Institut Mines-Télécom

Audio signal analysis, indexing and transformation

Roland Badeau
`roland.badeau@telecom-paris.fr`

M2 Mathématiques / Vision / Apprentissage

TELECOM Paris
IP PARIS

Contents

List of Figures

Acronyms

ACF *Auto-Covariance Function*

AIC *Akaike Information Criterion*

BSS *Blind Source Separation*

DFT *Discrete Fourier Transform*

DTFT *Discrete Time Fourier Transform*

DUET *Degenerate Unmixing Estimation Technique*

EDC *Efficient Detection Criteria*

EDS *Exponentially Damped Sinusoids*

ESM *Exponential Sinusoidal Model*

ESPRIT *Estimation of Signal Parameters via Rotational Invariance Techniques*

EVD *EigenValue Decomposition*

FFT *Fast Fourier Transform*

FIR *Finite Impulse Response*

FT *Fourier Transform*

HR *High Resolution*

ICA *Independent Component Analysis*

IID *Independent and Identically Distributed*

ITC *Information Theoretic Criteria*

JADE *Joint Approximate Diagonalization of Eigenmatrices*

LPC *Linear Predictive Coding*

MDCT *Modified Discrete Cosine Transform*

MDL *Minimum Description Length*

MMSE *Minimum Mean Square Error*

MSE *Mean Square Error*

OLA *OverLap-Add*

PDF *Probability Density Function*

PSD *Power Spectral Density*

PSOLA *Pitch-Synchronous OverLap-Add*

SNR *Signal to Noise Ratio*

SOBI *Second Order Blind Identification*

SOLA *Synchronized OverLap-Add*

STFT *Short Time Fourier Transform*

SVD *Singular Value Decomposition*

TD-PSOLA *Time-domain Pitch-Synchronous OverLap-Add*

TF *Time-Frequency*

WSS *Wide Sense Stationary*

Mathematical notation

$\mathbb{N}$ set of natural numbers

$\mathbb{Z}$ set of integers

$\mathbb{R}$ set of real numbers

$\mathbb{C}$ set of complex numbers

$\mathcal{R}e(.)$ real part

$\mathcal{I}m(.)$ imaginary part

x (normal font, lower case) scalar

$\boldsymbol{x}$ (bold font, lower case) vector

$\boldsymbol{A}$ (bold font, upper case) matrix

$\|.\|_2$ Euclidean norm of a real vector, or Hermitian norm of a complex vector

$\|.\|_F$ Frobenius norm of a matrix

$\overline{(.)}$ conjugate of a matrix / vector / number

$.^\top$ transpose of a matrix

$.^H$ conjugate transpose of a matrix

Span(.) range space of a matrix

Ker(.) kernel of a matrix

dim(.) dimension of a vector space

rank(.) rank of a matrix

trace(.) trace of a square matrix

det(.) determinant of a square matrix

$.^\dagger$ pseudo-inverse of a matrix (if $\boldsymbol{A} \in \mathbb{R}^{M\times K}$ with $M \geq K$, $\boldsymbol{A}^\dagger = (\boldsymbol{A}^\top \boldsymbol{A})^{-1}\boldsymbol{A}^\top$)

diag(.) diagonal matrix formed from a vector of diagonal coefficients, or from a matrix with same diagonal entries

$\boldsymbol{I}_K$ $K \times K$ identity matrix

$H(\nu) = \sum_{t\in\mathbb{Z}} h(t)e^{-2\imath\pi\nu t}$ discrete time Fourier transform

$L^\infty(\mathbb{R}^M)$ Lebesgue space of essentially bounded functions on $\mathbb{R}^M$

$*$ convolution product between two sequences (scalars, but also matrices and vectors of appropriate dimensions)

$\mathbf{1}_A$ is 1 if A is true, or 0 if A is false

$\mathbb{E}[.]$ expected value of a random variable or vector

$\mathbb{H}[.]$ entropy of a random variable or vector

$\mathbb{I}[.]$ mutual information of the entries of a random vector

$\widehat{(.)}$ estimator of a parameter

CRB$\{.\}$ Cramér-Rao bound

Chapter 1

Spectral and temporal modifications

This chapter is mostly a translation in English of a course handout by Bertrand David. It draws from passages of various documents and mainly from a work specifically dedicated to audio signal processing [KB98] (Chap. 7). It develops more particularly the phase vocoder-based methods and the temporal methods.

1 Introduction

The objective of these modifications, which correspond to usual needs in various fields of sound and speech processing, is to *independently* control the temporal, spectral and possibly formantic (slow variations of the spectrum) evolutions of the signal:

- temporal dilation: we want to modify duration scales without altering the spectral content and especially the pitch in the case of a harmonic signal,
- pitch variation: in the case of harmonic signals, we want to change the pitch of the sound, while retaining its temporal evolution (for example the prosodic flow in the case of speech) and especially its duration,
- formantic control: in the case of a pitch modification, one can choose to either modify the spectral scale as a whole (and therefore to move the formants or spectral envelope) or to keep the spectral envelope constant while transposing the line spectrum.

Applications where these types of independent modifications are numerous:

- synthesis by sampling of a wave table (musical sounds or speech segments [All91]),
- post-synchronization: to perform a synchronization of sound and image,
- data compression [MEJ86]
- reading for blind people: our inner reading is much faster than our diction. By shrinking durations we can allow blind people to increase their speed of browsing documents,
- learning foreign languages: slowing the speech flow is helpful,
- musical post-production: to mix several recordings it may be useful to speed up or slightly reduce the tempo. It may also be interesting to locally correct the precision of a voice or instrument.

We can classify in three types the methods that carry out these modifications:

- methods inspired by the circular reader head or modified radiocassette (we add / subtract portions of the signal [FEJ54]). These methods are called *temporal*,

- phase vocoder-based methods (*spectral* methods using the *Short Time Fourier Transform* (STFT) [SFL67, Por76]),
- methods based on signal models (*Linear Predictive Coding* (LPC) [Mak75], Sinus+Noise [SS90], Audio grains [JP88],. . .).

Temporal methods have resulted in many developments in digital signal processing: *Synchronized OverLap-Add* (SOLA) method [RW85], *Pitch-Synchronous OverLap-Add* (PSOLA) method [MC90]. This last one uses a synchronized copy/deletion technique on the glottal impulses. This achieves a very good quality modification of the time scale *without resampling the signal*.

The PSOLA method can be adapted to perform formantic modifications by modifying the duration of the segments without modifying the position of the glottal impulses. In this way, we can transpose the spectral envelope without modifying the pitch or the duration, and thus modify the timbre of a voice (transform a male voice into a more female voice by example). Other techniques of formantic modifications use cepstral representations [CLM95].

2 Signal models to define temporal and spectral distortions

A simple replay at 16 kHz of an audio signal sampled at 8 kHz is enough to convince us that the temporal and spectral expansions or compressions are interdependent. This dependence can be interpreted as a simple theoretical result on the *Fourier Transform* (FT): the Heisenberg uncertainty relation translates this dependence in terms of supports, and the high frequency decrease in the FT is related to the regularity of the time signal. Therefore the definition of *independent* temporal and spectral distortions can only be obtained for well-defined signal *models*.

2.1 McAuley-Quatieri model

Speech production model. The most common and widely used speech signal model is that of a time-varying linear filter, excited by a harmonic source (in the case of voiced sounds) or by a stationary random process with flat spectrum (in the case of unvoiced sounds). In the present case, we consider the voiced case, for which this source is a sum of sinusoidal components whose frequencies are multiple of a fundamental frequency $f_0(t)$. This representation is equivalent to writing the source as a Dirac comb whose period depends on time.

Let $g_t(\tau)$ be the impulse response of the system at time t. The signal is then simply written as a function of the excitation signal $e(t)$:

$$x(t) = \int_{-\infty}^{+\infty} g_t(\tau)e(t-\tau)d\tau. \tag{1.1}$$

This non time-invariant system can be represented by a time-dependent frequency response:

$$G(t,f) = M(t,f)\exp j\varphi(t,f).$$

The temporal variations of g_t are linked to the articulatory movements and are considered slow compared to the fundamental period of the signal. On the other hand, these variations are assumed to be weak over the duration of the filter memory. The system is *quasi-stationary*.

For voiced speech, that is to say involving a periodic vibration of the vocal cords, the excitation signal writes:

$$e(t) = \sum_{k=-\infty}^{+\infty} \exp j\xi_k(t) \tag{1.2}$$

with

$$\xi_k'(t) = 2\pi f_k(t).$$

The quasi-stationary nature of g_t leads to a practical limitation of the support of this function to a dimension of the order of the system memory. The integral in expression 1.1 is therefore well-defined in practice. In the same way, the frequency support of speech is limited in practice and the discrete sum of expression 1.2 is in fact a finite sum of $L(t)$ complex exponential terms. Taking into account the fact that f_0 varies little over the memory duration of the filter we can expand

$$\xi_k(t-\tau) \approx \xi_k(t) - 2\pi\tau k f_0(t)$$

in the vicinity of t (*i.e.* for τ lower than the memory of the filter). Then we obtain:

$$x(t) = \sum_{k=1}^{L(t)} M(t, f_k(t)) \exp j[\xi_k(t) + \varphi(t, f_k(t))] \tag{1.3}$$

Mc-Auley and Quatieri model. This model was introduced by McAulay and Quatieri around 1985 [MQ86], mainly for low rate speech coding. So it is related to the expression obtained in 1.3. It is however a little more general since it does not assume a necessarily harmonic relationship between the instantaneous frequencies. The signal is represented as a sum of sines whose frequencies, amplitudes and phases are controlled over time:

$$x(t) = \sum_{k=1}^{L(t)} A_k(t) \exp j\Psi_k(t) \quad \text{with} \quad \Psi_k'(t) = \omega_k(t) = 2\pi f_k(t) \tag{1.4}$$

where $A_k(t)$ is the amplitude at time t of sine k, $\Psi_k(t)$ is the *instantaneous phase* of this sine at time t and $f_k(t)$ is its *instantaneous frequency*. This decomposition is not unequivocal and we generally consider that the functions $A_k(t)$ and $\omega_k(t)$ have slow variations compared to the functions $\exp(j\Psi_k(t))$.

2.2 Serra-Smith model

This model was developed in the early 90s [SS90] in order to meet the need for an analysis/synthesis system accounting for the noisy component of music or speech. This component is very expensive to represent as a sum of sines. The proposed model is therefore an extension of that of MacAuley-Quatieri:

$$x(t) = \sum_{k=1}^{L(t)} A_k(t) \exp j\Psi_k(t) + b(t) \tag{1.5}$$

where $b(t)$ is a stationary random process filtered by a time-varying filter, like filter g_t presented above. Let h_t be this filter, we will then write, taking into account the causality of signals,

$$b(t) = \int_0^t h_t(\tau) u(t-\tau) d\tau \tag{1.6}$$

where $u(t)$ is a white stationary random process.

The complete analysis/modification/synthesis system includes

- an estimation phase of the deterministic components,
- a phase of linear interpolation of the amplitudes and cubic interpolation of the phases from one frame to another of the signal for these components,
- a subtraction of this deterministic part to get $b(t)$ for each frame,
- the application of a possibly distinct transformation algorithm for each of the two components,
- resynthesis.

3 Definitions and equivalences

All the definitions given here relate to a model of signal with sinusoidal components. They therefore apply to the McAuley-Quatieri model or to the deterministic part of the Serra-Smith model. The phases at $t = 0$ will be assumed to be zero for the sake of simplification (this term can be incorporated into the definition of the amplitudes).

3.1 Temporal distortion

We define the time distortion function using the new time scale τ and the original time scale t by:

$$\tau = T(t). \tag{1.7}$$

This function is continuous and bijective from $\mathbb{R}^+$ to $\mathbb{R}^+$. The modification of the signal's time scale $x(t)$ is then defined by

$$y(\tau) = \sum_{k=1}^{L(T^{-1}(\tau))} A_k(T^{-1}(\tau)) \exp(j\phi_k(\tau)) \tag{1.8}$$

The conservation of the frequency content then requires to maintain the values of the instantaneous frequencies, hence the relation:

$$\phi_k(\tau) = \int_0^{\tau} \omega_k(T^{-1}(u))du \tag{1.9}$$

3.2 Pitch modification

To modify the pitch of the signal $x(t)$ we build the signal:

$$y(t) = \sum_{k=1}^{L(t)} A_k(t) \exp(j\Phi_k(t)) \tag{1.10}$$

The alteration of the frequency content is defined using a function $\alpha(t)$ called frequency compression rate, according to the expression:

$$\Phi_k(t) = \int_0^{t} \alpha(u)\omega_k(u)du \tag{1.11}$$

3.3 Reciprocity

By a quick calculation we show that the operating sequence: $x \to x_1$ by time distortion ($\tau = T(t)$) followed a simple replay at a different temporal speed (*i.e.* without maintaining the frequency characteristics) $x_1(\tau) = y(v)$ with $v = T^{-1}(\tau)$ is equivalent to a frequency modification governed by the function $\alpha(t) = T'(t)$, that is to say:

$$y(v) = \sum_{k=1}^{L(v)} A_k(v) \exp j\Phi_k(v) \tag{1.12}$$

with

$$\Phi_k(t) = \int_0^{v} T'(u)\omega_k(u)du \tag{1.13}$$

This relationship is particularly useful in cases where the corresponding time distortion is a multiplying factor, like for instance $T(t) = 2t$. Then $T'(t)$ is constant and the replay operation is a simple replay of the signal obtained at a different rate (for example, for sampled signals, $F'_e = 2Fe$ in the previous case).

4 Short-term Fourier transform

The methods of analysis/synthesis and modification of sounds based on the use of the STFT are very common. The corresponding tool is usually called *phase vocoder*. It refers to the polar representation (module & phase) of the STFT.

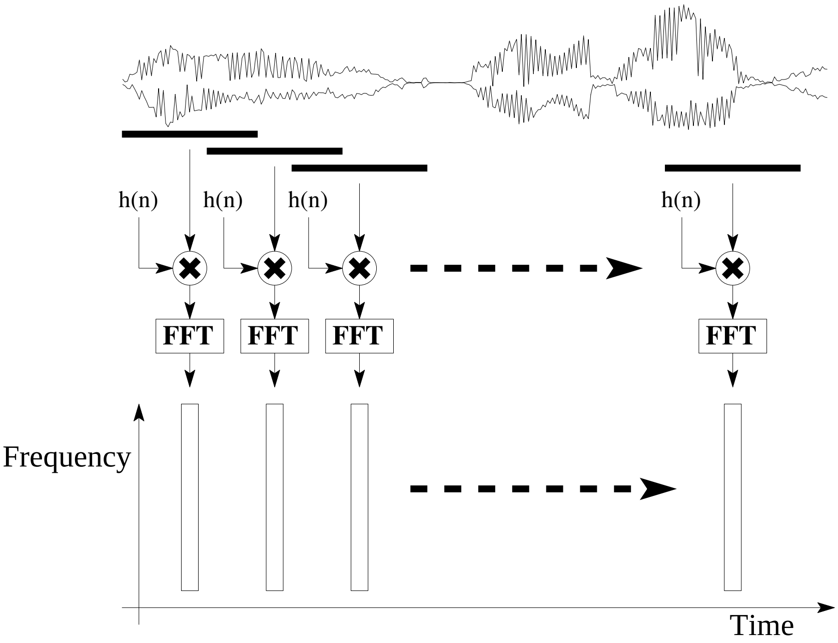


Figure 1.1: Short-term Fourier transform

4.1 Theoretical reminders

The block diagram of the STFT is represented in figure 1.1, as it is numerically computed. The principle is that of a sliding Fourier transform, performed on overlapping frames of the signal. Each frame is windowed by an analysis window. We will write the STFT of a digital signal $x(n)$ in the form

$$\tilde{X}(t_a, \nu) \triangleq \sum_{n\in\mathbb{Z}} x(n + t_a) w_a(n) e^{-j2\pi\nu n}. \tag{1.14}$$

w_a denotes the analysis window, most often of finished length, real and symmetrical. The analysis times are implicitly indexed by a natural integer u, that is $t_a = t_a(u), \quad u \in \mathbb{N}$. We preferred here a notation function of ν while a notation function of $e^{j2\pi\nu}$ would have been more consistent with the interpretation in terms of sliding Fourier transform, but this choice simplifies the expressions.

Interpretation. A quick calculation shows that, by defining $h(n) = w_a(-n)e^{j2\pi\nu_p n}$, the expression 1.14 can be written in the form of a convolution product:

$$\tilde{X}(t_a, \nu_p) = [x * h](t_a). \tag{1.15}$$

If $w_a(n)$ is a real and pair window of finite length, its FT $W_a(e^{j2\pi\nu})$ is real and even. The FT of h is then simply $H(e^{j2\pi\nu}) = W_a(e^{j2\pi(\nu-\nu_p)})$. An example of typical result is given in figure 1.2 for $\nu_p = 0.3$. This example shows that $\tilde{X}(t_a, \nu_p)$ performs a bandpass *Finite Impulse Response* (FIR) filtering around frequency ν_p. The characteristics of the filter are linked to that of the chosen analysis window. This interpretation is at the origin of the qualification of *bandpass convention* given to expression 1.14. There is another convention, called *low pass*, often used for its ease of handling calculations. We will, however, stick to the band pass convention because it corresponds to the practical realization.

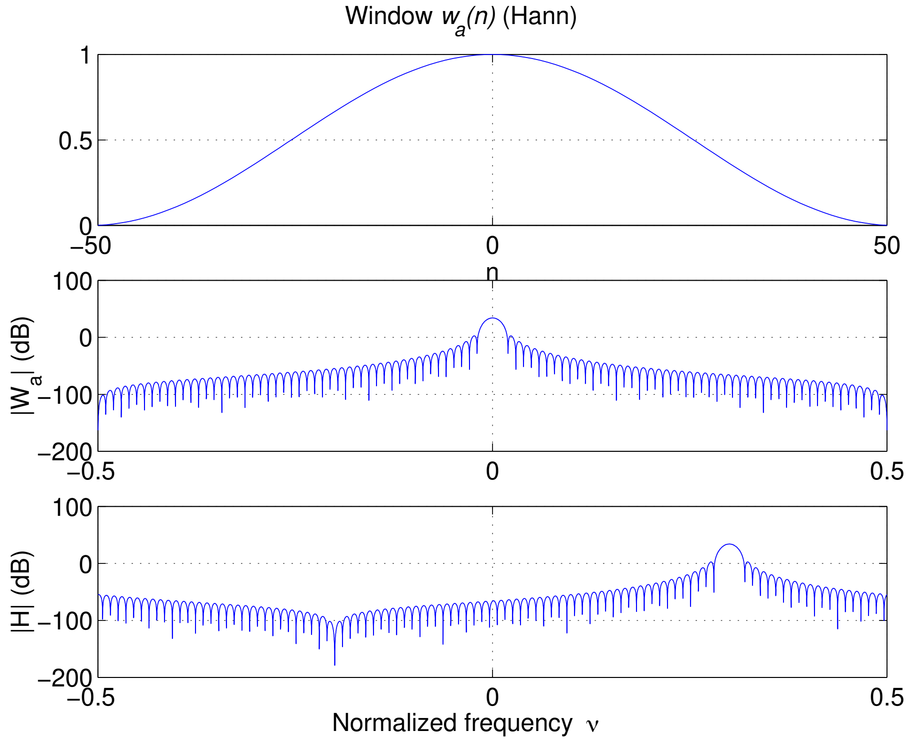


Figure 1.2: Bandpass filtering equivalent to an STFT channel

Discrete version of the STFT. In practice, the Fourier transform is evaluated using the *Discrete Fourier Transform* (DFT). This is equivalent to setting $\nu_p = p/N$ in the expression of $\tilde{X}(t_a, \nu_p)$. N is the order of the DFT. We thus obtain a discrete version of the STFT, i.e. sampled in frequency, i.e.

$$\tilde{X}(t_a, \nu_p) = \sum_{n=0}^{N-1} x(n + t_a) w_a(n) e^{-j2\pi \frac{pn}{N}}. \tag{1.16}$$

In order to avoid time aliasing, the length of the analysis windows will be less than or equal to N.

Modifications and problems posed. The modification of sounds by the phase vocoder involve obtaining a modified STFT from $\tilde{X}(t_a, \nu_p), \quad k = 0, \ldots, N-1$, then resynthesizing the signal. We denote by $t_s = t_s(u)$ the temporal synthesis marks. Hence the modification

$$\tilde{X}(t_a(u), \nu_p) \to Y(t_s(u), \nu_p).$$

The main difficulty encountered is that Y must satisfy strong conditions [Por76] to correspond to a well-defined original sequence. The solution to this problem is found in the least squares sense [ML95]. However, one can write the perfect reconstruction conditions in case no modification is performed (*i.e.* $t_s = t_a$ and $Y = \tilde{X}$).

Perfect reconstruction condition. The reverse operation of the analysis is carried out for the synthesis: from the stream of discrete spectra $\tilde{X}(t_a(u), \nu_p)$ we compute an inverse DFT and we reconstruct the signal by *OverLap-Add* (OLA). The result is given by

$$y(n) = \sum_u w_s(n - t_s(u)) y_w(n - t_s(u), t_s(u)) \tag{1.17}$$

with $t_s(u) = t_a(u)$ and

$$y_w(n, t_s(u)) = \frac{1}{N}\sum_{p=0}^{N-1} Y(t_s(u), \nu_p)e^{j2\pi\nu_p n}.$$

Taking $Y = \tilde{X}$ into account and substituting the expression 1.14 in 1.17, we show that $x(n) = y(n)$ is obtained by using the sufficient condition:

$$\sum_u w_a(n - t_a(u))w_s(n - t_a(u)) = 1 \tag{1.18}$$

5 Modifications using the phase-vocoder

5.1 Instantaneous frequency

The transformations presented in section 3 require the calculation of the instantaneous frequencies $\omega_k(t)$ of each of the components of the sum 1.4. This calculation is carried out from two successive short term spectra $\tilde{X}(t_a(u), \nu_p)$ and $\tilde{X}(t_a(u+1), \nu_p)$ $p = 0, \ldots, N-1$, under certain conditions that ensure the existence of a solution.

Narrow-band condition. This first condition ensures the presence of *at most* one component per channel of the STFT. The substitution of the expression 1.4,

$$x(t) = \sum_{k=1}^{L(t)} A_k(t) \exp j\Psi_k(t),$$

into expression 1.14 of the STFT gives:

$$\tilde{X}(t_a(u), \nu_p) = \sum_{n=0}^{N-1}\sum_{k=1}^{L(n+t_a)} A_k(n + t_a) \exp(j\Psi_k(n + t_a))w_a(n)e^{-j2\pi\nu_p n}$$

We then use the quasi-stationarity of the model, namely:

$$\begin{aligned} A_k(n + t_a) &\approx A_k(t_a) \\ \Psi_k(n + t_a) &\approx \Psi_k(t_a) + n\omega_k(t_a) \end{aligned}$$

and finally, by defining $\omega_k(t) = 2\pi f_k(t)$:

$$\tilde{X}(t_a(u), \nu_p) = \sum_{k=1}^{L(n+t_a)} A_k(t_a) \exp(j\Psi_k(t_a))W_a(e^{j2\pi(\nu_p - f_k(t_a))}) \tag{1.19}$$

The narrow band condition leads to non-negligible values of $W_a(e^{j2\pi(\nu_p - f_k(t_a))})$ for at most one value of k. Let $k = l$ be this value if it exists. If we note f_c the cutoff frequency of the low-pass filter whose impulse response is $w_a(n)$, then the existence of l implies

$$|\nu_p - f_l(t_a)| \le f_c\ ,$$

that is, the component number l is in the pass-band of the filter corresponding to the p-th channel of the STFT. The expression 1.19 is then reduced to the contribution of the l-th component alone:

$$\tilde{X}(t_a(u), \nu_p) = A_l(t_a) \exp(j\Psi_l(t_a))W_a(e^{j2\pi(\nu_p - f_l(t_a))}) \tag{1.20}$$

If we assume that w_a is real and even, and therefore W_a is real and even, this expression is interpreted as follows: the phase of the STFT gives access to the instantaneous phases of the components of $x(t)$, up to an indeterminacy of a multiple of 2π, and the module of the STFT gives access to instantaneous amplitudes of $x(t)$, up to an amplitude factor due to filtering. We can therefore deduce the instantaneous frequencies of each component from the phase of the flow of short-term spectra, provided that the indeterminacy of 2π is removed.

Example: for a Hann analysis window of length L, the narrow-band condition applied to a line spectrum of harmonics (case of a voiced speech segment for example) leads to a spacing of spectral peaks at least equal to the bandwidth of the Fourier transform of the window, that is $4/L$. That results in $f_0 < 4/L$, i.e. a window length at least equal to 4 times the fundamental period.

Overlap condition. We will see here that the removing of the indeterminacy leads to a condition of minimal recovery of the analysis windows. Indeed, the phase difference between two successive analysis times, for the p-th channel of STFT is written, by defining $\Phi(t_a(u), \nu_p) = \arg \tilde{X}(t_a(u), \nu_p)$,

$$\begin{aligned}
\Delta\Phi_p &= \Phi(t_a(u+1), \nu_p) - \Phi(t_a(u), \nu_p) = \Psi(t_a(u+1)) - \Psi(t_a(u)) \ [2\pi] \\
&= 2\pi f_l \Delta t_a(u) + 2n\pi \\
&= 2\pi(f_l - \nu_p)\Delta t_a(u) + 2\pi\nu_p\Delta t_a(u) + 2n\pi
\end{aligned}$$

where n is a relative integer and $\Delta t_a(u) = t_a(u+1) - t_a(u)$. By taking $|\nu_p - f_l(t_a)| \leq f_c$ into account, the previous equation leads, if the condition 1.21 below is verified

$$f_c \Delta t_a(u) < 1/2, \tag{1.21}$$

to the inequality

$$|\Delta\Phi_p - 2\pi\nu_p\Delta t_a(u) - 2n\pi| < \pi \ ,$$

however there is one and only one value of n that verifies this property. This has removed the indeterminacy. In summary, we can thus get the value of the instantaneous frequency *in each STFT channel* by the following algorithm:

1. Calculation of the STFT at two successive times of analysis, which gives $\Delta\Phi_p$ for each channel ($p = 0, \ldots, N-1$),
2. For each channel, we look for the value $Q(n_0)$ of $Q(n) = \Delta\Phi_p - 2\pi\nu_p\Delta t_a - 2n\pi$ such that $|Q(n_0)| < \pi$,
3. we deduce the instantaneous frequencies by $f_l = \nu_p + \dfrac{Q(n_0)}{2\pi\Delta t_a}$.

Interpretation: inequality 1.21 leads to a minimum overlap condition between the analysis windows. Indeed, if we take for example a Hann window, for which we can estimate $f_c = 2/L$ where L is the window length, it becomes

$$\Delta t_a < \frac{L}{4}$$

which corresponds to a minimum recovery of 75% in analysis.

5.2 Temporal distortion

Once the instantaneous frequencies in each channel are deduced[1], the temporal signal distortion may be considered. In particular, instantaneous phases can be "unwound" so as to synchronize the modified STFT on the synthesis times. We then obtain the following modification algorithm, assuming that the analysis STFT $\tilde{X}(t_a(u), \nu_p)$ and the synthesis STFT $\tilde{Y}(t_s(u), \nu_p)$ are calculated for the index u, and given the time distortion law $T(t)$:

1. calculation of the STFT at time $t_a(u+1)$ and deduction of the instantaneous frequency $f_k(t_a(u))$ in each channel,
2. calculation of the new synthesis time $t_s(u) = T(t_a(u))$; in practice we take the whole part of this new instant
3. iteration of the synthesis instantaneous phase
$$\Phi_s(t_s(u+1), \nu_p) = \Phi_s(t_s(u), \nu_p) + 2\pi f_p(t_a(u))(t_s(u+1) - t_s(u))$$
4. calculation of the synthesis STFT for the index $u+1$ according to
$$\tilde{Y}(t_s(u+1), \nu_p) = A_p(t_a(u+1)) \exp j\Phi_s(t_s(u+1), \nu_p)$$

[1] in doing so, we assume that there is one and only one component by channel and therefore we can identify the indexes p (STFT channel) and k (components).

5.3 Pitch modification

The modification of pitch, or more generally of the frequency scale, is obtained either by temporal resampling or by spectral resampling.

Temporal resampling. This method is based on the reciprocity properties as seen in paragraph 3.3.

In the case of a constant frequency compression ratio $\alpha(t) = \alpha_0$, we obtain the desired modification by

1. a time stretch of factor α_0,
2. a replay at sampling frequency $\alpha_0 F_e$

where F_e is the original sampling frequency. This technique is equivalent to performing a resampling of factor $\alpha_0 = F'_e/Fe$ and playing at F_e. In this last case, it should however be noted that the time support is divided by α_0.

An extension of this resampling technique can be applied to obtain compression ratios $\alpha(t)$ variable over time. We use the canonical resampling method of digital signals by approaching α by a rational fraction at each analysis time $\alpha(t_a) = L(t_a)/M(t_a)$ and by performing the processing chain of figure 1.3 where $H(z)$ is a low-pass filter of cutoff frequency $\nu_c = \min(1/2L, 1/2M)$. We can therefore apply this processing to each frame of the signal and use synchronized analysis and synthesis times $t_a(u) = t_s(u)$. It should be noted that in this case, the phase vocoder is not used. This method can be quite demanding because it requires the calculation of a new interpolation filter H at each analysis step.

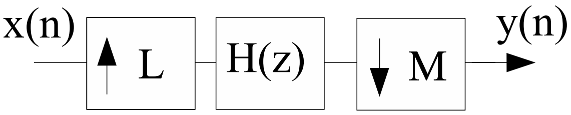


Figure 1.3: Canonical resampling chain of factor L/M

Spectral resampling. The phase vocoder allows a less greedy solution than the resampling in the case of variable compression ratios, by performing resampling in the frequency domain. This frequency resampling is performed by linear interpolation of the analysis short-term spectrum, that is

$$\begin{aligned} &q = \lfloor p/\alpha(t_a(u)) \rfloor \\ &\mu_p = p/\alpha(t_a(u)) - q \\ &\tilde{Y}(t_s(u), \nu_q) = (1-\mu_p)\tilde{X}(t_a(u), \nu_p) + \mu_p \tilde{X}(t_a(u), \nu_{p+1}) \end{aligned}$$

p and q are natural integers that index the channels of the STFT, they therefore vary from 0 to $N-1$. We note that this interpolation, if it presents no difficulty for compression rates greater than unity (pitch is increased), however requires completion of the high frequency spectrum for rates lower than unity (the synthesized sound is lower-pitched). One way to achieve this completion was suggested in [Sen82] and simply consists of copying the low frequency part of the spectrum into the missing part. This spectral copy gives good rendering for sampling frequencies of at least 16 kHz. In this case, the completion occurs at high frequencies where the sound has mainly unvoiced characteristics.

Finally, to carry out the modification, it is necessary to take into account the local modification of the time scale caused by the frequency modification. Indeed the phases of the synthesis STFT $\Phi_s(t_s(u), \nu_p) = \Phi_a(t_a(u), \nu_p)$ are now synchronized on synthesis times different from the analysis times:

$$\begin{aligned} \Phi_s(t_s(u+1), \nu_p) &= \Phi_s(t_s(u), \nu_p) + 2\pi f_p(t_a(u))\Delta t_a(u) \\ &= \Phi_s(t_s(u), \nu_p) + 2\pi\alpha(t_a(u)) f_p(t_a(u))\Delta t_s(u) \end{aligned}$$

We therefore see that the local analysis duration $\Delta t_a(u)$ has been divided by α in the synthesis. Let $\Delta t_s(u) = \Delta t_a(u)/\alpha(t_a(u))$. This corresponds to a virtual time distortion

$$T(t) = \int_0^t \alpha(w)^{-1} dw$$

To carry out the time scale modification, it is therefore necessary to finally apply a compensatory temporal distortion $D(t) = T^{-1}(t)$.

Note regarding the processing of spoken or singing voice. In the case of pitch modifications in spoken or singing voice, a direct transposition of the signal leads to the "Donald Duck" effect. Indeed, the transposition of the overall spectrum leads to a transposition of its envelope and therefore of the formants. The timbre is then severely modified and the voice acquires a nasal characteristic evoking the sound of the duck. This effect is also produced by the modification of the characteristic impedance of the medium caused by the mixed gaz breathed in by divers. A solution to overcome this defect consists in estimating the spectrum envelope before the processing (by LPC or direct modeling [EJM91]). The processing is then applied to the source signal (LPC residual for instance) then the obtained result is filtered to find the original spectral envelope, unchanged.

6 Pitch synchronous temporal method

This method, called *Time-domain Pitch-Synchronous OverLap-Add* (TD-PSOLA), assumes that we process a speech signal whose period is known.

The idea [MC90] is based on the assumption that the speech signal is made up of glottal pulses filtered by the vocal tract. We thus observe a succession of impulse responses, positioned at multiple times of the period (assumption of the time comb convoluted with the impulse response of the vocal tract).

We then define "analysis marks" synchronous with the fundamental frequency for the voiced parts, positioned on the waveform at each period. Scale modifications are then carried out as follows:

6.1 Modification of the time scale.

In order to modify the signal duration without altering the fundamental frequency, we will simply duplicate (for time stretching) or eliminate (for time compression) periods of the waveform, depending on the desired modification rate. So we are led to define synthesis marks also synchronous with the fundamental period, associated with the analysis marks (in a non-bijective way since some marks are duplicated or eliminated).

Short-term signals around each analysis mark are then extracted (by the use of a time window, by example a Hann window, of duration equal to two periods and centered on the analysis mark) and 'copied' around the corresponding synthesis marks, and the modified signal is obtained by a simple OLA method. Figure 1.4 illustrates the principle of this method for a local time stretch rate of 1.5.

We see that two periods of the original signal gave birth to three periods in the modified signal, which corresponds well to a time stretch but the duration of the period is not not modified (the spacing of the synthesis marks is the same as that of the analysis marks), the fundamental frequency of the signal is preserved. Figure 1.6 gives an application example to the sentence "il s'est" whose original is given in figure 1.5. We notice the unvoiced part in the center of the window (the sound 's'), separating the two voiced parts /i/ and /e/.

6.2 Modification of the frequency scale.

If we are able to position the analysis marks in the signal exactly at the start of each glottal wave (impulse response of the vocal tract occurring at each glottal closure), we can see that decreasing (resp. increasing) the time interval separating two consecutive analysis marks will increase (resp. decrease) the fundamental frequency, without the formants being modified (the impulse response is not modified, in particular its temporal decay and its resonance frequencies - the formants).

