in a room, under some assumptions, can be mathematically idealised as to be linear and time-invariant (LTI) so that the sound as it would be measured at the receiver can be calculated directly by convolution of the RIR and the source signal.

Assume that the target signal, say s, is a point source. Let h_m denote the RIR from the source s to microphone m. In that case, the signal x_m (the noise-free source signal received at the mth microphone) is given by

$$x_m(n) = (h_m * s)(n), \quad n \in \mathbb{Z}.$$

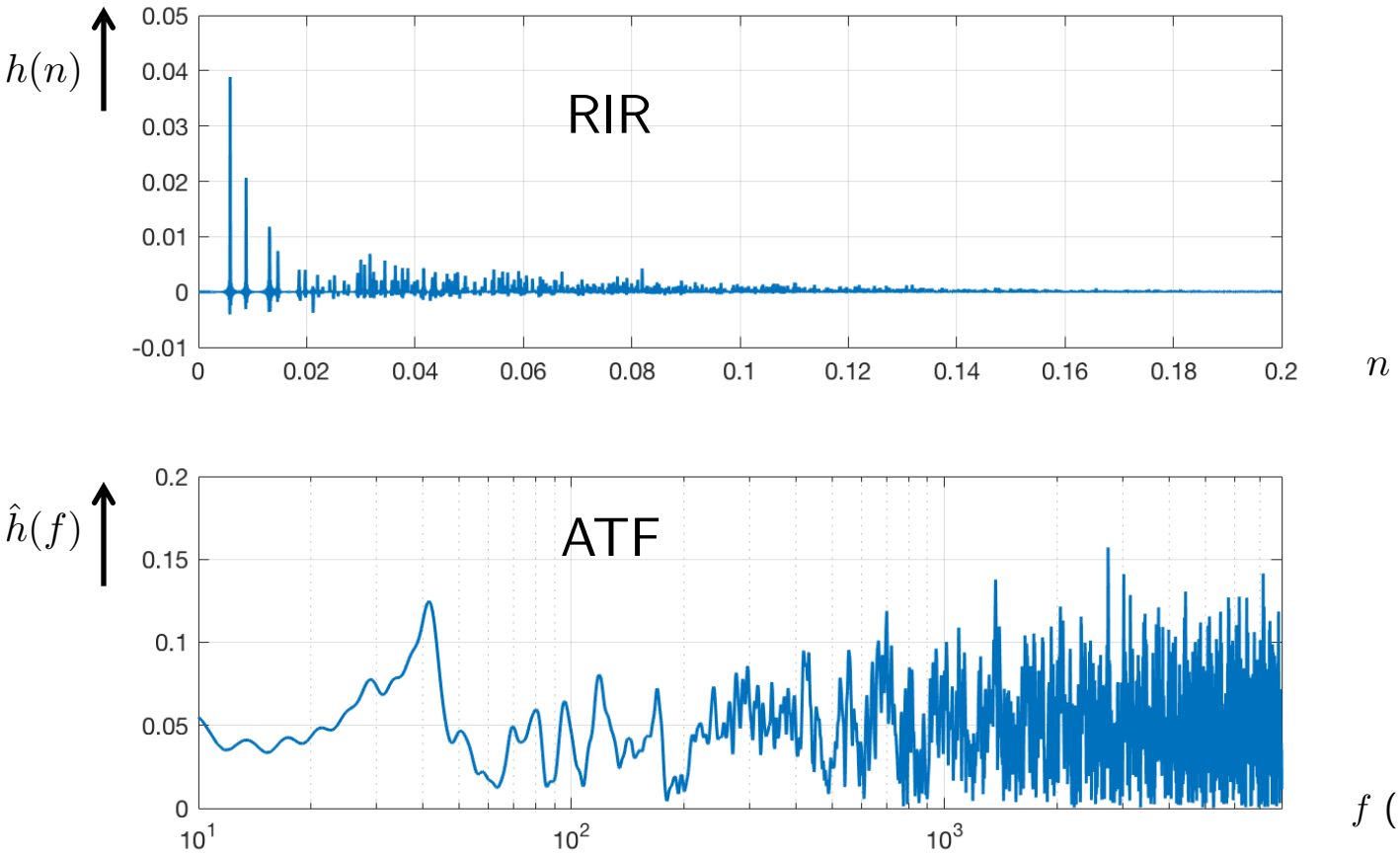
Acoustic transfer function

We can equivalently compute the effect of the RIR in the frequency domain, where we have that

$$\hat{x}_m(\omega) = \left(\hat{h}_m \cdot \hat{s}\right)(\omega), \quad \omega \in [0, 2\pi),$$

and $\hat{\cdot}$ denotes (temporal) Fourier transformation. The function $\hat{h}_m(\omega)$ is called the *acoustic transfer function (ATF)* from the source to the mth microphone and is the temporal-frequency domain representation of the RIR.

Acoustic transfer function



Acoustic transfer function

Given a fixed frequency ω, we can collect all M microphone signals $\hat{x}_m(\omega)$ in a vector $\hat{x}(\omega) = (\hat{x}_1(\omega), \ldots, \hat{x}_M(\omega))^T \in \mathbb{C}^M$, so that

$$\hat{x}(\omega) = d(\omega)\hat{s}(\omega),$$

where $d(\omega) = (\hat{h}_1(\omega), \ldots, \hat{h}_M(\omega))^T \in \mathbb{C}^M$. The vector d is called the *steering vector*. Note that d is frequency dependent!

