Published in Image Processing On Line on YYYY–MM–DD.
ISSN 2105–1232 © YYYY IPOL CC–BY–NC–SA
This article is available online with supplement materials,
implementation, datasets and demonstration at
http://dx.doi.org/10.5201/ipol.YYYY.XXXXXXXX

DRAFT

Interpolation of missing samples in audio signals based on autoregressive modeling

Laurent Oudre

Abstract

This article proposes an implementation and a study of the paper *Adaptive Interpolation of Discrete-Time Signals That Can Be Modeled as Autoregressive Processes* by *Janssen et al* [10]. The algorithm presented in this paper allows one to reconstruct an audio signal which presents localized degradations by interpolating the missing samples. This method assumes that the signal can locally be modeled as a realization of an autoregressive process and iteratively estimate the model parameters and the interpolated samples by minimizing a quadratic criterion. We investigate the limits and the algorithmic aspects of this method on several audio examples.

Source Code

The source code can be found at http://www.laurentoudre.fr/IPOL/ARInterpolation-1.0.zip (login = demo, password =demo).

Supplementary Material

An online demo can be found at http://dev.ipol.im/~oudre/ipol_demo/lo_ar/ (login = demo, password =demo).

1 Introduction

The restoration of audio signals containing localized degradations (known as clicks, scratches, bursts) consists in a detection phase (finding the locations of the degraded samples) and a reconstruction phase (replacing the degraded samples by more suitable values). In the case where some parts of the signal are completely missing, the first step is omitted and only the second phase is necessary. In this article, we assume in either way that the locations of the degraded or missing samples are known (for example thanks to a preliminary detection step) and we only focus on the reconstruction phase.

Numerous methods have been developed for interpolation of missing samples in music or speech signals. While some interpolation techniques, such as median filtering, are completely blind (no hypothesis on the signal is made) [17], they are often too crude to reconstruct gaps larger than a few samples. Most of the efficient methods dealing with audio signals are therefore model-based and introduce some prior information on the characteristics of the signal [7]. The choice of the model (sinusoidal [14, 12], autoregressive (AR) [10, 18, 19, 5, 6, 4, 3], Gabor decomposition [21]...) is guided by the type of data treated (speech, instrumentals, songs...) and by the type of degradation (long

or short gaps, number of missing samples...). For example, the physiological production of speech is often described by a source-filter model where the speech signal is assumed to have been produced by a source or excitation signal (the glottal airflow) and a linear filter (the vocal track). Assuming that the excitation is a white noise and that the filter is all-pole, this model writes as an autoregressive (AR) model [16], which makes AR-based interpolation methods relevant for speech signals. More generally, the AR model is an appropriate choice for many locally stationary signals. Indeed, the position of the poles permits to deal with both noise-like and harmonic signals (respectively by setting poles close to the origin or to the unit circle) [8]. Also, the order p of the model allows one to adapt to the complexity of the considered signal.

The good fit between AR models and audio signals has been widely used for interpolation of missing samples. The paper implemented in this article was the first to propose a method for interpolation of bursts containing several contiguous missing samples and based on AR models [10]. This principle has simultaneously been applied by [18, 19] for detection of impulsive noise and restoration of gramophone recordings. These works have been followed by a large number of variants. A Bayesian extension of this problem has been formulated in [5, 6], while some specific techniques have been developed for processing long gaps [4, 3]. Interestingly, a recent study on interpolation methods [1] shows that, compared to more recent or sophisticated methods, the *Janssen et al* [10] method already obtains satisfactory results on several basic interpolation tasks for gaps comprised between 1 milliseconds and 10 milliseconds (for comparison, real life clicks are often assumed to have durations ranging from less than 20 microseconds to 4 milliseconds [7]).

We propose in this article a rigorous implementation and testing of *Janssen et al* method for interpolation of missing samples with known locations. Section 2 gives some introductory results on AR processes. Section 3 describes the method and the algorithmic concepts behind it. Finally, Section 4 presents the performances of the algorithm on several audio examples, as well as a study on the limits of the method.

2 Some results on autoregressive processes

2.1 Definition

A real wide sense stationary stochastic process $(y_k)_{k\in\mathbb{Z}}$ is an autoregressive process of order $p \in \mathbb{N}^*$ if it satisfies

$$\forall k \in \mathbb{Z}, y_k + a_1\, y_{k-1} + \ldots + a_p\, y_{k-p} = e_k, \tag{1}$$

where $\mathbf{a} = [a_1, \ldots, a_p]^t \in \mathbb{R}^p$, $a_p \neq 0$ and $(e_k)_{k\in\mathbb{Z}}$ is a zero-mean white noise process of variance σ^2. By convention, we set $a_0 = 1$.

2.2 Yule-Walker equations

Let $R(\tau) = E\left[y_k\, y_{k+\tau}\right]$ be the autocorrelation function of process $(y_k)_{k\in\mathbb{Z}}$. From (1) we have

$$\begin{aligned} R(\tau) &= E\left[y_k\, y_{k+\tau}\right] \\ &= E\left[y_k \left(-\sum_{l=1}^{p} a_l\, y_{k-l+\tau} + e_{k+\tau}\right)\right] \\ &= -\sum_{l=1}^{p} a_l\, R(\tau - l) + E\left[y_k\, e_{k+\tau}\right]. \end{aligned} \tag{2}$$

From (1), (2) and by using the fact that $R(-\tau) = R(\tau)$ (since the process is real and stationary),

$$\begin{bmatrix} R(0) & R(1) & \cdots & R(p) \\ R(1) & R(0) & \cdots & R(p-1) \\ \vdots & \vdots & \ddots & \vdots \\ R(p) & R(p-1) & \cdots & R(0) \end{bmatrix} \begin{bmatrix} 1 \\ a_1 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} \sigma^2 \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{3}$$

This relationship is known as the Yule-Walker equations and is characteristic of an AR process.