We are thus led to define synthesis marks corresponding to the modified value of the fundamental, and to associate them with the analysis marks, as previously. Since the synthesis marks are closer (elevation of the fundamental) or

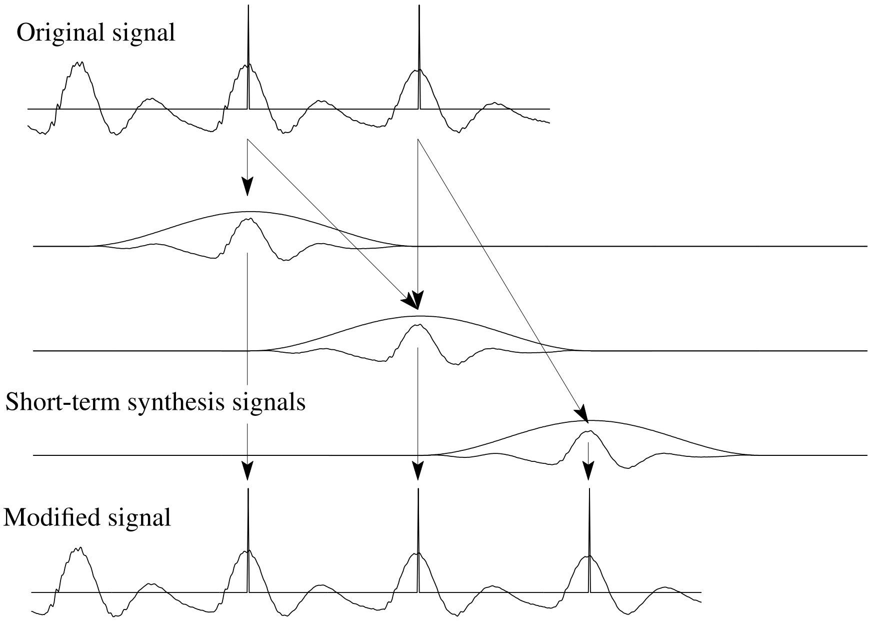


Figure 1.4: Modification of the duration of the signal by the TD-PSOLA method. At the top, the original signal, in the middle three short-term signals generated from two short-term signals centered around the first two analysis marks. At the bottom, the modified signal.

farther (lowering of the fundamental) than in the original signal, we have to duplicate or eliminate some marks in order to keep the duration of the signal. Figure 1.7 illustrates the principle of this method.

It can be seen that the synthesis marks being more spaced out than the analysis marks, the signal period is lengthened. In order to avoid an elongation of the signal, it is necessary to periodically eliminate some short-term signals.

When the signal no longer has a precise fundamental frequency (case of consonants for instance), the modification is carried out non-synchronously, until we find a region with a sharper fundamental.

The method described above is mainly applied to speech, and makes very good quality modifications. By its simplicity, it can be subject to a real-time implementation. However, its application to more complex sounds, or sounds devoid of "pitch" (case of music in general) poses serious problems.

The modifications of fundamental frequency are however very sensitive to the position of the analysis marks. To make the method more robust, the modifications of frequency scale can be performed in the frequency domain (FD-PSOLA method) [MC90, ML95].

For other methods based on very similar ideas, we can refer to [SG72, WW88, Mal79, Har90].

6.3 The circular memory technique

The circular memory technique is the simplest and most ancient time and frequency scale modification technique [Ben88]. It is also a method operating in the time domain.

6.3.1 The analog origin

This technique is derived from an analog system proposed in the 1950s [FEJ54]. It consists in using a tape recorder equipped with a rotating head. The closed loop tape wraps around half of the cylinder (as for the VCR and the

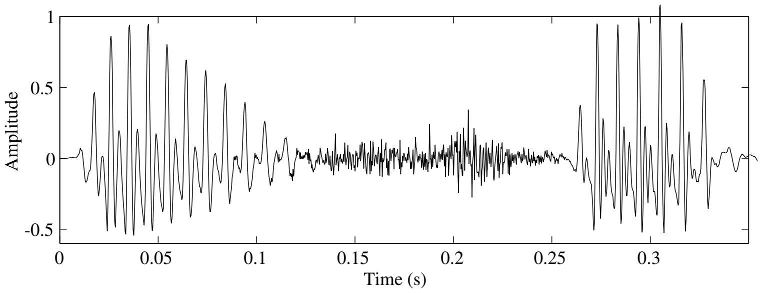


Figure 1.5: Original: "il s'est".

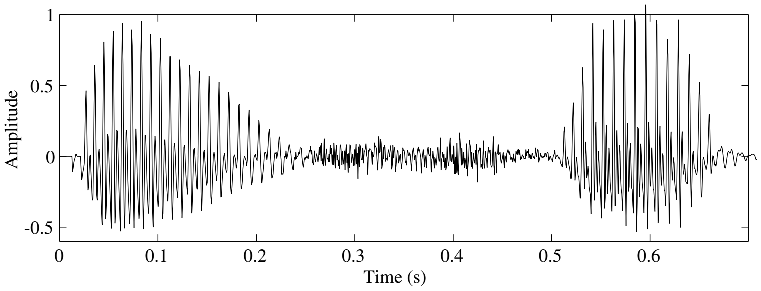


Figure 1.6: Signal stretched by factor 2.

DAT) and scrolls at constant speed. The cylinder is provided with two diametrically opposed reader heads whose signals are mixed with an identical gain. It is possible to control the direction of rotation and the speed of the cylinder.

When the cylinder is motionless, the tape scrolls in an identical manner in front of the recording head and in front of one of the reader heads. The signal read is therefore identical to the signal recorded (up to recording errors).

When the cylinder rotates in the opposite direction of the scrolling of the tape, the relative speed V_r of the scrolling tape with respect to the reader head is faster than its absolute scroll speed V_a. During the period of contact between the reader head and the tape, the signal is thus read faster than it has been recorded, which corresponds to a dilation of the frequency axis. The presence of two heads ensures the continuity thanks to a natural cross-fade (when a head leaves the band, the other approaches it, so that the total signal does not decrease). Note that some portions of the signal can be read *two or more times*, depending on the speed of rotation of the head. It is this rereading that keeps the signal duration.

Conversely, when the cylinder rotates in the scrolling direction of the band, the frequency content of the signal is contracted towards the origin since the tape is read at a slower speed than it is recorded. In this case, some portions of the signal may not be read at all.

The ratio of frequency homothety is expressed as:

$$\alpha = \frac{V_r}{V_a} = \frac{V_a + R\,\Omega_{cylinder}}{V_a}$$

where V_a is the tape scrolling speed in front of the recording head, V_r is the relative speed of the tape with respect to the reader head, $\Omega_{cylinder}$ is the speed of rotation of the cylinder in radians s^{-1}, and R is the radius of the cylinder. In all cases, the regular alternation of the two heads induces a periodic "noise" of frequency $\Omega_{cylinder}/\pi$.

Modifications in the signal time scale are obtained for instance by recording the signal a first time on the tape, then by replaying it with a tape scrolling speed multiplied by factor α. In the absence of rotation of the reader head,

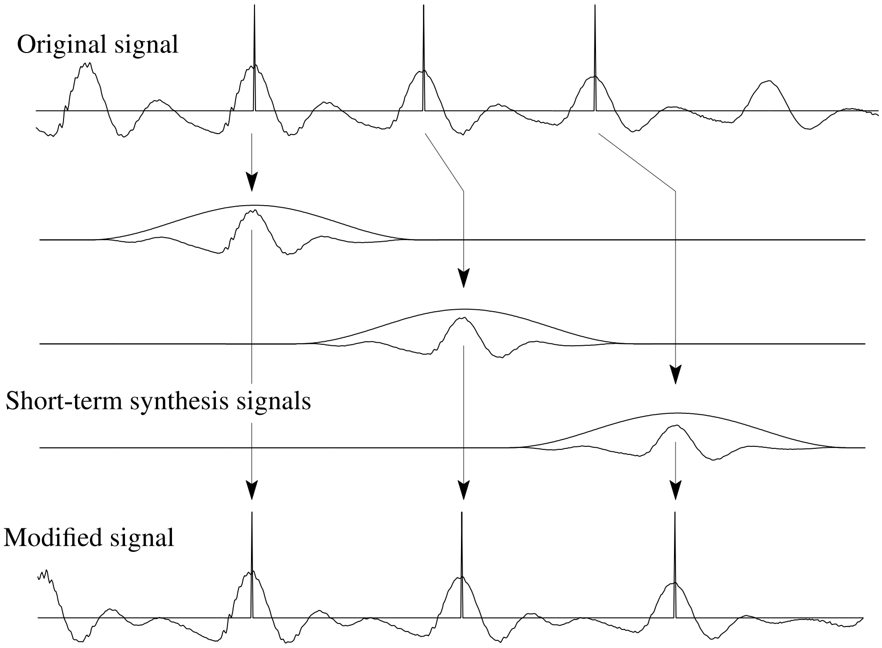


Figure 1.7: Modification of the pitch of the signal by the TD-PSOLA method. At the top, the original signal, in the middle three short-term signals generated from the three first analysis marks. At the bottom, the modified signal. The spacing of the synthesis marks is not identical to that of the analysis marks.

the signal pitch is of course multiplied by factor α, which we try to avoid. We therefore compensate for the pitch change by a proper rotation of the reader head.

6.3.2 Digital implementation

Most commercially available pitch modifiers are based on a digital realization of the system described above. The magnetic tape is replaced by a circular memory in which the input signal samples are placed. This circular memory is read by two diametrically opposite pointers.

For each sample written in the memory (every ΔT seconds), we advance the reading pointers by $\alpha \Delta T$ seconds, where α is the rate of change, and then we read a sample in memory. In general (for non-integer values of α), we find ourselves between two samples, and as in the case of the "flanger", it is necessary to calculate the signal value at this time. Here too, a simple linear interpolation is suitable.
Thus, the signal is read with a sampling frequency different from the one it was recorded with, which causes a

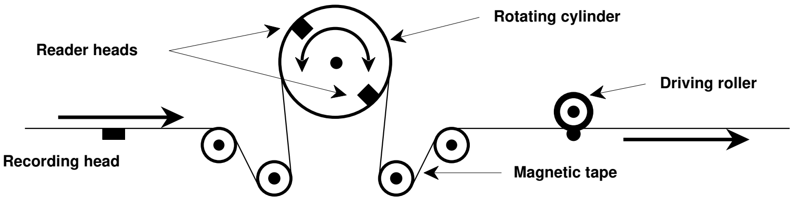


Figure 1.8: The circular memory technique

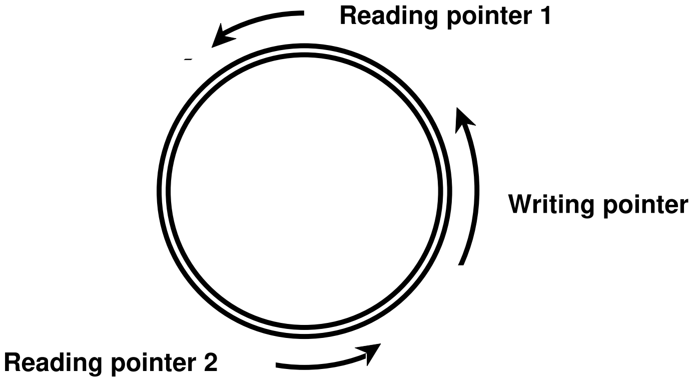


Figure 1.9: Digital implementation

modification of the frequency scale of rate α. A problem arises when the reading pointer catches up (when $\alpha > 1$) or is caught (when $\alpha < 1$) by the writing pointer. As in the analog equivalent, continuity is ensured by a mixture of the two pointers at the time when the encounter occurs ("cross-fade"): the sample read by the current reader pointer (e.g. pointer 1) is lowered while the one read by the other reader pointer (pointer 2) is increased. Finally, the second pointer becomes the current pointer, and keeps its maximum weighting, until the writing pointer gets close to it.

Implemented in this way, the pitch shifter has a behavior substantially equivalent to its analog counterpart (except that it is more easily configurable). Its implementation in real-time is not a particular problem, since it requires very few calculations.
Unfortunately, it produces artificial noise which comes from the periodic mixing of the two reader pointers. To try to improve the obtained quality, we seek to better combine the signals read by the two reader pointers, somewhat similarly to what is done in the synchronous methods. We can for example use the signal autocorrelation function to determine the most suitable location for the "cross-fading" [Dat87, Lar93a].

6.3.3 Modification of the duration by the technique of circular memory

Like its analog counterpart, the circular buffer technique can also be used for temporal scale modification: if you have a pitch shifter with circular memory, and want to perform a "time scaling" of parameter α, just change the signal sampling frequency by a rate α, then process it with the pitch shifter. So, in order to slow down the signal twice, it suffices to oversample it twice. If we listen to the signal obtained at the original frequency, it will be twice as much long, but also at a lower octave. So just listen to it at the original frequency by inserting a pitch shifter at rate $\alpha = 2$.

We quickly realize that it is easier to do both operations jointly: the technique then consists in repeating or periodically eliminating signal portions so as to increase (or decrease) the duration. Viewed from this angle, this technique (which is called "splicing method") approximates a TD-PSOLA technique in which we would not know the value of the fundamental frequency. The artifacts inherent in this method, which come from the breaks in the periodicity of the signal during the repetitions or eliminations, can be considerably reduced by the use of methods based on the autocorrelation of the signal in order to optimize the length and location of signal portions to be duplicated or destroyed [Lee72, Dat87, Lar93a, RW85, SK92, VR93].

Chapter 2

Audio source separation

1 Introduction

Source separation is the art of estimating *source* signals, which are assumed statistically independent, from the observation of one or several *mixtures* of these signals. It is useful in many audio signal processing tasks, including *denoising* applications:

- separation of the instruments in polyphonic music;
- karaoke: remove the singer voice in music recordings;
- cocktail party problem: isolate the voice of the person you are speaking to from many other voices;
- suppression of vuvuzela in TV broadcasting of football matches during the 2010 FIFA world cup.

Besides, the separated audio tracks can be used for remixing purposes, possibly including transformations (e.g. pitch shifting, time scaling, etc.) or re-spatialization of the separated audio sources.

1.1 Typology of the mixture models

Formally, the observed data is made of M mixture signals $x_m(t)$, concatenated in a vector $\boldsymbol{x}(t)$. The unknowns are the K (possibly different from M) source signals $s_k(t)$, concatenated in a vector $\boldsymbol{s}(t)$.

The mixture is modeled as a function $\mathcal{A}$ which transforms the source signals $\boldsymbol{s}(t)$ into the mixture signals $\boldsymbol{x}(t)$. Generally, some simplifying assumptions are introduced regarding the mixture model [VVG18, chap. 1]:

- *Stationarity*: function $\mathcal{A}$ is translation invariant.
- *Linearity*: function $\mathcal{A}$ is a linear map.
- *Memory*:
  - Transformations that are both stationary and linear can be modeled with convolution products in the time domain (linear filtering).
  - The *memory* of such transformations corresponds to the length of the impulse response.
  - If there is no memory (i.e. the length is zero), the mixture is called *instantaneous* and $\mathcal{A}$ is characterized by a *mixing matrix* $\boldsymbol{A}$ (of dimension $M \times K$): $\boldsymbol{x}(t) = \boldsymbol{A}\,\boldsymbol{s}(t)$. Instantaneous mixture models are suitable e.g. for some biomedical applications (electroencephalography (EEG) or magnetoencephalography (MEG)), but generally not for audio applications, because of reverberation.

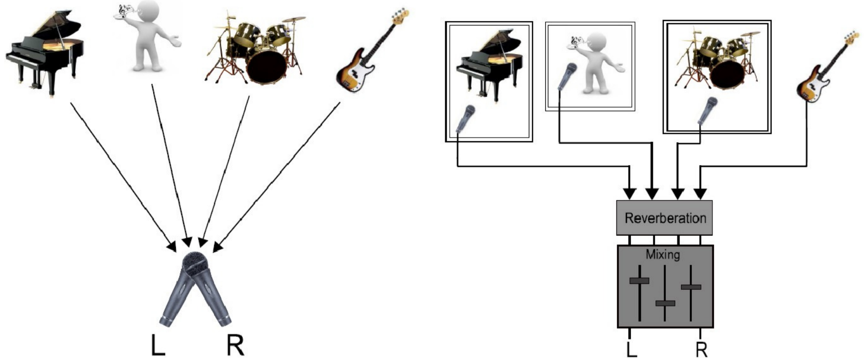
Depending on the respective values of M and K, the mixture may or may not be invertible:

- If $M = K$, the mixture is called *determined*: it is generally invertible.
- If $M > K$, the mixture is called *over-determined*: a unique solution can be found in the least squares sense.
- If $M < K$, the mixture is called *under-determined*: there are infinitely many solutions. Without additional information about the mixture or the source signals, it is impossible to retrieve the original sources from the mixture signals.

1.2 Instantaneous linear mixtures

Examples of instantaneous linear mixtures are given in Figure 2.1:

- In a real audio environment, an approximately instantaneous linear mixture can be obtained with the X-Y stereo recording technique, by putting two directional microphones at the same place, typically oriented at 90 degrees or more from each other (Figure 2.1-(a)). However the audio mixture obtained in this way is never perfectly instantaneous.
- Otherwise, truly instantaneous linear mixtures can of course be created artificially by using a mixing deck or a computer (Figure 2.1-(b)).



(a) XY Stereo configuration

(b) Direct injection to the mixer

Figure 2.1: Instantaneous linear mixtures

1.3 Anechoic linear mixtures

Anechoic linear mixtures are a particular case of convolutive mixtures that can be recorded in an *anechoic chamber*: because the sound reflections on the room walls are greatly attenuated, every impulse response is formed only of a single pulse, characterized by its delay and its magnitude, which corresponds to the direct propagation path from every source to every microphone (Figure 2.2).

1.4 Convolutive mixtures

In the general case, audio mixtures are convolutive: in a room, the sound waves are reflected on the walls, so the impulse response is formed of infinitely many pulses, which correspond to the direct propagation path and the various reflections, whose density grows quadratically with time. This phenomenon is called *reverberation* (Figure 2.3-(a)). Convolutive mixtures can also be created artificially, e.g. to simulate a 3-D stereo sound sensation for the listener using headphones (*binaural* mixture, illustrated in Figure 2.3-(b)).

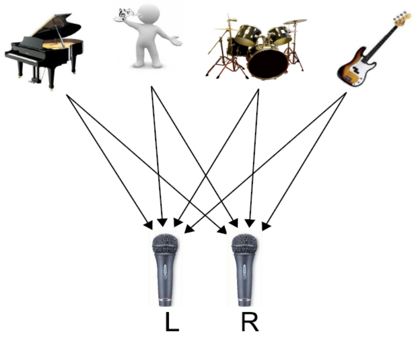
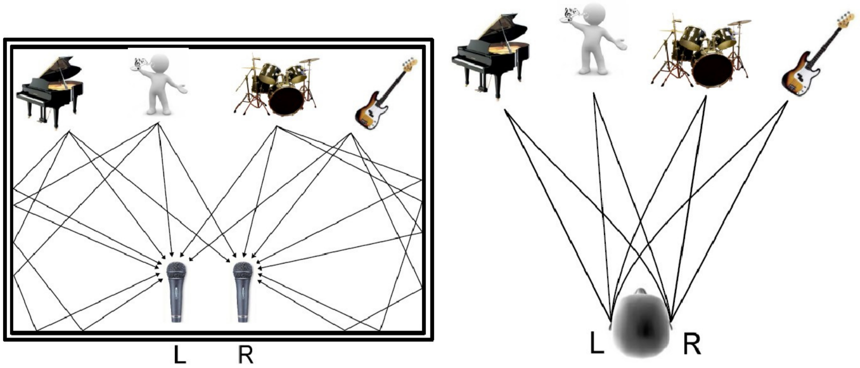


Figure 2.2: Anechoic linear mixtures



(a) Convolutive mixture

(b) Binaural mixture

Figure 2.3: Convolutive mixtures

2 Mathematical reminders

Because most source separation techniques involve probabilistic models, we first start with some mathematical reminders from probability theory and statistical signal processing.

2.1 Real random vectors

Let $\boldsymbol{x} \in \mathbb{R}^M$ denote a real random vector. In the rest of this document, we will use the following notation: $\phi[\boldsymbol{x}]$ (with square brackets) denotes a function of the distribution of the random vector $\boldsymbol{x}$, whereas a random variable defined as a function of $\boldsymbol{x}$ would be denoted $\psi(\boldsymbol{x})$ (with parentheses). In particular, we will consider:

- the mean vector: $\boldsymbol{\mu}_x = \mathbb{E}[\boldsymbol{x}] \in \mathbb{R}^M$ (where $\mathbb{E}$ denotes the *mathematical expectation*, a.k.a the *expected value*);
- the covariance matrix: $\boldsymbol{\Sigma}_{xx} = \mathbb{E}[(\boldsymbol{x} - \boldsymbol{\mu}_x)(\boldsymbol{x} - \boldsymbol{\mu}_x)^\top] \in \mathbb{R}^{M \times M}$, which is always symmetric (i.e. $\boldsymbol{\Sigma}_{xx}^\top = \boldsymbol{\Sigma}_{xx}$ where $.^\top$ denotes the transpose of a matrix) and positive semi-definite (i.e. $\forall \boldsymbol{v} \in \mathbb{R}^M$, $\boldsymbol{v}^\top \boldsymbol{\Sigma}_{xx} \boldsymbol{v} \geq 0$);
- the characteristic function: $\phi_x(\boldsymbol{f}) = \mathbb{E}[e^{-2\imath\pi \boldsymbol{f}^\top \boldsymbol{x}}] \in L^\infty(\mathbb{R}^M)$ (where $L^\infty(\mathbb{R}^M)$ denotes the Lebesgue space of essentially bounded functions on $\mathbb{R}^M$);
- when the inverse Fourier transform of ϕ_x is a measurable function on $\mathbb{R}^M$, $p(\boldsymbol{x}) = \int_{\mathbb{R}} \phi_x(\boldsymbol{f}) e^{+2\imath\pi \boldsymbol{f}^\top \boldsymbol{x}} d\boldsymbol{f}$ is called the *Probability Density Function* (PDF) of the random vector $\boldsymbol{x}$.

Some of the oldest source separation methods are based on the notion of *cumulants*. The cumulants of the random vector $\boldsymbol{x}$ will be denoted $\kappa^n_{k_1 \dots k_n}[\boldsymbol{x}] \in \mathbb{R}$ for all orders $n \in \mathbb{N}$ and entries $k_i \in \{1 \dots M\}$, and they are defined as the coefficients of the Taylor expansion of the *cumulant generating function*, which is the natural logarithm of the characteristic function: when ϕ_x is an analytic function, we can write

$$\ln(\phi_x(\boldsymbol{f})) = \sum_{n=1}^{+\infty} \frac{(-2\imath\pi)^n}{n!} \sum_{k_1=1}^{M} \sum_{k_n=1}^{M} \kappa^n_{k_1 \dots k_n}[\boldsymbol{x}] f_{k_1} \dots f_{k_n}.$$

The cumulants satisfy the following properties:

- $\forall n \in \mathbb{N}^*$, $\kappa^n[\boldsymbol{x}]$ is an n-th order tensor of coefficients $\kappa^n_{k_1 \dots k_n}[\boldsymbol{x}]$;
- $\kappa^1[\boldsymbol{x}]$ is the mean vector $\boldsymbol{\mu}_x$ and $\kappa^2[\boldsymbol{x}]$ is the covariance matrix $\boldsymbol{\Sigma}_{xx}$;
- If the PDF $p(\boldsymbol{x})$ is symmetric ($p(-\boldsymbol{x}) = p(\boldsymbol{x})$), then $\kappa^n[\boldsymbol{x}] = 0$ for any odd value n;
- The ratio between the fourth order cumulant $\kappa^4_{k,k,k,k}[\boldsymbol{x}]$ and the squared variance $(\kappa^2_{k,k}[\boldsymbol{x}])^2$ plays a special role in independent component analysis (*cf.* Section 3.2). It is called the *excess kurtosis*.

2.2 Real Gaussian random vectors

Among all probability distributions with well-defined cumulants of all orders, the Gaussian distribution is the one such that all cumulants of order $n > 2$ are zero. The characteristic function of a Gaussian random vector $\boldsymbol{x} \in \mathbb{R}^M$ can thus be expressed as

$$\phi_x(\boldsymbol{f}) = \exp\left(-2\imath\pi \boldsymbol{f}^\top \boldsymbol{\mu}_x - 2\pi^2 \boldsymbol{f}^\top \boldsymbol{\Sigma}_{xx} \boldsymbol{f}\right).$$

When the covariance matrix $\boldsymbol{\Sigma}_{xx}$ is invertible, then the PDF is defined as:

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{M}{2}} \det(\boldsymbol{\Sigma}_{xx})^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_x)^\top \boldsymbol{\Sigma}_{xx}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_x)\right).$$

2.3 WSS vector processes

A discrete vector process is a sequence of random vectors $\boldsymbol{x}(t) \in \mathbb{R}^M$ indexed by time $t \in \mathbb{Z}$. A second order vector process is a discrete vector process with well-defined second order moments. Finally, a *Wide Sense Stationary* (WSS) vector process $\boldsymbol{x}(t)$ is a second order vector process whose cumulants of orders 1 and 2 are invariant under any translation of time:

- $\mathbb{E}[\boldsymbol{x}(t)] = \boldsymbol{\mu}_x \ \forall t \in \mathbb{Z}$ where $\boldsymbol{\mu}_x \in \mathbb{R}^M$ is the *mean vector* of the vector process $\boldsymbol{x}(t)$;
- $\forall t \in \mathbb{Z}$, $\mathbb{E}\left[(\boldsymbol{x}(t+\tau) - \boldsymbol{\mu}_x)(\boldsymbol{x}(t) - \boldsymbol{\mu}_x)^\top\right] = \boldsymbol{R}_{xx}(\tau)$, where $\forall \tau \in \mathbb{Z}$ $\boldsymbol{R}_{xx}(\tau) \in \mathbb{R}^{M \times M}$ defines the *autocovariance function* of the vector process $\boldsymbol{x}(t)$. When $\tau = 0$, $\boldsymbol{R}_{xx}(0) = \boldsymbol{\Sigma}_{xx}$ is the covariance matrix of the random vector $\boldsymbol{x}(t)$ $\forall t \in \mathbb{Z}$, and as such it is symmetric and positive semi-definite.

Finally, given two jointly WSS vector processes $\boldsymbol{x}(t) \in \mathbb{R}^M$ and $\boldsymbol{y}(t) \in \mathbb{R}^N$ of mean zero, we define their *intercovariance function* $\boldsymbol{R}_{xy}(\tau) \in \mathbb{R}^{M \times N}$:

$$\forall \tau \in \mathbb{Z}, \ \boldsymbol{R}_{xy}(\tau) = \mathbb{E}\left[\boldsymbol{x}(t+\tau)\boldsymbol{y}(t)^\top\right].$$

When the *Discrete Time Fourier Transform* (DTFT) of the autocovariance function $\boldsymbol{R}_{xx}(\tau)$ of a WSS vector process $\boldsymbol{x}(t)$ is a measurable function $\boldsymbol{S}_{xx}(\nu) \in \mathbb{C}^{M \times M}$, this function is called the *Power Spectral Density* (PSD) of $\boldsymbol{x}(t)$:

$$\forall \nu \in \mathbb{R}, \ \boldsymbol{S}_{xx}(\nu) = \sum_{\tau \in \mathbb{Z}} \boldsymbol{R}_{xx}(\tau) e^{-2\imath\pi\nu\tau}.$$

The PSD is always periodic of period 1, and $\forall \nu \in \mathbb{R}$, matrix $\boldsymbol{S}_{xx}(\nu)$ is always Hermitian symmetric (i.e. $\boldsymbol{S}_{xx}(\nu)^H = \boldsymbol{S}_{xx}(\nu)$ where $.^H$ denotes the conjugate transpose of a matrix) and positive semi-definite (i.e. $\forall \boldsymbol{v} \in \mathbb{C}^M$, $\boldsymbol{v}^H \boldsymbol{S}_{xx}(\nu) \boldsymbol{v} \geq 0$).

2.4 Information theory

Information theory is a fundamental tool in *blind source separation* (*cf.* Section 3.1), because it makes it possible to measure the amount of information shared between several random variables.

We first consider the notion of *entropy*, which measures the degree of uncertainty in a probability distribution. For a discrete random variable x with probability distribution p, the *Shannon entropy* is defined as $\mathbb{H}[x] = -\mathbb{E}[\ln(p(x))]$. It is always a non-negative real number. The higher this number, the more "uncertain" the outcome of x is. For a continuous random vector $\boldsymbol{x} \in \mathbb{R}^M$ with PDF $p(\boldsymbol{x})$, the *differential entropy* is defined in the same way: $\mathbb{H}[\boldsymbol{x}] = -\mathbb{E}[\ln(p(\boldsymbol{x}))]$. However the differential entropy $\mathbb{H}[\boldsymbol{x}]$ is not necessarily non-negative.

For continuous random vectors $\boldsymbol{x} \in \mathbb{R}^M$, the *Kullback-Leibler divergence* measures the degree of dissimilarity between two probability distributions characterized by their PDFs p and q:

$$D_{KL}(p\|q) = \int_{\boldsymbol{x}\in\mathbb{R}^M} p(\boldsymbol{x}) \ln\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right) d\boldsymbol{x}.$$

As a *divergence*, it is always nonnegative, and $D_{KL}(p\|q) = 0$ if and only if $p = q$. However, the Kullback-Leibler divergence is not a *distance*, because it is not symmetric (in general $D_{KL}(p\|q) \neq D_{KL}(q\|p)$), and it does not satisfy the triangle inequality.

Finally, the *mutual information* measures the mutual dependence between several random variables. For instance if $\boldsymbol{x} \in \mathbb{R}^M$ is a continuous random vector, then the mutual information between the entries of $\boldsymbol{x}$ is defined as:

$$\mathbb{I}[\boldsymbol{x}] = \mathbb{E}\left[\ln\left(\frac{p(\boldsymbol{x})}{p(x_1)\dots p(x_M)}\right)\right] = D_{KL}(p(\boldsymbol{x})\|p(x_1)\dots p(x_M)).$$

Since D_{KL} is a divergence, $\mathbb{I}[\boldsymbol{x}]$ is always nonnegative, and $\mathbb{I}[\boldsymbol{x}] = 0$ if and only if $p(\boldsymbol{x}) = p(x_1)\dots p(x_M)$, i.e. if and only if the random variables $x_1 \dots x_M$ are mutually independent. The mutual information is related to the differential entropy through the equality

$$\mathbb{I}[\boldsymbol{x}] = \left(\sum_{m=1}^{M} \mathbb{H}[x_m]\right) - \mathbb{H}[\boldsymbol{x}]. \tag{2.1}$$

In *independent component analysis*, the mutual information is an objective function to be minimized, in order to make several random variables as independent as possible (*cf.* Section 3.2.2). Equation (2.1) shows that minimizing $\mathbb{I}[\boldsymbol{x}]$ is equivalent to minimizing the sum of the individual entropies $\mathbb{H}[x_m]$ when the joint entropy $\mathbb{H}[\boldsymbol{x}]$ is fixed.

3 Linear instantaneous mixtures

Even though we have seen in Section 1.2 page 24 that the linear instantaneous mixture model cannot accurately represent real acoustic mixtures, the oldest separation techniques which paved the way for modern audio source separation methods are based on this model. We thus first address this over-simplified mixture model, which will permit us to introduce several useful concepts and methods, that will then be extended to the more realistic convolutive mixture model in Section 4.

3.1 Blind source separation (BSS) model

Blind Source Separation (BSS) techniques [Car98] assume that we know very little about the source signals: they are only assumed to be statistically independent. This is the case for instance in many denoising applications, where the signal of interest (e.g. speech) is independent from the source of noise (e.g. background environmental noise). This hypothesis is the funding principle of most multichannel source separation methods.

Actually, BSS methods also rely on a generative source model, but this model is chosen as little informative as possible: the samples of each source signal are assumed *Independent and Identically Distributed* (IID). The IID source model thus ignores any temporal dynamics (i.e. power variations over time), or spectral dynamics (i.e. temporal correlations), that might be present in the source signals:

Definition 1 (IID source model)**.** *We consider K independent source signals $s_k(t) \in \mathbb{R}$ with $t \in \mathbb{Z}$. For all $k \in \{1 \dots K\}$, s_k is modeled as an IID random process: the samples $s_k(t)$ are independent random variables, of same probability distribution p_k (which depends on source k).*

The possibility of performing source separation in such a blind way may seem to be an incredible feat. Actually, the trick is that, contrary to the source model, the mixture model is *very* constraining: at first we will only consider linear instantaneous mixtures, characterized by a mixing matrix $\boldsymbol{A}$:

Definition 2 (Linear instantaneous mixture model)**.** *We consider K source signals $s_k(t) \in \mathbb{R}$ with $t \in \mathbb{Z}$. Then the samples of the M mixture signals $x_m(t) \in \mathbb{R}$ for $m \in \{1 \dots M\}$ are defined as the entries of the M-dimensional vector*

$$\boldsymbol{x}(t) = \boldsymbol{A}\boldsymbol{s}(t), \tag{2.2}$$

where $\boldsymbol{A} \in \mathbb{R}^{M \times K}$ is called the mixing matrix*, and $\boldsymbol{s}(t)$ is the K-dimensional vector of entries $s_k(t)$ for $k \in \{1 \dots K\}$.*

Given the source model in Definition 1 and the mixture model in Definition 2, the purpose of BSS is to estimate the source signals $s_k(t)$ given the observed mixture signals $x_m(t)$, *without knowing* the mixing matrix $\boldsymbol{A}$. When the mixture is *determined* ($M = K$) and matrix $\boldsymbol{A}$ is invertible, we will show that this is generally feasible.

3.1.1 Identifiability

Suppose that the mixture is determined ($M = K$) . Before investigating how source separation can be performed, we first need to study the *identifiability* of the linear instantaneous BSS model: is it really possible to retrieve both the source signals $s_k(t)$ and the mixing matrix $\boldsymbol{A}$ from only the observed mixture signals $x_m(t)$?

Clearly, if $\boldsymbol{P}$ is a permutation matrix (i.e. it has a unique 1 entry in each row and each column, all other entries being 0), then matrix $\tilde{\boldsymbol{A}} = \boldsymbol{A}\,\boldsymbol{P}^{-1}$ and vector $\tilde{\boldsymbol{s}}(t) = \boldsymbol{P}\boldsymbol{s}(t)$ lead to the same observations $\boldsymbol{x}(t)$, while satisfying all the properties of the linear instantaneous BSS model in Definitions 1 and 2. So the source signals can only be retrieved up to a permutation: at best we can retrieve the source signals, but we cannot *identify* them.