(Relative) acoustic transfer function

In many applications, we are only interested in the *relative acoustic transfer function*, which is the normalised ATF with respect to a reference microphone, which we will assume to be microphone 1.

The relative acoustic transfer function is given by

$$d'(\omega) = (1, \hat{h}_2(\omega)/\hat{h}_1(\omega), \ldots, \hat{h}_M(\omega)/\hat{h}_1(\omega))^T$$

In the case of a *far field* scenario, we have $|\hat{h}_m(\omega)| = |\hat{h}_1(\omega)|$ for all m, so that

$$d' = (1, e^{j\tau_2}, \ldots, e^{j\tau_M}),$$

where $\tau_m(\omega) = \angle\hat{h}_m(\omega) - \angle\hat{h}_1(\omega)$, the phase difference between microphone m and the reference microphone.

Beamforming

We will consider the received signals y, x, v and s as realisations of zero mean, wide-sense stationary processes Y, X, V, and S, respectively. We have

$$Y = X + V.$$

In addition, we will assume that X and V are mutually uncorrelated. That is, we have

$$\mathrm{E}(XV^H) = O,$$

so that

$$\mathrm{E}(YY^H) = \mathrm{E}(XX^H) + \mathrm{E}(VV^H),$$

or, equivalently,

$$R_Y = R_X + R_V.$$

Beamforming

The matrices $R_Y, R_X,$ and R_V are the auto-correlation matrices of the processes $Y, X,$ and V, respectively.

In many cases, we will apply a time-to-frequency (Fourier) transform so that the processes $Y, X, V,$ and S contain spectral data at a particular frequency ω.

In that case the matrices $R_Y, R_X,$ and R_V denote *cross-power spectral density (CPSD)* matrices.

In the remaining we will assume that all data is frequency-domain data

Beamforming

Assume that the target signal s is a point source, so that

$$x = dx_1,$$

where $d \in \mathbb{C}^M$ is the steering vector containing the (relative) acoustic transfer functions from the source to the microphones. With this, R_X can be expressed as

$$R_X = \mathrm{E}(XX^H) = \sigma_{X_1}^2 dd^H,$$

where $\sigma_{X_1}^2 = \mathrm{E}|X_1|^2$, the variance of the clean signal as received at the reference microphone 1.

Note that in this case $\mathrm{rank}(R_X) = 1$.

Beamforming

Beamformers seen so far:

- Delay-and-sum: $w_{DS} = \alpha d, \quad \alpha = 1/d^H d$
- MVDR: $w_{MVDR} = \frac{R_V^{-1} d}{d^H R_V^{-1} d}$
- Multi-channel Wiener filter: $w_{MWF} = \sigma_{X_1}^2 R_Y^{-1} d$

Note that these results only hold for a point (target) source where $R_X = \sigma_{X_1}^2 d d^H$.

Beamforming

The beamformers derived so far assume that the (relative) acoustic transfer function d is known a-priori

- In practice, d is unknown and needs to be estimated
- Estimation errors in d generally lead to severe performance degradation of the beamformer
- When there are multiple sources, the beamformers will be a function of a general correlation matrix R_X

In the following, we will focus on estimating R_X and give expressions for beamformers in terms of a general R_X, not necessarily of rank 1

Eigenvalues and eigenvectors

Let $A \in \mathbb{C}^{n \times n}$. A (non-zero) vector $x \in \mathbb{C}^n$ is called and *eigenvector* of A if it satisfies the linear equation

$$Ax = \lambda x$$

where $\lambda \in \mathbb{C}$ is called the *eigenvalue* corresponding to x.

The eigenvalues are found by finding the roots of the *characteristic polynomial*:

$$p(\lambda) = \det(A - \lambda I) = 0$$

Eigenvalues and eigenvectors

We can factor $p(\lambda)$ as

$$p(\lambda) = (\lambda - \lambda_1)^{n_1}(\lambda - \lambda_2)^{n_2} \cdots (\lambda - \lambda_k)^{n_k}$$

where n_i is the *algebraic multiplicity* of λ_i and $\sum_{i=1}^{k} n_i = n$.

For each λ_i, we have a specific eigenvalue equation

$$(A - \lambda_i I)x = 0$$

Hence, we need to find $x \in \ker(A - \lambda_i I)$

Eigenvalues and eigenvectors

There will be $1 \leq m_i \leq n_i$ linearly independent solutions, where $m_i = \dim \ker(A - \lambda_i I)$. The integer m_i is called the *geometric multiplicity* of λ_i.

A matrix A is diagonalisable if and only if, for every eigenvalue λ_i, its geometric and algebraic multiplicities coincide. That is, if and only if $m_i = n_i$ for all $i = 1, \ldots, k$

Not all matrices are diagonalisable. For example

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \qquad \text{but } A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \text{ is}$$

Eigenvalues and eigenvectors

If there are n linearly independent eigenvectors, we can factorise A as

$$A = X\Lambda X^{-1},$$

where

$$\Lambda = \begin{pmatrix} \lambda_1 I_{n_1} & & \\ & \ddots & \\ & & \lambda_k I_{n_k} \end{pmatrix},$$

which is called the *eigenvalue decomposition (EVD)*

Without proof we state that if A is Hermitian ($A^H = A$), then it is diagonalisable.