3 Method

3.1 Principle

Consider a finite audio signal $\mathbf{s} \in \mathbb{R}^N$, which can be modeled as a realization of an AR process of order p. Let $T \subset \{1, \ldots, N\}$ be the set containing the indexes of the missing samples (which is supposed to be known). In order to restore the signal, we assume that the first and last p samples are not missing, i.e. $\forall t \in T, p < t \leq N - p$. We define the set of known samples $\widetilde{T} = \{1, \ldots, N\} \setminus T$. Our aim is to estimate $\mathbf{s}(T)$ by using only the reliable information $\mathbf{s}(\widetilde{T})$ and the AR process hypothesis. Note that in the following, no other assumption on the signal is made.

The method presented in this paper is iterative and works as follows:

- **Step 1:** Given p, an estimate of $\mathbf{s}(T)$ and $\mathbf{s}(\widetilde{T})$, calculate the parameters $\mathbf{a}$ of the AR process of order p which best fits the corrupted signal $\mathbf{s}$ in a quadratic sense.
- **Step 2:** Given p, T, $\mathbf{s}(\widetilde{T})$ and an estimate of $\mathbf{a}$, calculate the missing samples $\mathbf{s}(T)$ which best fit the AR process estimated in Step 1 in a quadratic sense.
- Repeat Steps 1 & 2 until the reconstruction gets satisfactory.

The criterion minimized in Step 1 & 2 writes

$$Q(\mathbf{a}, \mathbf{s}) = \sum_{k=p+1}^{N} \left| s_k + \sum_{l=1}^{p} a_l \, s_{k-l} \right|^2 = \sum_{k=p+1}^{N} |e_k|^2 . \tag{4}$$

This criterion represents the squared error of reconstruction (difference between original and reconstructed signal) but also an estimation of the variance of the excitation signal (up to a multiplicative term). It is successively minimized w.r.t. $\mathbf{a}$ (Step 1) and $\mathbf{s}$ (Step 2). Note that the optimization performed in Step 2 actually only concerns $\mathbf{s}(T)$, as the samples $\mathbf{s}(\widetilde{T})$ are supposed to be known and shall not be updated.

3.2 Step 1: Estimation of the AR parameters

Given an estimate of the missing samples (a rough choice for initialization is to set the values of the missing samples to 0) and an order $p \in \mathbb{N}^*$, we want to estimate the AR parameters $\hat{\mathbf{a}} \in \mathbb{R}^p$ which minimize the criterion $Q(\mathbf{a}, \mathbf{s})$ defined in (4) with respect to $\mathbf{a}$.

3.2.1 Autocovariance method

Let $c_{i,j}$ be an unbiased estimate of the autocovariance function from samples $s_1, \ldots, s_N$ for delay $i - j$

$$c_{i,j} = \frac{1}{N-p} \sum_{k=p+1}^{N} s_{k-i}\, s_{k-j}. \tag{5}$$

Then, by setting $\mathbf{C} = (c_{i,j})_{1 \leq i,j \leq p}$ and $\mathbf{c_0} = [c_{0,1} \ldots c_{0,p}]^t$, we can rewrite criterion (4) as:

$$Q(\mathbf{a}, \mathbf{s}) = \mathbf{a}^t\, \mathbf{C}\, \mathbf{a} + 2\, \mathbf{a}^t\, \mathbf{c_0} + c_{0,0}. \tag{6}$$

The optimal AR parameters are obtained by setting $\nabla_{\mathbf{a}} Q(\mathbf{a}, \mathbf{s}) = 0$ where $\nabla_{\mathbf{a}}$ denotes the gradient with respect to $\mathbf{a}$, and thus are obtained by solving

$$\mathbf{C}\hat{\mathbf{a}} = -\mathbf{c_0}. \tag{7}$$

The problem of estimating AR parameters from data samples has been given much attention [11] and the method described here is only one among many others. As it requires the use of an estimate of the autocovariance function, it often referred to as the *autocovariance method*. There exist some fast methods for solving this problem [15], which provide a resolution in $O(p^2)$.

3.2.2 Autocorrelation method

Nevertheless, there are variants to this method which are easier to implement and with a lower complexity. Variants consist in using different estimators of the autocovariance function (biased or unbiased) or different index ranges (by padding the signal with zeros). Provided that N is large enough (compared to p), it is fair to assume that all these methods give similar results, as the eventual bias becomes undetectable [16]. Yet, the choice of the variant may be crucial when dealing with shorter signals [2]. As we work here with audio signals samples at $F_s = 44100$ Hz, which are likely to contain a large number of samples, we are able to choose the estimation method by focusing on its computational time.

Among these variants the *autocorrelation method* uses the relationship between the AR parameters and the autocorrelation function (known or estimated), previously defined as the Yule-Walker equations (3). This method is based on a biased estimator of the autocorrelation function

$$\hat{R}(\tau) = \frac{1}{N} \sum_{k=\tau+1}^{N} s_k\, s_{k-\tau}. \tag{8}$$

The estimation of the AR parameters (7) is rewritten as

$$\begin{bmatrix} \hat{R}(0) & \hat{R}(1) & \cdots & \hat{R}(p-1) \\ \hat{R}(1) & \hat{R}(0) & \cdots & \hat{R}(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{R}(p-1) & \hat{R}(p-2) & \cdots & \hat{R}(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = - \begin{bmatrix} \hat{R}(1) \\ \hat{R}(2) \\ \vdots \\ \hat{R}(p) \end{bmatrix}. \tag{9}$$

Just like the autocovariance method, there exist a fast algorithm (Levinson-Durbin algorithm [13]) dedicated to Toeplitz matrices which allows to solve (9) in $O(p^2)$ operations. As this algorithm is easy to understand and to implement, we have chosen to use the autocorrelation method for the estimation of AR parameters. Furthermore, this algorithm is faster than the one provided for the autocovariance method.