In the same way, if $\boldsymbol{D}$ is an invertible diagonal matrix, then matrix $\tilde{\boldsymbol{A}} = \boldsymbol{A}\,\boldsymbol{D}^{-1}$ and vector $\tilde{\boldsymbol{s}}(t) = \boldsymbol{D}\boldsymbol{s}(t)$ lead to the same observations $\boldsymbol{x}(t)$, while satisfying all the properties of the linear instantaneous BSS model in Definitions 1 and 2. Therefore the source signals can only be retrieved up to a multiplicative factor (which in most applications is not a problem: e.g. audio signals are generally scaled during playback).

So the linear instantaneous BSS model in Definitions 1 and 2 has at least *permutation* and *scale* indeterminacies. Actually, it can be proved that there is no other one. These two indeterminacies are summarized by the concept of *non-mixing matrices*:

Definition 3 (Non-mixing matrix)**.** *A matrix $\boldsymbol{C} \in \mathbb{R}^{K \times K}$ is* non-mixing *if and only if it has a unique non-zero entry in each row and each column.*

A non-mixing matrix can always be decomposed as the product of a permutation matrix and an invertible diagonal matrix.

3.1.2 Linear separation of sources

When the mixture is linear instantaneous, it may seem natural to estimate the source signals as linear instantaneous combinations of the mixture signals:

$$\boldsymbol{y}(t) = \boldsymbol{B}\boldsymbol{x}(t), \tag{2.3}$$

where the entries of vector $\boldsymbol{y}(t)$ are the source signal estimates, and $\boldsymbol{B} \in \mathbb{R}^{K \times M}$ is referred to as the *separation matrix*. Then the source separation problem amounts to finding an optimal separation matrix.

Linear source separation is generally feasible in the case of determined and over-determined mixtures:

- if $M = K$ and if matrix $\boldsymbol{A}$ is invertible, then the separation matrix $\boldsymbol{B} = \boldsymbol{A}^{-1}$ leads to $\boldsymbol{y}(t) = \boldsymbol{s}(t)$;
- more generally, if $M \geq K$ and if matrix $\boldsymbol{A}$ has full rank, then the separation matrix $\boldsymbol{B} = \boldsymbol{A}^{\dagger}$ leads to $\boldsymbol{y}(t) = \boldsymbol{s}(t)$, where $.^{\dagger}$ denotes the matrix the pseudo-inverse: $\boldsymbol{A}^{\dagger} = (\boldsymbol{A}^{\top}\boldsymbol{A})^{-1}\boldsymbol{A}^{\top}$, which is such that $\boldsymbol{A}^{\dagger}\boldsymbol{A} = \boldsymbol{I}_K$.

However, in the under-determined case ($M < K$), linear source separation is generally not feasible (*cf.* Section 5).

3.2 Independent component analysis (ICA)

Independent Component Analysis (ICA) [CJ10] is a linear source separation technique which consists in looking for a separation matrix $\boldsymbol{B}$ that makes the signals $y_k(t)$ independent.

Since $\boldsymbol{x}(t) = \boldsymbol{A}\boldsymbol{s}(t)$ (equation (2.2)) and $\boldsymbol{y}(t) = \boldsymbol{B}\boldsymbol{x}(t)$ (equation (2.3)), we have $\boldsymbol{y}(t) = \boldsymbol{C}\boldsymbol{s}(t)$ with $\boldsymbol{C} = \boldsymbol{B}\boldsymbol{A}$. According to the identifiability analysis in Section 3.1.1, the BSS problem is solved if and only if matrix $\boldsymbol{C}$ is non-mixing (*cf.* Figure 2.4).

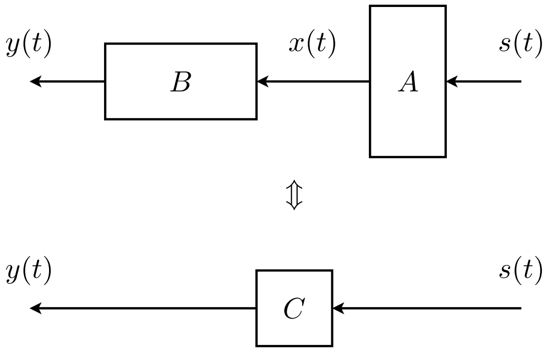


Figure 2.4: Identifiability theorem: signals $y_k(t)$ are independent if and only if matrix $\boldsymbol{C} = \boldsymbol{B}\boldsymbol{A}$ is non-mixing

The following identifiability theorem due to P. Comon [Com94] proves the feasibility of ICA under mild conditions about the source signals:

Theorem 1 (Identifiability theorem)**.** *Consider the linear instantaneous BSS model in Definitions 1 and 2 in the determined case ($M = K$). Among the K IID sources s_k, suppose that at most one is Gaussian-distributed. Let $\boldsymbol{C} \in \mathbb{R}^{K\times K}$ and $\forall t \in \mathbb{Z}$, $\boldsymbol{y}(t) = \boldsymbol{C}\boldsymbol{s}(t)$. Then the random processes $y_k(t)$ for $k \in \{1 \dots K\}$ are independent if and only if matrix $\mathbf{C}$ is non-mixing.*

Theorem 1 proves that finding a separation matrix $\boldsymbol{B}$ that makes signals $y_k(t)$ independent solves the BSS problem: the estimated signals $y_k(t)$ are equal to the source signals $s_k(t)$ up to permutation and scale indeterminacies.

Here, pay attention to the non-Gaussianity assumption in Theorem 1: in Section 3.2.1, we will show that indeed, if two (or more) sources are Gaussian, then the BSS problem cannot be solved.

3.2.1 Whitening

Independent component analysis can be performed in two steps; the first one consists in *whitening*[1], i.e. *decorrelating* the observed mixture signals. Remember that independence implies decorrelation, but decorrelation does generally not imply independence. Therefore the second step will consist in making the whitened signals independent.

To simplify the problem, we will address the case of determined mixtures $M = K$ (even though whitening could also be performed in the over-determined case $M > K$), and we will assume that matrix $\boldsymbol{A}$ is invertible and that the source signals are centered: $\mathbb{E}[\boldsymbol{s}(t)] = \boldsymbol{0}$ (which is always the case of audio signals).

To further simplify, we will focus on the *canonical BSS problem*: without loss of generality, we will assume that the random vectors $\boldsymbol{s}(t)$ are spatially white, i.e. their covariance matrix is $\boldsymbol{\Sigma}_{ss} = \mathbb{E}[\boldsymbol{s}(t)\boldsymbol{s}(t)^\top] = \boldsymbol{I}_K$. Indeed, since the source signals are independent, we already know that matrix $\boldsymbol{\Sigma}_{ss}$ is diagonal; since in addition the source

[1] Whitening is performed in the spatial domain, i.e. over channels, not in the time domain, i.e. over time samples.

signals can only be retrieved up to a multiplicative factor, we can also assume without loss of generality that the diagonal entries of matrix $\boldsymbol{\Sigma}_{ss}$ are 1.

Since $\boldsymbol{x}(t) = \boldsymbol{A}\boldsymbol{s}(t)$ (equation (2.2)), the covariance matrix of the mixture vectors $\boldsymbol{x}(t)$ is $\boldsymbol{\Sigma}_{xx} = \boldsymbol{A}\boldsymbol{\Sigma}_{ss}\boldsymbol{A}^\top = \boldsymbol{A}\,\boldsymbol{A}^\top$: we say that $\boldsymbol{A}$ is a *matrix square root* of $\boldsymbol{\Sigma}_{xx}$. This property is interesting because $\boldsymbol{\Sigma}_{xx}$ can be estimated from the observed data, and it carries information about the mixing matrix $\boldsymbol{A}$. Unfortunately, we will see that this property is not sufficient to fully characterize $\boldsymbol{A}$. Nevertheless, it allows us to make a first step towards the estimation of $\boldsymbol{A}$. For the moment, just note that since matrix $\boldsymbol{A}$ is invertible, matrix $\boldsymbol{\Sigma}_{xx}$ is also invertible, thus positive definite.

The whitening of the mixture signals can then be performed as follows [CS93]:

- Since matrix $\boldsymbol{\Sigma}_{xx}$ is positive definite, the spectral theorem in matrix theory shows us that it is diagonalizable in an orthonormal basis: there is an orthonormal matrix $\boldsymbol{Q} \in \mathbb{R}^{K\times K}$ (i.e. such that $\boldsymbol{Q}^{-1} = \boldsymbol{Q}^\top$), and a diagonal matrix $\boldsymbol{\Lambda} \in \mathbb{R}^{K\times K}$ with positive diagonal entries, such that
$$\boldsymbol{\Sigma}_{xx} = \boldsymbol{Q}\boldsymbol{\Lambda}^2\boldsymbol{Q}^\top. \tag{2.4}$$

- Then let $\boldsymbol{S} = \boldsymbol{Q}\,\boldsymbol{\Lambda} \in \mathbb{R}^{K\times K}$; matrix $\boldsymbol{S}$ is also a matrix square root of $\boldsymbol{\Sigma}_{xx}$, since $\boldsymbol{S}\,\boldsymbol{S}^\top = \boldsymbol{Q}\boldsymbol{\Lambda}^2\boldsymbol{Q}^\top = \boldsymbol{\Sigma}_{xx}$.

- Finally, let
$$\boldsymbol{W} = \boldsymbol{S}^{-1} \tag{2.5}$$
and
$$\forall t \in \mathbb{Z},\; \boldsymbol{z}(t) = \boldsymbol{W}\boldsymbol{x}(t). \tag{2.6}$$

Then the random vector process $\boldsymbol{z}(t)$ is spatially white, in the sense that on the one hand it is centered: $\mathbb{E}[\boldsymbol{z}(t)] = \boldsymbol{0}$ (since $\boldsymbol{z}(t) = \boldsymbol{W}\boldsymbol{A}\boldsymbol{s}(t)$ and $\mathbb{E}[\boldsymbol{s}(t)] = \boldsymbol{0}$), and on the other hand its covariance matrix is $\boldsymbol{\Sigma}_{zz} = \boldsymbol{W}\boldsymbol{\Sigma}_{xx}\boldsymbol{W}^\top = \boldsymbol{W}\boldsymbol{S}\boldsymbol{S}^\top\boldsymbol{W}^\top = \boldsymbol{I}_K$). Matrix $\boldsymbol{W}$ will thus be referred to as the *whitening matrix* and $\boldsymbol{z}(t)$ is the *whitened data*.

Then let us define matrix $\boldsymbol{U} = \boldsymbol{W}\boldsymbol{A}$. We have $\boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{W}\boldsymbol{A}\boldsymbol{A}^\top\boldsymbol{W}^\top = \boldsymbol{W}\boldsymbol{\Sigma}_{xx}\boldsymbol{W}^\top = \boldsymbol{I}_K$, therefore $\boldsymbol{U}$ is an orthonormal matrix. In particular, $|\det(\boldsymbol{U})| = 1$: if $\det(\boldsymbol{U}) = 1$, $\boldsymbol{U}$ is a *rotation matrix*, otherwise if $\det(\boldsymbol{U}) = -1$, $\boldsymbol{U}$ is a *reflection matrix*. However, remember that the source signals can only be retrieved up to a multiplicative factor, which might as well be negative. Therefore, by changing the sign of the k-th source signal $s_k(t)$, the product $\boldsymbol{A}\boldsymbol{s}(t)$ is left unchanged by changing the sign of the k-th column of matrix $\boldsymbol{A}$, which changes the sign of $\det(\boldsymbol{U})$. Therefore, without loss of generality, we can assume that $\boldsymbol{U}$ *is a rotation matrix* ($\det(\boldsymbol{U}) = 1$).

Finally, let
$$\boldsymbol{y}(t) = \boldsymbol{U}^\top\boldsymbol{z}(t). \tag{2.7}$$

Then $\boldsymbol{y}(t) = \boldsymbol{U}^\top\boldsymbol{W}\boldsymbol{x}(t) = (\boldsymbol{W}\boldsymbol{A})^{-1}\boldsymbol{W}(\boldsymbol{A}\boldsymbol{s}(t)) = \boldsymbol{s}(t)$. Therefore matrix $\boldsymbol{B} = \boldsymbol{U}^\top\boldsymbol{W}$ is a separation matrix. In practice of course, matrix $\boldsymbol{A}$ is unknown, thus so is matrix $\boldsymbol{U}$. But it remains that ICA can be performed in two steps (*cf.* Figure 2.5):

Step 1 : compute the whitening matrix $\boldsymbol{W}$ and the whitened data $\boldsymbol{z}(t)$ from an estimate of the covariance matrix $\boldsymbol{\Sigma}_{xx}$;

Step 2 : look for a rotation matrix $\boldsymbol{U}$ such that the entries of vector $\boldsymbol{y}(t) = \boldsymbol{U}^\top\boldsymbol{z}(t)$ are independent.

To summarize, the whiteness property (based on second order cumulants) determines matrix $\boldsymbol{W}$ and leaves the rotation matrix $\boldsymbol{U}$ unknown.

Note that in the Gaussian case, decorrelation implies independence. Therefore if the source signals are Gaussian-distributed, then the whitened signals $z_k(t)$ are independent, and $\boldsymbol{U}$ cannot be determined. This explains the assumption made in Theorem 1 page 29 that at most one source can be Gaussian-distributed (if they are two of them, they cannot be separated).

Therefore if we want to determine rotation $\boldsymbol{U}$, we will need to explicitly exploit the non-Gaussianity of the source signals. To do so, we will characterize the independence property by using cumulants of order greater than 2 [CS93].

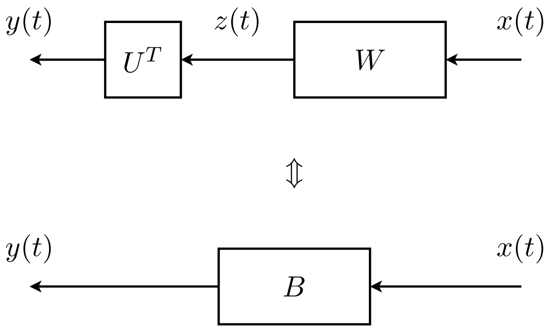


Figure 2.5: Pre-whitening for independent component analysis: $\boldsymbol{B} = \boldsymbol{U}^\top \boldsymbol{W}$ where $\boldsymbol{U}$ is a rotation matrix

3.2.2 Contrast functions

In Section 2.4, we have introduced the concept of *mutual information*: the mutual information between several random variables is always non-negative, and it is zero if and only if these random variables are independent. Therefore the mutual information between the entries of vector $\boldsymbol{y}(t)$ can be used as an objective function to be minimized in order to perform ICA.

More generally, the concept of *contrast functions* has been introduced in order to formulate ICA as an optimization problem, the mutual information being only one example of such functions. Formally, still in the determined case ($M = K$), Theorem 1 leads to the following definition of *contrast functions* [Car98]:

Definition 4 (Contrast function)**.** *For all $k \in \{1 \dots K\}$, we consider source signals $s_k(t)$ as defined in Definition 1. Then a function $\phi : \mathbb{R}^K \to \mathbb{R}$ is a* contrast function *when $\phi[\boldsymbol{C}\boldsymbol{s}(t)] \geq \phi[\boldsymbol{s}(t)]$ for any matrix $\boldsymbol{C} \in \mathbb{R}^{K\times K}$, and $\phi[\boldsymbol{C}\boldsymbol{s}(t)] = \phi[\boldsymbol{s}(t)]$ if and only if matrix $\boldsymbol{C}$ is non-mixing.*

Following Definition 4 and considering linear instantaneous mixtures $\boldsymbol{x}(t) = \boldsymbol{A}\boldsymbol{s}(t)$ (equation (2.2)) as in Definition 2 page 28, linear source separation $\boldsymbol{y}(t) = \boldsymbol{B}\boldsymbol{x}(t)$ (equation (2.3) page 28), and matrix $\boldsymbol{C} = \boldsymbol{B}\boldsymbol{A}$, the BSS problem is solved by minimizing the contrast function $\phi[\boldsymbol{y}(t)]$ with respect to (w.r.t.) the separation matrix $\boldsymbol{B}$, or w.r.t. the rotation matrix $\boldsymbol{U}$ if the data has been whitened. Among all possible contrast functions, the mutual information $\phi_{IM}[\boldsymbol{y}(t)] = \mathbb{I}[\boldsymbol{y}(t)]$ is considered as the *canonical* contrast function.

If the observed data has already been whitened, it is possible to consider only *orthogonal contrast functions*, which are such that ICA is performed by minimizing an orthogonal contrast function subject to the constraint $\mathbb{E}[\boldsymbol{y}(t)\boldsymbol{y}(t)^\top] = \boldsymbol{I}_K$. For instance, by using equation (2.1) page 27, it can be shown that an orthogonal contrast function associated to the mutual information is $\phi^\circ_{IM}[\boldsymbol{y}(t)] = \sum_{k=1}^K H(y_k(t))$, because $H(\boldsymbol{y}(t))$ is left unchanged under the constraint $\mathbb{E}[\boldsymbol{y}(t)\boldsymbol{y}(t)^\top] = \boldsymbol{I}_K$.

In practice, the orthogonal contrast function ϕ°_{IM} can be expressed as a function of the cumulants of $\boldsymbol{y}(t)$, and approximated by considering only the cumulants up to order 4 [CS93]:

$$\phi^\circ_{ICA}[\boldsymbol{y}(t)] = \sum_{ijkl \neq iiii} (\kappa^4_{ijkl}[\boldsymbol{y}(t)])^2. \tag{2.8}$$

Then the minimization of ϕ°_{ICA} with respect to the rotation matrix $\boldsymbol{U}$ can e.g. be performed by factorizing $\boldsymbol{U}$ as a product of Givens rotations (i.e. 2-dimensional rotation matrices parameterized by a single angle in $[0, 2\pi]$, which are applied iteratively to every pair of entries of vector $\boldsymbol{y}(t)$), and by performing a coordinate descent, also known as (a.k.a.) Jacobi technique, w.r.t. the angles of these Givens rotations [CJ10].

Compared to equation (2.8), the independence can also be tested on a smaller subset of cumulants, as

$$\phi^{\circ}_{JADE}[\boldsymbol{y}(t)] = \sum_{ijkl \neq ijkk} (\kappa^4_{ijkl}[\boldsymbol{y}(t)])^2. \tag{2.9}$$

The motivation for using this specific subset is that $\phi^{\circ}_{JADE}[\boldsymbol{y}(t)]$ can also be seen as a joint diagonalization criterion[2]. This approach leads to the celebrated *Joint Approximate Diagonalization of Eigenmatrices* (JADE) method [CS93] summarized in Algorithm 1.

Algorithm 1 JADE method

```
Estimation of the covariance matrix $\boldsymbol{\Sigma}_{xx}$ and diagonalization: $\boldsymbol{\Sigma}_{xx} = \boldsymbol{Q}\boldsymbol{\Lambda}^2\boldsymbol{Q}^\top$ (equation (2.4))
Computation of $\boldsymbol{S} = \boldsymbol{Q}\boldsymbol{\Lambda}$ and of the whitening matrix $\boldsymbol{W} = \boldsymbol{S}^{-1}$ (equation (2.5))
Data whitening: $\boldsymbol{z}(t) = \boldsymbol{W}\boldsymbol{x}(t)$ (equation (2.6))
Estimation of $\boldsymbol{U}$ by minimizing the contrast function $\phi^{\circ}_{JADE}$ in equation (2.9)
Estimation of source signals via $\boldsymbol{y}(t) = \boldsymbol{U}^\top\boldsymbol{z}(t)$ (equation (2.7))
```

3.3 Second order methods

The JADE method is dedicated to the linear instantaneous BSS model in Definitions 1 and 2 page 28. It makes use of higher order statistics (i.e. of order greater than 2), because the identifiability of the model requires that at most one source signal be Gaussian distributed (*cf.* Theorem 1 page 29). However, it is well known that the estimating higher order statistics is more sensitive (e.g. in terms of mean square error) than estimating second order statistics. Therefore it would be interesting to develop source separation methods that only make use of second order statistics. That will require to relax the source model in Definition 1, so that the model become identifiable from its second order statistics only. In Sections 3.3.1 and 3.3.2, we will show two different ways of relaxing the source model.

3.3.1 Temporal coherence of source signals

In this section, we keep the same mixture model as in Definition 2, and we consider the source model in Definition 1 in the determined case ($M = K$), except that each source signal $s_k(t)$ is no longer assumed to be IID, but rather WSS, with a non-flat power spectral density. Therefore, as in Definition 1, the samples $s_k(t)$ for $t \in \mathbb{Z}$ can still follow the same distribution p_k, but now they are assumed to be mutually dependent (the source model is relaxed by removing the first "I" of "IID"):

Definition 5 (WSS source model)**.** *We consider K independent source signals $s_k(t) \in \mathbb{R}$ with $t \in \mathbb{Z}$, concatenated in a vector $\boldsymbol{s}(t)$. For all $k \in \{1 \dots K\}$, s_k is modeled as a centered WSS random process. So $\boldsymbol{s}(t)$ is a WSS vector process of mean $\mathbb{E}[\boldsymbol{s}(t)] = \boldsymbol{0}$ and of autocovariance function $\boldsymbol{R}_{ss}(\tau) = \mathbb{E}\left[\boldsymbol{s}(t+\tau)\boldsymbol{s}(t)^\top\right]$.*

Since the source signals are independent, $\forall \tau \in \mathbb{Z}$ the covariance matrix $\boldsymbol{R}_{ss}(\tau)$ is diagonal: $\boldsymbol{R}_{ss}(\tau) = \text{diag}(r_{s_k}(\tau))$, where $r_{s_k}(\tau) \in \mathbb{R}$ is the autocovariance function of the scalar WSS process $s_k(t)$, and diag(.) denotes a diagonal matrix formed from a vector of diagonal coefficients.

As in Section 3.2.1 page 29, we can still consider the canonical BSS problem and assume that $\boldsymbol{\Sigma}_{ss} = \boldsymbol{R}_{ss}(0) = \boldsymbol{I}_K$. Then, still as in Section 3.2.1, we can *spatially* whiten the mixture signals:

- compute a matrix square root $\boldsymbol{S}$ of $\boldsymbol{\Sigma}_{xx}$;
- compute $\boldsymbol{W} = \boldsymbol{S}^{-1}$ and the whitened data $\boldsymbol{z}(t) = \boldsymbol{W}\boldsymbol{x}(t)$ $\forall t \in \mathbb{Z}$.

Again, since $\boldsymbol{\Sigma}_{xx} = \boldsymbol{A}\,\boldsymbol{A}^\top$, matrix $\boldsymbol{U} = \boldsymbol{W}\boldsymbol{A}$ is a rotation matrix. The novelty, compared with the mathematical developments in Section 3.2.1, is that we can now consider matrices $\boldsymbol{R}_{zz}(\tau)$ $\forall \tau \in \mathbb{Z}$, since they are no longer

[2] Joint diagonalization of matrices will be addressed in Section 3.3.1.

assumed to be zero. On the contrary, we have $\forall \tau \in \mathbb{Z}$, $\boldsymbol{R}_{zz}(\tau) = \boldsymbol{W}\boldsymbol{R}_{xx}(\tau)\boldsymbol{W}^\top = \boldsymbol{W}\boldsymbol{A}\boldsymbol{R}_{ss}(\tau)\boldsymbol{A}^\top\boldsymbol{W}^\top = \boldsymbol{U}\boldsymbol{R}_{ss}(\tau)\boldsymbol{U}^\top$. This equation shows that matrices $\boldsymbol{R}_{zz}(\tau)$ are jointly diagonalized by the same set of eigenvectors, which are the columns of matrix $\boldsymbol{U}$, the eigenvalues being the diagonal entries of matrices $\boldsymbol{R}_{ss}(\tau)$.

Therefore the estimation of matrix $\boldsymbol{U}$ will no longer require the use of higher order statistics: $\boldsymbol{U}$ can be uniquely determined as the only rotation matrix (up to a non-mixing matrix) that jointly diagonalizes matrices $\boldsymbol{R}_{zz}(\tau)$ for different values of τ, as shown by the following theorem:

Theorem 2 (Unicity theorem)**.** *Let us consider a set of matrices $\boldsymbol{R}_{zz}(\tau) \in \mathbb{R}^{K\times K}$ indexed by $\tau \in \mathbb{Z}$, of the form $\boldsymbol{R}_{zz}(\tau) = \boldsymbol{U}\boldsymbol{R}_{ss}(\tau)\boldsymbol{U}^\top$, where matrix $\boldsymbol{U} \in \mathbb{R}^{K\times K}$ is orthonormal and matrices $\boldsymbol{R}_{ss}(\tau) \in \mathbb{R}^{K\times K}$ are diagonal: $\boldsymbol{R}_{ss}(\tau) = \mathrm{diag}(r_{s_k}(\tau))$. Then $\boldsymbol{U}$ is unique (up to a non-mixing matrix) if and only if $\forall k \neq l \in \{1\ldots K\}$, there is $\tau \in \mathbb{Z}$ such that $r_{s_k}(\tau) \neq r_{s_l}(\tau)$.*

In order to compute matrix $\boldsymbol{U}$, we can thus use any joint diagonalization method [CS96]. For instance, we can numerically minimize the following objective function:

$$J(\boldsymbol{U}) = \sum_{\tau} \|\boldsymbol{U}^\top\boldsymbol{R}_{zz}(\tau)\boldsymbol{U} - \mathrm{diag}(\boldsymbol{U}^\top\boldsymbol{R}_{zz}(\tau)\boldsymbol{U})\|_F^2 \tag{2.10}$$

where $\|.\|_F$ denotes the Frobenius norm of a matrix (i.e. the Euclidean norm of a vector made of all its entries) and diag(.) denotes a diagonal matrix formed from a matrix with same diagonal entries. The criterion $J(\boldsymbol{U})$ is zero if and only if all matrices $\boldsymbol{U}^\top\boldsymbol{R}_{zz}(\tau)\boldsymbol{U}$ are diagonal.

As in Section 3.2.2 page 31, this minimization can be performed by factorizing $\boldsymbol{U}$ as a product of Givens rotations, and by performing a coordinate descent w.r.t. the angles of these Givens rotations [CJ10]. The resulting BSS method is known as the *Second Order Blind Identification* (SOBI) technique [BAMCM97], and summarized in Algorithm 2.

Algorithm 2 SOBI method for WSS sources

```
Estimation of the covariance matrix $\boldsymbol{\Sigma}_{xx}$ and diagonalization: $\boldsymbol{\Sigma}_{xx} = \boldsymbol{Q}\boldsymbol{\Lambda}^2\boldsymbol{Q}^\top$ (equation (2.4))
Computation of $\boldsymbol{S} = \boldsymbol{Q}\boldsymbol{\Lambda}$ and of the whitening matrix $\boldsymbol{W} = \boldsymbol{S}^{-1}$ (equation (2.5))
Data whitening: $\boldsymbol{z}(t) = \boldsymbol{W}\boldsymbol{x}(t)$ (equation (2.6))
Estimation of covariance matrices $\boldsymbol{R}_{zz}(\tau)$ for various delays $\tau$
Approximate joint diagonalization of matrices $\boldsymbol{R}_{zz}(\tau)$ in a common basis $\boldsymbol{U}$ by minimizing (2.10)
Estimation of source signals via $\boldsymbol{y}(t) = \boldsymbol{U}^\top\boldsymbol{z}(t)$ (equation (2.7))
```

3.3.2 Non-stationarity of source signals

In this section, we keep the same mixture model as in Definition 2, and we consider the source model in Definition 1 in the determined case ($M = K$), except that each source signal $s_k(t)$ is no longer assumed to be IID, but rather non-stationary. More precisely, as in Definition 1, the samples $s_k(t)$ for $t \in \mathbb{Z}$ can still be assumed independent, but now they no longer follow the same distribution p_k (the source model is relaxed by removing the last letters "ID" of "IID"):

Definition 6 (Non-stationary source model)**.** *We consider K independent source signals $s_k(t) \in \mathbb{R}$ with $t \in \mathbb{Z}$, concatenated in a vector $\boldsymbol{s}(t)$. For all $k \in \{1\ldots K\}$, s_k is modeled as a centered random process with uncorrelated samples $s_k(t)$ for $t \in \mathbb{Z}$, of time-varying variance $\sigma_k^2(t)$. So $\boldsymbol{s}(t)$ is a random vector process of mean $\mathbb{E}[\boldsymbol{s}(t)] = \boldsymbol{0}$ and of time-varying covariance matrix $\boldsymbol{\Sigma}_{ss}(t) = \mathbb{E}\left[\boldsymbol{s}(t)\boldsymbol{s}(t)^\top\right]$.*

Since the source signals are independent, $\forall t \in \mathbb{Z}$ matrix $\boldsymbol{\Sigma}_{ss}(t)$ is diagonal: $\boldsymbol{\Sigma}_{ss}(t) = \mathrm{diag}(\sigma_k^2(t))$.

Then, still as in Section 3.2.1 page 29, we can *spatially* whiten the mixture signals:

- compute a matrix square root $\boldsymbol{S}$ of $\boldsymbol{\Sigma}_{xx} = \sum_t \boldsymbol{\Sigma}_{xx}(t)$;
- compute $\boldsymbol{W} = \boldsymbol{S}^{-1}$ and the whitened data $\boldsymbol{z}(t) = \boldsymbol{W}\boldsymbol{x}(t)$ $\forall t \in \mathbb{Z}$.

Again, as in Section 3.2.1 we can consider a canonical BSS problem and assume $\mathbf{\Sigma}_{xx} = \boldsymbol{A}\,\boldsymbol{A}^\top$, therefore matrix $\boldsymbol{U} = \boldsymbol{W}\boldsymbol{A}$ is a rotation matrix.

Then if we consider the covariance matrices of the whitened data: $\forall t \in \mathbb{Z}$, $\mathbf{\Sigma}_{zz}(t) = \mathbb{E}\left[z(t)z(t)^\top\right]$, we get $\forall t \in \mathbb{Z}$, $\mathbf{\Sigma}_{zz}(t) = \boldsymbol{W}\mathbf{\Sigma}_{xx}(t)\boldsymbol{W}^\top = \boldsymbol{W}\boldsymbol{A}\mathbf{\Sigma}_{ss}(t)\boldsymbol{A}^\top\boldsymbol{W}^\top = \boldsymbol{U}\mathbf{\Sigma}_{ss}(t)\boldsymbol{U}^\top$. Therefore, as in Section 3.3.1, matrix $\boldsymbol{U}$ can be determined by solving a joint diagonalization problem [CS96], e.g. by minimizing the following objective function:

$$J(\boldsymbol{U}) = \sum_t ||\boldsymbol{U}\mathbf{\Sigma}_{zz}(t)\boldsymbol{U}^\top - \mathrm{diag}(\boldsymbol{U}\mathbf{\Sigma}_{zz}(t)\boldsymbol{U}^\top)||_F^2. \tag{2.11}$$

We thus get a variant of the SOBI algorithm [BAMCM97], summarized in Algorithm 3.

Algorithm 3 SOBI method for non-stationary sources

```
Estimation of the covariance matrix $\mathbf{\Sigma}_{xx}$ and diagonalization: $\mathbf{\Sigma}_{xx} = \boldsymbol{Q}\mathbf{\Lambda}^2\boldsymbol{Q}^\top$ (equation (2.4))
Computation of $\boldsymbol{S} = \boldsymbol{Q}\mathbf{\Lambda}$ and of the whitening matrix $\boldsymbol{W} = \boldsymbol{S}^{-1}$ (equation (2.5))
Data whitening: $\boldsymbol{z}(t) = \boldsymbol{W}\boldsymbol{x}(t)$ (equation (2.6))
Segmentation of whitened data and estimation of covariance matrices $\mathbf{\Sigma}_{zz}(t)$ on the different time frames
Approximate joint diagonalization of matrices $\mathbf{\Sigma}_{zz}(t)$ in a common basis $\boldsymbol{U}$ by minimizing (2.11)
Estimation of source signals via $\boldsymbol{y}(t) = \boldsymbol{U}^\top\boldsymbol{z}(t)$ (equation (2.7))
```

3.4 Time-frequency methods

So far, we have seen that:

- the use of higher order cumulants is only necessary for the non-Gaussian IID source model in Definition 1;
- second order statistics are sufficient for separating source signals that are:
  - either WSS but not IID, as in Definition 5, which amounts to exploit their *spectral* dynamics (through the autocovariance function $\boldsymbol{R}_{ss}(\tau)$);
  - or uncorrelated but not stationary, as in Definition 6, which amounts to exploit their *temporal* dynamics (through the time-varying covariance matrices $\mathbf{\Sigma}_{ss}(t)$).

The take-home message is that classical signal processing tools based on second order statistics are appropriate for performing blind separation of independent (and possibly Gaussian) sources, provided that the spectral and/or temporal source dynamics are taken into account.

However, a very simple way of highlighting the spectral and temporal dynamics of a signal is to use a *Time-Frequency* (TF) representation. In this section, we will show how TF representations allow us to easily perform source separation of determined linear instantaneous mixtures. Then in Sections 4 and 5, we will see that TF representations reveal their full potential when processing convolutive and/or under-determined mixtures.

3.4.1 Time-frequency representations

Here we use the expression *time-frequency representation* to refer to complex or real-valued linear time-frequency transforms that can be implemented by means of perfect-reconstruction filterbanks [Vai93]. Classical examples of perfect reconstruction filterbanks include the STFT (which is complex-valued) and the *Modified Discrete Cosine Transform* (MDCT) (which is real-valued) [VVG18, chap. 2].

Every mixture signal $x_m(t)$ is thus filtered by F *analysis filters* h_f corresponding to each frequency channel $f \in \{1 \dots F\}$. The output signals are decimated by a factor $T \leq F$ to produce the F sub-band signals:

$$x_m(f, n) = (h_f * x_m)(nT), \tag{2.12}$$

where $n \in \mathbb{Z}$ is the *time frame index* and T is the *hop-size*.

Since we consider perfect-reconstruction filterbanks, we assume that there exist F *synthesis filters* g_f so that every signal $x_m(t)$ can be perfectly reconstructed from the sub-band signals: $x_m(t) = \sum_{f=1}^{F} g_f(t - nT)x_m(f,n)$.

