PERCEPTUAL LINEAR FILTERS: LOW-ORDER ARMA APPROXIMATION FOR SOUND SYNTHESIS

Rémi Mignot , Vesa Välimäki*

Aalto University,
Department of Signal Processing and Acoustics
Otakaari 5A, 02150 Espoo, Finland

ABSTRACT

This paper deals with the approximation of a given frequency response by a low-order linear ARMA filter (Auto-Regressive Moving Average). The aim of this work is the audio synthesis, then to improve the perceptual quality, a criterion based on human listening is defined and minimized. Two complementary approaches are proposed here for solving this non-linear and non-convex problem: first, a weighted version of the Iterative Prefiltering, second, an adaptation of the Gauss-Newton method. This algorithm is adapted to guarantee the causality/stability of the obtained filter, and eventually its minimum phase property. The benefit of the new method is illustrated and evaluated.

1. INTRODUCTION

The goal of this paper is the approximation of a given frequency response by a low-order linear ARMA filter (Auto-Regressive Moving Average), with a high sampling rate, $F_s \geq 44.1$ kHz. The context of this work is the low-cost sound synthesis of musical tones using the *Source-Filter* principle which consists of the filtering of an excitation signal. Then, because the aim is an audio application, the obtained filter must be as close as possible to the original one in a perceptual sense, rather than using a physical or signal-based criterion.

It is known that in a general case a spectral envelope has a sparser representation with an ARMA model than a purely AR or MA model. It is especially the case for nasal speech, and for musical instruments. For example, even if an ARMA(q, p) filter and an AR(q+p) filter have approximately the same complexity for the time simulation, the ARMA modeling will be more efficient in most of the cases. Some ARMA approximations exist, cf. e.g.: Prony's method [1], Shanks's method [2], the Iterative Prefiltering [3], Durbin's method [4] or the Inverse Linear Prediction [5] (or cf. e.g. [6] for a partial review). Nevertheless, with these methods the cost function is adapted to facilitate the algorithm, and is never adapted to the perception.

A usual idea is to adapt the model to the frequency resolution of the ear. In [7, 8, 9] a warped frequency scale is used to fit the Bark scale, cf. [10, 11], and a warped AR filter is obtained. Unfortunately, first we have shown in [12] that for low-orders, the warped modeling is not satisfying in a perceptual sense. This observation can be explained because the optimization criterion is not fully perceptually based. Moreover, the time-domain implementation of the warped AR filter is two or three times more expensive than a linear AR filter with the same order, cf. e.g. [8].

* This work is funded by the Marie Curie Action project ESUS 299781.

In this work, we propose to directly estimate a linear ARMA filter on the linear frequency scale using the minimization of a perceptually-based criterion. In the context of the Source-Filter principle, the target frequency response is obtained by a spectral envelope estimation of an original sound, which can be periodic. This estimation can be done by the DAP method of [13], the True Envelope of [14, 15], or the True Discrete Cepstrum of [16]. Note that it is also possible to use a post-processing, MTELPC [17] or PCF [18], which provide a "quasi-perceptual" pre-smoothing. These points are not detailed in this work.

This paper is organized as follows: in Sec. 2, the ARMA model is given, and the perceptually-based criterion is defined step by step in Sec. 3. Then, the two parts of the algorithm are given in Sec. 4. Section 5 gives one practical example, and presents a perceptual comparison of the proposed method with other standard methods. Finally, section 6 concludes this paper and gives some perspectives.

2. MODEL

Given a complex frequency response $H(f)$, where f is the frequency in [Hz], this work deals with its approximation by the following ARMA(Q, P) filter

$$\widetilde{H}(z) = \frac{B(z)}{A(z)} = \frac{b_0 + \sum_{q=1}^{Q} b_q z^{-q}}{1 + \sum_{p=1}^{P} a_p z^{-p}}, \qquad (1)$$

where Q and P are the orders of the numerator B and the denominator A respectively. z is the complex variable of the z-transform, which is $z = \mathrm{e}^{j2\pi f/F_s}$ on the unit circle, with f the frequency variable and F_s the sampling rate in [Hz]. The polynomial coefficients b_q and a_p are the variables to optimize.

3. PERCEPTUAL CRITERION

3.1. First criterion

Let us define the following criterion which provides a distance between the target $H(f)$ and the model $\widetilde{H}(f)$:

$$\mathcal{C}_1 = \int_0^{\frac{F_s}{2}} \frac{\left[\sigma(H(f), f) - \sigma(\widetilde{H}(f), f)\right]^2}{\sigma(H(f), f)^2} \, M(df). \qquad (2)$$

This cost function is perceptually meaningful because of the following reasons.

Loudness conversion First, the function $\sigma(X, f)$ is the conversion of the (physical) sound pressure level X in pascals [Pa], to the (perceptual) loudness in sones, depending on the frequency f. The conversion σ is here calculated with the consecutive conversions: $\sigma(X, f) = s(\ell(\delta(X), f))$, where $X_{db} = \delta(X) = 20\log_{10}(|X|/p_0)$ is the standard scale in [dB SPL], with $p_0 = 2\times 10^{-5}$ Pa the reference sound level, $L_p = \ell(X_{db}, f)$ is the conversion from the decibel scale to the phon scale, relative to the equal loudness curves cf. e.g. [19, 20], and $L_s = s(L_p) = 2^{(L_p-40)/10}$ is the conversion to the sone scale, cf. e.g. [21].