3.2.3 Levinson-Durbin algorithm for solving the autocorrelation problem

The Levinson-Durbin algorithm is a procedure which iteratively evaluates the model parameters $\{a_1, \ldots, a_{l+1}, \sigma^2_{l+1}\}$ of order $l+1$ from those of order l, until the desired order p. Let us write

$$\hat{\mathbf{R}}_l = \begin{bmatrix} \hat{R}(0) & \hat{R}(1) & \cdots & \hat{R}(l) \\ \hat{R}(1) & \hat{R}(0) & \cdots & \hat{R}(l-1) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{R}(l) & \hat{R}(l-1) & \cdots & \hat{R}(0) \end{bmatrix}, \tag{10}$$

then we have

$$\hat{\mathbf{R}}_{l+1} = \begin{bmatrix} & & & \hat{R}(l+1) \\ & \hat{\mathbf{R}}_l & & \hat{R}(l) \\ & & & \vdots \\ \hat{R}(l+1) & \hat{R}(l) & \cdots & \hat{R}(0) \end{bmatrix} = \begin{bmatrix} \hat{R}(0) & \hat{R}(1) & \cdots & \hat{R}(l+1) \\ \hat{R}(1) & & & \\ \vdots & & \hat{\mathbf{R}}_l & \\ \hat{R}(l+1) & & & \end{bmatrix}. \tag{11}$$

One obtains $\hat{\mathbf{R}}_{l+1}$ from $\hat{\mathbf{R}}_l$ by adding a row and a column. We can therefore combine (3) and (11) so as to get

$$\begin{bmatrix} & & & \hat{R}(l+1) \\ & \hat{\mathbf{R}}_l & & \hat{R}(l) \\ & & & \vdots \\ & & & \vdots \\ \hat{R}(l+1) & \hat{R}(l) & \cdots & \hat{R}(0) \end{bmatrix} \begin{bmatrix} 1 \\ a_1^l \\ \vdots \\ a_l^l \\ 0 \end{bmatrix} = \begin{bmatrix} \sigma_l^2 \\ 0 \\ \vdots \\ 0 \\ k_{l+1}\,\sigma_l^2 \end{bmatrix}. \tag{12}$$

$a_1^l, \ldots, a_l^l$ and σ_l^2 are respectively the estimated AR parameters and the estimated variance of excitation process e_k at step l. By definition, we set $k_{l+1} = \frac{1}{\sigma_l^2} \sum_{j=1}^{l} a_j^l \, \hat{R}(l+1-j)$: in the literature k_{l+1} is called the reflection coefficient.

By using the second formulation of $\hat{\mathbf{R}}_{l+1}$ in (11) and using the circulant form of matrix $\hat{\mathbf{R}}_{l+1}$, we get another formulation for (12)

$$\begin{bmatrix} \hat{R}(0) & \hat{R}(1) & \cdots & \cdots & \hat{R}(l+1) \\ \hat{R}(1) & & & & \\ \vdots & & \hat{\mathbf{R}}_l & & \\ \vdots & & & & \\ \hat{R}(l+1) & & & & \end{bmatrix} \begin{bmatrix} 0 \\ a_l \\ \vdots \\ a_1 \\ 1 \end{bmatrix} = \begin{bmatrix} k_{l+1}\,\sigma_l^2 \\ 0 \\ \vdots \\ 0 \\ \sigma_l^2 \end{bmatrix}. \tag{13}$$

Calculating (12) - $k_{l+1}\times$ (13) gives:

$$\hat{\mathbf{R}}_{l+1} \begin{bmatrix} 1 \\ a_1 - k_{l+1}\,a_l \\ \vdots \\ a_l - k_{l+1}\,a_1 \\ -k_{l+1} \end{bmatrix} = \begin{bmatrix} \sigma_l^2(1-k_{l+1}^2) \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix} \tag{14}$$

which can be interpreted as a Yule-Walker equation of order $l+1$ by setting:

- $k_{l+1} = \frac{1}{\sigma_l^2} \sum_{j=1}^{l} a_j^l \, \hat{R}(l+1-j)$
- $a_{l+1}^{l+1} = -k_{l+1}$

- $a_j^{l+1} = a_j^l - k_{l+1}\, a_{l+1-j}^l \quad for\ 1 \leq j \leq l$
- $\sigma_{l+1}^2 = \sigma_l^2(1 - k_{l+1}^2)$

Algorithm 1 describes the Levinson-Durbin algorithm following from these recursive relations.

Algorithm 1: Levinson-Devinson algorithm

Input: $\hat{R}(0), \ldots, \hat{R}(p)$
Output: $a_1, \ldots, a_p$

Initialization : $a_1^{old} = -\frac{\hat{R}(1)}{\hat{R}(0)}$, $\sigma_1^2 = (1 - (\frac{\hat{R}(1)}{\hat{R}(0)})^2)\hat{R}(0)$;
for $l = 1$ **to** $p - 1$ **do**
  $k_{l+1} = \frac{1}{\sigma_l^2} \sum_{j=1}^{l} a_j^{old}\, \hat{R}(l + 1 - j)$;
  $a_{l+1} = -k_{l+1}$;
  $\sigma_{l+1}^2 = \sigma_l^2(1 - k_{l+1}^2)$;
  for $j = 1$ **to** l **do**
    $a_j = a_j^{old} - k_{l+1}\, a_{l+1-j}^{old}$;
  end
  $\mathbf{a}^{old} = \mathbf{a}$;
end

3.3 Step 2: Estimation of the missing samples

Given an estimate of the AR parameters $\hat{\mathbf{a}}$, we want to interpolate the missing samples $\mathbf{s}(T)$ so as to minimize the criterion $Q(\mathbf{a}, \mathbf{s})$ defined in (4). We can rewrite criterion (4) as

$$Q(\mathbf{a}, \mathbf{s}) = \mathbf{s}(T)^t\, \mathbf{B}\, \mathbf{s}(T) + 2\, \mathbf{s}(T)^t\, \mathbf{d} + \Gamma\left(\mathbf{s}(\widetilde{T})\right), \tag{15}$$

where $\Gamma\left(\mathbf{s}(\widetilde{T})\right)$ is a term only depending on $\mathbf{s}(\widetilde{T})$. The expressions of $\mathbf{B}$ and $\mathbf{d}$ can be obtained by identification. Intuitively, $\mathbf{B}$ rules the dependencies between the unknown samples, while $\mathbf{d}$ links the unknown samples with the known ones.