In the same way, the source signals are decomposed in F sub-band signals $s_k(f,n) = (h_f * s_k)(nT)$ and reconstructed as

$$s_k(t) = \sum_{f=1}^{F} \sum_{n\in\mathbb{Z}} g_f(t - nT)s_k(f,n). \tag{2.13}$$

An interesting property of such a time-frequency representation is that it leaves the linear instantaneous mixture model in Definition 2 page 28 unchanged: if $\boldsymbol{x}(f,n) \in \mathbb{C}^M$ (resp. $\boldsymbol{s}(f,n) \in \mathbb{C}^K$) denotes the vector of coefficients $x_m(f,n)$ (resp. $s_k(f,n)$), then the equality $\boldsymbol{x}(t) = \boldsymbol{A}\,\boldsymbol{s}(t)\ \forall t \in \mathbb{Z}$ is equivalent to

$$\forall f \in \{1\dots F\}, \forall n \in \mathbb{Z},\ \boldsymbol{x}(f,n) = \boldsymbol{A}\,\boldsymbol{s}(f,n). \tag{2.14}$$

Indeed, if $a_{m,k}$ denotes the entries of the mixing matrix $\boldsymbol{A}$, we have $\forall m \in \{1\dots M\}$,

$$x_m(f,n) = (h_f * x_m)(nT) = \left(h_f * \sum_{k=1}^{K} a_{m,k}s_k\right)(nT) = \sum_{k=1}^{K} a_{m,k}(h_f * s_k)(nT) = \sum_{k=1}^{K} a_{m,k}s_k(f,n).$$

3.4.2 Time-frequency source model

Let us now introduce a general non-stationary source model. As usual we assume that the K sub-band source signals $s_k(f,n)$ are centered and independent.

In Section 3.3.1 page 32, we have presented the SOBI BSS method, that exploits the *spectral dynamics* of the source signals, by modeling them as WSS processes. Remember that the well-known *spectral representation theorem* of WSS processes [BD87] shows that the Fourier transforms of WSS processes are formed of uncorrelated random elements whose variances vary over frequency.

In a similar way, in Section 3.3.2 page 33, we have presented a variant of the SOBI method, that exploits the *temporal dynamics* of the source signals, by modeling them as sequences of uncorrelated random variables whose variances vary over time.

Now, the use of a time-frequency representation allows use to jointly exploit the spectral and the temporal dynamics, just by modeling the samples of the sub-band source signals $s_k(f,n)$ for $f \in \{1\dots F\}$ and $n \in \mathbb{Z}$ as uncorrelated random variables whose variance $\sigma_k^2(f,n)$ depends both on the frequency channel f and the time frame n:

Definition 7 (Non-stationary TF source model)**.** *We consider K independent source signals $s_k(t) \in \mathbb{R}$ with $t \in \mathbb{Z}$ and their TF representations $s_k(f,n)$ as defined in Section 3.4.1, concatenated in a vector $\mathbf{s}(f,n)$. For all $k \in \{1\dots K\}$, the sub-band source signals $s_k(f,n)$ for $f \in \{1\dots F\}$ and $n \in \mathbb{Z}$ are modeled as uncorrelated random variables of mean* 0 *and whose variance $\sigma_k^2(f,n)$ depends both on f and n. So $\mathbf{s}(f,n)$ is a random vector process of mean $\mathbb{E}[\boldsymbol{s}(f,n)] = \mathbf{0}$ and of TF-varying covariance matrix $\boldsymbol{\Sigma}_{ss}(f,n) = \mathbb{E}\left[\boldsymbol{s}(f,n)\boldsymbol{s}(f,n)^H\right]$.*

3.4.3 Separation method

This leads us to a new variant of the SOBI method in the determined case ($M = K$), based on the TF source model in Definition 7. Let us define the mixture covariance matrices $\boldsymbol{\Sigma}_{xx}(f,n) = \mathbb{E}[\boldsymbol{x}(f,n)\boldsymbol{x}(f,n)^H]$. Since $\boldsymbol{x}(f,n) = \boldsymbol{A}\,\boldsymbol{s}(f,n)$ (equation (2.14)), we have $\boldsymbol{\Sigma}_{xx}(f,n) = \boldsymbol{A}\boldsymbol{\Sigma}_{ss}(f,n)\boldsymbol{A}^\top$. Moreover, since the source signals are independent, matrix $\boldsymbol{\Sigma}_{ss}(f,n)$ is diagonal: $\boldsymbol{\Sigma}_{ss}(f,n) = \mathrm{diag}(\sigma_k^2(f,n))$.

Then, as in Section 3.2.1 page 29, we can *spatially* whiten the mixture signals:

- compute a matrix square root $\boldsymbol{S}$ of $\boldsymbol{\Sigma}_{xx} = \sum_{f,n} \boldsymbol{\Sigma}_{xx}(f,n)$;
- compute $\boldsymbol{W} = \boldsymbol{S}^{-1}$ and the whitened data

$$\forall f \in \{1\dots F\}, \forall n \in \mathbb{Z},\ \boldsymbol{z}(f,n) = \boldsymbol{W}\boldsymbol{x}(f,n). \tag{2.15}$$

Again, as in Section 3.2.1 we can consider a canonical BSS problem and assume $\boldsymbol{\Sigma}_{xx} = \boldsymbol{A}\,\boldsymbol{A}^\top$, therefore matrix $\boldsymbol{U} = \boldsymbol{W}\boldsymbol{A}$ is a rotation matrix.

Then if we consider the covariance matrices of the whitened data: $\forall f, n$, $\boldsymbol{\Sigma}_{zz}(f,n) = \mathbb{E}\left[z(f,n)z(f,n)^H\right]$, we get $\boldsymbol{\Sigma}_{zz}(f,n) = \boldsymbol{W}\boldsymbol{\Sigma}_{xx}(f,n)\boldsymbol{W}^H = \boldsymbol{W}\boldsymbol{A}\boldsymbol{\Sigma}_{ss}(f,n)\boldsymbol{A}^H\boldsymbol{W}^H = \boldsymbol{U}\boldsymbol{\Sigma}_{ss}(f,n)\boldsymbol{U}^H$. Therefore, as in Section 3.3.1 page 32, matrix $\boldsymbol{U}$ can be determined by solving a joint diagonalization problem [CS96], e.g. by minimizing the following objective function:

$$J(\boldsymbol{U}) = \sum_{f,n} \|\boldsymbol{U}\boldsymbol{\Sigma}_{zz}(f,n)\boldsymbol{U}^H - \operatorname{diag}(\boldsymbol{U}\boldsymbol{\Sigma}_{zz}(f,n)\boldsymbol{U}^H)\|_F^2. \tag{2.16}$$

Finally, the source sub-band signals can be estimated as

$$\boldsymbol{y}(f,n) = \boldsymbol{U}^\top \boldsymbol{z}(f,n). \tag{2.17}$$

We thus get a variant of the SOBI algorithm [BAMCM97], summarized in Algorithm 4.

Algorithm 4 SOBI method in the TF domain

```
TF analysis of mixture signals: $x_m(f,n) = (h_f * x_m)(nT)$ (equation (2.12))
Estimation of the covariance matrix $\boldsymbol{\Sigma}_{xx}$ and diagonalization: $\boldsymbol{\Sigma}_{xx} = \boldsymbol{Q}\boldsymbol{\Lambda}^2\boldsymbol{Q}^H$ (equation (2.4))
Computation of $\boldsymbol{S} = \boldsymbol{Q}\boldsymbol{\Lambda}$ and of the whitening matrix $\boldsymbol{W} = \boldsymbol{S}^{-1}$ (equation (2.5))
Data whitening: $z(f,n) = \boldsymbol{W}\boldsymbol{x}(f,n)$ (equation (2.15))
Estimation of covariance matrices $\boldsymbol{\Sigma}_{zz}(f,n)$ on all time-frequency bins
Approximate joint diagonalization of matrices $\boldsymbol{\Sigma}_{zz}(f,n)$ in a common basis $\boldsymbol{U}$ by minimizing (2.16)
Estimation of source signals via $\boldsymbol{y}(f,n) = \boldsymbol{U}^H\boldsymbol{z}(f,n)$ (equation (2.17))
TF synthesis of source signals: $y_k(t) = \sum_{f=1}^F \sum_{n\in\mathbb{Z}} g_f(t-nT)y_k(f,n)$ (equation (2.13))
```

4 Convolutive mixtures

As already mentioned in Section 1.2 page 24, linear instantaneous mixtures cannot accurately model real acoustic mixtures, since reverberation in a room involves convolutive effects. For this reason, we now address the extension of the BSS methods presented in Section 3 page 27 to convolutive mixtures.

4.1 Source images

First, suppose that K source signals $s_k(t)$ are simultaneously emitted in a room, and that M microphones receive the observed data vector $\boldsymbol{x}(t) \in \mathbb{R}^M$. The raw source separation problem would consist in estimating the *image* of each source k, i.e. the data vector $\boldsymbol{x}_k(t) \in \mathbb{R}^M$ that would be received by the M microphones if only source k was active. These images are such that $\boldsymbol{x}(f,n) = \sum_{k=1}^K \boldsymbol{x}_k(f,n)$. Then the task that consists in estimating the scalar source signals $s_k(t)$ from each vector image $\boldsymbol{x}_k(t)$ is called *deconvolution* or *dereverberation*.

In this way, the source separation problem is decomposed in two steps:

- **separation**: estimate the image $\boldsymbol{x}_k(f,n)$ from the mixture $\boldsymbol{x}(f,n)$
- **deconvolution**: estimate the source signal $s_k(f,n)$ from $\boldsymbol{x}_k(f,n)$

4.2 Convolutive mixture model

Let us now introduce the convolutive mixture model in the time domain:

Definition 8 (Convolutive mixture model)**.** *We consider K source signals $s_k(t) \in \mathbb{R}$ with $t \in \mathbb{Z}$, concatenated in a vector $\mathbf{s}(t)$. The samples of the M mixture signals $x_m(t) \in \mathbb{R}$ for $m \in \{1 \dots M\}$ are then defined as*

$$x_m(t) = \sum_{k=1}^K (a_{mk} * s_k)(t).$$

where $\forall k \in \{1 \dots K\}$, $\forall m \in \{1 \dots M\}$, a_{mk} *is the impulse response of a stable*[3] *filter. In vector form, we will write* $\boldsymbol{x}(t) = \boldsymbol{A} * \boldsymbol{s}(t)$ *where* $\boldsymbol{A}$ *is an* $M \times K$ *matrix of stable impulse responses* a_{mk}*, and* $*$ *denotes the convolution product between a sequence of matrices and a sequence of vectors.*

The following identifiability theorem [NTJ95] generalizes Theorem 1 page 29 to the convolutive case: it proves the feasibility of ICA under mild conditions about the source signals:

Theorem 3 (Identifiability theorem)**.** *Consider the source model in Definition 1, with the convolutive mixture model in Definition 8 in the determined case* ($M = K$)*. Among the* K *IID sources* s_k*, suppose that at most one is Gaussian-distributed. Let* $\boldsymbol{C}$ *be a* $K \times K$ *matrix of stable impulse responses, and* $\forall t \in \mathbb{Z}$, $\boldsymbol{y}(t) = \boldsymbol{C} * \boldsymbol{s}(t)$*. Then the random processes* $y_k(t)$ *for* $k \in \{1 \dots K\}$ *are independent if and only if matrix* $\boldsymbol{C}$ *is non-mixing.*

Theorem 3 shows that the source signals can be retrieved up to an unknown permutation and an unknown scale factor.

4.3 Time-frequency approach

In order to simplify the problem and to make it possible to reuse the source separation methods introduced in Section 3, let us now rewrite the mixture model in Definition 8 in the time-frequency domain. More precisely, signals will be represented by their STFT, using the filterbank notation introduced in Section 3.4.1 page 34.

Moreover, we will consider the *narrow-band* approximation: we will assume that the impulse response of each mixing filter a_{mk} is short w.r.t. the time frame length of the STFT [4]. As a consequence, the spectral variations of the frequency responses $A_{mk}(\nu)$ are slow compared to those of $H_f(\nu)$ $\forall f \in \{1 \dots F\}$. Since h_f is a very narrow band-pass filter, we will even assume that $A_{mk}(\nu)$ is approximately constant in the pass-band of $H_f(\nu)$ (see Figure 2.6). Consequently, we can make the approximation $H_f(\nu) A_{mk}(\nu) \approx H_f(\nu)\, a_{mk}(f)$, where $a_{mk}(f)$ is the average value of $A_{mk}(\nu)$ in frequency channel f. Back in the time domain, this approximation can be rewritten $(h_f * a_{mk})(t) \approx a_{mk}(f)\, h_f(t)$.

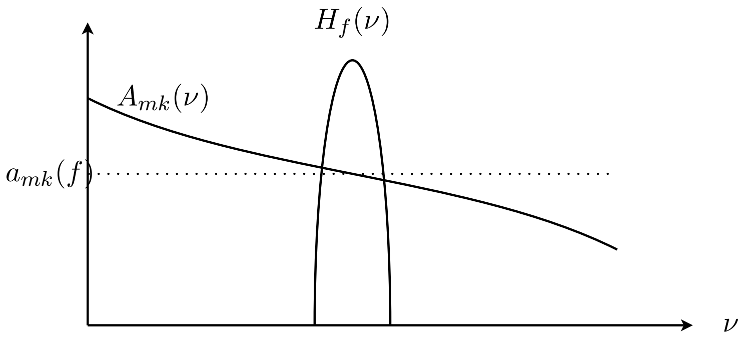


Figure 2.6: Narrow-band approximation

[3] Stability is defined in the *bounded-input, bounded-output* (BIBO) sense: if the input signal is bounded, then the output signal is also bounded.

[4] Note that this assumption is not realistic: the length of the impulse response a_{mk} corresponds to the reverberation time, which is usually several hundreds of miliseconds, while the typical time frame length in an STFT is a few tens of miliseconds. Nevertheless, this approximation is often used in audio source separation methods because it leads to a very simple mixture model in the TF domain (*cf.* Definition 9), which proves to perform well in various applications [OF10].

If we now consider the STFT of the m-th mixture signal, we get

$$\begin{aligned} x_m(f,n) &= (h_f * x_m)(nT) \\ &= \left(h_f * \left(\textstyle\sum_{k=1}^{K} a_{mk} * s_k\right)\right)(nT) \\ &= \left(\textstyle\sum_{k=1}^{K} \left(h_f * a_{mk}\right) * s_k\right)(nT) \\ &\approx \textstyle\sum_{k=1}^{K} a_{mk}(f)\left(h_f * s_k\right)(nT) \\ &= \textstyle\sum_{k=1}^{K} a_{mk}(f) s_k(f,n). \end{aligned}$$

Hence the following approximate convolutive mixture model in the time-frequency domain:

Definition 9 (TF mixture model)**.** *We consider K source signals $s_k(t) \in \mathbb{R}$ with $t \in \mathbb{Z}$ and their TF representations $s_k(f,n)$ as defined in Section 3.4.1, concatenated in a vector $\boldsymbol{s}(f,n)$. For all $m \in \{1 \dots M\}$, the sub-band mixture signals $x_m(f,n)$ for $f \in \{1 \dots F\}$ and $n \in \mathbb{Z}$ are then defined as*

$$x_m(f,n) = \sum_{k=1}^{K} a_{mk}(f) s_k(f,n),$$

or in matrix form,

$$\boldsymbol{x}(f,n) = \boldsymbol{A}(f)\boldsymbol{s}(f,n), \tag{2.18}$$

where $\boldsymbol{A}(f) \in \mathbb{C}^{M \times K}$ is the matrix of entries a_{mk}.

It can be noted that in each frequency channel f, (2.18) is a linear instantaneous mixture model (to be compared to equation (2.2) page 28), parameterized by the mixing matrix $\boldsymbol{A}(f)$. Therefore we are tempted to apply any ICA method designed for the linear instantaneous mixture model, such as those described in Section 3, in every frequency channel of the STFT, in order to estimate the sub-band signals $s_k(f,n)$.

4.4 Independent component analysis

As in Section 3.1.2 page 28, it may seem natural to estimate the sub-band source signals as linear instantaneous combinations of the sub-band mixture signals: $\boldsymbol{y}(f,n) = \boldsymbol{B}(f)\,\boldsymbol{x}(f,n)$, where the entries of vector $\boldsymbol{y}(f,n)$ are the sub-band source signal estimates, and $\boldsymbol{B}(f) \in \mathbb{C}^{K \times M}$ is referred to as the *separation matrix*. Then the source separation problem amounts to finding an optimal separation matrix.

Linear source separation is generally feasible in the case of determined and over-determined mixtures:

- if $M = K$ and if matrix $\boldsymbol{A}(f)$ is invertible, then the separation matrix $\boldsymbol{B}(f) = \boldsymbol{A}(f)^{-1}$ leads to $\boldsymbol{y}(f,n) = \boldsymbol{s}(f,n)$;
- more generally, if $M \geq K$ and if matrix $\boldsymbol{A}(f)$ has full rank, then the separation matrix $\boldsymbol{B}(f) = \boldsymbol{A}(f)^{\dagger}$ leads to $\boldsymbol{y}(f,n) = \boldsymbol{s}(f,n)$.

However, in the under-determined case ($M < K$), linear source separation is generally not feasible (*cf.* Section 5).

In the determined case ($M = K$), independent component analysis [CJ10] can be applied in each frequency channel f. It aims to find a *separation matrix* $\boldsymbol{B}(f)$ that makes the K sub-band signals $y_k(f,n)$ independent. Then, as in Section 3.2 page 29, we get $\boldsymbol{y}(f,n) = \boldsymbol{C}(f)\boldsymbol{s}(f,n)$, where $\boldsymbol{C}(f) = \boldsymbol{B}(f)\boldsymbol{A}(f)$ is a non-mixing matrix (*cf.* Definition 3 page 28).

4.5 Indeterminacies

While the indeterminacies induced by the non-mixing matrix $\boldsymbol{C}$ were acceptable when we were considering linear instantaneous mixtures in Section 3, we now encounter an unexpected issue, because with the TF mixture model in Definition 9, there are F possibly different non-mixing matrices $\boldsymbol{C}(f)$. The problem is that, for instance, the permutations can be different in two different frequency channels f. If we choose to ignore that, and to just reconstruct the source signals $y_k(t)$ from the separated sub-band signals $y_k(f,n)$ with the synthesis filters g_f, it is very likely that the resulting signals $y_k(t)$ will be formed of different sources s_k in different frequency channels.

In other words, reconstructing the source signals from the separated sub-band signals would amount to remix the estimated sources!

In order to avoid this problem, we need to solve the permutation indeterminacy in the frequency channels of the STFT. Note that this multiple permutation issue is inherent to the time-frequency approach and to the narrow-band approximation introduced in Section 4.3: if BSS was performed in the time domain instead, Theorem 3 page 37 proves that all sources can theoretically be retrieved up to a unique permutation.

Even though, assuming that we do have a method that allows us to solve the multiple permutation indeterminacy, the other indeterminacy remains: there is an unknown multiplicative factor associated to each source in each frequency channel f.

Actually, both kinds of indeterminacies can be solved jointly, by introducing additional assumptions about the mixing filters a_{mk} and/or the source signals s_k:

- regarding the source signals, we can assume that the temporal dynamics (over n) of $\sigma_k^2(f,n)$ are similar between different frequency channels f for a same source k (nonparametric approach), or we can also exploit a parametric model, such as the *Nonnegative Matrix Factorization (NMF)* [VVG18, chap. 8] [OF10].

- regarding the mixing filters, we can assume that their frequency responses $a_{mk}(f)$ are slowly varying w.r.t. f (nonparametric approach), or we can also exploit a parametric mixture model, such as the beamforming model or the anechoic model. The beamforming model [VVG18, chap. 10] relies on the plane wave and far field hypotheses (no reverberation) and assumes that the microphone antenna is linear. In this case, we get $a_{mk}(f) = e^{-2\imath\pi f\tau_{mk}}$ where $\tau_{mk} = \frac{d_m}{c}\sin(\theta_k)$, where parameters d_m denote the positions of the sensors on the linear antenna and parameters θ_k denote the angles of the sources (see Figure 2.7). The anechoic model is a bit more general: it assumes that the sources are punctual and that there is no reverberation. In this case, we get $a_{mk}(f) = \alpha_{mk}e^{-2\imath\pi f\tau_{mk}}$ where $\alpha_{mk} = \frac{1}{\sqrt{4\pi}r_{mk}}$, $\tau_{mk} = \frac{r_{mk}}{c}$, and parameters r_{mk} denote the distances between the sensors and sources. In practice, none of these two mixture models is able to accurately represent real acoustic mixtures; nevertheless they can be helpful to solve the multiple permutation problem.

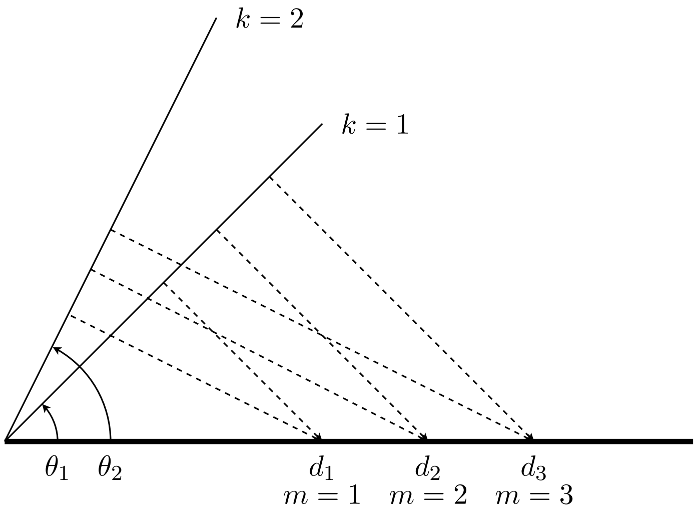


Figure 2.7: Beamforming mixture model

5 Under-determined mixtures

As mentioned in Section 1.1 page 23, linear source separation is generally not feasible in the under-determined case, because there are infinitely many solutions. Without additional information about the mixture or the source signals, it is impossible to retrieve the original sources from the mixture signals.

Unfortunately, the under-determined case is often encountered in audio signal processing: indeed, many audio signals are either monophonic ($M = 1$) or stereophonic ($M = 2$), whereas the number of sources K is generally greater than 2. In this section, we will see how additional information can be taken into account to perform source separation in such a challenging scenario.

5.1 Under-determined convolutive mixtures

We still consider the TF mixture model in Definition 9 page 38 : $\boldsymbol{x}(f,n) = \boldsymbol{A}(f)\boldsymbol{s}(f,n)$ (equation (2.18)), and the TF source model in Definition 7 page 35: the samples of the sub-band source signals $s_k(f,n)$ for $f \in \{1 \dots F\}$ and $n \in \mathbb{Z}$ are uncorrelated random variables whose variance $\sigma_k^2(f,n)$ depends both on the frequency channel f and the time frame n, so that the covariance matrix of $\boldsymbol{s}(f,n)$ is $\boldsymbol{\Sigma}_{ss}(f,n) = \text{diag}(\sigma_k^2(f,n))$.

As in Section 4.4 page 38, it may seem natural to estimate the sub-band source signals as linear instantaneous combinations of the sub-band mixture signals: $\boldsymbol{y}(f,n) = \boldsymbol{B}(f)\,\boldsymbol{x}(f,n)$, where the entries of vector $\boldsymbol{y}(f,n)$ are the sub-band source signal estimates, and $\boldsymbol{B}(f) \in \mathbb{C}^{K\times M}$ is referred to as the *separation matrix*. Then the source separation problem amounts to finding an optimal separation matrix.

However if $M < K$, even if the mixing matrix $\boldsymbol{A}(f)$ and the source model $\boldsymbol{\Sigma}_{ss}(f,n)$ were known, the exact separation would not be feasible: there is no matrix $\boldsymbol{B}(f)$ such that $\boldsymbol{B}(f)\,\boldsymbol{A}(f) = \boldsymbol{I}_K$, because the maximum rank of matrix $\boldsymbol{B}(f)\,\boldsymbol{A}(f)$ is $M < K$ (*cf.* Figure 2.8).

However, we can still try to find an approximate solution $\boldsymbol{y}(f,n)$ in the least squares sense.

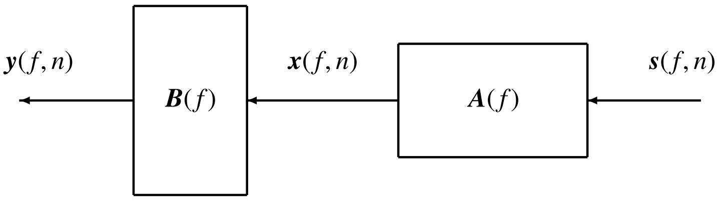


Figure 2.8: Under-determined mixtures: there is no matrix $\boldsymbol{B}(f)$ such that $\boldsymbol{B}(f)\,\boldsymbol{A}(f) = \boldsymbol{I}_K$

5.2 Separation via non-stationary filtering

First, we suppose that the mixing matrix $\boldsymbol{A}(f)$ and the source model $\boldsymbol{\Sigma}_{ss}(f,n)$ are known. Even though only an approximate solution $\boldsymbol{y}(f,n)$ can be obtained, this solution can be improved by adding a degree of freedom to the separation matrix $\boldsymbol{B}$: we will now make it depend on time n, so that the estimate $\boldsymbol{y}(f,n)$ is obtained by *non-stationary filtering*:

$$\boldsymbol{y}(f,n) = \boldsymbol{B}(f,n)\,\boldsymbol{x}(f,n) \tag{2.19}$$

where $\boldsymbol{B}(f,n) \in \mathbb{C}^{K\times M}$. The approximate solution will be found in the least squares sense, by considering the *Minimum Mean Square Error* (MMSE) estimator [SAK+13]: we will look for the separation matrix $\boldsymbol{B}(f,n)$ which minimizes the *Mean Square Error* (MSE) $\mathbb{E}[\|\boldsymbol{y}(f,n) - \boldsymbol{s}(f,n)\|_2^2]$. This MSE is such that

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{y}(f,n) - \boldsymbol{s}(f,n)\|_2^2] &= \mathbb{E}\left[(\boldsymbol{B}(f,n)\,\boldsymbol{x}(f,n) - \boldsymbol{s}(f,n))^H\,(\boldsymbol{B}(f,n)\,\boldsymbol{x}(f,n) - \boldsymbol{s}(f,n))\right] \\
&= \text{trace}\left(\mathbb{E}\left[(\boldsymbol{B}(f,n)\,\boldsymbol{x}(f,n) - \boldsymbol{s}(f,n))\,(\boldsymbol{B}(f,n)\,\boldsymbol{x}(f,n) - \boldsymbol{s}(f,n))^H\right]\right) \\
&= \text{trace}\left(\boldsymbol{B}(f,n)\boldsymbol{\Sigma}_{xx}(f,n)\boldsymbol{B}(f,n)^H - \boldsymbol{B}(f,n)\boldsymbol{\Sigma}_{xs}(f,n) - \boldsymbol{\Sigma}_{sx}(f,n)\boldsymbol{B}(f,n)^H + \boldsymbol{\Sigma}_{ss}(f,n)\right).
\end{aligned}$$

The MSE is minimized when the Wirtinger matrix gradient [GR65] w.r.t. $\boldsymbol{B}(f,n)$ is zero:

$$\boldsymbol{B}(f,n)\boldsymbol{\Sigma}_{xx}(f,n) - \boldsymbol{\Sigma}_{sx}(f,n) = \boldsymbol{0}.$$

Therefore the solution is given by $\boldsymbol{B}(f,n) = \boldsymbol{\Sigma}_{sx}(f,n)\,\boldsymbol{\Sigma}_{xx}(f,n)^{-1}$, where $\boldsymbol{\Sigma}_{xx}(f,n) = \boldsymbol{A}(f)\,\boldsymbol{\Sigma}_{ss}(f,n)\,\boldsymbol{A}(f)^H$ and $\boldsymbol{\Sigma}_{sx}(f,n) = \boldsymbol{\Sigma}_{ss}(f,n)\,\boldsymbol{A}(f)^H$. We can finally express the MMSE estimator as (2.19), with

$$\boldsymbol{B}(f,n) = \boldsymbol{\Sigma}_{ss}(f,n)\,\boldsymbol{A}(f)^H\left(\boldsymbol{A}(f)\,\boldsymbol{\Sigma}_{ss}(f,n)\,\boldsymbol{A}(f)^H\right)^{-1}. \tag{2.20}$$

We remark that the MMSE estimator guarantees the perfect reconstruction of the mixture signals from the estimated source signals: $\boldsymbol{A}(f)\,\boldsymbol{y}(f,n) = \boldsymbol{x}(f,n)$.

This MMSE estimator is also known as the *generalized* or *multichannel* Wiener filter, because in the particular case of monophonic mixtures ($M = 1$), it boils down to the well-known Wiener filter. Indeed, because of the scale indeterminacy of the model, we can assume without loss of generality that $\boldsymbol{A}(f) = [1, \dots, 1]$. Then the MMSE estimator defined by (2.19) and (2.20) can be rewritten as $y_k(f,n) = \frac{\sigma_k^2(f,n)}{\sum_{l=1}^K \sigma_l^2(f,n)}\, x(f,n)$, which is the usual form of the Wiener filter.

In practice of course, the mixing matrix $\boldsymbol{A}(f)$ and the source model $\boldsymbol{\Sigma}_{ss}(f,n)$ are unknown; they thus have to be estimated from the observed data. For instance, $\boldsymbol{A}(f)$ can be assumed slowly varying over f (nonparametric approach), or parameterized according to the beamforming or the anechoic model introduced in Section 4.5 page 38, and $\boldsymbol{\Sigma}_{ss}(f,n)$ can be assumed sparse in the TF domain as in Section 5.3.1 (nonparametric approach), or parameterized according to an NMF model [VVG18, chap. 8] [OF10].

The resulting algorithm is sketched in Algorithm 5.

Algorithm 5 Under-determined source separation in the TF domain

```
TF analysis of mixture signals: $x_k(f,n) = (h_f * x_k)(nT)$ (equation (2.12))
Estimation of $\boldsymbol{A}(f)$ and $\sigma_k^2(f,n)$
Computation of $\boldsymbol{B}(f,n) = \boldsymbol{\Sigma}_{ss}(f,n)\,\boldsymbol{A}(f)^H\left(\boldsymbol{A}(f)\,\boldsymbol{\Sigma}_{ss}(f,n)\,\boldsymbol{A}(f)^H\right)^{-1}$ (equation (2.20))
Estimation of source sub-band signals as $\boldsymbol{y}(f,n) = \boldsymbol{B}(f,n)\,\boldsymbol{x}(f,n)$ (equation (2.19))
TF synthesis of source signals: $y_k(t) = \sum_{f=1}^F \sum_{n\in\mathbb{Z}} g_f(t-nT) y_k(f,n)$ (equation (2.13))
```

5.3 Stereophonic mixtures: separation based on sparsity

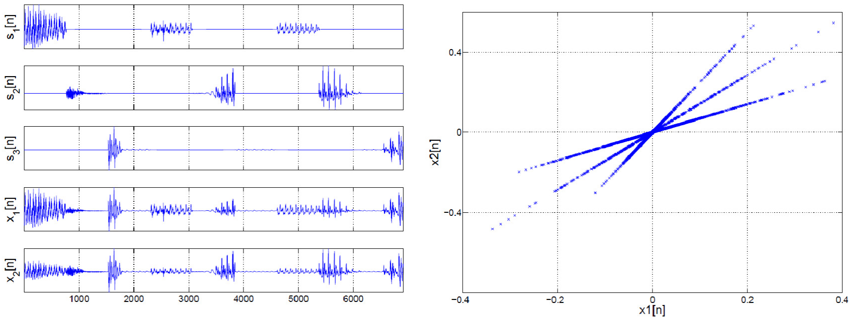
5.3.1 Temporal sparsity

We now consider the particular case of stereophonic ($M = 2$) linear instantaneous mixtures as in Definition 2 page 28 (defined by a unique mixing matrix $\boldsymbol{A}$ in order to simplify, but this approach would also work with the TF mixture model in Definition 9 page 38), so that the mixture model in the time domain is $\boldsymbol{x}(t) = \boldsymbol{A}\,\boldsymbol{s}(t)$.

We consider the example in Figure 2.9-(a): the $K = 3$ source signals represented in the three top lines of the figure are never active at the same time. We say that they are *sparse* in the time domain, in the sense that most of their temporal samples are zero. The $M = 2$ mixture signals are represented in the two bottom lines. Clearly, in this particular case of non-overlapping source signals, the source separation problem is a simple *classification* problem: it amounts to *segment* the mixture signals, and label the successive segments as "source 1", "source 2", etc.

In this simple scenario, the classification of the time samples can be easily performed by plotting the *dispersion diagram*, represented in Figure 2.9-(b): for every time t, the point of coordinates $(x_1(t), x_2(t))$ is drawn in the plane. This diagram clearly makes appear three straight lines, which correspond to the three sources. Indeed, when only source k is active, the linear instantaneous mixture model in Definition 2 yields $\boldsymbol{x}(t) = \boldsymbol{a}_k\, s_k(t)$, where $\boldsymbol{a}_k$ is the k-th column of matrix $\boldsymbol{A}$ (remember that because of the scale indeterminacy, we can assume without loss of generality that $\boldsymbol{a}_k$ is a unit vector). Therefore all points of coordinates $\boldsymbol{x}(t)$ in the plane that are generated by source k belong to the straight line passing through the origin and defined by the direction vector $\boldsymbol{a}_k$, and their position on this straight line corresponds to the value of the time sample $s_k(t)$.

Therefore in this simple case, source separation can be very easily performed by detecting the lines in the dispersion diagram (e.g. by using the Hough transform [SS01]), which makes it possible to jointly estimate the column vectors $\boldsymbol{a}_k$ and the number of sources K. Then the images of the sources (defined in Section 4.1 page 36) can be retrieved by selecting the points $\boldsymbol{x}(t)$ that are the closest to each line, and finally the source signals $s_k(t)$ can be estimated by calculating the positions of these points on the line.



(a) Temporal source signals and corresponding stereo mixture

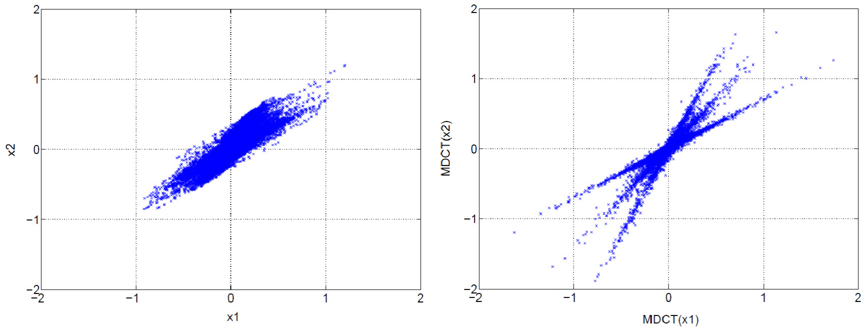
(b) Dispersion diagrams (x_1, x_2) over time

Figure 2.9: Sparsity in time domain

5.3.2 Sparsity in a transformed domain

Unfortunately, the example in Figure 2.9 is not very realistic: it rarely happens, especially in music, that the different source signals do not overlap in the time domain. Figure 2.10-(a) shows an example of dispersion diagram obtained with a mixture of overlapping music sources: no straight line emerges from the cloud of points.