Properties of eigenvalues

- The eigenvalues are unique:

  Since the eigenvalues are the roots of the characteristic polynomial (at most n), they are unique.

Suppose $A \in \mathbb{C}^{n \times n}$ has n linearly independent eigenvectors. Then

- The eigenspaces are unique, *not* the eigenvectors itself:

  For each eigenvalue λ_k (multiplicity n_k), there exists an n_k-dimensional subspace $\mathbb{E}_k$ of vectors such that

  $$\forall x \in \mathbb{E}_k : \; Ax = \lambda_k x$$

  This space $(\ker(A - \lambda_k I))$ is unique, but not the eigenvectors since we are free to choose any basis for $\mathbb{E}_k$.

Properties of eigenvalues

- A is *invertible* if and only if $\forall i : \lambda_i \neq 0$

$$A^{-1} = X\Lambda^{-1}X^{-1}$$

- If $B = TAT^{-1}$ *(similarity transform)*, then A and B share the same eigenvalues:

$$B = TAT^{-1} = T\left(X\Lambda X^{-1}\right)T^{-1} = (TX)\Lambda(TX)^{-1}$$

and since the eigenvalues are unique, Λ is the eigenvalue matrix of B.

Properties of eigenvalues

Assume A is *Hermitian*.

- Every eigenvalue is real:

$$\lambda\|x\|^2 = \lambda x^H x = x^H A x = x^H A^H x = \overline{\lambda} x^H x = \overline{\lambda}\|x\|^2$$

- The eigenvectors are mutually orthogonal ($X^{-1} = X^H$):

  Let x and y be eigenvectors of A corresponding to distinct eigenvalues λ and μ, respectively. Then

$$\lambda y^H x = y^H A x = y^H A^H x = \overline{\mu} y^H x = \mu y^H x$$

  Hence $(\lambda - \mu) y^H x = 0$ and thus $x \perp y$.

Properties of eigenvalues

If, in addition, A is *positive semi-definite* ($x^H Ax \geq 0$ for all $x \in \mathbb{C}^n$), then

- Every eigenvalue is nonnegative:

$$\lambda \|x\|^2 = \lambda x^H x = x^H Ax \geq 0 \;\Rightarrow\; \lambda \geq 0$$

If A is *unitary* ($AA^H = A^H A = I$), then

- Every eigenvalue has magnitude 1:

$$\|x\|^2 = x^H x = x^H A^H Ax = \|Ax\|^2 = |\lambda|^2 \|x\|^2 \;\Rightarrow\; |\lambda| = 1$$

Eigenvalue decomposition

Since $R_X = \mathrm{E}(XX^H)$ is Hermitian and positive semi-definite, there exists a unitary matrix $U = (u_1, \ldots, u_M)$, $u_i \in \mathbb{C}^M$, such that the *eigenvalue decomposition (EVD)* of R_X is given by

$$R_X = U\Lambda U^{-1} = U\Lambda U^H,$$

where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_M)$, $\lambda_i \geq 0$ for all i.

Suppose that V is a *spatially white* (uncorrelated) process with covariance matrix

$$R_V = \sigma_V^2 I_M,$$

where I_M denotes the indentity operator (matrix) in $\mathbb{C}^M$.

Eigenvalue decomposition

The covariance matrix of the noisy process $Y = X + V$ is then given by

$$\begin{aligned} R_Y &= R_X + \sigma_V^2 I_M \\ &= U\Lambda U^H + \sigma_V^2 I_M \\ &= U\Lambda U^H + \sigma_V^2 U U^H \\ &= U(\Lambda + \sigma_V^2 I_M)U^H \end{aligned}$$

which is the eigenvalue decomposition of R_Y.

Eigenvalue decomposition

Conclusions:

- R_X (which is not available) and R_Y (which we can estimate from the observed data) share the same eigenvectors
- Adding (spatially uncorrelated) noise to the desired speech data *only* affects the eigenvalues of R_X