- $\mathbf{B}$ is a $|T| \times |T|$ matrix with entries

$$\forall (t, t') \in T^2 \quad b_{t,t'} = \begin{cases} \sum_{l=0}^{p-|t-t'|} a_l\, a_{l+|t-t'|} & \text{if } 0 \leq |t - t'| \leq p \\ 0 & \text{else} \end{cases} ; \tag{16}$$

- $\mathbf{d}$ is a $|T|$ vector with entries:

$$\forall t \in T \quad d_t = \sum_{\substack{-p \leq k \leq p \\ t-k \in \widetilde{T}}} b_{|k|}\, s_{t-k}. \tag{17}$$

The solution of the minimization of (15) w.r.t. $\mathbf{s}(T)$ is:

$$\mathbf{B}\, \mathbf{s}(T) = -\mathbf{d} \tag{18}$$

It can easily be proved that matrix $\mathbf{B}$ is symmetric and positive definite. There exist some general algorithms for solving (18) in $O(|T|^3)$ operations: we used a method based on the Cholesky decomposition of matrix $\mathbf{B}$. The first step consists in calculating the Cholesky decomposition (symmetric indefinite factorization) $\mathbf{B} = \mathbf{L}\mathbf{D}\mathbf{L}^t$ with $\mathbf{L}$ is a lower triangular matrix with constant values 1 on its main diagonal and $\mathbf{D}$ being a diagonal matrix [20]. Equation (18) is solved by successively solving the two triangular systems $\mathbf{L}\, \mathbf{x} = -\mathbf{d}$ and $\mathbf{L}^t\, \mathbf{s}(T) = \mathbf{D}^{-1}\, \mathbf{x}$

3.4 Application to long audio signals

Let us consider a real life scenario for our system. A 30 seconds excerpt with a sampling frequency of 44.1 kHz contains $N = 1323000$ samples. If we introduce 10 bursts per second, each with a duration of 4 milliseconds (176 samples), it gives a total of $|T| = 52800$ missing samples. While one could think of directly applying the method previously described to the whole signal, the number of samples considered is in practice too high to allow an efficient processing:

- It is unrealistic to model several seconds of music with one single AR model. Indeed, the stationarity assumed by the model is only valid on a local scale. Also, the complexity of the considered signal is limited by the order p: in particular, if several instruments and several successive notes are played it is unlikely that we can find p large enough to capture the time progression.
- The computation of autocovariances and autocorrelations in (5) and (8) depends on N and becomes time-consuming, as well as the reconstruction of the missing samples, which implies a $O(|T|^3)$ complexity.

For these reasons, it is proposed to divide the audio signal into frames before processing it with the interpolation algorithm. More precisely, the signal is divided into overlapping frames of length N_w with an hop size of N_h samples. In practice, we choose $N_h = N_w/4$, corresponding to a 75% overlap. In order to reconstruct the information provided by the different overlapping frames which correspond to the same period, we use the overlap-add (OLA) procedure [9]:

1. Pad the input signal with zeros: N_w zeros are added before and after the signal samples
2. Add $\left(\lceil \frac{N+N_w}{N_h} \rceil N_h - N_w\right) - N$ zeros at the end of the signal so as to round up the number of frames
3. Divide the signal into overlapping frames of length N_w with 75% overlap
4. Apply the algorithm to each frame (in particular, the p first and last samples of the frame are not processed)
5. Multiply each frame with a Hamming window of size N_w.
6. Add iteratively all the frames with a 75% overlap so as to reconstruct the signal
7. Remove the N_w first and the $\left(\lceil \frac{N+N_w}{N_h} \rceil N_h\right) - N$ last samples

Note that due to the chosen overlap and window, this procedure allows a perfect reconstruction: if no processing is applied on Step 4, the output signal is exactly the input one. The Hamming window used in the process is defined as

$$\forall 1 \leq k \leq N_w \quad w(k) = \frac{1}{4 \times 0.54}\left(0.54 - 0.46\cos\left(2\pi\frac{k-1}{N_w}\right)\right). \tag{19}$$

4 Results

4.1 Evaluation

For our experiments, we consider 30 seconds excepts of audio signals sampled at 44.1 kHz, thus containing $N = 1323000$ samples. As our method assumes that the locations of the missing samples is known, the easiest way to test our system is to randomly create artificial bursts. To this aim, two parameters are introduced:

- The number of desired groups of missing samples: N_{bursts}et
- The maximum length for a burst (in samples): N_{max}

The locations of the missing samples are randomly chosen so as to form N_{bursts} bursts of size randomly chosen between 1 and N_{max} samples. We make sure that the bursts are at least separated with one sample (i.e. there is really N_{bursts} distinct bursts).

Another advantage of using artificial bursts is that the ground truth is available for assessing the performances of the reconstruction system. Given the ground truth signal $\mathbf{s}$ of length N and the reconstructed signal $\hat{\mathbf{s}}$, the signal-to-noise ratio is defined as

$$SNR = 10 \log_{10} \frac{\sum_{k=1}^{N} |s_k|^2}{\sum_{k=1}^{N} |s_k - \hat{s_k}|^2}. \tag{20}$$

Since by definition $\mathbf{s}(\widetilde{T}) = \hat{\mathbf{s}}(\widetilde{T})$, it seems more relevant to evaluate the SNR only on the reconstructed samples, *i.e.* :

$$SNR_{\ T} = 10 \log_{10} \frac{\sum_{k \in T} |s_k|^2}{\sum_{k \in T} |s_k - \hat{s_k}|^2}. \tag{21}$$

4.2 Parameters

There are three parameters that need to be set by the user:

- **The order p of the model.** Intuitively, the order of the model depends on the complexity of the signal on the current frame. In order to get an acceptable interpolation, *Janssen et al* [10] propose to use $p = 3N_{max} + 2$. Indeed, it seems fair to assume that p should at least be greater than N_{max}, so as to only use known samples for reconstruction.
- **The hop size N_h / window size N_w (with $N_w = 4N_h$).** The choice of the window size depends on the order p and on the maximum burst length N_{max}. The estimation of the AR parameters (Step 1) implies the calculation of autocovariances or autocorrelations, which requires a large number of samples (compared to p) to be significant. Intuitively, the larger the order p, the larger the number of samples needed. Furthermore, the number of samples should be greater than N_{max}, as we must insure that within a frame, the number of known samples is at least larger than the number of missing ones, since otherwise the learnt model becomes irrelevant. Also due to the 75% overlap and to the fact that the first and last p samples of the frame are not processed, we must have $N_w > \frac{8}{3}p$ in order to insure that each sample is processed at least in one frame.
- **The number of iterations.** While no proof of convergence for the iterative process described in Section 3.1 has been proposed, in practice we shall see that only a few iterations are needed to obtain acceptable results.

4.3 First results

In this section, we propose to investigate the influence of the order p within the reconstruction process. To this aim, we used a 5 seconds excerpt of pop music (The Beatles) sampled at 44.1 kHz and removed $N_{max} = 200$ samples (4.5 ms). The goal is to interpolate the missing samples with different values of p. We chose the hop size as $N_h = 2\max(p, N_{max})$, which imposed that the window size was greater to both p and N_{max}. The window size was set as $N_w = 4N_h$.

Figure 1 displays the $SNR_{\ T}$ obtained with different values for order p and different numbers of iterations.

We remark that:

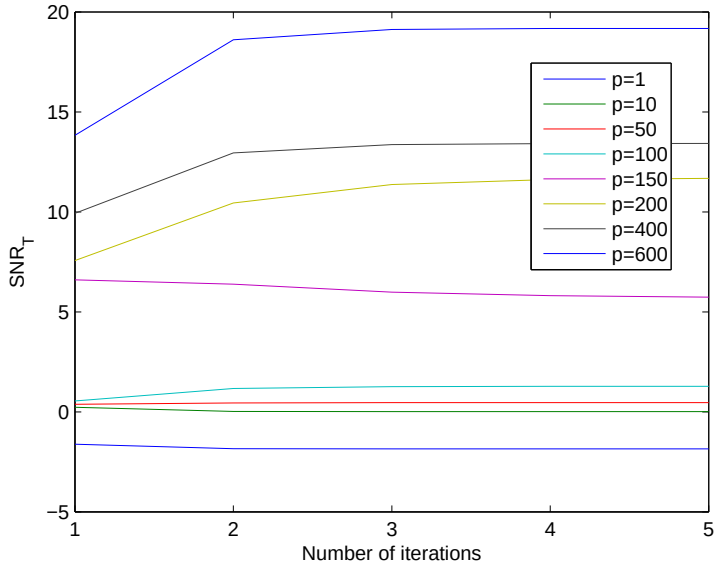

Figure 1: SNR_T obtained for the interpolation of 200 missing samples in a 5 seconds excerpt of pop music for different values of the order p and for different numbers of iterations.

- The SNR_T increases when the order p increases.
- The algorithm converges after 2 or 3 iterations in all cases: after 2 iterations, the variations in SNR_T are limited.
- Increasing the number of iterations does not always increase the SNR_T. Intuitively, if the first reconstruction is irrelevant (for example, this can happen when p is too small), then the iterative process works as a vicious circle: the AR parameters learnt in the next iteration are biased and subsequently produce a bad reconstruction, etc... As a matter of fact, for some values of p, the SNR_T slowly decreases after the first iteration.

In the remainder of this article, the number of iterations will be fixed to two. This is a good compromise between performances and calculation time.

The reconstructed signals obtained after 2 iterations for different values of p are displayed on Figure 2. This figure confirms the intuitive idea that a large value for p yields a finer model for the signal. For $p = 1$ the model is a simple linear interpolation, and when p increases, the reconstructed signal becomes more complex and thus better fits the original signal. This observation is also reflected by the SNR_T previously shown on Figure 1.

Nevertheless, according to Figure 3, this phenomenon seems to occur until p reaches a certain value (here $p = 1000$) and then the SNR_T slowly decreases. Thus the choice of order p seems to be a tricky problem, as it significantly influences the performances in term of SNR_T. Intuitively, we want p to be *large enough* to approximate the signal, but in practice we shall choose a moderate value for p as:

- If p is too large, the SNR_T may decrease.
- As seen in Section 3.2 the computation cost of the AR parameters estimation phase depends on the order p.
- The interpolated samples do not need to perfectly fit the original signal in order to give good perceptual results, as the human ear is not able to detect such small differences. This fact is in particular true when the gaps are small.