However, even when they do overlap in time, audio signals are often sparse in the TF domain: the spectrum of various sounds (especially in music) is made of a discrete set of frequencies. Figure 2.10-(b) shows another dispersion diagram obtained from the same mixture of music sources as in Figure 2.10-(a), except that the coordinates of the points are not obtained from the time samples $\boldsymbol{x}(t)$, but from the MDCT TF transform[5] $\boldsymbol{x}(f, n)$. The resulting dispersion diagram is not as clean as that of Figure 2.9-(b), but again three straight lines clearly emerge from the cloud of points, which shows that the mixture is made of $K = 3$ sources, which can be separated in the same way as in Section 5.3.1, but in the TF domain.



(a) Time samples

(b) Time-frequency coefficients after MDCT decomposition

Figure 2.10: Sparsity in TF domain

[5] The MDCT is known to produce very sparse TF representations, which is why it is widely used in lossy audio data compression [Luo09].

5.3.3 DUET method

We can now introduce the celebrated *Degenerate Unmixing Estimation Technique* (DUET) method [JRY00, Ric07], which is dedicated to stereophonic ($M = 2$) mixtures, in the linear instantaneous mixture case: $\boldsymbol{x}(f,n) = \boldsymbol{A}\,\boldsymbol{s}(f,n)$ (equation 2.14 page 35). Without loss of generality, the column vectors of the mixing matrix $\boldsymbol{A}$ are parameterized as $\boldsymbol{a}_k = \begin{bmatrix} \cos(\theta_k) \\ \sin(\theta_k) \end{bmatrix}$, where $\theta_k \in \mathbb{R}$. Regarding the source signals, we consider a *sparse* source model in the TF domain:

Definition 10 (Sparse TF source model)**.** *We consider the same source model as in Definition 7. In addition, we assume that* $\forall f, n$, *there is a unique* $k_{(f,n)} \in \{1 \dots K\}$ *such that* $\sigma^2_{k_{(f,n)}}(f,n) > 0$, *and* $\forall l \neq k_{(f,n)}$, $\sigma^2_l(f,n) = 0$.

This additional assumption means that only one source can be active at any TF bin (f,n) (sources do not overlap in the TF domain). Therefore $\forall f, n$, $\boldsymbol{x}(f,n) = \boldsymbol{a}_{k_{(f,n)}} s_{k_{(f,n)}}(f,n)$.

The DUET method then consists of two steps: parameter estimation and source separation.

In the first step, the TF representation of the mixture signals is first computed by using the analysis filters h_f as in equation (2.12) page 34. Then, in order to estimate the mixture parameters θ_k, the histogram of the angles of vectors $\boldsymbol{x}(f,n)$ is first computed. The peaks of this histogram theoretically correspond to the angles θ_k, which can thus be estimated by performing peak detection. Then the active source $k_{(f,n)}$ at frequency bin (f,n) is estimated by selecting the angle θ_k which is the closest to the angle of the observed vector $\boldsymbol{x}(f,n)$.

In the second step, the source images (defined in Section 4.1 page 36) are estimated via binary masking [YR04]:

$$\forall k \in \{1 \dots K\},\ \boldsymbol{y}_k(f,n) = \begin{cases} \boldsymbol{x}(f,n)\ \forall (f,n) \text{ such that } k_{(f,n)} = k, \\ \boldsymbol{0} \text{ for the other time-frequency bins } (f,n). \end{cases} \tag{2.21}$$

Then the sub-band source signals are estimated with the MMSE estimator introduced in equation (2.20) page 41, which here boils down to a zero separation matrix $\boldsymbol{B}(f,n)$, except its k-th row which is[6]

$$\boldsymbol{a}_k(f)^{\dagger} = \frac{\boldsymbol{a}_k(f)^H}{\|\boldsymbol{a}_k(f)\|_2^2}. \tag{2.22}$$

Therefore the estimate of the k-th sub-band source signal as defined in (2.19) page 40 is

$$y_k(f,n) = \boldsymbol{a}_k(f)^{\dagger} \boldsymbol{y}_k(f,n). \tag{2.23}$$

Finally, the source signals are reconstructed in the time domain by using the synthesis filters g_f, as in equation (2.13) page 35. The DUET method is summarized in Algorithm 6.

Algorithm 6 DUET method

```
TF analysis of mixture signals: $x_k(f,n) = (h_f * x_k)(nT)$ (equation (2.12))
Estimation of parameters $\theta_k$ and of the active source $k_{(f,n)}$
    Computation of the histogram of the angles of vectors $\boldsymbol{x}(f,n)$
    Peak detection in order to estimate parameters $\theta_k$
    Determination of the active source at $(f,n)$ by proximity with $\theta_k$
Source separation:
    Estimation of source images $\boldsymbol{y}_k(f,n)$ via binary masking (equation (2.21))
    MMSE estimation of sub-band source signals: $y_k(f,n) = \boldsymbol{a}_k(f)^{\dagger}\boldsymbol{y}_k(f,n)$ (equation (2.23))
TF synthesis of source signals: $y_k(t) = \sum_{f=1}^{F} \sum_{n\in\mathbb{Z}} g_f(t-nT) y_k(f,n)$ (equation (2.13))
```

[6] In equation (2.20), matrix $\boldsymbol{\Sigma}_{ss}$ is singular, so the matrix inverse is replaced by the matrix pseudo-inverse, leading to equation (2.22).

6 Conclusion

In this chapter, we have reviewed several source separation models and methods, dedicated to determined linear instantaneous mixtures, determined convolutive mixtures, and under-determined mixtures. All these methods exploit the spatial diversity of the observed mixture signals (they require that $M > 1$), and their funding principle is that all source signals are statistically independent.

Source separation requires to make assumptions about the mixture and about the source signals, which are generally expressed in terms of probability distributions. For (over-)determined linear mixtures, we have seen that assuming independent sources is generally sufficient to make the separation possible (under mild conditions on the source probability distributions). When the mixture is under-determined however, it is necessary to make additional assumptions about the mixture and about the source signals, that can be formulated either in a non-parametric way (via regularization), or by exploiting parametric models.

This chapter forms an introduction to audio source separation, with a selection of models and methods; several topics that have been investigated in the literature could not been addressed here, for instance:

- The separation of non-stationary mixtures requires to develop adaptive separation algorithms [CL96];
- *Informed source separation* techniques exploit some possibly available extra information about the mixture or the sources, such as the spatial positions of the sources and microphones (e.g. via *beamforming*), or the transcription of the source signals (speech or music) [EM12];
- Deep learning techniques are able to automatically learn how to perform separation from a large database of source and mixture signals [VVG18].
- Criteria for the objective assessment of audio source separation are required in order to compare the performance of various separation methods [VGF06].

Chapter 3

High resolution methods

In the context of speech and music signal processing, the tonal part of a wide variety of sounds is accurately modeled as a sum of sinusoids with slowly varying parameters. For example, the sounds that produce a well-defined perception of pitch have a quasi-periodic waveform (over a duration greater than a few tens of miliseconds). Fourier analysis shows that these signals are composed of sinusoids satisfying a relation of *harmonicity*, which means that their frequencies are multiples of the fundamental frequency, defined as the inverse of the period. This is the case of voiced speech signals produced by the quasi-periodic vibration of the vocal cords, such as vowels. Many wind or string instruments also produce harmonic or quasi-harmonic sounds. However, in a polyphonic music signal, the sounds emitted simultaneously by one or more instruments overlap; thus the harmonic relationship is no longer verified, but the signal remains essentially made up of sinusoids.

The estimation of sinusoids is a classic problem, more than two hundred years old. In this area, the Fourier transform is a privileged tool because of its robustness, the simplicity of its implementation, and the existence of fast algorithms (*Fast Fourier Transform* (FFT)). However, it has a number of drawbacks. First, its frequency *precision*, that is, the precision with which the frequency of a sinusoid can be estimated, is limited by the number of samples used to calculate it. This first limitation can be circumvented by extending the signal by a series of zeros (this operation is called *zero-padding*). However, its frequency *resolution*, that is to say its ability to distinguish two close sinusoids, remains limited by the duration of the observed signal. Despite these drawbacks, the Fourier transform remains the most used tool in spectral analysis. It has given rise to numerous frequency estimation methods [KM02].

The *High Resolution* (HR) methods, which find their applications both in antenna processing and in spectral analysis [MSAB$^+$98], have the advantage of overcoming the natural limitations of Fourier analysis. Indeed, in the absence of noise, their frequency precision and resolution are virtually infinite (although in practice limited by the finite precision of computers). This is made possible by exploiting a parametric signal model. Thus, unlike the Fourier analysis which consists in representing the signal in a transformed domain, the HR methods are parametric estimation methods. In the context of audio signal processing, despite their superiority in terms of spectral resolution (in particular over short time windows), they remain little used because of their high computational complexity. Nevertheless, the HR methods are well suited for estimating the parameters of a sum of sinusoids whose amplitudes vary exponentially (*Exponential Sinusoidal Model* (ESM) model). This type of modulation makes it possible to describe the natural damping of free vibratory systems, such as the vibration of a plucked string [JHJ04]. On the other hand, it has been shown in [Lar93b] that the HR methods prove to be particularly effective in the case of strongly attenuated signals. More generally, the ESM model makes it possible to describe signals with large amplitude variations [HVW02]. In addition, the music signals often contain pairs or triplets of very close frequencies which generate a beat phenomenon. These beats strongly contribute to the natural appearance of the sound. They often result from the special properties of vibration systems. For example, a minor asymmetry in the geometry of a bell leads to pairs of vibration modes. In the case of a guitar, the coupling between the strings and the bridge can be represented by a so-called mobility matrix, from which it is possible to deduce frequency pairs [LC93]. In the case of the piano, the coupling of the horizontal and vertical vibration modes of each string and the presence of pairs or triplets of strings for most notes explain the presence of four or six neighboring frequencies at the level of each harmonic [Wei77]. The Fourier analysis generally does not make it possible to distinguish all these frequencies. Studies in [Lar93b] on piano and guitar sounds have shown the superiority of HR methods in this area. The same technique was used to estimate physical parameters, such as the radiation factor of a guitar [Dav99], and to study the propagation of mechanical waves in solid materials [JMP98].

1 Introduction

This chapter is devoted to the parametric estimation of a signal composed of a sum of exponentially modulated sinusoids, perturbed by an additive noise. The maximum likelihood principle then reduces the estimation of amplitudes and phases to a simple least squares problem, while the estimation of frequencies and damping factors requires more sophisticated methods, called *high resolution methods*, because they overcome the limits of Fourier analysis in terms of spectral resolution.

The origin of the HR methods dates back to Prony's work published in 1795, which aims to estimate a sum of exponentials by linear prediction techniques [RdP95]. More recently, this approach was further developped by Pisarenko to estimate sinusoids of constant amplitude [Pis73]. In comparison, modern HR methods are based on the particular properties of the signal covariance matrix. Thus, the study of its rank makes it possible to separate the data space into two subspaces, the signal subspace spanned by the sinusoids, and the noise subspace which is its orthogonal complement. The HR methods resulting from this decomposition into subspaces are known to be more robust than linear prediction techniques. This is the case of the MUSIC [Sch86] and root-MUSIC [Bar83] methods (which are based on the noise subspace), of the *Toeplitz Approximation Method* (TAM) algorithm [KAR83], as well as the *Estimation of Signal Parameters via Rotational Invariance Techniques* (ESPRIT) algorithm [RPK86] and its variants TLS-ESPRIT [RK87] and PRO-ESPRIT [ZS89] (which are based on the signal subspace). All of these estimation methods can be applied to the ESM (Exponential Sinusoidal Model), which represents the signal as a sum of exponentially modulated sinusoids. This model is also called *Exponentially Damped Sinusoids* (EDS) when the modulation is decreasing [NHD98]. Other estimation techniques have been specifically developed for the ESM model, such as the *Kumaresan and Tufts* (KT) algorithm, also called Min-Norm method [KT82], its modified version [LLR97] (based on linear prediction), and the *Matrix Pencil* method [HS90] (based on subspaces). A more complete list of methods can be found in [VdVDS93].

This chapter is not intended to present the HR methods exhaustively, but rather to familiarize the reader with the concepts on which they are based. This is why only some of them are presented here: the Prony, Pisarenko, MUSIC and ESPRIT methods. This presentation will start with the definition of the signal model (section 2). Then the maximum likelihood method, which makes it possible to establish a link with the Fourier transform, will be presented in section 3. Then the high resolution methods that estimate the complex poles will be introduced in section 4, and techniques for estimating the other parameters of the model will be presented in section 5. Section 6 will be devoted to the analysis of the performance of HR methods. Finally, the results of this chapter will be summarized in section 7.

2 Signal model

Consider the discrete signal model (defined for all $t \in \mathbb{Z}$)

$$\boxed{s(t) = \sum_{k=0}^{K-1} \alpha_k {z_k}^t} \tag{3.1}$$

where $K \in \mathbb{N}^*$, $\forall k \in \{0 \dots K-1\}$, $\alpha_k \in \mathbb{C}^*$, and the poles $z_k \in \mathbb{C}^*$ are pairwise distinct. In the particular case where all the poles belong to the unit circle, the signal is represented as a sum of complex sinusoids. Thus, each pole z_k is written in the form $z_k = e^{i2\pi f_k}$ where $f_k \in \mathbb{R}$ is the frequency of the sinusoid. More generally, if the poles are not on the unit circle, the sinusoids are exponentially modulated (ESM). In this case, each pole z_k is written in polar form $z_k = e^{\delta_k}\, e^{i2\pi f_k}$, where $\delta_k \in \mathbb{R}$ is the damping factor (or attenuation rate) of the sinusoid. In particular, poles with the same polar angle and different modules are associated with the same frequency. The complex amplitudes α_k are also written in polar form $\alpha_k = a_k\, e^{i\phi_k}$, where $a_k \in \mathbb{R}_+^*$ and $\phi \in \mathbb{R}$.

In addition, the observed signal $x(t)$ can be modeled as the sum of the deterministic signal $s(t)$ defined above and of a complex centered white Gaussian noise $b(t)$ of variance σ^2. Remember that a complex centered white Gaussian noise is a sequence of *i.i.d* random variables with complex values, of probability density $p(b) = \frac{1}{\pi\sigma^2}\, e^{-\frac{|b|^2}{\sigma^2}}$. We thus obtain the relation

$$x(t) = s(t) + b(t). \tag{3.2}$$

The signal is observed over time windows of length $N \geq K$. Thus, for all $t \in \mathbb{Z}$, we consider the time window $\{t-l+1 \ldots t+n-1\}$, where the integers n and l are such that $N = n+l-1$, and we define the vector $\boldsymbol{s}(t) = [s(t-l+1), \ldots, s(t+n-1)]^\top$, of dimension N. For all $z \in \mathbb{C}$, let us define $\boldsymbol{v}(z) = [1, z, \ldots, z^{N-1}]^\top$. However $\boldsymbol{s}(t) = \sum_{k=0}^{K-1} \alpha_k\, z_k^{t-l+1}\, \boldsymbol{v}(z_k)$. This equality can be rewritten in the form of a product: $\boldsymbol{s}(t) = \boldsymbol{V}^N\, \boldsymbol{D}^{t-l+1}\, \boldsymbol{\alpha}$, where $\boldsymbol{\alpha} = [\alpha_0, \ldots, \alpha_{K-1}]^\top$ is a vector of dimension K, $\boldsymbol{D} = \operatorname{diag}(z_0, \ldots, z_{(K-1)})$ is a diagonal matrix of dimension $K \times K$, and $\boldsymbol{V}^N = [\boldsymbol{v}(z_0), \ldots, \boldsymbol{v}(z_{(K-1)})]$ is a Vandermonde matrix of dimensions $N \times K$: (*cf.* definition 12 in appendix 8.2 page 62)

$$\boldsymbol{V}^N = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ z_0 & z_1 & \ldots & z_{K-1} \\ \vdots & \vdots & \vdots & \vdots \\ {z_0}^{N-1} & {z_1}^{N-1} & \ldots & {z_{K-1}}^{N-1} \end{bmatrix}.$$

Then define the vector of amplitudes at time t, $\boldsymbol{\alpha}(t) = \boldsymbol{D}^{t-l+1}\boldsymbol{\alpha}$, so that $\boldsymbol{s}(t) = \boldsymbol{V}^N\boldsymbol{\alpha}(t)$. It is known that the determinant of the square Vandermonde matrix $\boldsymbol{V}^K$ extracted from the first K rows of $\boldsymbol{V}^N$ (remember that $N \geq K$) is (*cf.* proposition 12 in appendix 8.2 page 62)

$$\boxed{\det(\boldsymbol{V}^K) = \prod_{0 \leq k_1 < k_2 \leq K-1} (z_{k_2} - z_{k_1}).} \tag{3.3}$$

Thus, matrix $\boldsymbol{V}^N$ has full rank if and only if all the poles are distinct. The relation $\boldsymbol{s}(t) = \boldsymbol{V}^N\boldsymbol{\alpha}(t)$ therefore shows that for each time t the vector $\boldsymbol{s}(t)$ lies in the range space of matrix $\boldsymbol{V}^N$, of dimension less than or equal to K in the general case, and equal to K if all poles are distinct.

Let $\boldsymbol{b}(t) = [b(t-l+1), \ldots, b(t+n-1)]^\top$ be the vector containing the samples of the additive noise. It is a centered Gaussian random vector, whose covariance matrix is $\boldsymbol{R}_{bb} = \sigma^2\boldsymbol{I}_N$. Finally, let $\boldsymbol{x}(t) = [x(t-l+1), \ldots, x(t+n-1)]^\top$ be the vector of observed data. This vector therefore satisfies $\boldsymbol{x}(t) = \boldsymbol{s}(t) + \boldsymbol{b}(t)$. The model being posed, the analysis of the signal $s(t)$ will consist in estimating the parameters σ^2, $z_0, \ldots, z_{(K-1)}$ and $\boldsymbol{\alpha}(t)$. A classical parametric estimation technique, the maximum likelihood method, is applied to this model in the next section.

3 Maximum likelihood method

The maximum likelihood principle is a general parametric estimation method. It provides asymptotically efficient and unbiased estimators. This is why it is often preferred to other estimation techniques when it has a simple closed-form solution.

3.1 Application of the maximum likelihood principle to the ESM model

The maximum likelihood principle consists in maximizing the conditional probability of observing the signal x over the interval $\{t-l+1, \ldots, t+n-1\}$, given the parameters σ^2, $z_0, \ldots, z_{(K-1)}$ and $\boldsymbol{\alpha}(t)$ (or the natural logarithm of this probability, called *log-likelihood* of the observations). Since $\boldsymbol{x}(t) = \boldsymbol{s}(t) + \boldsymbol{b}(t)$, where $\boldsymbol{s}(t) = \boldsymbol{V}^N\boldsymbol{\alpha}(t)$ is a deterministic vector and $\boldsymbol{b}(t)$ is a centered complex Gaussian random vector of covariance matrix $\boldsymbol{R}_{bb} = \sigma^2\boldsymbol{I}_N$, $\boldsymbol{x}(t)$ is itself a complex Gaussian random vector with expected value $\boldsymbol{s}(t)$ and covariance matrix $\boldsymbol{R}_{bb}$. Remember that the probability density of such a random vector is

$$p(\boldsymbol{x}(t)) = \frac{1}{\pi^N \det(\boldsymbol{R}_{bb})}\, e^{-(\boldsymbol{x}(t)-\boldsymbol{s}(t))^H \boldsymbol{R}_{bb}^{-1}(\boldsymbol{x}(t)-\boldsymbol{s}(t))}.$$

So the log-likelihood of the observations is

$$L(\sigma^2, z_0 \dots z_{K-1}, \boldsymbol{\alpha}(t)) = -N \ln(\pi\sigma^2) - \frac{1}{\sigma^2}\, g(z_0 \dots z_{K-1}, \boldsymbol{\alpha}(t))$$

where

$$g(z_0 \dots z_{K-1}, \boldsymbol{\alpha}(t)) = \left(\boldsymbol{x}(t) - \boldsymbol{V}^N \boldsymbol{\alpha}(t)\right)^H \left(\boldsymbol{x}(t) - \boldsymbol{V}^N \boldsymbol{\alpha}(t)\right).$$

Maximizing this log-likelihood with respect to the parameters $(\sigma^2, z_0 \dots z_{K-1}, \boldsymbol{\alpha}(t))$ can be done by first minimizing g with respect to the pair $(z_0 \dots z_{K-1},\ \boldsymbol{\alpha}(t))$, then by maximizing L with respect to σ. We thus obtain $\sigma^2 = \frac{1}{N}\, g(z_0 \dots z_{K-1}, \boldsymbol{\alpha}(t))$, i.e.

$$\boxed{\sigma^2 = \tfrac{1}{N} \left\|\boldsymbol{x}(t) - \boldsymbol{V}^N \boldsymbol{\alpha}(t)\right\|^2.} \tag{3.4}$$

It appears that σ^2 is estimated by calculating the power of the residual obtained by subtracting the exponentials from the observed signal.

The matrix $\boldsymbol{V}^N$ has full rank, since it has been assumed in section 2 that the poles are pairwise distinct. Thus, matrix ${\boldsymbol{V}^N}^H \boldsymbol{V}^N$ is invertible. In order to minimize g with respect to the pair $(z_0 \dots z_{K-1},\ \boldsymbol{\alpha}(t))$, we just use the decomposition

$$\begin{aligned} g(z_0 \dots z_{K-1}, \boldsymbol{\alpha}(t)) &= \boldsymbol{x}(t)^H \boldsymbol{x}(t) - \boldsymbol{x}(t)^H \boldsymbol{V}^N \left({\boldsymbol{V}^N}^H \boldsymbol{V}^N\right)^{-1} {\boldsymbol{V}^N}^H \boldsymbol{x}(t) \\ &+ \left(\boldsymbol{\alpha}(t) - \left({\boldsymbol{V}^N}^H \boldsymbol{V}^N\right)^{-1} {\boldsymbol{V}^N}^H \boldsymbol{x}(t)\right)^H \left({\boldsymbol{V}^N}^H \boldsymbol{V}^N\right) \left(\boldsymbol{\alpha}(t) - \left({\boldsymbol{V}^N}^H \boldsymbol{V}^N\right)^{-1} {\boldsymbol{V}^N}^H \boldsymbol{x}(t)\right). \end{aligned}$$

The last term of this equation is always non-negative, and can be zeroed by defining

$$\boxed{\boldsymbol{\alpha}(t) = \left({\boldsymbol{V}^N}^H \boldsymbol{V}^N\right)^{-1} {\boldsymbol{V}^N}^H \boldsymbol{x}(t).} \tag{3.5}$$

It appears that the vector of complex amplitudes $\boldsymbol{\alpha}(t)$ is estimated in the same way as with the ordinary least squares method.

Function g is therefore minimal when the K -tuple $(z_0 \dots z_{K-1})$ maximizes function $\mathcal{J}$ defined by

$$\boxed{\mathcal{J}(z_0, \dots, z_{(K-1)}) = \boldsymbol{x}(t)^H \boldsymbol{V}^N \left({\boldsymbol{V}^N}^H \boldsymbol{V}^N\right)^{-1} {\boldsymbol{V}^N}^H \boldsymbol{x}(t).} \tag{3.6}$$

As this optimization problem does not have a closed-form solution in the general case, it must be solved numerically. In summary, the maximum likelihood principle leads to estimating the parameters of the model in three stages:

complex poles are obtained by maximizing function $\mathcal{J}$ (equation (3.6)),

complex amplitudes are obtained by calculating the right side of equation (3.5),

the standard deviation is then given by equation (3.4).

Unfortunately, it turns out that the first step of this estimation method, which requires the optimization of a function of K complex variables, is difficult to implement, because the function to be maximized has many local maxima. In addition, it is extremely costly in terms of computation time. This is why we generally use more reliable and faster methods to estimate complex poles. However, once the poles are estimated, the maximum likelihood principle can be used to determine the complex amplitudes and the standard deviation of the noise.

3.2 Maximum likelihood and Fourier resolution

Let us now take a look at the particular case where all the poles are on the unit circle ($\forall k,\ \delta_k = 0$). The results of section 3.1 showed that the maximum likelihood principle leads to an optimization problem which does not have a simple closed-form solution in the general case. However, such a solution exists in the particular case where $K = 1$, as well as an approximate solution if $K > 1$.

Let us first examine the case of a single complex exponential ($K = 1$). Then equation (3.6) is simplified as $\mathcal{J}(z_0) = \widehat{R}_x(z_0)$, where $\widehat{R}_x$ is the periodogram of the signal $x(t)$ observed on the time window $\{t-l+1 \dots t+n-1\}$:

$$\widehat{R}_x(e^{i2\pi f_0}) = \frac{1}{N}\left|X(e^{i2\pi f_0})\right|^2$$

where $X(e^{i2\pi f_0}) = \boldsymbol{v}(e^{i2\pi f_0})^H \boldsymbol{x}(t) = \sum\limits_{\tau=0}^{N-1} x(t-l+1+\tau)\, e^{-i2\pi f_0 \tau}$. Similarly, equation (3.5) is simplified as $\alpha_0(t) = \frac{1}{N}\, X(e^{i2\pi f_0})$. Finally, equation (3.4) is simplified as $\sigma^2 = \frac{1}{N}\left(\|\boldsymbol{x}(t)\|^2 - \widehat{R}_x(e^{i2\pi f_0})\right)$.

These results lead to the following conclusion:

The maximum likelihood principle leads in the case of a complex sinusoid to detect the frequency at which the periodogram reaches its maximum. The corresponding complex amplitude is proportional to the value of the DFT of the signal at this frequency. The noise variance is estimated as the signal average power after subtracting the sinusoid.

Let us now address the general case $K \geq 1$, for which the maximization of function $\mathcal{J}(z)$ no longer has an exact closed-form solution. We then introduce the following hypothesis:

$$N >> \frac{1}{\min\limits_{k_1 \neq k_2} \left|f_{k_2} - f_{k_1}\right|}.$$

Matrix $\boldsymbol{V}^{N^H}\boldsymbol{V}^N$ is a positive definite Hermitian matrix of dimension $K \times K$, whose entries can be calculated in closed-form: $\left\{\boldsymbol{V}^{N^H}\boldsymbol{V}^N\right\}_{(k_1,\,k_2)} = \sum\limits_{\tau=0}^{N-1} (\overline{z_{k_1}} z_{k_2})^\tau$. We then obtain

$$\begin{array}{lcll} \frac{1}{N}\left\{\boldsymbol{V}^{N^H}\boldsymbol{V}^N\right\}_{(k_1,\,k_2)} & = & e^{i\pi(N-1)(f_{k_2}-f_{k_1})}\, \frac{\sin(\pi N(f_{k_2}-f_{k_1}))}{N\sin(\pi(f_{k_2}-f_{k_1}))} & \text{if } k_1 \neq k_2 \\ \frac{1}{N}\left\{\boldsymbol{V}^{N^H}\boldsymbol{V}^N\right\}_{(k,\,k)} & = & 1 & \text{if } k_1 = k_2 = k \end{array}$$

Therefore, when $N >> \frac{1}{\min\limits_{k_1 \neq k_2}|f_{k_2}-f_{k_1}|}$, $\frac{1}{N}\boldsymbol{V}^{N^H}\boldsymbol{V}^N = \boldsymbol{I}_K + O\left(\frac{1}{N}\right)$, thus

$$\left(\boldsymbol{V}^{N^H}\boldsymbol{V}^N\right)^{-1} = \frac{1}{N}\boldsymbol{I}_K + O\left(\frac{1}{N^2}\right).$$

Then equation (3.6) is simplified as

$$\mathcal{J}(z_0, \dots, z_{K-1}) = \frac{1}{N}\left\|\boldsymbol{V}^{N^H}\boldsymbol{x}(t)\right\|^2 + O\left(\frac{1}{N^2}\right) = \sum_{k=0}^{K-1} \widehat{R}(z_k) + O\left(\frac{1}{N^2}\right).$$

Similarly, equation (3.5) is simplified as $\boldsymbol{\alpha}(t) = \frac{1}{N}\, \boldsymbol{V}^{N^H}\boldsymbol{x}(t) + O\left(\frac{1}{N^2}\right)$, hence

$$\alpha_k(t) = \frac{1}{N}\, X(e^{i2\pi f_k}) + O\left(\frac{1}{N^2}\right).$$

Finally, equation (3.4) is simplified as $\sigma^2 = \frac{1}{N}\left(\|\boldsymbol{x}(t)\|^2 - \sum\limits_{k=0}^{K-1} \widehat{R}(e^{i2\pi f_k})\right) + O\left(\frac{1}{N^2}\right)$.

Thus, the joint maximization of $\mathcal{J}$ with respect to $z_0, \dots, z_{K-1}$ leads to determine the K frequencies corresponding to the K largest values of the periodogram. The corresponding complex amplitudes are proportional to the value of the DFT of the signal at these frequencies. Remember that these results are only valid if all the poles are on the unit circle and are based on the assumption $N >> \frac{1}{\min\limits_{k_1 \neq k_2}|f_{k_2}-f_{k_1}|}$.

We thus observe the limit of the Fourier analysis in terms of spectral resolution: the parameters are estimated correctly provided that the length of the observation window is sufficiently large compared to the inverse of the smallest frequency difference between two neighboring poles. It is this limit that the HR methods presented in section 4 allow to overcome. So, HR methods are able to distinguish two close sinusoids, that Fourier analysis does not allow to distinguish. In applications, HR methods can be used with shorter windows than those usually used with Fourier analysis.

Figure 3.1: Jean Baptiste Joseph FOURIER (1768-1830)

4 High resolution methods

We begin here by introducing the oldest high-resolution methods, which are based on linear prediction techniques (section 4.1), before addressing in section 4.2 the more recent subspace methods.

4.1 Linear prediction techniques

The first two high-resolution methods presented in this chapter are based on a fundamental result related to linear recurrence equations, presented in section 4.1.1.

4.1.1 Linear recurrence equations

Let $p_0 \in \mathbb{C}^*$, $K \in \mathbb{N}^*$, and $\{z_0, \ldots, z_{K-1}\}$ be K distinct and non-zero complex numbers. We define the polynomial of degree K whose dominant coefficient is p_0 and whose roots are z_k:

$$P[z] = p_0 \prod_{k=0}^{K-1} (z - z_k) = \sum_{\tau=0}^{K} p_{K-\tau}\, z^{\tau}.$$

The following theorem characterizes the signal model.

Theorem 4. *A complex discrete signal $\{s(t)\}_{t\in\mathbb{Z}}$ satisfies the recurrence equation*

$$\boxed{\sum_{\tau=0}^{K} p_\tau\, s(t-\tau) = 0} \tag{3.7}$$

for all $t \in \mathbb{Z}$ if and only if there are scalars $\alpha_0, \ldots, \alpha_{K-1} \in \mathbb{C}$ such that $s(t) = \sum_{k=0}^{K-1} \alpha_k\, {z_k}^t$.

Proof. First of all, it is straightforward to check that the set of signals which satisfy the relation (3.7) forms a vector space E over $\mathbb{C}$. Next, we will prove that this vector space has dimension less than or equal to K. Consider the application

$$\begin{array}{rccc} f: & E & \to & \mathbb{C}^K \\ & s[t] & \mapsto & [s[0], \ldots, s[K-1]]^\top \end{array}$$

We can notice that f is a linear map. Let $s \in E$ be a signal such that $f(s) = \mathbf{0}$. Then s is zero over the interval $[0, K-1]$. By using the recurrence (3.7), we deduce that s is also zero over the interval $[K, +\infty[$. Finally, using the recurrence (3.7) and the fact that $p_K \neq 0$, we show that s is also zero over the interval $]-\infty, -1]$. Consequently, $s \equiv 0$, so the linear map f is injective. We conclude that the vector space E is at most of dimension K.

Now we will show that any signal of the form $s[t] = {z_k}^t$ where $k \in \{0, \ldots, K-1\}$ belongs to the vector space E. Indeed, if $s[t] = {z_k}^t$, then $\forall t \in \mathbb{Z}$, $\sum_{k=0}^{K} p_k\, s[t-k] = {z_k}^{t-K} \sum_{k=0}^{K} p_k\, {z_k}^{K-k} = {z_k}^{t-K} P[z_k] = 0$, therefore $s[t]$ satisfies the relationship (3.7).