Eigenvalue decomposition

Let us assume that $\mathrm{rank}(R_X) = r < M$. We can partition R_Y as

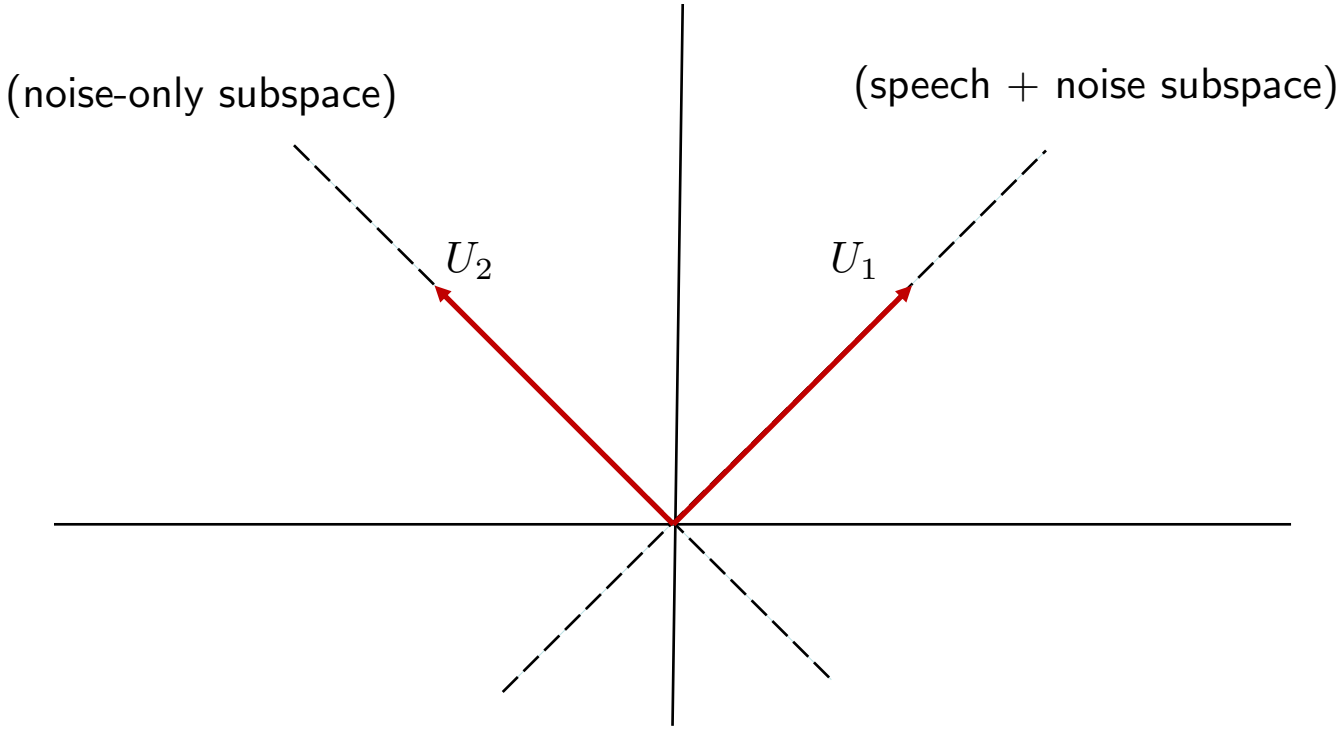
$$R_Y = (U_1\ U_2) \begin{pmatrix} \Lambda_1 + \sigma_V^2 I_r & O \\ O & \sigma_V^2 I_{M-r} \end{pmatrix} \begin{pmatrix} U_1^H \\ U_2^H \end{pmatrix},$$

where $U_1 \in \mathbb{C}^{M\times r}, U_2 \in \mathbb{C}^{M\times(M-r)}$ and $\Lambda_1 \in \mathbb{C}^{r\times r}$.

Since $R_Y = U_1(\Lambda_1 + \sigma_V^2 I_r)U_1^H + \sigma_V^2 U_2 U_2^H$, we conclude that the eigenvectors $u_1, \ldots, u_r$ span the *speech (+ noise) subspace*, whereas $u_{r+1}, \ldots, u_M$ span the *noise only* subspace.

Since U is unitary, we have $U_1^H U_2 = O$ (orthogonal subspaces).

Geometric interpretation



Estimation of R_X

Despite the fact that we do not know what the signal subspace is a priori (R_X is unknown), we can compute (estimate) R_X from the EVD of R_Y.

Least-squares estimate is obtained by approximating R_X by

$$\hat{R}_X = \arg \min_{\text{rank}(P)=r} \|R_Y - P\|_F^2$$

The solution is a classical result and follows by truncating the $M - r$ smallest eigenvalues. That is,

$$\hat{R}_X = U_1(\Lambda_1 + \sigma_V^2 I_r)U_1^H$$

Estimation of R_X

Since the last $M - r$ eigenvalues are given by σ_V^2, we can even do better by subtracting σ_V^2 from the largest r eigenvalues (results in a minimum variance estimator). That is,

$$\hat{R}_X = U_1 \Lambda_1 U_1^H$$

Note that in practice we have to estimate U and Λ (and thus U_1 and Λ_1) from the noisy observations and for that reason the resulting estimator is *not* identical to R_X although the above equation suggests so.

Pre-whitening

If the noise process V is *not* white ($R_V \neq \alpha I_M$ for some $\alpha > 0$), we can pre-whiten the data, assuming that $R_V \succ 0$ (positive definite)

Since R_V is Hermitian and positive definite, we have

$$R_V = U\Lambda U^H = U\Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}}U^H = (U\Lambda^{\frac{1}{2}}U^H)(U\Lambda^{\frac{1}{2}}U^H) = R_V^{\frac{1}{2}}R_V^{\frac{1}{2}}$$

where $R_V^{\frac{1}{2}}$ is the (unique) Hermitian *square root* of R_V.

Consider the transformed process $\tilde{V} = R_V^{-\frac{1}{2}}V$. The process $\tilde{V}$ is spatially white:

$$R_{\tilde{V}} = \mathrm{E}(\tilde{V}\tilde{V}^H) = R_V^{-\frac{1}{2}}\mathrm{E}(VV^H)R_V^{-\frac{1}{2}} = I_M$$

Pre-whitening

Next consider the transformed process $\tilde{Y} = R_V^{-\frac{1}{2}} Y$. Since this transformation transforms the original noise process into a spatially uncorrelated one, we have

$$R_{\tilde{Y}} = \mathrm{E}(\tilde{Y}\tilde{Y}^H) = R_V^{-\frac{1}{2}} E(YY^H) R_V^{-\frac{1}{2}} = R_V^{-\frac{1}{2}} R_X R_V^{-\frac{1}{2}} + I_M.$$

Hence, we can apply the same techniques as discussed previously to the transformed process $\tilde{Y}$ and de-whiten the result thus obtained.