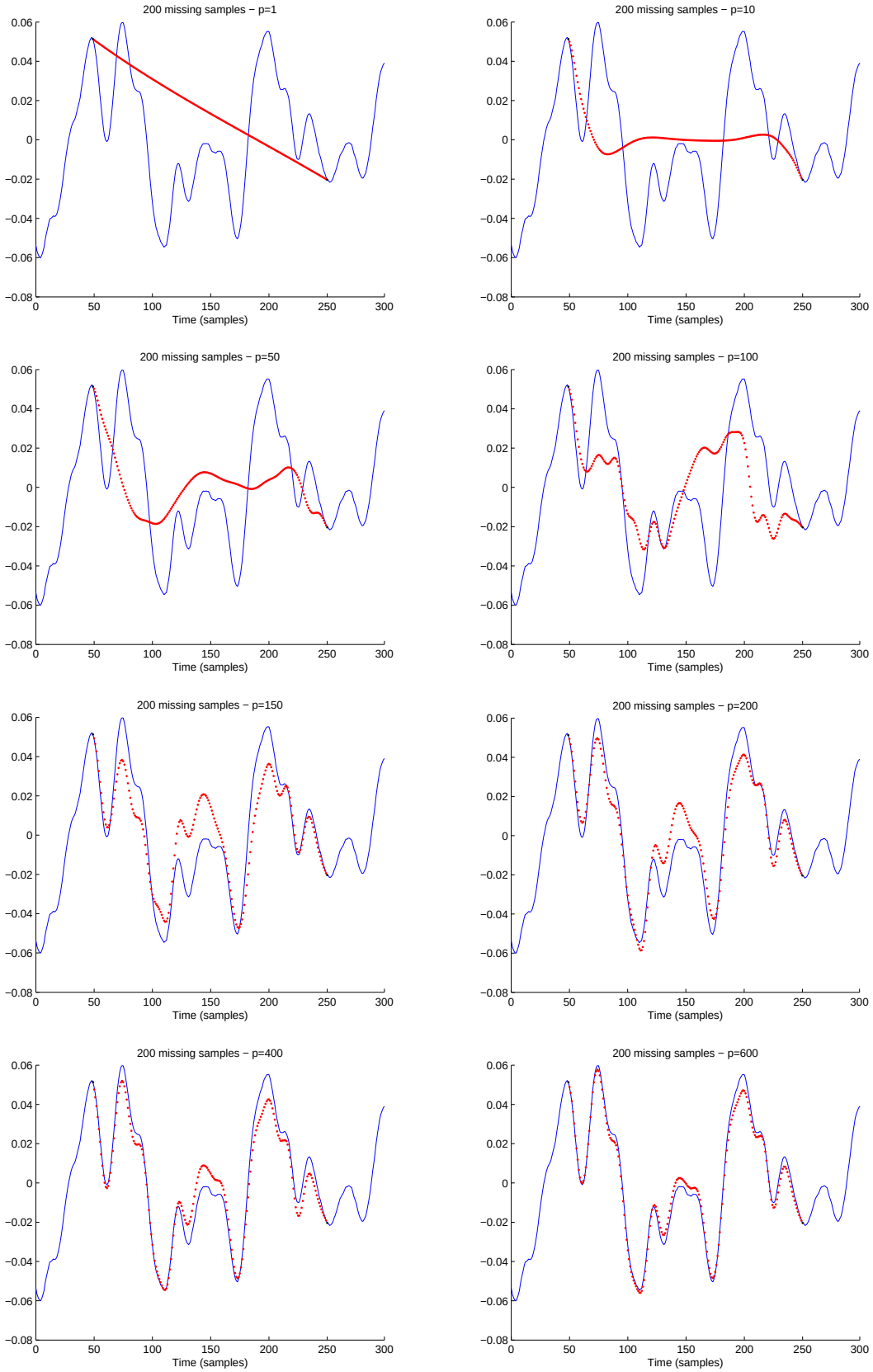

Figure 2: Interpolation of 200 missing samples in a 5 seconds excerpt of pop music for different values of order p. The window size is set to $N_w = 8\max(p, N_{max})$ and the frames have a 75% overlap. The results are obtained after 2 iterations of the algorithm. The blue line stands for the original signal and the red dots to the reconstructed one.

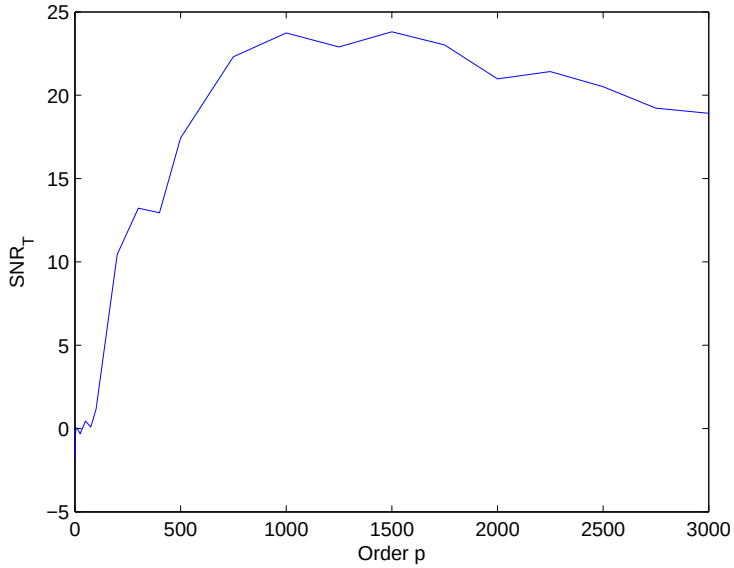

Figure 3: SNR_T obtained for the interpolation of 200 missing samples in a 5 seconds excerpt of pop music for different values of order p after 2 iterations.

To conclude, we must keep in mind that as no prior information on the signal is known (in particular we have no idea of the complexity of the signal on the given frame), the choice of p remains heuristic and that we can only seek for an acceptable value for p, which will provide good perceptual results. In their article, *Janssen et al* [10] propose to use $p = 3N_{max} + 2$. According to our simulations, this choice seems reasonable, even though we have no way to prove or explain it: in particular it is unclear why the authors chose $3N_{max} + 2$ instead of $3N_{max}$.

4.4 Real-life examples

This section describes tests of the reconstruction algorithm in several realistic situations. Three audio signals were considered with a duration of 30 seconds and sampled at 44.1kHz[1]:

- **Speech**: Woman speaking with no audible noise
- **Music**: Artificially synthesized instrumental piece
- **Noise**: Cocktail situation with many speakers and background noise

300 bursts of length comprised between 1 sample and 175 samples (which corresponds to the range $20\mu s$ - 4 ms) were generated. The missing samples were interpolated with the default parameters of the algorithm (2 iterations, $p = 3N_{max} + 2$ and $N_h = 2\max(p, N_{max})$). For each excerpt, the experiment was repeated 100 times to calculate an average SNR_T.

Speech	Music	Noise
11.0 ($\pm$ 1.0)	13.8 ($\pm$ 0.4)	4.0 ($\pm$ 0.2)

Table 1: SNR_T (in dB) obtained by reconstructing 300 bursts of maximum size 175 samples. We represent the average SNR_T of 100 simulations ($\pm$ its standard deviation).