Finally, consider the family of vectors $\{{z_k}^t\}_{\{k\in\{0,\ldots,K-1\}\}}$. The square matrix whose columns are extracted from these vectors and whose rows correspond to times $\{0 \ldots K-1\}$ is a Vandermonde matrix (*cf.* definition 12 of the appendix 8.2 page 62). According to proposition 12 in appendix 8.2, it is invertible, since the poles z_k are pairwise distinct. Therefore, the family $\{{z_k}^t\}_{\{k\in\{0,\ldots,K-1\}\}}$ is linearly independent. However it contains precisely K vectors of E. This vector space is therefore exactly of dimension K, and this family forms a basis of it. Thus, a signal $s[t]$ belongs to E if and only if it is of the form (3.2). □ □

4.1.2 Prony method

The work of Baron de Prony is at the origin of the development of high resolution methods. He proposed an estimation method inspired by the previous result on the linear recurrence equations [RdP95]. This method was originally intended to estimate noiseless real exponentials; however we apply it here to the estimation of noisy complex exponentials. Prony's method consists in first determining the polynomial $P[z]$ using linear prediction techniques, then extracting the roots of this polynomial. We define the prediction error

$$\varepsilon(t) \triangleq \sum_{\tau=0}^{K} p_\tau\, x(t-\tau). \tag{3.8}$$

In particular, by substituting equations (3.2) and (3.7) into equation (3.8), we get $\varepsilon(t) = \sum_{\tau=0}^{K} p_\tau\, b(t-\tau)$. The prediction error therefore characterizes only the noise which is superimposed on the signal. Let us address the particular case $n = K+1$, and suppose that $l \geq K+1$. Thus, the signal is observed on the window $\{t-l+1 \ldots t+K\}$. By applying equation (3.8) at times $\{t-l+K+1,\ t-l+K+2,\ \ldots,\ t+K\}$, we get the system of equations

$$\left\{\begin{array}{ccccccccc} p_0\, x(t-l+K+1) & + & p_1\, x(t-l+K) & + & \ldots & + & p_K\, x(t-l+1) & = & \varepsilon(t-l+K+1) \\ p_0\, x(t-l+K+2) & + & p_1\, x(t-l+K+1) & + & \ldots & + & p_K\, x(t-l+2) & = & \varepsilon(t-l+K+2) \\ \vdots & + & \vdots & + & \ldots & + & \vdots & = & \vdots \\ p_0\, x(t+K) & + & p_1\, x(t+K-1) & + & \ldots & + & p_K\, x(t) & = & \varepsilon(t+K) \end{array}\right. \tag{3.9}$$

Figure 3.2: Gaspard-Marie RICHE de PRONY (1755-1839)

Then define $\boldsymbol{p} = [p_K,\ p_{(K-1)}, \dots, p_0]^H$, $\boldsymbol{\varepsilon}(t) = [\varepsilon(t-l+K+1), \varepsilon(t-l+K+2), \dots, \varepsilon(t+K)]^H$ and

$$\boldsymbol{X}(t) = \begin{bmatrix} x(t-l+1) & \cdots & x(t-1) & x(t) \\ x(t-l+2) & \cdots & x(t) & x(t+1) \\ \vdots & \cdots & \vdots & \vdots \\ x(t-l+K+1) & \cdots & x(t+K-1) & x(t+K) \end{bmatrix} \tag{3.10}$$

so that the system of equations (3.9) can be condensed in the form $\boldsymbol{p}^H \boldsymbol{X}(t) = \boldsymbol{\varepsilon}(t)^H$.

Prony's method consists in minimizing the power of the prediction error $\frac{1}{l}\,||\boldsymbol{\varepsilon}||^2$ with respect to $\boldsymbol{p}$, subject to the constraint $p_0 = 1$. However it is possible to write $\frac{1}{l}||\boldsymbol{\varepsilon}||^2 = \boldsymbol{p}^H \widehat{\boldsymbol{R}}_{xx}(t)\,\boldsymbol{p}$, where matrix $\widehat{\boldsymbol{R}}_{xx}(t) = \frac{1}{l}\,\boldsymbol{X}(t)\,\boldsymbol{X}(t)^H$ has dimension $(K+1) \times (K+1)$. As matrix $\boldsymbol{X}(t)$ has $K+1$ rows and $l \geq K+1$ columns, we can assume that matrix $\widehat{\boldsymbol{R}}_{xx}(t)$ is invertible.

Theorem 11 in appendix 8.1 page 61 allows to prove [1] that the solution of this optimization problem is [2]

$$\boldsymbol{p} = \frac{1}{\boldsymbol{e}_1^H \widehat{\boldsymbol{R}}_{xx}(t)^{-1} \boldsymbol{e}_1}\,\widehat{\boldsymbol{R}}_{xx}(t)^{-1} \boldsymbol{e}_1$$

where $\boldsymbol{e}_1 \triangleq [1,\ 0 \dots 0]^\top$ is a vector of dimension $K+1$. Thus, the Prony estimation method includes the following steps:

- Construct matrix $\boldsymbol{X}(t)$ and calculate $\widehat{\boldsymbol{R}}_{xx}(t)$;
- Compute $\boldsymbol{p} = \frac{1}{\boldsymbol{e}_1^H \widehat{\boldsymbol{R}}_{xx}(t)^{-1} \boldsymbol{e}_1}\,\widehat{\boldsymbol{R}}_{xx}(t)^{-1} \boldsymbol{e}_1$;
- Determine the poles $\{z_0,\ \dots,\ z_{K-1}\}$ as the roots of the polynomial $P[z] = \sum_{k=0}^{K} p_k\, z^{K-k}$.

[1] As the data are complex, it is necessary to decompose the vector $\boldsymbol{p}$ into its real part and its imaginary part to be able to apply theorem 11, which deals exclusively with the real data case.

[2] The scalar $\boldsymbol{e}_1^H \widehat{\boldsymbol{R}}_{xx}(t)^{-1} \boldsymbol{e}_1$ is non-zero, since the vector $\boldsymbol{e}_1$ is unitary and matrix $\widehat{\boldsymbol{R}}_{xx}(t)$ is positive definite.

4.1.3 Pisarenko method

The Pisarenko method is a variant of the Prony method. It consists in minimizing the power of the prediction error $\frac{1}{l}\|\boldsymbol{\varepsilon}\|^2 = \boldsymbol{p}^H \widehat{\boldsymbol{R}}_{xx}(t)\, \boldsymbol{p}$ subject to the constraint that vector $\boldsymbol{p}$ has norm 1. Theorem 11 in appendix 8.1 page 61 allows us to prove that the solution of this optimization problem is the eigenvector of matrix $\widehat{\boldsymbol{R}}_{xx}(t)$ associated with the smallest eigenvalue.

Thus the Pisarenko method [Pis73] consists of the following stages:

- calculate and diagonalize $\widehat{\boldsymbol{R}}_{xx}(t)$;
- determine $\boldsymbol{p}$ as the eigenvector associated with the smallest eigenvalue;
- extract the roots of polynomial $P[z]$.

The Prony and Pisarenko methods are the oldest HR methods. As we will show in section 6.2, they do not prove to be very robust in practice, this is why the subspace methods, proposed more recently, are generally preferred to them.

4.2 Subspace methods

In the same spirit as the Pisarenko method, modern HR methods (for example [Sch86, RPK86, HS90]) are based on a decomposition of matrix $\widehat{\boldsymbol{R}}_{xx}(t)$.

4.2.1 Singular structure of the data matrix

Suppose now that $n \geq K+1$ and $l \geq K+1$, and construct the data matrix of the noiseless signal $s(t)$ on the same model as matrix $\boldsymbol{X}(t)$ in equation (3.10), according to a Hankel structure:

$$\boldsymbol{S}(t) = \begin{bmatrix} s(t-l+1) & \cdots & s(t-1) & s(t) \\ s(t-l+2) & \cdots & s(t) & s(t+1) \\ \vdots & \cdots & \vdots & \vdots \\ s(t-l+n) & \cdots & s(t+n-2) & s(t+n-1) \end{bmatrix}. \tag{3.11}$$

The following proposition characterizes the signal model.

Proposition 5 (Factorization of the data matrix)**.** *The following assertions are equivalent:*

1. *The signal $s(t)$ satisfies the model defined in equation* (3.1) *on the interval* $\{t-l+1,\dots,t+n-1\}$*;*
2. *The matrix $\boldsymbol{S}(t)$ defined in equation* (3.11) *can be factorized as*

$$\boxed{\boldsymbol{S}(t) = \boldsymbol{V}^n\, \boldsymbol{A}(t)\, \boldsymbol{V}^{l\top}} \tag{3.12}$$

where the diagonal matrix $\boldsymbol{A}(t) = \mathrm{diag}(z_0^{t-l+1}\alpha_0,\dots,z_{(K-1)}^{t-l+1}\alpha_{(K-1)})$ has dimension $K\times K$, $\boldsymbol{V}^n$ has dimension $n\times K$, and $\boldsymbol{V}^l$ has dimensions $l\times K$.

Proof of Proposition 5: Let us prove each of the two implications.

Proof of 1. ⇒ 2. The signal $s(t)$ is defined as a sum of complex exponentials:$s(t) = \sum\limits_{k=0}^{K-1} s_k(t)$, where $s_k(t) = \alpha_k\, {z_k}^t$. Consequently, matrix $\boldsymbol{S}(t)$ defined in equation (3.11) can be decomposed in the same way: $\boldsymbol{S}(t) = \sum\limits_{k=0}^{K-1} \boldsymbol{S}_k(t)$, where matrices $\boldsymbol{S}_k(t)$ are constructed in the same way as $\boldsymbol{S}(t)$ in equation (3.11), from the signals $s_k(t)$. However it is straightforward to check that $\boldsymbol{S}_k(t) = \alpha_k\, z_k^{t-l+1}\, \boldsymbol{v}^n(z_k)\, \boldsymbol{v}^l(z_k)^\top$, where $\boldsymbol{v}^n(z) = [1, z, \dots, z^{n-1}]^\top$ and $\boldsymbol{v}^l(z) = [1, z, \dots, z^{l-1}]^\top$. Equation (3.12) follows directly.

Proof of 1. ⇒ 2. If matrix $\boldsymbol{S}(t)$ is defined by equation (3.12), the reverse reasoning allows to show that $\boldsymbol{S}(t)$ can be written in the form (3.11), where $s(t)$ is defined in equation (3.1). □ □

Proposition 5 induces a result on which all subspace methods are based:

Corollary 6. *Matrix $\boldsymbol{S}(t)$ defined in equation* (3.11) *has rank less than or equal to K. More precisely, it has rank K if and only if $n \geq K$, $l \geq K$, all the poles z_k are distinct and non-zero, and all the amplitudes α_k are non-zero. In this case, its range space is spanned by matrix $\boldsymbol{V}^n$.*

Proof of Corollary 6: It turns out that matrices $\boldsymbol{V}^n$ and $\boldsymbol{V}^l$ have full rank equal to K. Indeed, if $\boldsymbol{V}_0$ is the Vandermonde matrix made up of the first K rows of $\boldsymbol{V}^n$ or $\boldsymbol{V}^l$, proposition 12 shows that $\boldsymbol{V}_0$ is invertible, since the poles z_k are pairwise distinct. Consequently, $\operatorname{rank}(\boldsymbol{V}^n) \geq K$ and $\operatorname{rank}(\boldsymbol{V}^l) \geq K$. However $\dim(\boldsymbol{V}^n) = n \times K$ and $\dim(\boldsymbol{V}^l) = l \times K$, therefore $\operatorname{rank}(\boldsymbol{V}^n) = \operatorname{rank}(\boldsymbol{V}^l) = K$. Furthermore, matrix $\boldsymbol{A}(t)$, of dimensions $K \times K$, is invertible, hence of rank K.

We can deduce from this remark that matrix $\boldsymbol{S}(t)$ is also of rank K. To do this, let us first show that $\operatorname{Ker}(\boldsymbol{S}(t)) = \operatorname{Ker}({\boldsymbol{V}^l}^\top)$. Indeed,

- $\forall \boldsymbol{y} \in \mathbb{C}^n$, ${\boldsymbol{V}^l}^\top \boldsymbol{y} = \boldsymbol{0} \Rightarrow \boldsymbol{V}^n\, \boldsymbol{A}(t) {\boldsymbol{V}^l}^\top \boldsymbol{y} = \boldsymbol{0}$;
- $\forall \boldsymbol{y} \in \mathbb{C}^n$, $\boldsymbol{V}^n\, \boldsymbol{A}(t) {\boldsymbol{V}^l}^\top \boldsymbol{y} = \boldsymbol{0} \Rightarrow \left({\boldsymbol{V}^n}^H \boldsymbol{V}^n\right) \boldsymbol{A}(t) {\boldsymbol{V}^l}^\top \boldsymbol{y} = \boldsymbol{0}$. However matrix $\left({\boldsymbol{V}^n}^H \boldsymbol{V}^n\right) \boldsymbol{A}(t)$ is invertible, hence ${\boldsymbol{V}^l}^\top \boldsymbol{y} = \boldsymbol{0}$.

The rank-nullity theorem then implies: $\operatorname{rank}(\boldsymbol{S}(t)) = n - \dim(\operatorname{Ker}(\boldsymbol{S}(t))) = n - \dim(\operatorname{Ker}({\boldsymbol{V}^l}^\top) = \operatorname{rank}({\boldsymbol{V}^l}^\top) = K$. □ □

The singular structure of the data matrix induces an equivalent structure for the correlation matrix, defined below.

4.2.2 Singular structure of the correlation matrix

The subspace methods are based on the particular structure of the signal correlation matrix $\boldsymbol{C}_{ss}(t) = \boldsymbol{S}(t)\, \boldsymbol{S}(t)^H$, and in particular on its eigensubspaces, that we will now study. Let us define $\boldsymbol{R}_{ss}(t) = \frac{1}{l}\, \boldsymbol{C}_{ss}(t)$. Equation (3.12) shows that

$$\boxed{\boldsymbol{R}_{ss}(t) = \boldsymbol{V}^n\, \boldsymbol{P}(t)\, {\boldsymbol{V}^n}^H} \tag{3.13}$$

where

$$\boldsymbol{P}(t) = \frac{1}{l}\, \boldsymbol{A}(t)\, {\boldsymbol{V}^l}^\top \overline{\boldsymbol{V}^l}\, \boldsymbol{A}(t)^H \tag{3.14}$$

is a symmetric positive definite matrix. Thus, equation (3.13) shows that under the same assumptions as for $\boldsymbol{S}(t)$, matrix $\boldsymbol{R}_{ss}(t)$ has rank K.

> The range space of matrix $\boldsymbol{R}_{ss}(t)$, of dimension K, is spanned by matrix $\boldsymbol{V}^n$. This vector space is called *signal subspace* in the literature.

Then let $\{\boldsymbol{w}_m\}_{m=0\ldots n-1}$ be an orthonormal basis of eigenvectors of matrix $\boldsymbol{R}_{ss}(t)$, associated to the eigenvalues $\lambda_0 \geq \lambda_1 \geq \ldots \geq \lambda_{n-1} \geq 0$. Since matrix $\boldsymbol{R}_{ss}(t)$ has only of rank K, we actually have $\lambda_m = 0\ \forall m \geq K$. We denote by $\boldsymbol{W}(t)$ matrix $[\boldsymbol{w}_0 \ldots \boldsymbol{w}_{K-1}]$, and by $\boldsymbol{W}_\perp(t)$ matrix $[\boldsymbol{w}_K \ldots \boldsymbol{w}_{n-1}]$. We can then check that ${}^\top\mathcal{I}\mathrm{m}(\boldsymbol{W}(t)) = {}^\top\mathcal{I}\mathrm{m}(\boldsymbol{V}^n)$. Indeed, vectors $\{\boldsymbol{w}_m\}_{m=0\ldots K-1}$ are eigenvectors of matrix $\boldsymbol{R}_{ss}(t) = \boldsymbol{V}^n \boldsymbol{P}(t) {\boldsymbol{V}^n}^H$ associated with nonzero eigenvalues. So $\forall\, k \in \{0 \ldots K-1\}$, $\boldsymbol{w}_m \in {}^\top\mathcal{I}\mathrm{m}(\boldsymbol{V}^n)$. Thus ${}^\top\mathcal{I}\mathrm{m}(\boldsymbol{W}(t)) \subset {}^\top\mathcal{I}\mathrm{m}(\boldsymbol{V}^n)$. However matrices $\boldsymbol{W}(t)$ and $\boldsymbol{V}^n$ have the same rank K, therefore ${}^\top\mathcal{I}\mathrm{m}(\boldsymbol{W}(t)) = {}^\top\mathcal{I}\mathrm{m}(\boldsymbol{V}^n)$.

> Matrix $\boldsymbol{W}(t)$ is another basis of the signal subspace, generally distinct from $\boldsymbol{V}^n$.

We then define matrix $\boldsymbol{X}(t)$ from the samples of the noisy signal $x(t)$, in the same way as matrix $\boldsymbol{S}(t)$ in equation (3.11), and we consider the correlation matrix

$$\boldsymbol{C}_{xx}(t) = \boldsymbol{X}(t)\, \boldsymbol{X}(t)^H. \tag{3.15}$$

Then let $\widehat{\boldsymbol{R}}_{xx}(t) = \frac{1}{l}\boldsymbol{C}_{xx}(t)$ (as in section 4.1.2). Since the additive noise $b(t)$ is white and centered, of variance σ^2, matrix $\boldsymbol{R}_{xx}(t) = \mathbb{E}[\widehat{\boldsymbol{R}}_{xx}(t)]$ is such that

$$\boldsymbol{R}_{xx}(t) = \boldsymbol{R}_{ss}(t) + \sigma^2 \boldsymbol{I}_n. \tag{3.16}$$

Using equation (3.16), we show that the family $\{\boldsymbol{w}_m\}_{m=0\dots n-1}$ defined above is also an orthonormal basis of eigenvectors of matrix $\boldsymbol{R}_{ss}(t)$, associated with the eigenvalues

$$\overline{\lambda}_m = \begin{cases} \lambda_m + \sigma^2 \ \forall m \in \{0, \dots K-1\} \\ \sigma^2 \ \forall m \in \{K, \dots n-1\} \end{cases} .$$

Thus, all the eigenvectors of matrix $\boldsymbol{R}_{ss}(t)$ are also eigenvectors of $\boldsymbol{R}_{xx}(t)$, and the corresponding eigenvalues of $\boldsymbol{R}_{xx}(t)$ are equal to those of $\boldsymbol{R}_{ss}(t)$ plus σ^2. Consequently, the signal subspace, defined as the range space of matrix $\boldsymbol{R}_{ss}(t)$, is also the principal subspace of dimension K of matrix $\boldsymbol{R}_{xx}(t)$, i.e. the eigensubspace of $\boldsymbol{R}_{xx}(t)$ associated with the K largest eigenvalues, all strictly greater than σ^2. The $n-K$ eigenvalues associated with the orthogonal complement of the signal subspace, called *noise subspace*, are all equal to σ^2. Thus, it is possible to estimate the signal subspace and the noise subspace by calculating the *EigenValue Decomposition* (EVD) of matrix $\widehat{\boldsymbol{R}}_{xx}(t)$, or even the *Singular Value Decomposition* (SVD) of $\boldsymbol{X}(t)$. By concatenating the K principal eigen or singular vectors of one of these matrices, we thus obtain matrix $\boldsymbol{W}(t) = [\boldsymbol{w}_0 \dots \boldsymbol{w}_{K-1}]$ of dimensions $n \times K$ spanning the signal subspace, and by concatenating the $n-K$ other vectors, we obtain matrix $\boldsymbol{W}_\perp(t) = [\boldsymbol{w}_K \dots \boldsymbol{w}_{n-1}]$ of dimensions $n \times (nK)$ spanning the noise subspace.

The idea of decomposing the data space into two subspaces (signal and noise) is the source of several high-resolution methods, including the MUSIC method, presented in section 4.2.4, and the ESPRIT method, presented in section 4.2.5.

4.2.3 Complement: analogy between the spectrum in the matrix sense and in the Fourier sense

We examine here the particular case where all poles are on the unit circle ($\forall k$, $\delta_k = 0$) and where all frequencies f_k are both multiple of $\frac{1}{n}$ and $\frac{1}{l}$ (we will consider the results that we will get as asymptotic results). We denote by $X\left(e^{i2\pi\frac{\nu}{n}}\right)$ the DFT of the signal observed on the window $\{t, \dots, t+n-1\}$:

$$X\left(e^{i2\pi\frac{\nu}{n}}\right) = \sum_{\tau=0}^{n-1} x(t+\tau)\, e^{-i2\pi\frac{\nu}{n}}.$$

We then define the periodogram of the signal as follows

$$\widehat{R}_x\left(e^{i2\pi\frac{\nu}{n}}\right) = \frac{1}{n}\left|X\left(e^{i2\pi\frac{\nu}{n}}\right)\right|^2.$$

Since all frequencies f_k are multiple of $\frac{1}{n}$, the discrete spectrum $R_x\left(e^{i2\pi\frac{\nu}{n}}\right) \triangleq \mathbb{E}\left[\widehat{R}_x\left(e^{i2\pi\frac{\nu}{n}}\right)\right]$ is such that

$$R_x\left(e^{i2\pi\frac{\nu}{n}}\right) = \sigma^2 + \sum_{k=0}^{K-1} a_k^2\, \mathbf{1}_{\left\{e^{i2\pi\frac{\nu}{n}} = z_k\right\}}. \tag{3.17}$$

Furthermore, since all frequencies f_k are multiple of $\frac{1}{l}$, we can verify that

$$\boldsymbol{P}(t) = \mathrm{diag}(a_0^2, \dots, a_{(K-1)}^2).$$

Therefore, $\boldsymbol{R}_{xx}(t) = \sigma^2\boldsymbol{I}_n + \sum_{k=0}^{K-1} a_k^2 \boldsymbol{v}(z_k)\,\boldsymbol{v}(z_k)^H$. It is then assumed without loss of generality that the a_k are sorted in decreasing order. However, the family $\{\boldsymbol{v}(z_k)\}_{k=0\dots K-1}$ is orthogonal. Therefore $\forall m \in \{0 \dots K-1\}$, $\boldsymbol{R}_{xx}(t)\boldsymbol{v}(z_k) = (a_k^2 + \sigma^2)\boldsymbol{v}\boldsymbol{b}(z_k)$. Thus, $\{\boldsymbol{v}(z_k)\}_{k=0\dots K-1}$ forms a family of eigenvectors of matrix $\boldsymbol{R}_{xx}(t)$, associated to the eigenvalues $\{a_k^2 + \sigma^2\}_{k=0\dots K-1}$. By completing this family with the other vectors of the n-th order Fourier basis, we obtain a basis of eigenvectors of matrix $\boldsymbol{R}_{xx}(t)$ (all other eigenvectors being associated with the same eigenvalue σ^2).

Consequently, the spectrum of the eigenvalues of matrix $\boldsymbol{R}_{xx}(t)$ matches the discrete spectrum given in equation (3.17) (whose periodogram forms an estimator).

4.2.4 MUltiple SIgnal Characterization (MUSIC)

The MUSIC method, developed by R. O. Schmidt [Sch81], is based on the following remark: the poles $\{z_k\}_{k=0\dots K-1}$ are the only solutions of equation

$$||\boldsymbol{W}_\perp(t)^H \boldsymbol{v}(z)||^2 = 0 \tag{3.18}$$

where $\boldsymbol{v}(z) = [1, z, , \dots, z^{n-1}]^\top$. Indeed, z is solution if and only if $\boldsymbol{v}(z) \in \text{Span}(\boldsymbol{W}(t)) = \text{Span}(\boldsymbol{V}^n)$. So every pole z_k is a solution, and there can be no other because otherwise the signal subspace would be of dimension strictly larger than K. So the *root-MUSIC* [Bar83] method consists of the following stages:

- calculate and diagonalize matrix $\widehat{\boldsymbol{R}}_{xx}(t)$;
- deduce a basis of the noise subspace $\boldsymbol{W}_\perp(t)$;
- extract the roots of equation (3.18).

In the particular case where the noise subspace has dimension 1, it is equivalent to the Pisarenko method presented in section 4.1.3.

In practice, real signals do not strictly correspond to the model, and equation (3.18) is not rigorously verified. This is why the *spectral-MUSIC* [Sch86] method rather consists in searching for the K highest peaks of function $\widehat{S}(z) = \frac{1}{||\boldsymbol{W}_\perp^H \boldsymbol{v}(z)||^2}$. The *spectral-MUSIC* method is illustrated in figure I.3, where it is applied to a piano note.

The ESPRIT method, presented below, avoids the optimization of function $\widehat{S}(z)$, or the resolution of equation (3.18), and provides the values of the complex poles in a more direct way.

4.2.5 Estimation of Signal Parameters via Rotational Invariance Techniques

The ESPRIT [RPK86] method is based on a particular property of the signal subspace: the rotational invariance. Let $\boldsymbol{V}^n_\downarrow$ be the matrix of dimensions $(n-1) \times K$ which contains the first $n-1$ rows of $\boldsymbol{V}^n$, and $\boldsymbol{V}^n_\uparrow$ the matrix of dimensions $(n-1) \times K$ which contains the last $n-1$ rows of $\boldsymbol{V}^n$. Similarly, let $\boldsymbol{W}(t)_\downarrow$ be the matrix of dimensions $(n-1) \times K$ which contains the first $n-1$ rows of $\boldsymbol{W}(t)$, and $\boldsymbol{W}(t)_\uparrow$ the matrix of dimensions $(n-1) \times K$ which contains the $n-1$ last rows of $\boldsymbol{W}(t)$. Then we have

$$\boxed{\boldsymbol{V}^n_\uparrow = \boldsymbol{V}^n_\downarrow \boldsymbol{D}} \tag{3.19}$$

where $\boldsymbol{D} = \text{diag}(z_0, \dots, z_{(K-1)})$. Now the columns of $\boldsymbol{V}^n$ and those of $\boldsymbol{W}(t)$ form two bases of the same vector space of dimension K. Thus, there is an invertible matrix $\boldsymbol{G}(t)$ of dimension $K \times K$ such that

$$\boldsymbol{V}^n = \boldsymbol{W}(t)\,\boldsymbol{G}(t) \tag{3.20}$$

where $\boldsymbol{G}(t)$ is defined as the transition matrix from the first basis to the second one. By substituting equation (3.20) into equation (3.19), we show that

$$\boxed{\boldsymbol{W}(t)_\uparrow = \boldsymbol{W}(t)_\downarrow\,\boldsymbol{\Phi}(t)} \tag{3.21}$$

where $\boldsymbol{\Phi}(t)$, called *spectral matrix*, is defined by its EVD:

$$\boxed{\boldsymbol{\Phi}(t) = \boldsymbol{G}(t)\,\boldsymbol{D}\,\boldsymbol{G}(t)^{-1}.} \tag{3.22}$$

In particular, the eigenvalues of $\boldsymbol{\Phi}(t)$ are the poles $\{z_k\}_{k=0\dots K-1}$.

By multiplying equation (3.21) on the left by $\boldsymbol{W}(t)_\downarrow^H$, we get

$$\boldsymbol{W}(t)_\downarrow^H \boldsymbol{W}(t)_\uparrow = \boldsymbol{W}(t)_\downarrow^H \boldsymbol{W}(t)_\downarrow\,\boldsymbol{\Phi}(t). \tag{3.23}$$

However, if $\text{rank}(\boldsymbol{W}(t)_\downarrow) = K$, matrix $\boldsymbol{W}(t)_\downarrow^H \boldsymbol{W}(t)_\downarrow$ is invertible. Indeed, we trivially note that $\forall \boldsymbol{x} \in \mathbb{C}^n$, $\boldsymbol{W}(t)_\downarrow^H \boldsymbol{W}(t)_\downarrow \boldsymbol{x} = \boldsymbol{0} \Leftrightarrow \boldsymbol{W}(t)_\downarrow \boldsymbol{x} = 0$. So $\dim(\ker(\boldsymbol{W}(t)_\downarrow^H \boldsymbol{W}(t)_\downarrow)) = \dim(\ker(\boldsymbol{W}(t)_\downarrow))$. The rank-nullity theorem allows us to conclude that $\text{rank}(\boldsymbol{W}(t)_\downarrow^H \boldsymbol{W}(t)_\downarrow) = \text{rank}(\boldsymbol{W}(t)_\downarrow) = K$, so $\boldsymbol{W}(t)_\downarrow^H \boldsymbol{W}(t)_\downarrow$ is invertible. Therefore equation (3.23) implies $\boldsymbol{\Phi}(t) = \left(\boldsymbol{W}(t)_\downarrow^H \boldsymbol{W}(t)_\downarrow\right)^{-1} \boldsymbol{W}(t)_\downarrow^H \boldsymbol{W}(t)_\uparrow$.

Finally, the ESPRIT algorithm is composed of four steps:

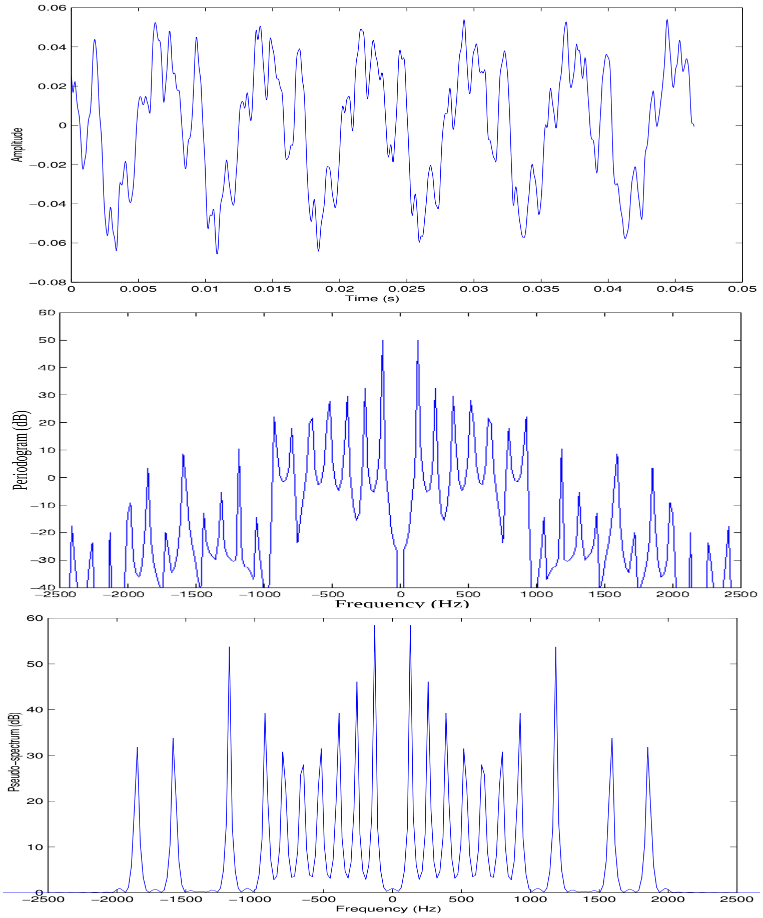


Figure 3.3: Waveform ($t \mapsto x(t)$), periodogram ($f \mapsto 20\log_{10}|S(e^{i2\pi f})|$) and pseudo-spectrum ($f \mapsto 20\log_{10}|\widehat{S}(e^{i2\pi f})|$ with $K = 20$ and $n = 256$) of a piano note

- calculate and diagonalize matrix $\widehat{\boldsymbol{R}}_{xx}(t)$;
- deduce a basis of the signal subspace $\boldsymbol{W}(t)$;
- extract from $\boldsymbol{W}(t)$ matrices $\boldsymbol{W}(t)_{\downarrow}$ and $\boldsymbol{W}(t)_{\uparrow}$;
- compute the spectral matrix $\boldsymbol{\Phi}(t) = \left(\boldsymbol{W}(t)_{\downarrow}^{H}\boldsymbol{W}(t)_{\downarrow}\right)^{-1} \boldsymbol{W}(t)_{\downarrow}^{H}\boldsymbol{W}(t)_{\uparrow}$
- diagonalize $\boldsymbol{\Phi}(t)$ and deduce the estimated poles.

Theoretical and experimental studies have shown that the ESPRIT method is the most efficient of the HR methods presented above. (*cf.* section 6.2).

5 Estimation of the other parameters

The high resolution methods exposed in the previous sections estimate only the poles z_k. We are now interested in the estimation of the other model parameters.

5.1 Estimation of the modeling order

Until now, the order of the ESM model was assumed to be known, which is generally not the case in practice. Many methods have been proposed in the literature for estimating the number of sinusoids present in white noise. The most classic ones are the maximum likelihood method [BK83] and the *Information Theoretic Criteria* (ITC) [WK85], among which the *Akaike Information Criterion* (AIC) criterion [Aka73] and the *Minimum Description Length* (MDL) criterion by Schwartz [Sch78] and Rissanen [Ris78]. Another technique in the context of ITC is the *Efficient Detection Criteria* (EDC) criterion [ZKB86a], which is also robust to a multiplicative white noise [GB03]. These various ITC criteria are based on the similarity of the eigenvalues in the noise subspace, and not on the existence of a gap between the signal and noise subspaces [LR01]. A criterion for selecting the modeling order based on this gap, formulated in terms of *maximally stable* decomposition, was developed in [LRD99]. Other approaches are based on Wishart matrices [GLC96] and on the cross validation method [KM00].

However, in the case where the noise is colored, all these methods tend to overestimate the order of the model. Thus, specific methods have been designed to deal with the case of colored noise, among which new ITC criteria [ZKB86b, ZW93], a technique based on a model of noise *Auto-Covariance Function* (ACF) of finite support [Fuc92], and a maximum a posteriori criterion [BD96].

Among all these methods, we present here the most classic ones, namely the three main ITC criteria: AIC,MDL and EDC (which is a robust generalization of AIC and MDL). These methods consist in minimizing a cost function composed of a first common term and a second term which forms a penalizing factor:

$$\boxed{\mathrm{ITC}(p) = -(n-p)\,l\,\ln\left(\frac{\left(\prod\limits_{q=p+1}^{n}\sigma_q^2\right)^{\frac{1}{n-p}}}{\frac{1}{n-p}\sum\limits_{q=p+1}^{n}\sigma_q^2}\right) + p\,(2n-p)\,C(l)}$$

where the scalars σ_q^2 are the eigenvalues of matrix $\widehat{\boldsymbol{R}}_{xx}(t)$ sorted in decreasing order, and $C(l)$ is a function of variable l. The AIC criterion is defined by setting $C(l) = 1$, and the MDL criterion is defined by setting $C(l) = \frac{1}{2}\ln(l)$. The EDC criteria are obtained for all functions $l \mapsto C(l)$ such that $\lim\limits_{l\to+\infty}\frac{C(l)}{l} = 0$ and $\lim\limits_{l\to+\infty}\frac{C(l)}{\ln(\ln(l))} = +\infty$. These criteria lead to maximizing the ratio of the geometric mean of the eigenvalues of the noise subspace to their arithmetic mean. However this ratio is maximum and equal to 1 when all these eigenvalues are equal; it therefore measures the whiteness of the noise (in theory the eigenvalues are all equal to σ^2). The penalty term $C(l)$ avoids overestimating p. In practice, these methods are relatively satisfactory when processing signals that fit well the signal model, but their performance collapses when this model is less well fitted, in particular when the noise is colored.

5.2 Estimation of amplitudes, phases and standard deviation of noise

The maximum likelihood principle developed in section 3.1 suggests using the least squares method to estimate the complex amplitudes (*cf.* equation (3.5)):

$$\boldsymbol{\alpha}(t) = \boldsymbol{V}^{N\dagger}\boldsymbol{x}(t),$$

from which $a_k = |\alpha_k|$ and $\phi_k = \arg(\alpha_k)$ are deduced. Remember that according to the Gauss-Markov theorem, the least squares estimator is an unbiased linear estimator, with minimum variance among all unbiased linear estimators, since the additive noise is white. In the case where the additive noise is colored, the optimal estimator is obtained by the weighted least squares method (we can refer to [SL100] for detailed information on the estimation of amplitudes by the weighted least squares method) .