Pre-whitening

Estimation of R_X:

1. Compute $R_V^{\frac{1}{2}}$ and pre-whiten the data: $\tilde{Y} = R_V^{-\frac{1}{2}} Y$
2. Compute the EVD $R_{\tilde{Y}} = \tilde{U}\left(\tilde{\Lambda} + I_M\right)\tilde{U}^H$, truncate the $M - r$ smallest eigenvalues and reduce the remaining ones by one
3. Estimate $\hat{R}_{\tilde{X}} = \tilde{U}_1 \tilde{\Lambda}_1 \tilde{U}_1^H$
4. De-whiten the result thus obtained so that

$$\hat{R}_X = R_V^{\frac{1}{2}} \tilde{U}_1 \tilde{\Lambda}_1 \tilde{U}_1^H R_V^{\frac{1}{2}}$$

Generalised eigenvalue decomposition

Remarks:

- The explicit use of $R_V^{\frac{1}{2}}$ may result in a loss of accuracy in the data
- Can be avoided by working directly with R_Y and R_V
- In addition, when R_V and/or R_X are updated in a recursive way, it is generally very complicated to update $R_{\tilde{Y}}$

Solution is given by the *generalised eigenvalue decomposition*

Generalised eigenvalue decomposition

Given the Hermitian matrices $A, B \in \mathbb{C}^{n\times n}$ with $B \succ 0$, there exists a non-singular $U = (u_1, \ldots, u_n), u_i \in \mathbb{C}^n$, such that

$$U^H A U = \mathrm{diag}(a_1, \ldots, a_n) \quad \text{and} \quad U^H B U = \mathrm{diag}(b_1, \ldots, b_n).$$

Hence, we have $BU = U^{-H}\Lambda_B$ so that

$$AU = U^{-H}\Lambda_A = U^{-H}\Lambda_B\Lambda_B^{-1}\Lambda_A = BU\Lambda$$

That is, $Au_i = \lambda_i B u_i$ for $i = 1, \ldots, n$ where $\lambda_i = a_i/b_i$.

This decomposition is known as the *generalised eigenvalue decomposition (GEVD)*.

Generalised eigenvalue decomposition

Note that since $B \succ 0$ (B is invertible), we have

$$B^{-1}Au_i = \lambda_i u_i$$

Hence, the generalised eigenvalues and eigenvectors of (A, B) are the (ordinary) eigenvalues and eigenvectors of the matrix $B^{-1}A$.

Application to R_X and R_V, and setting $b_i = 1$ for all i, we have

$$U^H R_X U = \Lambda \quad \text{and} \quad U^H R_V U = I_M,$$

where $\Lambda \succeq 0$. Hence, the pair (Λ, U) are the eigenvalues/vectors of the matrix $R_V^{-1}R_X$.

Generalised eigenvalue decomposition

Again, since $R_Y = R_X + R_V$, we have

$$U^U R_Y U = \Lambda + I_M \quad \Leftrightarrow \quad R_Y = U^{-H}(\Lambda + I_M)U^{-1}$$

Note that the matrix $R_V^{-1}R_X$ is *not* Hermitian, and as such $U^{-1} \neq U^H$.

However, since $R_V^{-1}R_X = R_V^{-1/2}SR_V^{1/2}$ with $S = R_V^{-1/2}R_X R_V^{-1/2}$ is Hermitian and positive semi-definite, we conclude that $R_V^{-1}R_X$ is similar to a Hermitian positive semi-definite matrix and, therefore, has (real) nonnegative eigenvalues.

Generalised eigenvalue decomposition

Let $Q = U^{-H} = (q_1, \ldots, q_M)$, $q_i \in \mathbb{C}^M$. With this we can express R_Y as