Results are presented on Table 1: while the algorithm obtains good results for speech and music, the SNR_T is lower on the noise signal. Intuitively, these results can be explained by the fact that

[1] http://www.phon.ucl.ac.uk/resource/hearloss/

the AR model is a predictive model which is particularly efficient when a linear relationship can be learnt between samples. In the case of random noise, while the source/filter model is still valid, the excitation/source is prominent over the filter and learning the AR parameters does not help much for the reconstruction. Furthermore, the algorithm is based on the minimization of criterion (4), which can be interpreted as the variance of the excitation. If the signal mainly consists in random noise, this variance is naturally large and attempting to minimize it shall somehow *de-noise* the signal but also prevents reconstructing the original noise signal. On the contrary, harmonic signals such as speech or music are well modeled as the minimization of criterion (4) tends to reconstruct an harmonic structure by reducing the influence of noise.

4.5 Limits of the algorithm

The algorithm proposed in this paper can *a priori* be applied to any type of degradation. Yet, we shall seen that, in practice, the performance of the reconstruction strongly depends on the number of missing samples.

4.5.1 Number of missing samples

This point was investigated by taking a 30 seconds excerpt of pop music (Beatles song), containing voice, drums, guitar, piano, etc... One random burst of size N_{max} was created and the missing samples interpolated. Since the SNR_T should strongly depends on the burst position, the experiment was repeated 1000 times. The average SNR_T was calculated for size N_{max}. Since the default value for p (equal to $3N_{max}+2$) also depends on N_{max} and can influence the results, a second variant was tested for this experiment with the same p for all experiments.

The algorithms were tested with $N_{max} = 1, 50, 100, 150, \ldots, 500$. In Experiment A, p was fixed to $3 \times 500 + 2 = 1502$ for all N_{max} (which corresponds to the default order for $N_{max} = 500$), while in Experiment B, p depends on N_{max} such as $p = 3N_{max} + 2$. In both experiments, the hop and window size were respectively equal to $N_h = 2\max(p, N_{max})$ and $N_w = 4N_h$. Results are presented in Figure 4. For Experiment A, we see that the average SNR_T decreases when the burst length increases and that most of the loss in dB occurs between $N_{max} = 1$ and $N_{max} = 50$. From $N_{max} = 50$, the SNR_T slowly decreases (note that the value used for p is large enough to deal with all the considered burst lengths, which can explain those good results). Interestingly, the evolution is not the same for Experiment B: in particular, under 400 samples, the value of p influences the performances as much as the burst length. It seems that on this example, the default choice for parameter p is appropriate for small bursts (< 50 samples) and large bursts (> 400 samples), but is not optimal for average length bursts. This confirms the fact that the choice of order p is a tricky question and that in real life situations, performances will both depend on the number or missing samples and on the correct estimation of appropriate order p.

4.5.2 Types of bursts

Another question arises from these experiments: the SNR_T obtained for very small bursts is significantly larger than those obtained for long bursts. Is it due to the burst type (i.e. N_{max}) or to the number of missing samples (i.e. $N_{bursts} \times N_{max}$). A second experiment investigated the ability of the algorithm to interpolate different types of bursts. In a 30 seconds excerpt of pop music, vairous types of bursts were tested while keeping a fixed number (2000) of missing samples. It seems natural to think that the algorithm should perceive better 2000 disjoint bursts of 1 sample, than 1 burst of 2000 samples. Various burst lengths were tested N_{max} (1, 2, 4, 5, 8, 10, 16, 20, 25, 40, 50, 80, 100, 125, 200, 250, 400, 500, 1000, 2000 samples). For each considered N_{max}, the number of burst N_{bursts} is adapted to have $N_{max} \times N_{bursts} = 2000$. This experiment was repeated 100 times and the results are presented on Figure 5.

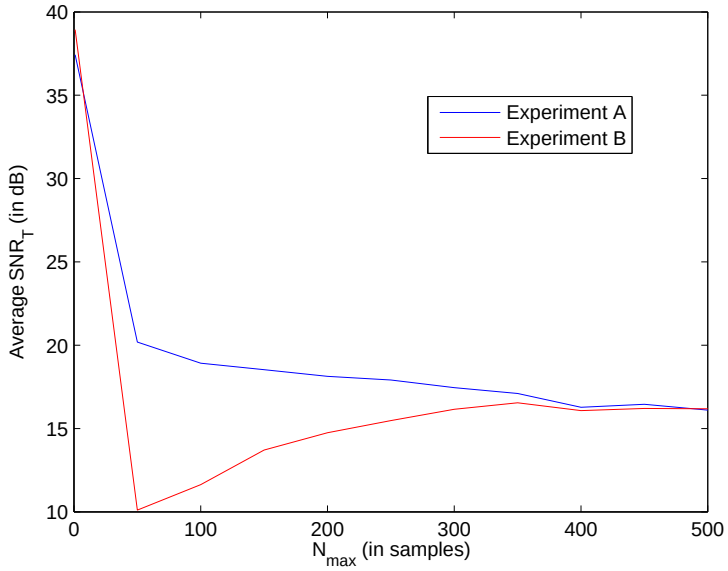

Figure 4: Average SNR_T obtained for the interpolation of one burst of N_{max} missing samples after two iterations. In Experiment A, p is constant and equal to 1502 while in Experiment B, $p = 3N_{max} + 2$.

Interestingly, while the number of missing samples $N_{max} \times N_{bursts}$ is the same in all these simulations, the SNR_T varies much according to N_{max}. The general behaviour observed on this experiment is indeed comparable with the one of Experiment B in Figure 4 for the range $1 \leq N_{max} \leq 500$ (even though the number of bursts N_{bursts} is no longer equal to 1): again, the order p seems to influence the results, as well as N_{max}. The degradation due to presence of multiple bursts is limited (approximately -2dB for $100 \leq N_{max} \leq 500$) and the number of bursts N_{bursts} does not have a strong influence on the performance.