Finally, the maximum likelihood principle suggests estimating the standard deviation by calculating the power of the residual (*cf.* equation (3.4)):

$$\sigma^2 = \frac{1}{N}\left\|\boldsymbol{x}(t) - \boldsymbol{V}^N\boldsymbol{\alpha}(t)\right\|^2.$$

6 Performance of the estimators

6.1 Cramer-Rao bound

The Cramér-Rao bound is a fundamental tool in probability theory, because it makes it possible to analyze the performance of an estimator, by comparing its variance to an optimal value, which in a way acts as a quality benchmark. In the particular case of the ESM signal model, a study of the Cramér-Rao bound was proposed in [HS90]. The general Cramér-Rao bound theorem is summarized below (*cf.* [Kay93]). It is based on the assumption of a regular statistical model.

Definition 11 (Regular statistical model)**.** *Let us consider a statistical model dominated by a measure μ and parameterized by $\boldsymbol{\theta} \in \Theta$, where Θ is an open set of $\mathbb{R}^q$. Let $\boldsymbol{x}$ denote the random vector of dimension N. Then the parameterization is said to be regular if the following conditions are satisfied:*

1. *the probability density $p(\boldsymbol{x};\, \boldsymbol{\theta})$ is continuously differentiable, μ-almost everywhere, with respect to $\boldsymbol{\theta}$.*
2. *the Fisher information matrix*
$$\boldsymbol{F}(\boldsymbol{\theta}) \triangleq \int_H \boldsymbol{l}(\boldsymbol{x};\, \boldsymbol{\theta})\, \boldsymbol{l}(\boldsymbol{x};\, \boldsymbol{\theta})^\top\, p(\boldsymbol{x};\, \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{x}$$
defined from the score function $\boldsymbol{l}(\boldsymbol{x};\, \boldsymbol{\theta}) \triangleq \boldsymbol{\nabla}_{\boldsymbol{\theta}} \ln p(\boldsymbol{x};\, \boldsymbol{\theta})\, \mathbf{1}_{p(\boldsymbol{x};\,\boldsymbol{\theta})>0}$ is positive definite for any value of parameter $\boldsymbol{\theta}$, and continuous with respect to $\boldsymbol{\theta}$.

Theorem 7 (Cramér-Rao bound)**.** *Let us consider a regular statistical model parameterized by $\boldsymbol{\theta} \in \Theta$. Let $\widehat{\boldsymbol{\theta}}$ be an unbiased estimator of $\boldsymbol{\theta}$ ($\forall \boldsymbol{\theta} \in \Theta$, $\mathbb{E}_{\boldsymbol{\theta}}[\widehat{\boldsymbol{\theta}}] = \boldsymbol{\theta}$). Then the dispersion matrix $\boldsymbol{D}(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) \triangleq \mathbb{E}_{\boldsymbol{\theta}}\left[\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)^\top\right]$ is such that matrix $\boldsymbol{D}(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) - \boldsymbol{F}(\boldsymbol{\theta})^{-1}$ is positive semidefinite.*

In particular, the diagonal entries of matrix $\boldsymbol{D}(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) - \boldsymbol{F}(\boldsymbol{\theta})^{-1}$ are non-negative. Consequently, the variances of the coefficients of $\widehat{\boldsymbol{\theta}}$ are greater than the diagonal entries of matrix $\boldsymbol{F}(\boldsymbol{\theta})^{-1}$. Thus the Cramér-Rao estimation bounds for all the scalar parameters are obtained in three stages:

- calculation of the Fisher information matrix;
- inversion of this matrix;
- extraction of its diagonal entries.

As mentioned in section 3.1, the vector $\boldsymbol{x}(t)$ containing the N samples of the observed signal is a Gaussian random vector of expected value $\boldsymbol{s}(t)$ and covariance matrix $\boldsymbol{R}_{bb}$. Below, the dependence of $\boldsymbol{s}(t)$ and $\boldsymbol{R}_{bb}$ on the parameters of the model will be mentioned explicitly. On the other hand, in order to simplify the notation, we will omit the dependence of $\boldsymbol{s}(t)$ with respect to time. It is known that the Fisher information matrix of a Gaussian random vector is expressed simply as a function of the model parameters, as shown in the following proposition [Kay93, pp. 525].

Proposition 8 (Fisher's information matrix for a Gaussian density)**.** *For a family of complex Gaussian probability laws of covariance matrix $\boldsymbol{R}_{bb}(\boldsymbol{\theta})$ and of mean $\mathbf{s}(\boldsymbol{\theta})$, where $\boldsymbol{R}_{bb} \in C^1(\Theta, \mathbb{C}^{N\times N})$ and $\boldsymbol{s} \in C^1(\Theta, \mathbb{C}^N)$, the entries of the Fisher information matrix $\left\{\boldsymbol{F}_{(i,j)}(\boldsymbol{\theta})\right\}_{1\leq i,\, j\leq k}$ are given by the extended Bangs-Slepian formula:*

$$\boxed{\boldsymbol{F}_{(i,j)}(\boldsymbol{\theta}) = \operatorname{trace}\left(\boldsymbol{R}_{bb}^{-1}\frac{\partial \boldsymbol{R}_{bb}(\boldsymbol{\theta})}{\partial\theta_i}\boldsymbol{R}_{bb}^{-1}\frac{\partial \boldsymbol{R}_{bb}(\boldsymbol{\theta})}{\partial\theta_j}\right) + 2\mathcal{R}e\left(\frac{\partial \boldsymbol{s}(\boldsymbol{\theta})}{\partial\theta_i}^H \boldsymbol{R}_{bb}^{-1}\,\frac{\partial \boldsymbol{s}(\boldsymbol{\theta})}{\partial\theta_j}\right).} \tag{3.24}$$

By applying formula (3.24) to the ESM model, we obtain a closed-form expression of the Fisher information matrix. We deduce the following theorem, proved in [HS90]:

Proposition 9. *The Cramér-Rao bounds for parameters* (ϕ_k, δ_k, f_k) *are independent of* $a_{k'}$ *for all* $k' \neq k$, *but proportional to* $\frac{1}{a_k^2}$. *The bound for parameter* a_k *is independent of all* $a_{k'}$. *Finally, the bounds for all parameters are independent of all the phases* $\phi_{k'}$, *and are unchanged by a translation of the set of frequencies* $f_{k'}$.

In addition, the Cramér-Rao bounds can be calculated in closed-form under certain assumptions, as was done in [RZ93].

Proposition 10. *Suppose that all the damping factors are zero, and let us make* N *tend towards* $+\infty$. *Then the Cramér-Rao bounds for the parameters of the ESM model admit the following first-order expansions:*

- $\mathrm{CRB}\{\sigma\} = \frac{\sigma^2}{4N} + O\left(\frac{1}{N^2}\right)$;
- $\mathrm{CRB}\{f_k\} = \frac{6\sigma^2}{4\pi^2 N^3 a_k^2} + O\left(\frac{1}{N^4}\right)$;
- $\mathrm{CRB}\{a_k\} = \frac{2\sigma^2}{N} + O\left(\frac{1}{N^2}\right)$;
- $\mathrm{CRB}\{\phi_k\} = \frac{2\sigma^2}{N a_k^2} + O\left(\frac{1}{N^2}\right)$.

We note in particular that the Cramér-Rao bounds related to the frequencies f_k are of order $\frac{1}{N^3}$, which is unusual in parametric estimation. Furthermore, it is known that the maximum likelihood principle provides asymptotically efficient estimators [Kay93]. Thus, the variances of the estimators given in section 3.1 are asymptotically equivalent to the Cramér-Rao bounds given in proposition 10. The case of HR methods is discussed below.

6.2 Performance of HR methods

The performance of an estimator is generally expressed in terms of bias and variance. It is also possible to measure its efficiency, defined as the ratio between its variance and the Cramér-Rao bound. In particular, an estimator is said to be efficient if its efficiency is equal to 1.

In the case of HR methods, calculating the bias and variance in closed-form unfortunately turns out to be impossible, because the extraction of the roots of a polynomial, or of the eigenvalues of a matrix, induces a complex relationship between the statistics of the signal and those of the estimators. However, asymptotic results have been obtained thanks to the perturbation theory. These results are based either on the hypothesis $N \to +\infty$ (in the case where all poles are on the unit circle), or on the hypothesis of a high *Signal to Noise Ratio* (SNR) (SNR $\to +\infty$). Under each of these two hypotheses, it has been shown that all HR methods presented in this chapter are unbiased. Furthermore, under the assumption $N \to +\infty$, the variances of the Prony and Pisarenko methods were calculated in [SN88], and those of MUSIC and ESPRIT in [SS91]. Under the hypothesis SNR$\to +\infty$, the variance of Prony's method was calculated in [KPTV87], that of MUSIC in [ESS93], and that of ESPRIT in [HS91, ESS93].

The mathematical developments proposed in all these articles are quite complex, and are strongly related to the estimation method considered, this is why they are not reproduced as part of this document. Only the main results are summarized here. First of all, it has been proved in [KPTV87, SN88] that the Prony and Pisarenko methods are very inefficient, in the statistical sense: their variances are much greater than the Cramér-Rao bounds. In addition, they increase faster than the Cramér-Rao bounds when the SNR decreases. On the other hand, the MUSIC and ESPRIT methods have an asymptotic efficiency close to 1. More precisely, it has been proved in [SS91, ESS93] (in the case of unmodulated sinusoids) that these two methods achieve almost identical performances, but ESPRIT is slightly better than MUSIC. The study carried out in [HS91] (in the more general case of exponentially modulated sinusoids) goes in the same direction: ESPRIT is less sensitive to noise than MUSIC.

7 Conclusion

In this chapter, we have shown that the estimation of frequencies and damping factors by the maximum likelihood method leads to a difficult optimization problem. When all the poles of the signal are on the unit circle, it can be approximated by detecting the K main peaks of the periodogram. This result is only valid when the length of the observation window is sufficiently large compared to the inverse of the smallest frequency difference between neighboring poles. The main interest of HR methods is that they overcome this limit of Fourier analysis in terms of spectral resolution. The first methods of this family, proposed by Prony and Pisarenko, are based on the linear recurrence equations which characterize the signal model. On the other hand, more modern techniques, including the MUSIC and ESPRIT methods, rely on the decomposition of the data space into two eigensubspaces of the covariance matrix, called signal subspace and noise subspace. The statistical study of these various estimation techniques has shown that the ESPRIT method is the most efficient. The amplitudes and phases of the complex exponentials can then be estimated by the least squares method.

8 Appendices

8.1 Constrained optimization

Let V be a Hilbert space and F be a closed subset of V defined by $F = \{\boldsymbol{y} \in V / f_1(\boldsymbol{y}) = 0 \dots f_M(\boldsymbol{y}) = 0\}$, where functions $\{f_m\}_{m=1\dots M}$ are continuous from V to $\mathbb{R}$. Let J be a real function on V and $\boldsymbol{p}$ be a local minimum of J over F. We also assume that functions J and $\{f_m\}_{m=1\dots M}$ are differentiable at $\boldsymbol{p}$.

Theorem 11 (Lagrange multipliers)**.** *There is $\mu_0, \dots \mu_M \in \mathbb{R}$ not all zero such that $\mu_0 J'(\boldsymbol{p}) + \sum\limits_{m=1}^{M} \mu_m f_m'(\boldsymbol{p}) = 0$. If in addition the vectors $\{f_m'(\boldsymbol{p})\}_{m=1\dots M}$ are linearly independent, then there are coefficients $\lambda_1 \dots \lambda_M \in \mathbb{R}$ called Lagrange multipliers such that $J'(\boldsymbol{p}) + \sum\limits_{m=1}^{M} \lambda_m f_m'(\boldsymbol{p}) = 0$.*

Figure 3.4: Joseph-Louis LAGRANGE (1736-1813)

8.2 Vandermonde matrices

Definition 12 (Vandermonde matrix)**.** *Let $K \in \mathbb{N}^*$ and $z_0, \dots, z_{K-1} \in \mathbb{C}$. We call Vandermonde matrix a matrix* $\boldsymbol{V}$ *of dimension $K \times K$ of the form*

$$\boldsymbol{V} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ z_0 & z_1 & \dots & z_{K-1} \\ \vdots & \vdots & \ddots & \vdots \\ {z_0}^{K-1} & {z_1}^{K-1} & \dots & {z_{K-1}}^{K-1} \end{bmatrix}.$$

Proposition 12 (Vandermonde determinant)**.** *The determinant of the Vandermonde matrix is*

$$\det(\boldsymbol{V}) = \prod_{0 \leq k_1 < k_2 \leq K-1} (z_{k_2} - z_{k_1}).$$

Bibliography

[Aka73] H. Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki, editors, *Proc. of the 2nd International Symposium on Information Theory*, pages 267–281, Budapest, Hongrie, 1973. Akademia Kiado.

[All91] J. Allen. Overwiew of text-to-speech systems. In S. Furui and M. Sondhi, editors, *Advances in Speech Signal Processing*, chapter 23, pages 741–790. Marcel Dekker, 1991.

[BAMCM97] Adel Belouchrani, Karim Abed-Meraim, Jean-Fran¸ois Cardoso, and Eric Moulines. A blind source separation technique using second-order statistics. *IEEE Transactions on Signal Processing*, 45(2):434–444, 1997.

[Bar83] A. J. Barabell. Improving the resolution performance of eigenstructure-based direction-finding algorithms. In *Proc. of ICASSP'83*, pages 336–339, Boston, MA, USA, 1983. IEEE.

[BD87] Peter J. Brockwell and Richard A. Davis. *The Spectral Representation of a Stationary Process*, pages 112–158. Springer New York, New York, NY, 1987.

[BD96] W. B. Bishop and P. M. Djuric. Model order selection of damped sinusoids in noise by predictive densities. 44(3):611–619, March 1996.

[Ben88] J. Benson. *Audio Engineering Handbook*. mcGraw-Hill, New York, 1988.

[BK83] G. Bienvenu and L. Kopp. Optimality of high-resolution array processing using the eigensystem method. 31(5):1235–1245, October 1983.

[Car98] Jean-Fran¸ois Cardoso. Blind signal separation: statistical principles. *Proceedings of the IEEE*, 86(10):2009–2025, 1998.

[CJ10] Pierre Comon and Christian Jutten. *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Academic Press, Inc. (Elsevier), USA, 1st edition, 2010.

[CL96] Jean-Fraņois Cardoso and B. H. Laheld. Equivariant adaptive source separation. *IEEE Transactions on Signal Processing*, 44(12):3017–3030, 1996.

[CLM95] O. Cappé, J. Laroche, and E. Moulines. Regularized estimation of cepstrum envelope from discrete frequency points. *IEEE ASSP Workshop on app. of sig. proc. to audio and acous.*, 1995.

[Com94] Pierre Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287 – 314, April 1994. Special issue on Higher-Order Statistics.

[CS93] Jean-Fran¸ois Cardoso and Antoine Souloumiac. Blind beamforming for non Gaussian signals. *IEE Proceedings-F*, 140(6):362–370, 1993.

[CS96] Jean-Fran¸ois Cardoso and Antoine Souloumiac. Jacobi angles for simultaneous diagonalization. *SIAM Journal on Matrix Analysis and Applications*, 17(1):161–164, 1996.

[Dat87] J. Dattorro. Using digital signal processor chips in a stereo audio time compressor/expander. *Proc. 83rd AES Convention, New York*, Oct 1987. preprint 2500 (M-6).

[Dav99] B. David. *Caractérisations acoustiques de structures vibrantes par mise en atmosphère raréfiée*. PhD thesis, University of Paris VI, 1999.

[EJM91] A. El-Jaroudi and J. Makhoul. Discrete all pole modeling. *IEEE Trans. Acoust., Speech, Signal Processing*, 39(2):411–423, Fev 1991.

[EM12] Sebastian Ewert and Meinard Müller. *Multimodal Music Processing*, volume 3, chapter Score-Informed Source Separation for Music Signals, pages 73–94. January 2012.

[ESS93] A. Eriksson, P. Stoica, and T. Soderstrom. Second-order properties of MUSIC and ESPRIT estimates of sinusoidal frequencies in high SNR scenarios. *IEE Proceedings on Radar, Sonar and Navigation*, 140(4):266–272, August 1993.

[FEJ54] G. Fairbanks, W.L. Everitt, and R.P. Jaeger. Method for time or frequency compression-expansion of speech. *IEEE Trans. Audio Electroacoust.*, AU-2:7–12, Jan 1954.

[Fuc92] J. J. Fuchs. Estimation of the number of signals in the presence of unknown correlated sensor noise. 40(5):1053–1061, May 1992.

[GB03] F. Gini and F. Bordoni. On the behavior of information theoretic criteria for model order selection of InSAR signals corrupted by multiplicative noise. *Signal Processing*, 83:1047–1063, 2003.

[GLC96] J. Grouffaud, P. Larzabal, and H. Clergeot. Some properties of ordered eigenvalues of a Wishart matrix: application in detection test and model order selection. In *Proc. of ICASSP'96*, volume 5, pages 2465–2468. IEEE, 1996.

[GR65] Robert C. Gunning and Hugo Rossi. *Analytic Functions of Several Complex Variables*. AMS Chelsea Publishing. Prentice-Hall, Englewood Cliffs, N.J., USA, 1965.

[Har90] E. Hardam. High quality time scale modification of speech signals using fast synchronized overlap add algorithms. *Proc. IEEE ICASSP-90*, pages 409–412, 1990.

[HS90] Y. Hua and T. K. Sarkar. Matrix pencil method for estimating parameters of exponentially damped/undamped sinusoids in noise. 38(5):814–824, May 1990.

[HS91] Y. Hua and T. K. Sarkar. On SVD for estimating generalized eigenvalues of singular matrix pencil in noise. 39(4):892–900, April 1991.

[HVW02] K. Hermus, W. Verhelst, and P. Wambacq. Psychoacoustic modeling of audio with exponentially damped sinusoids. In *Proc. of ICASSP'02*, volume 2, pages 1821–1824. IEEE, 2002.

[JHJ04] J. Jensen, R. Heusdens, and S. H. Jensen. A perceptual subspace approach for modeling of speech and audio signals with damped sinusoids. 12(2):121–132, March 2004.

[JMP98] M. Jeanneau, P. Mouyon, and C. Pendaries. Sintrack analysis, application to detection and estimation of flutter for flexible structures. In *Proc. of EUSIPCO*, pages 789–792, Ile de Rhodes, Grèce, September 1998.

[JP88] D.L. Jones and T.W. Parks. On the generation and combination of grains for music synthesis. *Computer Music J.*, 12(2):27–34, Summer 1988.

[JRY00] A. Jourjine, Scott Rickard, and O. Yilmaz. Blind separation of disjoint orthogonal signals: demixing N sources from 2 mixtures. In *Proc. of 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 2985–2988, 2000.

[KAR83] S. Y. Kung, K. S. Arun, and D. B. Rao. State-space and singular value decomposition based approximation methods for harmonic retrieval problem. *J. of Opt. Soc. of America*, 73:1799–1811, December 1983.

[Kay93] S. M. Kay. *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice-Hall, Englewood Cliffs, NJ, USA, 1993.

[KB98] M. Kahrs and K. Brandenbourg. *Applications of Digital Signal Processing to Audio and Acoustics*. Kluwer Academic Press, Dortrecht, Netherland, 1998.

[KM00] D. Kundu and A. Mitra. Detecting the number of signals for an undamped exponential model using cross-validation approach. *Signal Processing*, 80(3):525–534, 2000.

[KM02] F. Keiler and S. Marchand. Survey on extraction of sinusoids in stationary sounds. In *Proc. of DAFx-02*, pages 51–58, Hambourg, Allemagne, September 2002.

[KPTV87] A. Kot, S. Parthasarathy, D. Tufts, and R. Vaccaro. The statistical performance of state-variable balancing and Prony's method in parameter estimation. In *Proc. of ICASSP'87*, volume 12, pages 1549–1552, April 1987.

[KT82] R. Kumaresan and D. W. Tufts. Estimating the parameters of exponentially damped sinusoids and pole-zero modeling in noise. 30(6):833–840, December 1982.

[Lar93a] J. Laroche. Autocorrelation method for high quality time/pitch scaling. *IEEE ASSP Workshop on app. of sig. proc. to audio and acous.*, 1993.

[Lar93b] J. Laroche. The use of the Matrix Pencil method for the spectrum analysis of musical signals. *Journal of the Acoustical Society of America*, 94(4):1958–1965, October 1993.

[LC93] C. Lambourg and A. Chaigne. Measurements and modeling of the admittance matrix at bridge in guitars. In *Proc. of SMAC'93*, pages 449–453, Stockholm, Suède, July 1993.

[Lee72] F. Lee. Time compression and expansion of speech by the sampling method. *J. Audio Eng. Soc.*, 20(9):738–742, 1972.

[LLR97] Y. Li, K. Liu, and J. Razavilar. A parameter estimation scheme for damped sinusoidal signals based on low-rank Hankel approximation. 45:481–486, February 1997.

[LR01] A. P. Liavas and P. A. Regalia. On the behavior of Information Theoretic Criteria for model order selection. 49(8):1689–1695, August 2001.

[LRD99] A. P. Liavas, P. A. Regalia, and J.-P. Delmas. Blind channel approximation: effective channel order determination. 47(12):3336–3344, December 1999.

[Luo09] Fa-Long Luo. *Mobile Multimedia Broadcasting Standards: Technology and Practice*. Springer Science & Business Media, January 2009.

[Mak75] J. Makhoul. Linear prediction: A tutorial review. *Proc. IEEE*, 63(11):1380–1418, Nov 1975.

[Mal79] D. Malah. Time-domain algorithms for harmonic bandwidth reduction and time scaling of speech signals. *IEEE Trans. Acoust., Speech, Signal Processing*, 27(2):121–133, 1979.

[MC90] E. Moulines and F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5/6):453–467, Dec 1990.

[MEJ86] J. Makhoul and A. El-Jaroudi. Time scale modification in medium to low rate speech coding. *Proc. IEEE ICASSP-86*, pages 1705–1708, 1986.

[ML95] E. Moulines and J. Laroche. Non parametric techniques for pitch-scale and time-scale modification of speech. *Speech Communication*, 16:175–205, Feb 1995.

[MQ86] R. J. McAulay and T. F. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-34(4):744–754, Aug 1986.

[MSAB$^+$98] S. Marcos, J. Sanchez-Araujo, N. Bertaux, P. Larzabal, and P. Forster. *Les Méthodes à haute résolution : traitement d'antenne et analyse spectrale. Chapitres 4 et 5*. Hermès, Paris, France, 1998. Ouvrage collectif sous la direction de S. Marcos.

[NHD98] J. Nieuwenhuijse, R. Heusens, and Ed. F. Deprettere. Robust exponential modeling of audio signals. In *Proc. of ICASSP'98*, volume 6, pages 3581–3584. IEEE, May 1998.

[NTJ95] Hoang-Lan Nguyen Thi and Christian Jutten. Blind source separation for convolutive mixtures. *Signal Processing*, 45(2):209 – 229, 1995.

[OF10] Alexey Ozerov and Cédric Févotte. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):550–563, 2010.

[Pis73] V. F. Pisarenko. The retrieval of harmonics from a covariance function. *Geophysical J. Royal Astron. Soc.*, 33:347–366, 1973.

[Por76] M. R. Portnoff. Implementation of the digital phase vocoder using the fast fourier transform. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-24(3):243–248, Jun 1976.

[RdP95] Gaspard-Marie Riche de Prony. Essai expérimental et analytique: sur les lois de la dilatabilité de fluides élastiques et sur celles de la force expansive de la vapeur de l'eau et de la vapeur de l'alcool à différentes températures. *Journal de l'école polytechnique*, 1(22):24–76, 1795.

[Ric07] Scott Rickard. *The DUET Blind Source Separation Algorithm*, pages 217–241. Springer, Dordrecht, Netherlands, 2007.

[Ris78] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

[RK87] R. Roy and T. Kailath. Total least squares ESPRIT. In *Proc. of 21st Asilomar Conference on Signals, Systems, and Computers*, pages 297–301, November 1987.

[RPK86] R. Roy, A. Paulraj, and T. Kailath. ESPRIT–A subspace rotation approach to estimation of parameters of cisoids in noise. 34(5):1340–1342, October 1986.

[RW85] S. Roucos and A. M. Wilgus. High quality time-scale modification of speech. *Proc. IEEE ICASSP-85, Tampa*, pages 493–496, Apr 1985.

[RZ93] C. R. Rao and L. C. Zhao. Asymptotic behavior of maximum likelihood estimates of superimposed exponential signals. 41(3):1461–1464, March 1993.

[SAK$^+$13] M. Souden, S. Araki, K. Kinoshita, T. Nakatani, and Hiroshi Sawada. A multichannel MMSE-based framework for speech source separation and noise reduction. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(9):1913–1928, 2013.

[Sch78] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[Sch81] R. O. Schmidt. *A signal subspace approach to multiple emitter location and spectral estimation*. PhD thesis, Stanford University, Stanford, Californie, USA, November 1981.

[Sch86] Ralph. O. Schmidt. Multiple emitter location and signal parameter estimation. 34(3):276–280, March 1986.

[Sen82] S. Seneff. System to independently modify excitation and/or spectrum of speech waveform without explicit pitch extraction. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-24:358–365, 1982.

[SFL67] M.R. Schroeder, J.L Flanagan, and E.A. Lundry. Bandwidth compression of speech by analytic-signal rooting. *Proc. IEEE*, 55:396–401, Mar 1967.

[SG72] R. Scott and S. Gerber. Pitch-synchronous time-compression of speech. *Proceedings of the Conference for Speech Communication Processing*, pages 63–65, Apr 1972.

[SK92] B. Sylvestre and P. Kabal. Time-scale modification of speech using an incremental time-frequency approach with waveform structure compensation. *Proc. IEEE ICASSP-92*, pages 81–84, 1992.

[SLl00] P. Stoica, H. Li, and J. li. Amplitude estimation of sinusoidal signals: survey, new results, and an application. 48(2):338–352, 2000.

[SN88] P. Stoica and A. Nehorai. Study of the statistical performance of the Pisarenko harmonic decomposition method. *IEE Proceedings Radar and Signal Processing*, 135(2):161–168, April 1988.

[SS90] X. Serra and J. Smith. Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition. *Computer Music J.*, 14(4):12–24, Winter 1990.

[SS91] P. Stoica and T. Söderström. Statistical Analysis of MUSIC and Subspace Rotation Estimates of Sinusoidal Frequencies. 39:1836–1847, August 1991.

[SS01] George Stockman and Linda G. Shapiro. *Computer Vision*. Prentice Hall PTR, USA, 1st edition, 2001.

[Vai93] P. P. Vaidyanathan. *Multirate Systems and Filter Banks*. Prentice-Hall, Inc., USA, 1993.

[VdVDS93] A-J. Van der Veen, ED. F. Deprettere, and A. L. Swindlehurst. Subspace based signal analysis using singular value decomposition. *Proc. of IEEE*, 81(9):1277–1308, September 1993.

[VGF06] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1462–1469, 2006.

[VR93] W. Verhelst and M. Roelands. An overlap-add technique based on waveform similarity (wsola) for high quality time-scale modification of speech. *Proc. IEEE ICASSP-93, Minneapolis*, pages 554–557, Apr 1993.

[VVG18] Emmanuel Vincent, Tuomas Virtanen, and Sharon Gannot. *Audio Source Separation and Speech Enhancement*. Wiley Publishing, 1st edition, 2018.

[Wei77] G. Weinreich. Coupled piano strings. *Journal of the Acoustical Society of America*, 62(6):1474–1484, 1977.

[WK85] M. Wax and T. Kailath. Detection of signals by information theoretic criteria. 33(2):387–392, April 1985.

[WW88] J.L. Wayman and D.L. Wilson. Some improvements on the sychronized-overlap-add method of time scale modification for use in real-time speech compression and noise filtering. *IEEE Trans. Acoust., Speech, Signal Processing*, 36(1):139–140, Jan 1988.

[YR04] O. Yilmaz and Scott Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, 52(7):1830–1847, 2004.

[ZKB86a] L. C. Zhao, P. R. Krishnaiah, and Z. D. Bai. On detection of the number of signals in presence of white noise. *Journal of Multivariate Analysis*, 20(1):1–25, 1986.

[ZKB86b] L. C. Zhao, P. R. Krishnaiah, and Z. D. Bai. On detection of the number of signals when the noise covariance matrix is arbitrary. *Journal of Multivariate Analysis*, 20(1):26–49, 1986.

[ZS89] M. Zoltawski and D. Stavrinides. Sensor array signal processing via a Procrustes rotations based eigen-analysis of the ESPRIT data pencil. 37(6):832–861, June 1989.

[ZW93] Q. T. Zhang and K. M. Wong. Information theoretic criteria for the determination of the number of signals in spatially correlated noise. 41(4):1652–1663, April 1993.

Exercises on high resolution methods

Roland Badeau
`roland.badeau@telecom-paristech.fr`

Let us consider the *Exponential Sinusoidal Model* (ESM):

$$s[t] = \sum_{k=0}^{K-1} a_k\, e^{\delta_k t} e^{i(2\pi f_k t + \phi_k)},$$

which, to each frequency $f_k \in \left]-\frac{1}{2}, \frac{1}{2}\right]$, associates a real amplitude $a_k > 0$, a phase $\phi_k \in]-\pi, \pi]$, and a damping factor $\delta_k \in \mathbb{R}$. By defining the complex amplitudes $\alpha_k = a_k e^{i\phi_k}$ and the complex poles $z_k = e^{\delta_k + i2\pi f_k}$, this model can be rewritten in the more compact form

$$s[t] = \sum_{k=0}^{K-1} \alpha_k\, {z_k}^t.$$

In practice, the observed signal $x[t]$ never exactly fits this model. It is rather modeled as the sum of signal $s[t]$ plus a complex Gaussian white noise $b[t]$ of variance σ^2:

$$x[t] = s[t] + b[t].$$

Remark: A complex Gaussian white noise of variance σ^2 is a complex process whose real part and imaginary part are two Gaussian white noises of same variance $\frac{\sigma^2}{2}$, independent from each other.

We assume that the signal $x[t]$ is observed on the time interval $\{0 \dots N-1\}$ of length $N > 2K$. We then consider two integers n and l such that $n > K$, $l > K$, and $N = n + l - 1$.

Finally, we define the $n \times l$ Hankel matrix which contains the N samples of the observed signal:

$$\boldsymbol{X} = \begin{bmatrix} x[0] & x[1] & \dots & x[l-1] \\ x[1] & x[2] & \dots & x[l] \\ \vdots & \vdots & \vdots & \vdots \\ x[n-1] & x[n] & \dots & x[N-1] \end{bmatrix}.$$

We define in the same way the Hankel matrices $\boldsymbol{S}$ and $\boldsymbol{B}$ of same dimension $n \times l$, from the samples of $s[t]$ and $b[t]$, respectively.

Notation:

- $\boldsymbol{X}^T$: transpose of matrix $\boldsymbol{X}$,
- $\boldsymbol{X}^*$: conjugate of matrix $\boldsymbol{X}$,
- $\boldsymbol{X}^H$: Hermitian transpose (conjugate transpose) of matrix $\boldsymbol{X}$.

1 *Mu*ltiple *S*ignal *C*lassification (MUSIC)

Question 1 For all $k \in \{0 \dots K-1\}$, we consider the component $s_k[t] = \alpha_k\, {z_k}^t$. We then define the $n \times l$ Hankel matrix

$$\boldsymbol{S_k} = \begin{bmatrix} s_k[0] & s_k[1] & \dots & s_k[l-1] \\ s_k[1] & s_k[2] & \dots & s_k[l] \\ \vdots & \vdots & \vdots & \vdots \\ s_k[n-1] & s_k[n] & \dots & s_k[N-1] \end{bmatrix}$$

For all $z \in \mathbb{C}$, let us define the n-dimensional vector $\boldsymbol{v}^n(z) = [1,\, z,\, z^2, \dots,\, z^{n-1}]^T$, and the l-dimensional vector $\boldsymbol{v}^l(z) = [1,\, z,\, z^2, \dots,\, z^{l-1}]^T$. Then prove that $\boldsymbol{S_k} = \alpha_k\, \boldsymbol{v}^n(z_k)\, \boldsymbol{v}^l(z_k)^T$.

Question 2 Use the result of question 1 to prove that $\boldsymbol{S} = \sum\limits_{k=0}^{K-1} \alpha_k\, \boldsymbol{v}^n(z_k)\, \boldsymbol{v}^l(z_k)^T$. Prove that this last equality can be rewritten in the form $\boldsymbol{S} = \boldsymbol{V}^n\, \boldsymbol{A}\, \boldsymbol{V}^{l^T}$, where

- $\boldsymbol{V}^n$ is an $n \times K$ Vandermonde matrix:

$$\boldsymbol{V}^n = \begin{bmatrix} 1 & 1 & \dots & 1 \\ z_0 & z_1 & \dots & z_{K-1} \\ {z_0}^2 & {z_1}^2 & \dots & {z_{K-1}}^2 \\ \vdots & \vdots & \vdots & \vdots \\ {z_0}^{n-1} & {z_1}^{n-1} & \dots & {z_{K-1}}^{n-1} \end{bmatrix}$$

- $\boldsymbol{V}^l$ is an $l \times K$ Vandermonde matrix,
- $\boldsymbol{A} = \mathrm{diag}(\alpha_0,\, \alpha_1, \dots,\, \alpha_{K-1})$ is a $K \times K$ diagonal matrix.

Question 3 Let us define matrix $\boldsymbol{R}_{ss} = \frac{1}{l}\, \boldsymbol{S}\, \boldsymbol{S}^H$. Prove that matrix $\boldsymbol{R}_{ss}$ is Hermitian and positive semidefinite. Prove that $\boldsymbol{R}_{ss}$ can be factorized in the form $\boldsymbol{R}_{ss} = \boldsymbol{V}^n\, \boldsymbol{P}\, \boldsymbol{V}^{n^H}$, where $\boldsymbol{P}$ is a $K \times K$ Hermitian and positive definite matrix. Conclude that matrix $\boldsymbol{R}_{ss}$ has rank K (we remind that the poles z_k are pairwise distinct).

Question 4 Prove that matrix $\boldsymbol{R}_{ss}$ is diagonalizable in an orthonormal basis, and that its eigenvalues $\{\lambda_i\}_{i=0\dots n-1}$ are non-negative. By assuming that they are sorted in decreasing order and by using the result of question 3, conclude that

- $\forall i \in \{0 \dots K-1\}$, $\lambda_i > 0$;
- $\forall i \in \{K \dots n-1\}$, $\lambda_i = 0$.