$$R_Y = Q(\Lambda + I_M)Q^H$$

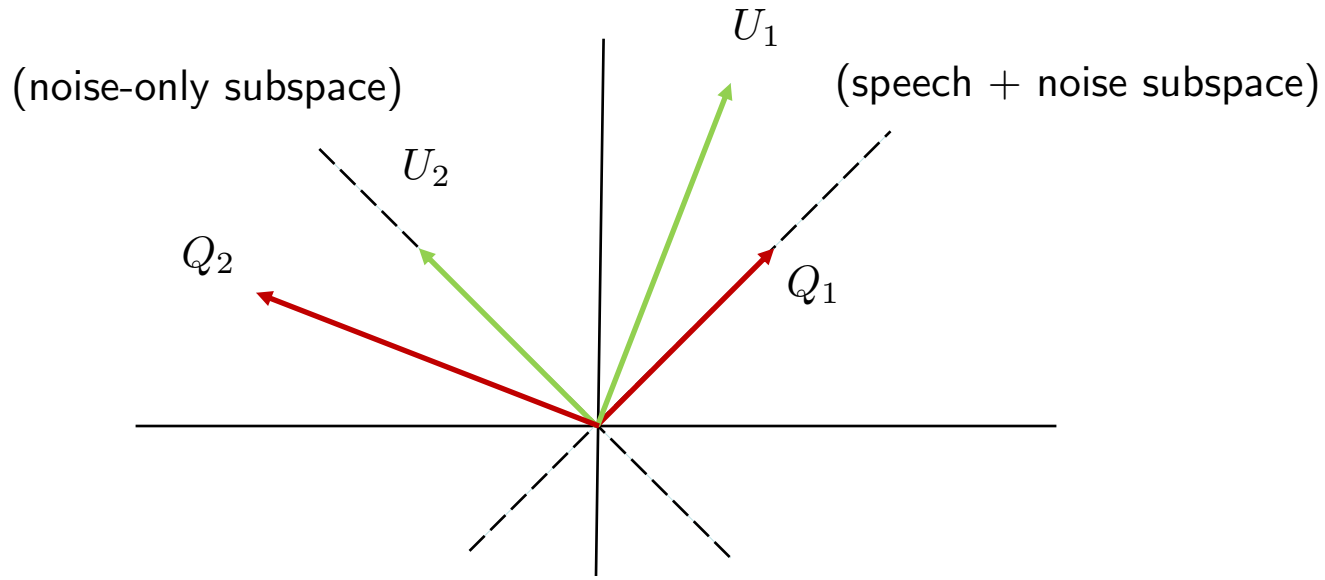
Again, if we assume that $\mathrm{rank}(R_X) = r < M$, we can partition R_Y as

$$R_Y = (Q_1\ Q_2) \begin{pmatrix} \Lambda_1 + I_r & O \\ O & I_{M-r} \end{pmatrix} \begin{pmatrix} Q_1^H \\ Q_2^H \end{pmatrix},$$

where $Q_1 \in \mathbb{C}^{M \times r}$ and $Q_2 \in \mathbb{C}^{M \times (M-r)}$.

Geometric interpretation

Since $R_Y = Q_1(\Lambda_1 + I_r)Q_1^H + Q_2Q_2^H$, the vectors $q_1, \ldots, q_r$ span the *speech (+ noise) subspace*. Since $Q^H U = I_M$ we conclude that $Q_1^H U_2 = O$ so that the vectors $u_{r+1}, \ldots, u_M$ span the orthogonal subspace containing *noise* only.



Estimation of R_X

Similar to what we did before, we can compute (estimate) R_X from the GEVD of R_Y as

$$\hat{R}_X = Q_1(\Lambda_1 + I_r)Q_1^H$$

or, by reducing the remaining eigenvalues by one,

$$\hat{R}_X = Q_1\Lambda_1 Q_1^H$$

GEVD versus pre-whitening

We have $R_Y = Q(\Lambda + I_M)Q^H$ so that

EVD of $R_{\tilde{Y}}$

$$R_{\tilde{Y}} = R_V^{-\frac{1}{2}} R_Y R_V^{-\frac{1}{2}} = R_V^{-\frac{1}{2}} Q(\Lambda + I_M)Q^H R_V^{-\frac{1}{2}} = \tilde{U}(\tilde{\Lambda} + I_M)\tilde{U}^H$$

from which we conclude that $\tilde{\Lambda} = \Lambda$ and $\tilde{U} = R_V^{-\frac{1}{2}} Q$, and thus $Q = R_V^{\frac{1}{2}} \tilde{U}$.

The approximation of R_X obtained by the GEVD is thus given by

$$\hat{R}_X = Q_1 \Lambda_1 Q_1^H = R_V^{\frac{1}{2}} \tilde{U}_1 \tilde{\Lambda}_1 \tilde{U}_1^H R_V^{\frac{1}{2}},$$

which is identical to the result obtained by pre-whitening.

Beamforming

Recall that if $\mathrm{rank}(R_X) = r < M$, we can express R_Y as

$$R_Y = Q_1(\Lambda_1 + I_r)Q_1^H + Q_2Q_2^H$$

Since the beamformer takes linear combinations of the microphone signals ($z = w^H y$), we have that

$$R_Z = w^H R_Y w = w^H Q_1(\Lambda_1 + I_r)Q_1^H w + w^H Q_2 Q_2^H w$$

Since we know that $U_1^H Q_1 = I_r$ and $U_1^H Q_2 = O$, we expect that a "good" beamformer can be expressed as a linear combination of the columns of U_1. That is, $w = U_1 a$, where $a \in \mathbb{C}^r$.

Beamformer performance measures

Beamformer performance measures:

- Output signal-to-noise ratio (SNR)
- Means square error (MSE)
- Noise reduction
- Speech distortion
- ...

Output SNR

We can consider the output SNR, given by

$$\mathrm{SNR}_{\mathrm{out}}(w) = \frac{E|w^H X|^2}{E|w^H V|^2} = \frac{w^H R_X w}{w^H R_V w}.$$

Note that the SNR is a real-valued function of the complex vector variable w.