From these two simulations we can conclude that:

- Our algorithm performance depends on the size of the bursts it has to process rather than on the number of observed bursts.
- The choice of p is a tricky problem that can influence the quality of the obtained results.

5 Conlusion

The algorithm presented in this article is a straightforward application of the autoregressive hypothesis to audio signals for the interpolation of missing or degraded samples. Efficient algorithms exist for the resolution of the two main steps of the method (the parameter estimation and the calculation of the missing samples), allowing to provide a good quality reconstruction in a limited time. Provided that the bursts present in the signal have a reasonable size, this method obtains good results on speech and music signals. Nevertheless, the heuristic choice of order p is a tricky problem which sometimes prevents from achieving optimal results.

References

[1] A. Adler, V. Emiya, M.G. Jafari, M. Elad, R. Gribonval, and M.D. Plumbley. Audio inpainting. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 20(3):922–932, 2012.

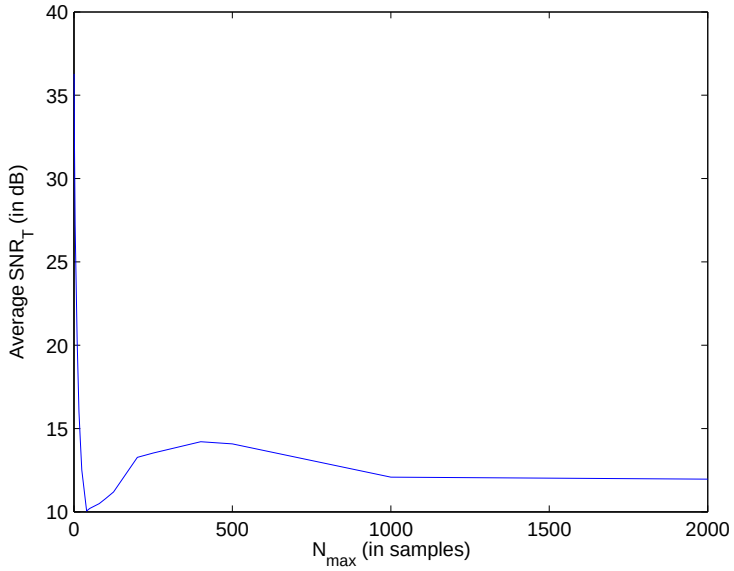

Figure 5: Average SNR_T obtained for the interpolation of N_{bursts} bursts of N_{max} missing samples after 2 iterations (such as $N_{max} \times N_{bursts} = 2000$).

[2] M.J.L. De Hoon, T. Van der Hagen, H. Schoonewelle, and H. Van Dam. Why Yule-Walker should not be used for autoregressive modelling. *Annals of Nuclear Energy*, 23(15):1219–1228, 1996.

[3] P.A.A. Esquef, V. Välimäki, K. Roth, and I. Kauppinen. Interpolation of long gaps in audio signals using the warped Burgs method. In *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, pages 08–11, London, UK, 2003.

[4] W. Etter. Restoration of a discrete-time signal segment by interpolation based on the left-sided and right-sided autoregressive parameters. *IEEE Transactions on Signal Processing*, 44(5):1124–1135, 1996.

[5] S.J. Godsill and P.J.W. Rayner. A Bayesian approach to the detection and correction of error bursts in audio signals. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 261–264. IEEE, 1992.

[6] S.J. Godsill and P.J.W. Rayner. A Bayesian approach to the restoration of degraded audio signals. *Speech and Audio Processing, IEEE Transactions on*, 3(4):267–278, 1995.

[7] S.J. Godsill and P.J.W. Rayner. *Digital Audio Restoration - A Statistical Model-Based Approach*, chapter 5 - Removal of clicks, pages 99–134. Springer-Verlag London, 1998.

[8] S.J. Godsill, P.J.W. Rayner, and O. Cappé. Digital audio restoration. *Applications of digital signal processing to audio and acoustics*, pages 133–194, 2002.

[9] D. Griffin and J. Lim. Signal estimation from modified short-time Fourier transform. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 32(2):236–243, 1984.

[10] A. Janssen, R. Veldhuis, and L. Vries. Adaptive interpolation of discrete-time signals that can be modeled as autoregressive processes. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34(2):317–330, 1986.

[11] S.M. Kay and S.L. Marple Jr. Spectrum analysis–a modern perspective. *Proceedings of the IEEE*, 69(11):1380–1419, 1981.

[12] M. Lagrange, S. Marchand, and J.B. Rault. Long interpolation of audio signals using linear prediction in sinusoidal modeling. *Journal of the Audio Engineering Society*, 53(10):891–905, 2005.

[13] N. Levinson. The Wiener RMS (root mean square) error criterion in filter design and prediction. *Journal of Mathematics and Physics*, 25:261–278, 1947.

[14] R.C. Maher. A method for extrapolation of missing digital audio data. *Journal of the Audio Engineering Society*, 42(5):350–357, 1994.

[15] M. Morf, B. Dickinson, T. Kailath, and A. Vieira. Efficient solution of covariance equations for linear prediction. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 25(5):429–433, 1977.

[16] M.B. Priestley. *Spectral Analysis and Time Series*. Academic press, 1994.

[17] J.W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1971.

[18] S.V. Vaseghi and P.J.W. Rayner. A new application of adaptive filters for restoration of archived gramophone recordings. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 2548–2551, New York, New York, USA, 1988.

[19] SV Vaseghi and PJW Rayner. Detection and suppression of impulsive noise in speech communication systems. In *Communications, Speech and Vision, IEE Proceedings I*, volume 137, pages 38–46, 1990.

[20] D. Watkins. *Fundamentals of Matrix Computations*. New York: Wiley, 1991.

[21] P.J. Wolfe and S.J. Godsill. Interpolation of missing data values for audio signal restoration using a gabor regression model. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 5, pages 514–517, Philadelphia, PA, USA, 2005.