Question 5 Let $\widehat{\boldsymbol{R}}_{xx} = \frac{1}{l} \boldsymbol{X}\, \boldsymbol{X}^H$ and $\boldsymbol{R}_{xx} = \mathbb{E}\left[\widehat{\boldsymbol{R}}_{xx}\right]$. Similarly, let $\widehat{\boldsymbol{R}}_{bb} = \frac{1}{l} \boldsymbol{B}\, \boldsymbol{B}^H$ and $\boldsymbol{R}_{bb} = \mathbb{E}\left[\widehat{\boldsymbol{R}}_{bb}\right]$. By using equality $\boldsymbol{X} = \boldsymbol{S} + \boldsymbol{B}$ and the fact that the noise is centered, prove that $\boldsymbol{R}_{xx} = \boldsymbol{R}_{ss} + \boldsymbol{R}_{bb}$. Then prove that for a complex Gaussian white noise, $\boldsymbol{R}_{bb} = \sigma^2\, \boldsymbol{I}_n$.

Question 6 For all $i \in \{0 \dots n-1\}$, let $\boldsymbol{w}_i$ denote the eigenvector of matrix $\boldsymbol{R}_{ss}$ corresponding to the eigenvalue λ_i. By using the result of question 5, prove that $\boldsymbol{w}_i$ is also an eigenvector of $\boldsymbol{R}_{xx}$ corresponding to the eigenvalue $\lambda_i' = \lambda_i + \sigma^2$. Conclude that

- $\forall i \in \{0 \dots K-1\}$, $\lambda_i' > \sigma^2$;
- $\forall i \in \{K \dots n-1\}$, $\lambda_i' = \sigma^2$.

Question 7 Let $\boldsymbol{W}$ denote the matrix $[\boldsymbol{w}_0 \dots \boldsymbol{w}_{K-1}]$, and $\boldsymbol{W}_\perp$ the matrix $[\boldsymbol{w}_K \dots \boldsymbol{w}_{n-1}]$. Prove that $\mathrm{Span}(\boldsymbol{W}) = \mathrm{Span}(\boldsymbol{V}^n)$ (you can start by proving that $\mathrm{Span}(\boldsymbol{W}) \subset \mathrm{Span}(\boldsymbol{V}^n)$).

Remark: The subspace spanned by $\boldsymbol{W}_\perp$ is an eigen-subspace of matrix $\boldsymbol{R}_{xx}$ corresponding to the eigenvalue σ^2. This is why it is called *noise subspace*. The orthonormal matrix $\boldsymbol{W}$ and the Vandermonde matrix $\boldsymbol{V}^n$ span the same subspace. It thus completely characterizes the K poles of the signal, This is why it is called *signal subspace*. However, all the eigenvalues of $\boldsymbol{R}_{xx}$ corresponding to the signal subspace are increased by σ^2, which means that this subspace also contains noise.

Question 8 Prove that the poles $\{z_k\}_{k\in\{0\ldots K-1\}}$ are the solutions of equation $\left\|\boldsymbol{W}_\perp^H \boldsymbol{v}^n(z)\right\|^2 = 0$.

Remark: In practice, real signals do not rigorously fit the model, and this equation does never hold exactly. This is why the "spectral-MUSIC" method for estimating the poles consists in detecting the K highest peaks of function $z \mapsto \frac{1}{\|\boldsymbol{W}_\perp^H \boldsymbol{v}^n(z)\|^2}$. It is thus easier to implement than the maximum likelihood method, which requires the numerical optimization of a cost function of K complex variables.

2 *E*stimation of *S*ignal *P*arameters via *R*otational *I*nvariance *T*echniques (ESPRIT)

Let $\boldsymbol{V}^n_\downarrow$ be the $(n-1)\times K$ matrix that contains the $n-1$ first rows of $\boldsymbol{V}^n$, and $\boldsymbol{V}^n_\uparrow$ the $(n-1)\times K$ matrix that contains the $n-1$ last rows of $\boldsymbol{V}^n$. Similarly, let $\boldsymbol{W}_\downarrow$ be the $(n-1)\times K$ matrix that contains the $n-1$ first rows of $\boldsymbol{W}$, and $\boldsymbol{W}_\uparrow$ the $(n-1)\times K$ matrix that contains the $n-1$ last rows of $\boldsymbol{W}$.

Question 1 Prove that matrices $\boldsymbol{V}^n_\downarrow$ and $\boldsymbol{V}^n_\uparrow$ are such that $\boldsymbol{V}^n_\uparrow = \boldsymbol{V}^n_\downarrow \boldsymbol{D}$, where $\boldsymbol{D}$ is a $K\times K$ diagonal matrix. What are its diagonal entries?

Question 2 Prove that there is a $K\times K$ invertible matrix $\boldsymbol{G}$ such that $\boldsymbol{V}^n = \boldsymbol{W}\boldsymbol{G}$ (we do not ask to compute $\boldsymbol{G}$, but only to prove its existence). Then prove that $\boldsymbol{V}^n_\downarrow = \boldsymbol{W}_\downarrow \boldsymbol{G}$ and $\boldsymbol{V}^n_\uparrow = \boldsymbol{W}_\uparrow \boldsymbol{G}$.

Question 3 Conclude that there is an invertible matrix $\boldsymbol{\Phi}$ such that $\boldsymbol{W}_\uparrow = \boldsymbol{W}_\downarrow \boldsymbol{\Phi}$. What are the eigenvalues of $\boldsymbol{\Phi}$?

Question 4 By assuming that matrix $\boldsymbol{W}_\downarrow^H \boldsymbol{W}_\downarrow$ is invertible, compute $\boldsymbol{\Phi}$ as a function of $\boldsymbol{W}_\downarrow$ and $\boldsymbol{W}_\uparrow$.

Question 5 Propose an estimation method of the poles $\{z_k\}_{k\in\{0\ldots K-1\}}$.

Remark: The principal advantage of this method with respect to spectral-MUSIC is that it does longer require to numerically optimize a cost function in order to determine the poles, which instead are obtained by a direct computation.

Pitch and temporal scale modifications

Roland Badeau

In this practical work, you will implement the PSOLA method for the analysis/synthesis of speech signals. This method will be tested on signals that can be downloaded on the website of the module "Audio Signal Analysis, Indexing and Transformation"of M2 MVA. These signals are sampled at `Fs`. You can load them with Matlab e.g. by typing up `load aeiou`; they will then be stocked in variable `s`. To listen to them, you can type up `soundsc(s,Fs)`. In Python, you can use the provided notebook template `template-TP-modifications.ipynb` that you can download on the website.

1 Extraction of the analysis marks

Firstly, you will program the following function

```
function A = AnalysisPitchMarks(s,Fs)
```

which extracts the analysis marks. The arguments `s` and `Fs` respectively are the signal to be analyzed and the sampling frequency. The returned matrix `A` will contain the times and pitches corresponding to each analysis mark. More precisely, `A` will be formed of three rows, such that `A(1,n)` $= t_a(n)$ is the time corresponding to the n^{th} analysis mark ($t_a(n) \in \mathbb{N}$ is expressed in number of samples), `A(2,n)` = voiced(n) is a Boolean which indicates whether the signal is voiced or unvoiced in the neighborhood of this mark, and `A(3,n)` $= P_a(n) \in \mathbb{N}$ describes the pitch corresponding to the same mark (i.e. the period expressed in number of samples) in the voiced case, or equals 10ms × `Fs` in the unvoiced case.

To do so, you will need a pitch estimator. In order to spare time, you can use function `period.m`, whose Matlab code is provided with the example signals, and whose Python code is included in the notebook template `template-TP-HR.ipynb`. This function requires two arguments: a short term signal `x` extracted from `s`, and the sampling frequency `Fs` (the other arguments are optional), and returns a couple `[P, voiced]` where `voiced` is a Boolean which indicates whether `x` is voiced or non, and `P` $\in \mathbb{N}$ is the period expressed in number of samples in the voiced case, or equals 10ms × `Fs` in the unvoiced case.

Let us now detail how to determine the analysis marks. For the sake of simplicity, we will not try to align the mark $t_a(n)$ on the beginning of a glottal pulse. To compute $P_a(n)$ and $t_a(n)$, we proceed by recursion on $n \geq 1$:

- extraction of a sequence `x` that starts at time $t_a(n-1)$, and whose duration is equal to $2.5\,P_a(n-1)$;
- computation of $P_a(n)$ and voiced(n) by means of function `period`;
- computation of $t_a(n) = t_a(n-1) + P_a(n)$.

The algorithm will be initialized by setting $t_a(0) = 1$ (in Matlab) or $t_a(0) = 0$ (in Python) and $P_a(0) =$ 10ms × `Fs`.

2 Synthesis and modification of the temporal and spectral scales

To perform the synthesis of the signal, we must start by defining the synthesis marks. They will be stocked in a matrix `B` formed of two rows, such that `B(1,k)` $= t_s(k)$ is the time corresponding to the k^{th} synthesis mark, and `B(2,k)` $= n(k)$ is the index of the analysis mark corresponding to this same synthesis mark. To start, you can perform a synthesis without modification, by setting:

- `B(1,:)` = `A(1,:)`;
- `B(2,:)` = $[1, 2, 3, \ldots]$.

2.1 Signal synthesis

You will now program the following function

```
function y = Synthesis(s,Fs,A,B)
```

which computes the synthesis signal `y` from the original signal `s`, the sampling frequency `Fs`, the analysis marks stocked in matrix `A` and the synthesis marks stocked in matrix `B`. The synthesis is very simply performed by recursion on $k \geq 1$ (vector `y` being initialized to the zero vector of dimension $t_s(k_{\text{end}}) + P_a(n(k_{\text{end}}))$):

- extraction of a sequence `x` centered at $t_a(n(k))$ and of length $2\,P_a(n(k)) + 1$;
- windowing of `x` by a Hann window (Matlab function `hann` or Python function `scipy.signal.hanning`);
- overlap-add of the sequence `x` windowed on $y(t_s(k) - P_a(n(k)) : t_s(k) + P_a(n(k)))$.

2.2 Modification of the temporal scale

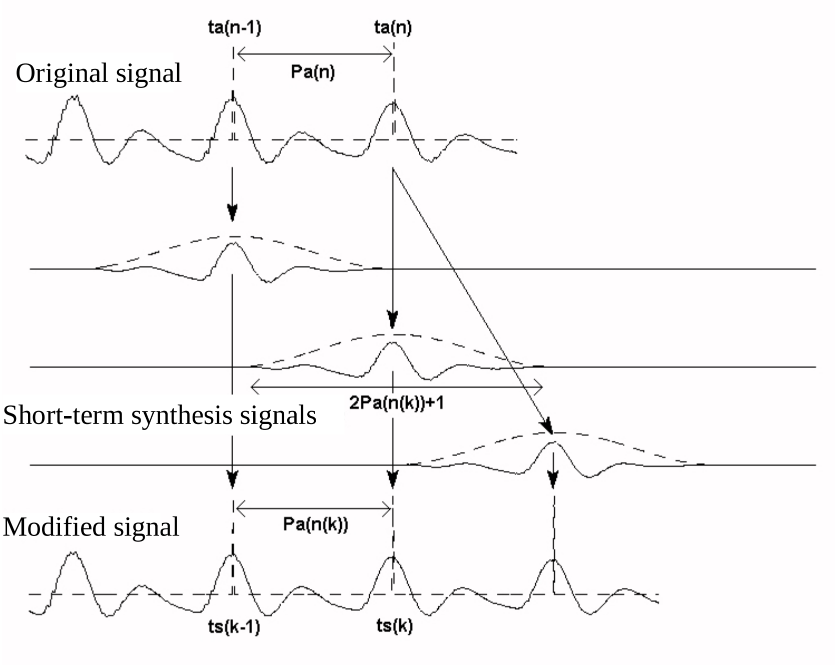


Figure 1: Modification of the temporal scale

We now want to determine the synthesis marks that will modify the temporal scale by a factor α, i.e. to determine a matrix `B` such that the duration of the signal synthesized by function `Synthesis` is equal to that of the original signal `s` multiplied by α. This operation will be performed by function

```
function B = ChangeTimeScale(alpha,A,Fs)
```

which computes matrix `B` from the factor α, the analysis marks stocked in `A`, and the sampling frequency `Fs`. You can proceed by recursion on $k \geq 1$, by using a non-integer index $n(k)$:

- $t_s(k) = t_s(k-1) + P_a(\lfloor n(k) \rfloor)$;
- $n(k+1) = n(k) + \frac{1}{\alpha}$.

The algorithm will be initialized by setting $t_s(0) = 1$ and $n(1) = 1$. You will take care of only stocking integer values in matrix `B`.

2.3 Modification of the spectral scale

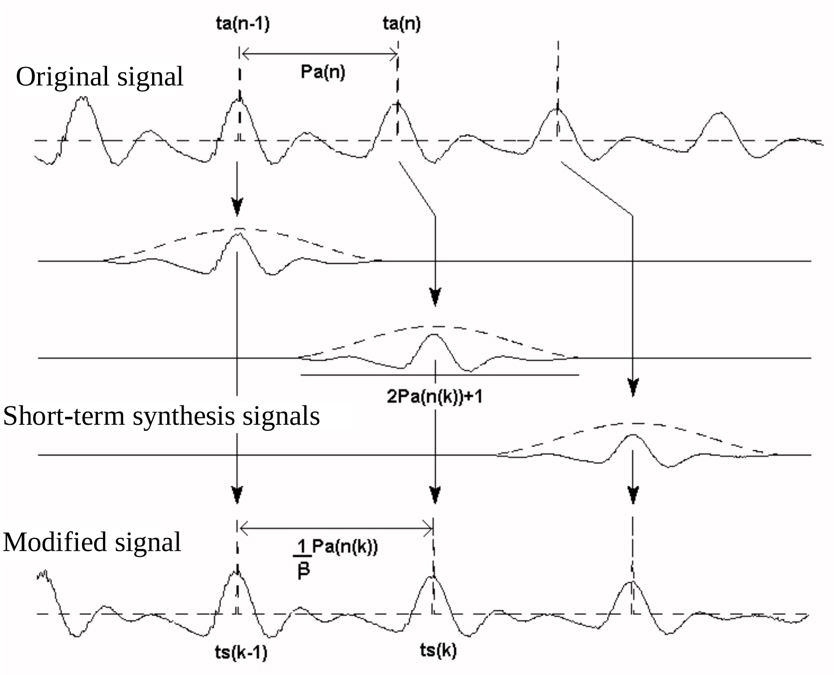


Figure 2: Modification of the spectral scale

You will now perform the dual operation of the previous one: determine the synthesis marks that will modify the spectral scale by a factor β, i.e. determine a matrix `B` such that the fundamental frequency of the signal synthesized by function `Synthesis` is equal to that of the original signal `s` multiplied by β. This operation will be performed by function

```
function B = ChangePitchScale(beta,A,Fs)
```

which computes matrix `B` from the factor β, the analysis marks stocked in `A`, and the sampling frequency `Fs`. As in the previous case, you can proceed by recursion on $k \geq 1$, by using a non-integer index $n(k)$ and non-integer synthesis times $t_s(k)$, and by making the difference between the voiced and unvoiced cases:

- if the analysis mark of index $\lfloor n(k) \rfloor$ is voiced, $\text{scale}(k) = \frac{1}{\beta}$, otherwise $\text{scale}(k) = 1$;
- $t_s(k) = t_s(k-1) + \text{scale}(k) \times P_a(\lfloor n(k) \rfloor)$;
- $n(k+1) = n(k) + \text{scale}(k)$.

Again, you will take care of only stocking integer values in matrix `B`.

2.4 Joint modification of the temporal and spectral scales

To finish, you will program a function that jointly modifies the two scales:

```
function B = ChangeBothScales(alpha,beta,A,Fs)
```

where the arguments are defined as previously. The content of this function will be almost identical to that of `ChangePitchScale`; you will just need to modify it properly.

Institut Mines-Télécom

Practical work on audio source separation

Roland Badeau

TELECOM Paris
IP PARIS

In this practical work, you will program a simple implementation of a variant of the DUET (*Degenerate Unmixing Estimation Technique*) method, which aims to separate sound sources in a stereophonic mixture. You will thus address the case of an under-determined ($M = 2$ sensors and $K > 2$ sources) instantaneous linear mixture. The separation is achieved by exploiting the spatial information, and by assuming that the sources are sparse in the time-frequency plane (a single source is active at each time-frequency bin). The transformation that you will use is the MDCT (*Modified Discrete Cosine Transform*), which presents the double advantage of having real values, and of producing a sparser representation than the STFT (short time Fourier transform). In order to compute it, you will use the *Linear Time/Frequency Toolbox* (`ltfat`) toolbox in Matlab, that you can download on the website of the module "Audio Signal Analysis, Indexing and Transformation"of M2 MVA, or you can use the provided Python notebook template `template-TP-separation.ipynb` which is based on the `mdct` toolkit. You will also find on this website the stereophonic sound file to be processed, named `mix.wav`. In order to use the functions of the `ltfat` toolbox in Matlab, you will first have to load it by calling function `ltfatstart`.

1 Mixture model and principle of the DUET method

The DUET method relies on the following mixture model: at every time-frequency bin (f, n),

$$\boldsymbol{X}(f, n) = \boldsymbol{S}(f, n)\, \boldsymbol{A}$$

where

- the row vectors $\boldsymbol{X}(f, n) = [X(f, n, 1), X(f, n, 2)]$ of dimension $M = 2$ contain the MDCT of the two stereophonic channels $x(t, m)$ of the observed mixture;
- the $K \times M$ matrix $\boldsymbol{A} = [\cos(\boldsymbol{\theta}), \sin(\boldsymbol{\theta})]$ is the mixing matrix;
- the K-dimensional column vector $\boldsymbol{\theta}$ contains the angles $\theta(k)$ of sources k;
- the K-dimensional row vectors $\boldsymbol{S}(f, n) = [S(f, n, 1), \ldots, S(f, n, K)]$ contain the MDCT of the K unknown source signals.

If only source k is active at (f, n), the point of affix $Z(f, n) = X(f, n, 1) + iX(f, n, 2) \in \mathbb{C}$ is such that $Z(f, n) = S(f, n, k)\, e^{i\theta(k)}$, where $S(f, n, k) \in \mathbb{R}$. We remark that its argument permits us both to identify the active source k at (f, n) and its angle $\theta(k)$. The magnitude of $Z(f, n)$ permits us to determine the value of $S(f, n, k)$, up to its sign. In order to remove the sign ambiguity of $S(f, n, k)$, we can assume that $\theta(k) \in [-\frac{\pi}{2}, \frac{\pi}{2}]$. Once source k is identified at every time-frequency bin, a binary mask $B(f, n, k)$ can be applied to $X(f, n, m)$, in order to obtain an estimation $Y(f, n, m, k)$ of the stereophonic image of source k. The source signal is finally reconstructed by means of the MMSE (*Minimum Mean Square Error*) estimator, which is such that $S(f, n, k) = Y(f, n, 1, k) \cos(\theta(k)) + Y(f, n, 2, k) \sin(\theta(k))$.

2 Work to do

1. Open file `mix.wav` and load it in a $T \times M$ matrix $x(t, m)$, where $M = 2$ and T is the number of samples. Use your headphones to listen to the mixture. What is the number K of instruments that you can hear? From which direction do you perceive them?

2. Plot the temporal dispersion diagram, defined as the set of points in the plane of coordinates $(x(t,1), x(t,2))$ for all t (in order to plot a set of points, you can use the Matlab function `plot` or the Python function `matplotlib.pyplot.plot` with parameter `'x'`, and you can normalize the axes with the Matlab instruction `axis equal` or the Python Matplotlib function `axis('equal')`). Can you distinguish the directions of the sources?

3. Compute the MDCT $X(f,n,m)$ of the two stereophonic channels $x(t,m)$ (you can use the Matlab function `wmdct`, with $F = 512$ frequency bands and the window `'sqrthann'`, or the Python function `mdct`). Plot the corresponding time-frequency representations $|X(f,n,m)|^2$ (you can use the Matlab function `plotwmdct` or the Python function `matplotlib.pyplot.imshow`).

4. Plot the time-frequency dispersion diagram, defined as the set of points in the plane of affix $Z(f,n)$ for all f and n. Can you distinguish the directions of the sources? How do you explain it?

5. Plot the histogram of the arguments of the points of affix $Z(f,n)$ for all f and n (you can use the Matlab function `atan` or the Python function `numpy.arctan` to compute the arguments modulo π, between $-\frac{\pi}{2}$ and $+\frac{\pi}{2}$, and the Matlab function `hist` or the Python function `matplotlib.pyplot.hist` to compute the histogram, whose number of classes has to be tuned so as to make the directions of the sources clearly visible). Estimate the angles $\theta(k)$ (you can determine these values graphically from the histogram).

6. In order to estimate the active source at every time-frequency bin (f,n), you can look for the source k whose angle $\theta(k)$ is closest to the argument of $Z(f,n)$, modulo π (you can use a deviation measure invariant modulo π, for instance $|\sin(\theta(k) - \angle Z(f,n))|$). Then generate the binary masks $B \in \{0,1\}$, such that $B(f,n,k)$ is equal to 1 if source k is active at (f,n), or 0 otherwise.

7. Apply masks B to the MDCT $X(f,n,m)$ in order to estimate the MDCT of the stereophonic images $Y(f,n,m,k)$. Then reconstruct the images $y(t,m,k)$ of the source signals by applying the inverse MDCT (you can use the Matlab function `iwmdct` or the Python function `imdct`).

8. Listen to the K reconstructed stereophonic images $y(:,:,k)$. What defects can you perceive?

9. Compute the MMSE estimator $S(f,n,k)$ of source k. Reconstruct the source signals $s(t,k)$ by applying the inverse MDCT to $S(f,n,k)$. Listen to the result.

10. We now wish to respatialize the sources, i.e. to resynthesize the mixture $x(t,m)$ by modifying the angles $\theta(k)$ (remark that it is not needed to switch back to the MDCT domain). For instance, try to permute the directions of the sources. Listen to the result. What audible defects can you notice?

Analysis and synthesis of bell sounds

Roland Badeau

The various files related to this practical work can be downloaded on the website of the module "Audio Signal Analysis, Indexing and Transformation"of M2 MVA. You can load a sound file with Matlab by using the command `[x,Fs] = wavread('clocheB.wav')`. In order to listen to it, you can type up `soundsc(x,Fs)`. In Python, you can use the provided notebook template `template-TP-HR.ipynb`.

1 Introduction

Bells are among the oldest music instruments and the sound they produce is often evocative because it has soothed the everyday life of generations for about 3000 years, accompanying minor and major events. This evocation is partly due to the structure of the sound spectrum: the eigenmodes of vibration are generally tuned by the bell makers so that their frequencies follow a particular series, which includes the minor third (E flat if the bell is tuned in C). This series is not harmonic, but the ratios between the eigenfrequencies are such that one can perceive a well-defined pitch. In particular, the presence of the series 2-3-4, strong at the beginning of the sound, reinforces the feeling of pitch in the neighborhood of the fundamental frequency. This feeling is related to a psychoacoustic effect (processing of the signal received by the brain).

Let f_p be the frequency corresponding to the perceived pitch. The analysis of the eigenfrequencies series leads to a table of about 15 ratios $\alpha_n = f_n/f_p$. Their orders of magnitude are as follows: 0.5 (hum / "*bourdon*"), 1 (prime / fundamental), 1.2 (minor third), 1.5 (fifth), 2 (nominal / octave), 2.5, 2.6, 2.7, 3, 3.3, 3.7, 4.2 (wrong double octave), 4.5, 5, 5.9. The timbre of the corresponding sound depends on the amplitude and on the decrease of each of these partials.

This practical work aims to develop a high resolution spectral estimation method in order to perform the analysis / synthesis of bell sounds. As can be noticed in figure 1, this type of sounds presents a strong temporal decrease.

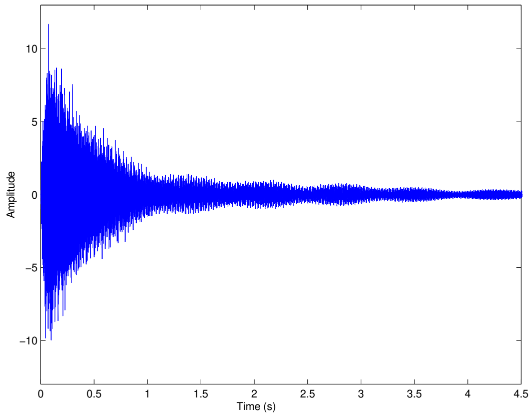


Figure 1: Bell sound

In order to take this damping into account, we consider the *Exponential Sinusoidal Model* (ESM):

$$s[t] = \sum_{k=0}^{K-1} a_k\, e^{\delta_k t} e^{i(2\pi f_k t + \phi_k)},$$

which to each frequency $f_k \in \left]-\frac{1}{2}, \frac{1}{2}\right]$ associates a real amplitude $a_k > 0$, a phase $\phi_k \in]-\pi, \pi]$, and a damping factor $\delta_k \in \mathbb{R}$. By defining the complex amplitudes $\alpha_k = a_k e^{i\phi_k}$ and the complex poles $z_k = e^{\delta_k + i2\pi f_k}$, this model can be rewritten in the form

$$s[t] = \sum_{k=0}^{K-1} \alpha_k \, {z_k}^t.$$

The case $\delta_k < 0$ corresponds to exponentially decreasing sinusoids, which are solutions of physical propagation equations. The model parameters are then $\{\delta_k, f_k, a_k, \phi_k\}_{k\in\{0...K-1\}}$. In order to estimate them, we will use the ESPRIT method presented in the course. Firstly, we will apply it to a synthetic signal in order to highlight the superiority of high resolution methods over Fourier analysis in terms of spectral resolution. Then this method will be applied to bell sounds.

2 Reminder about Matlab/Python

In Matlab:

- `A'`: Hermitian transpose (conjugate transpose) of matrix $\boldsymbol{A}$;
- `A.'`: transpose of matrix $\boldsymbol{A}$ (without complex conjugation);
- `A(l1:l2,c1:c2)`: matrix extracted from $\boldsymbol{A}$ between rows l_1 and l_2 (inclusive) and columns c_1 and c_2 (inclusive).

In Python:

- `A.conj().T`: Hermitian transpose (conjugate transpose) of matrix $\boldsymbol{A}$;
- `A.T`: transpose of matrix $\boldsymbol{A}$ (without complex conjugation);
- `A[l1:l2,c1:c2]`: matrix extracted from $\boldsymbol{A}$ between rows l_1 and l_2 (excluded) and columns c_1 and c_2 (excluded).

3 Synthetic signal

Here we consider a synthetic signal of length N, consisting of a sum of two complex exponentials, whose frequencies are separated by an interval $\Delta f = \frac{1}{N}$ (which corresponds to the resolution limit of Fourier analysis). The phases are drawn randomly, according to a uniform probability distribution on $(-\pi, \pi)$. We do not add noise to this signal, so that that the observed signal $x[t]$ is equal to $s[t]$. You can use the following parameters: $N = 63$, $f_0 = \frac{1}{4}$, $f_1 = f_0 + \frac{1}{N}$, $a_0 = 1$, $a_1 = 10$, $\delta_0 = 0$, $\delta_1 = -0.05$. In order to synthesize it, you can call the provided function

```
x = Synthesis(N,delta,f,a,phi);
```

whose arguments are N, the vector `delta` of damping factors δ_k, the vector `f` of frequencies f_k, the vector `a` of amplitudes a_k, and the vector `phi` of phases ϕ_k.

3.1 Spectral analysis by Fourier transform

Observe the periodogram of this signal. Briefly study the separability of the two spectral lines, without zero-padding ($N_{\text{fft}} = N$) and with zero-padding ($N_{\text{fft}} = 1024 > N$).

3.2 High resolution methods

Our goal is to write functions

`MUSIC(x,n,K)` and `[delta,f] = ESPRIT(x,n,K)`

which analyze the signal x of length N by using methods MUSIC and ESPRIT, with a signal subspace of dimension K and a noise subspace of dimension $n - K$, and the data vectors of length n ranging from $K+1$ to $N-K+1$. In order to process the synthetic signals, you can choose $n = 32$ and $K = 2$. The two methods share the following steps:

1. **Computation of the empirical covariance matrix**

   The empirical covariance matrix of the observed signal is defined by the equation

   $$\widehat{\boldsymbol{R}}_{xx} = \frac{1}{l}\boldsymbol{X}\,\boldsymbol{X}^H$$

   where $\boldsymbol{X}$ is an $n \times l$ Hankel matrix containing the $N = n + l - 1$ samples of the signal:

   $$\boldsymbol{X} = \begin{bmatrix} x[0] & x[1] & \dots & x[l-1] \\ x[1] & x[2] & \dots & x[l] \\ \vdots & \vdots & \ddots & \vdots \\ x[n-1] & x[n] & \dots & x[N-1] \end{bmatrix}$$

   Matrix $\boldsymbol{X}$ can be constructed with function `hankel`.

2. **Estimation of the signal subspace**

   Matrix $\widehat{\boldsymbol{R}}_{xx}$ can be diagonalized with the command `[U1,Lambda,U2] = svd(Rxx)`. Matrix $\widehat{\boldsymbol{R}}_{xx}$ being positive semidefinite, the column vectors of the $n \times n$ matrices $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ are the eigenvectors of $\widehat{\boldsymbol{R}}_{xx}$, corresponding to the n eigenvalues sorted in the diagonal matrix $\boldsymbol{\Lambda}$ in decreasing order (we thus have $\widehat{\boldsymbol{R}}_{xx} = \boldsymbol{U}_1\,\boldsymbol{\Lambda}\,\boldsymbol{U}_1^H = \boldsymbol{U}_2\,\boldsymbol{\Lambda}\,\boldsymbol{U}_2^H$). Therefore you can extract from $\boldsymbol{U}_1$ (or from $\boldsymbol{U}_2$) an $n \times K$ basis of the signal subspace $\boldsymbol{W}$.

3.2.1 ESPRIT method

In a first stage, the ESPRIT method consists in estimating the frequencies and damping factors:

3. **Estimation of the frequencies and damping factors**

   In order to estimate the frequencies, you can proceed in the following way:

   - extract from $\boldsymbol{W}$ the matrices $\boldsymbol{W}_{\downarrow}$ (obtained by removing the last row of $\boldsymbol{W}$) and $\boldsymbol{W}_{\uparrow}$ (obtained by removing the first row of $\boldsymbol{W}$);
   - compute $\boldsymbol{\Phi} = \left(\left(\boldsymbol{W}_{\downarrow}^H\boldsymbol{W}_{\downarrow}\right)^{-1}\boldsymbol{W}_{\downarrow}^H\right)\boldsymbol{W}_{\uparrow} = \boldsymbol{W}_{\downarrow}^{\dagger}\,\boldsymbol{W}_{\uparrow}$, where the symbol † denotes the pseudo inverse operator (Matlab function `pinv` or Python function `numpy.linalg.pinv`).
   - compute the eigenvalues of $\boldsymbol{\Phi}$ by using the Matlab function `eig` or the Python function `numpy.linalg.eig` (we remind that the eigenvalues of $\boldsymbol{\Phi}$ are the poles $z_k = e^{\delta_k + i2\pi f_k}$). Then compute $\delta_k = \ln(|z_k|)$ and $f_k = \frac{1}{2\pi}\,\text{angle}(z_k)$.

4. **Estimation of the amplitudes and phases**

   We now aim to write a function

   ```
   [a,phi] = LeastSquares(x,delta,f)
   ```

   which estimates the amplitudes a_k and phases ϕ_k by means of the least squares method, given the signal x, the damping factors δ_k and the frequencies f_k. The complex amplitudes are thus determined by the equation

   $$\boldsymbol{\alpha} = \left((\boldsymbol{V}^{N^H}\boldsymbol{V}^{N})^{-1}\boldsymbol{V}^{N^H}\right)\boldsymbol{x} = \boldsymbol{V}^{N^\dagger}\boldsymbol{x} \tag{1}$$

   where $\boldsymbol{x}$ is the vector $[x[0], \ldots, x[N-1]]^\top$ and $\boldsymbol{V}^N$ is the $N \times K$ Vandermonde matrix, whose entries are such that $\boldsymbol{V}^N_{(t,k)} = {z_k}^t$ for all $(t, k) \in \{0 \ldots N-1\} \times \{0 \ldots K-1\}$. In order to compute matrix $\boldsymbol{V}^N$, it is possible of avoid using a `for` loop by noting that $\ln\left(\boldsymbol{V}^N_{(t,k)}\right) = t(\delta_k + i2\pi f_k)$. Therefore, the matrix containing the coefficients $\ln\left(\boldsymbol{V}^N_{(t,k)}\right)$ can be expressed as the product of a column vector and a row vector. Then compute $a_k = |\alpha_k|$ and $\phi_k = \text{angle}(\alpha_k)$ for all $k \in \{0 \ldots K-1\}$.

5. **Application to synthetic signals**

   Apply functions `ESPRIT` and `LeastSquares` to the previously synthesized signal. Comment.

3.2.2 MUSIC method

We remind that the MUSIC pseudo-spectrum is defined as $P(z) = \dfrac{1}{||\boldsymbol{W}_\perp^H \boldsymbol{v}^n(z)||^2}$, where matrix $\boldsymbol{W}_\perp$ spans the noise subspace.

6. **MUSIC pseudo-spectrum**

   Write a function `MUSIC(x,n,K)` which plots the logarithm of the pseudo-spectrum as a function of the two variables $f \in [0, 1]$ and $\delta \in [-0.1, 0.1]$ (you can use the Matlab function `surf` or the Python Matplotlib function `plot_surface`). Apply function `MUSIC` to the previously synthesized signal, and check that the pseudo-spectrum makes the two poles $z_k = e^{\delta_k + i2\pi f_k}$ clearly visible.

4 Audio signals

We now propose to apply the functions developed in the previous part to bell sounds.

4.1 Spectral analysis by Fourier transform

Look at the periodogram of the signal `ClocheA.wav` or `ClocheB.wav`, and compare the series of its eigenfrequencies with the values given in the introduction.

4.2 High resolution method

We now want to apply the ESPRIT method to this signal. Let $K = 54$, $n = 512$ and $l = 2n = 1024$ (hence $N = n + l - 1 = 1535$).

In order to guarantee that the signal model holds on the analysis window (exponential damping), we will extract a segment of length N whose beginning is posterior to the maximum of the waveform envelope. We may thus start at the 10000th sample.

Apply function `ESPRIT` to the extracted signal in order to estimate the eigenfrequencies and the corresponding damping factors. Then estimate the amplitudes and phases by calling function `LeastSquares`. Finally, listen to the signal resynthesized with function `Synthesis` (on a duration longer than the extracted segment, in order to clearly highlight the sound resonances), and comment.