Output SNR

Theorem: Let $f : \mathbb{C}^n \mapsto \mathbb{R}$ be a real valued function of a complex variable z. Let $f(z) = g(z, \bar{z})$, where $g : \mathbb{C}^n \times \mathbb{C}^n \mapsto \mathbb{R}$ is a function of two complex variables such that $g(z, a)$ and $g(b, z)$, $a, b \in \mathbb{C}$, are analytic functions of z. Then a necessary and sufficient condition for f to have a stationary point is that $\nabla_z g = 0$, where the partial derivative with respect to z treats $\bar{z}$ as a constant, or $\nabla_{\bar{z}} g = 0$.

Theorem: Let f and g be defined as above. Then the gradient $\nabla_{\bar{z}} g(z)$ defines the direction of steepest descent of f at z.

[1] D.H. Brandwood, "A complex gradient operator and its application in adaptive array theory ", *IEE Proceedings*, vol. 130, no. 1, pp. 11-16, February 1983.

Output SNR

Taking the derivative of $\mathrm{SNR}_{\mathrm{out}}$ with respect to $\bar{w}$, we find that

$$\nabla_{\bar{w}} \mathrm{SNR}_{\mathrm{out}}(v) = R_X v - \frac{v^H R_X v}{v^H R_V v} R_V v = 0,$$

where v is a stationary point of $\mathrm{SNR}_{\mathrm{out}}$. Hence, we have $R_X v = \lambda R_V v$ where v is a generalised eigenvector with corresponding eigenvalue

$$\lambda = \frac{v^H R_X v}{v^H R_V v},$$

and we conclude that

$$\mathrm{SNR}_{\mathrm{out}}(w) \leq \max_v \frac{v^H R_X v}{v^H R_V v} = \lambda_1.$$

Output SNR

We conclude that the choice $w = u_1$ results in maximising the output SNR.

Note that this result is unique up to a scaling. Indeed, if $z = \alpha u_1$ for any $\alpha \neq 0$, we have

$$\frac{z^H R_X z}{z^H R_V z} = \frac{w^H R_X w}{w^H R_V w}$$

which is obvious since the eigenvectors are unique up to an arbitrary scaling $\alpha \neq 0$.

In addition, this result is independent of $r = \mathrm{rank}(R_X)$.

Mean squared-error

Consider the mean squared-error (MSE) between the beamformer output and the desired target signal at the reference microphone, which we will assume, without loss of generality, to be microphone 1.

We have

$$\begin{aligned} \mathrm{E}|w^H Y - X_1|^2 &= \mathrm{E}|w^H X + w^H V - X_1|^2 \\ &= \mathrm{E}|w^H X - X_1|^2 + \mathrm{E}|w^H V|^2, \end{aligned}$$

where we used the property $\mathrm{E}(XV^H) = 0$. The term $\mathrm{E}|w^H X - X_1|^2$ represents the *signal distortion*, whereas the term $\mathrm{E}|w^H V|^2$ represents the *residual noise variance*

Mean squared-error

We can compromise between signal distortion and noise reduction by defining the constraint optimisation problem

$$\begin{aligned} \text{minimise} \quad & \mathrm{E}|w^H X - X_1|^2 \\ \text{subject to} \quad & \mathrm{E}|w^H V|^2 \leq c, \end{aligned}$$

where $0 \leq c \leq \sigma_{V_1}^2$ and $\sigma_{V_1}^2$ the noise variance at the reference microphone before beamforming.

MMSE solution

In order to find the expressions for the different beamformers, we express the beamformers weights in terms of the generalised eigenvectors. That is, we have $w = Ua$ with $a \in \mathbb{C}^M$.

Let $e_1 = (1, 0, \ldots, 0)^T \in \mathbb{C}^M$. With this we have $x_1 = e_1^H x$ so that we can express the objective function as

$$\begin{aligned} \mathrm{E}|w^H X - X_1|^2 &= \mathrm{E}|w^H X - e_1^H X|^2 \\ &= a^H U^H R_X U a + \sigma_{X_1}^2 - 2\mathrm{Re}\{a^H U^H R_X e_1\} \\ &= a^H \Lambda a + \sigma_{X_1}^2 - 2\mathrm{Re}\{a^H U^H R_X e_1\}, \end{aligned}$$

and the feasible set becomes $\{a \in \mathbb{C}^M : a^H a \leq c\}$.

MMSE solution

The corresponding Lagrangian is given by

$$L(a, \mu) = a^H \Lambda a + \sigma_{X_1}^2 - 2\mathrm{Re}\{a^H U^H R_X e_1\} + \mu(a^H a - c),$$

with $\mu \geq 0$ a Lagrange multiplier.

Let a^* denote the (unique) minimiser. The optimality conditions (KKT conditions) for a^* to be optimal are then given by

$$\nabla_{\bar{a}} L(a^*, \mu) = \Lambda a^* - U^H R_X e_1 + \mu a^* = 0.$$

MMSE solution

Hence,

$$a^* = (\Lambda + \mu I_M)^{-1} U^H R_X e_1,$$

and thus

$$\begin{aligned} w^* &= U a^* \\ &= U(\Lambda + \mu I_M)^{-1} U^H R_X e_1 \end{aligned}$$

where the Lagrange multiplier $\mu \geq 0$ is chosen such that[1] $a^H a = c$.

[1] Since the minimum of our minimisation problem is attained on the boundary of the feasible set $\{a \in \mathbb{C}^m : a^H a \leq c\}$, we can replace the inequality constraint by an equality one.

MMSE solution

As mentioned before, in many applications we have $\mathrm{rank}(R_X) = r < M$ and we have $R_X = Q_1 \Lambda_1 Q_1^H$.

In those cases the optimal filter weights w^* become

$$\begin{aligned} w^* &= U(\Lambda + \mu I_M)^{-1} U^H Q_1 \Lambda_1 Q_1^H e_1 \\ &= U_1(\Lambda_1 + \mu I_r)^{-1} \Lambda_1 Q_1^H e_1 \end{aligned}$$

since $U_1^H Q_1 = I_r$ and $U_2^H Q_1 = O$.

We indeed conclude that MMSE optimal beamformers can be expressed as a linear combination of the columns of U_1

MMSE solution

Note that since $R_X = Q\Lambda Q^H$ and $R_V = QQ^H$ we have

$$\begin{aligned} U(\Lambda + \mu I_M)^{-1}U^H &= (U^{-H}(\Lambda + \mu I_M)U^{-1})^{-1} \\ &= (Q(\Lambda + \mu I_M)Q^H)^{-1} \\ &= (R_X + \mu R_V)^{-1} \end{aligned}$$

and we conclude that

$$w^* = (R_X + \mu R_V)^{-1} R_X e_1.$$

This solution is referred to as the *signal-distortion weighted* (SDW) Wiener filter

Multi-channel Wiener filter

The case $\mu = 1$ gives the classical multi-channel Wiener filter:

$$w_{MWF} = R_Y^{-1} R_X e_1.$$

In the case we have $R_X = \sigma_{X_1}^2 dd^H$ this reduces to

$$w_{MWF} = \sigma_{X_1}^2 R_Y^{-1} d.$$

In fact, the parameter μ can be seen as a trade-off parameter that controls the signal distortion and noise reduction.

MVDR beamformer

The choice $\mu = 0$ will lead to the MVDR beamformer.

Recall that

$$\mathrm{E}|w^H X - X_1|^2 = a^{*H}\Lambda_1 a^* + \sigma_{X_1}^2 - 2\mathrm{Re}\{a^{*H}U_1^H R_X e_1\},$$

where

$$\begin{aligned} a^* &= \Lambda_1^{-1} U_1^H R_X e_1 \\ &= \Lambda_1^{-1} U_1^H Q_1 \Lambda_1 Q_1^H e_1 \\ &= Q_1^H e_1 \end{aligned}$$

MVDR beamformer

With this we have

$$a^{*H}\Lambda_1 a^* = e_1^H \underbrace{Q_1\Lambda_1 Q_1^H}_{R_X} e_1 = \sigma_{X_1}^2,$$

and

$$a^{*H}U_1^H R_X e_1 = e_1^H Q_1 \cancel{U_1^H} Q_1\Lambda_1 Q_1^H e_1 = \sigma_{X_1}^2,$$

I_r

so that

$$\mathrm{E}|w^H X - X_1|^2 = a^{*H}\Lambda_1 a^* + \sigma_{X_1}^2 - 2\mathrm{Re}\{a^{*H}U_1^H R_X e_1\} = 0,$$

and we conclude that the response is distortionless.

MVDR beamformer

As a special case, consider $r = 1$ so that R_X can be expressed as $\sigma_{X_1}^2 dd^H$.

We have $w^* = u_1 a^* = u_1 q_1^H e_1$ so that

$$\begin{aligned} w^* &= u_1 u_1^H q_1 q_1^H e_1 \\ &= U U^H q_1 q_1^H e_1 \\ &= R_V^{-1} q_1 q_1^H e_1 \\ &= \lambda_1^{-1} R_V^{-1} R_X e_1 \\ &= \lambda_1^{-1} \sigma_{X_1}^2 R_V^{-1} dd^H e_1 \\ &= \lambda_1^{-1} \sigma_{X_1}^2 R_V^{-1} d \end{aligned}$$

MVDR beamformer

To find an expression for λ_1, we note that w^* is a scaled version of u_1 and, therefore, maximises the output SNR:

$$\begin{aligned}\lambda_1 &= \frac{w^{*H} R_X w^*}{w^{*H} R_V w^*} \\ &= \frac{d^H R_V^{-1} \left(\sigma_{X_1}^2 d d^H\right) R_V^{-1} d}{d^H R_V^{-1} d} \\ &= \sigma_{X_1}^2 d^H R_V^{-1} d\end{aligned}$$

and we conclude that

$$w^* = \frac{R_V^{-1} d}{d^H R_V^{-1} d}.$$

Multi-channel Wiener filter

Recall that in general we have $w^* = U(\Lambda + \mu I_M)^{-1} U^H R_X e_1$. Using the same arguments as before, we have for $r = 1$ that

$$
\begin{aligned}
w^* &= \frac{\sigma_{X_1}^2}{\lambda_1 + \mu} R_V^{-1} d \\
&= \frac{\sigma_{X_1}^2}{\sigma_{X_1}^2 d^H R_V^{-1} d + \mu} R_V^{-1} d \\
&= \frac{\sigma_{X_1}^2}{\sigma_{X_1}^2 + \mu (d^H R_V^{-1} d)^{-1}} \frac{R_V^{-1} d}{d^H R_V^{-1} d}
\end{aligned}
$$

which shows that the (SDW) MWF can be implemented as an MVDR beamformer, followed by a single-channel Wiener filter.