Frequency scale Second, the measure $M(df)$ takes the frequency resolution of the ear into account, as the standard warping mentioned earlier. With $m(f)$ the conversion from the linear frequency scale in [Hz] to any warped scale, we write $M(df) = dm(f) = m'(f)df$. For example, with the Mel scale of [22], $m(f) = 2595\log_{10}(1 + f/700)$.

Relative error Third, note that the sone and the phon scales are respectively linear and logarithmic scales in the loudness domain, such as the pascal and the decibel scales in the sound level domain respectively. Then, the relative error is computed in (2) in order to take into account the logarithmic sensitivity of the ear. Remark that it would be also possible to directly define $\mathcal{C}_1$ with the absolute error in the phon scale, logarithmic, but it is equivalent up to the first order and the denominator will be used in next section.

3.2. Modified criterion

For a numerical computation, first a new version of $\mathcal{C}_1$ is derived using a discrete sum. Second, the loudness conversion is simplified using a first-order limited development of $\sigma(\widetilde{H}, f)$ around H. Then $\sigma(\widetilde{H}, f) \approx \sigma(H, f)+\sigma'(H, f)(|\widetilde{H}|-|H|)$ with $\sigma'(X, f) = \partial\sigma(X, f)/\partial|X|$, and the criterion becomes:

$$\mathcal{C}_2 = \sum_{m=1}^{M} \frac{\left(|H_m| - |\widetilde{H}_m|\right)^2 \sigma'(H_m, f_m)^2}{\sigma(H_m, f_m)^2}\, m'(f_m), \qquad (3)$$

where the frequencies f_m uniformly sample the range $[0, F_s/2]$ and $H_m = H(f_m)$. Note that in this work σ and its derivative are computed using the analytical expression of [23].

If the phase of the target response H is known, we can replace $(|H_m| - |\widetilde{H}_m|)^2$ by $|H_m - \widetilde{H}_m|^2$. This actually simplifies the optimization procedure and facilitates the convergence. Note that, only knowing $|H|$, its phase can be recovered assuming a minimum phase system, cf. e.g. [24].

Since a sound with a level below the auditory threshold is imperceptible in principle, the function σ is not defined below this threshold which corresponds to 0 phon. Then with $X_0(f)$ the auditory threshold in pascals, such that $\sigma(X_0(f), f) = s(0) = 2^{-4}$ sons, we define the saturated function

$$\underline{\sigma}(X, f) = \begin{cases} \sigma(X, f) & \text{if } |X| \geq X_0(f) \\ 2^{-4} & \text{if } |X| < X_0(f) \end{cases} \qquad (4)$$

and the saturated derivative $\underline{\sigma'}$ in the same way. Finally, the criterion to minimize is written as

$$\mathcal{C} = \sum_{m=1}^{M} \left|H_m - \widetilde{H}_m\right|^2 W_m^2 \qquad (5)$$

$$\text{with} \quad W_m = \frac{\underline{\sigma}'(H_m, f_m)}{\underline{\sigma}(H_m, f_m)} \sqrt{m'(f_m)}. \qquad (6)$$

In consequence, the criterion $\mathcal{C}$ is just the weighted squared sum of the error, with a weight W_m which takes into account the sensitivity of the ear to the frequencies via $m(f)$, to the sound level via σ, and to the auditory threshold via the "saturated" $\underline{\sigma}$.

3.3. Remarks

Because most of the time the sensitivity of the recording device is not available, a possible way to adapt the unscaled recording sound to the pascal scale is just by applying a gain which gives the desired sound level. For example $X_{db} = 70$ dB SPL is a normal level for a single musical instrument.

In (4), $X_0(f)$ is the absolute auditory threshold. It is also possible to combine it with the simultaneous masking threshold, cf. e.g. [25], calculated from the target response $H(f)$. Nevertheless, this strategy seems hazardous because $H(f)$ and $\widetilde{H}(f)$ are not "concrete" spectra, but "abstract" spectral envelopes.

4. OPTIMIZATION ALGORITHM

With an ARMA modeling $\widetilde{H} = B/A$, the minimization of (5) is not trivial because the error is non-linear with the coefficients a_p of the denominator A and this optimization problem is not convex. In this section two complementary iterative algorithms are proposed to minimize the cost function $\mathcal{C}$. The first approach is based on the *Iterative Prefiltering* of [3]. It is referred as the Mode 1 because its result is used as initialization of the second one, the Mode 2, which is based on the Gauss-Newton algorithm, cf. e.g. [26].

4.1. Mode 1: Weighted Iterative Prefiltering

Instead of optimizing a non-linear problem, the Iterative Prefiltering method, initially proposed in [3], consists in iteratively solving linear sub-problems using the Least Mean Square optimization (LMS). For that, the criterion $\mathcal{C}$ is modified at every iteration using the previous estimation.

4.1.1. Secondary criterion

With A' the estimated denominator of the previous iteration, the multiplication of the error $e_m := (H_m - B_m/A_m)W_m$ of (5) by A_m/A'_m, leads to the secondary criterion which follows

$$\mathcal{C}' := \sum_{m=1}^{M} \left|A_m \frac{H_m W_m}{A'_m} - B_m \frac{W_m}{A'_m}\right|^2. \qquad (7)$$

Since A' is known, the new defined error is linear with the parameters a_p and b_q, and the minimization of $\mathcal{C}'$ can be solved using the standard LMS. This procedure is equivalent to the Iterative Prefiltering method of Steiglitz and McBride, cf. [3, 27], with an additional frequency weight $W(f)$. It is important to note that at the convergence, if it happens, A/A' goes toward 1, consequently the secondary criterion $\mathcal{C}'$ gets closer to the primary criterion $\mathcal{C}$.

4.1.2. Linear optimization

In (7), $\mathcal{C}'$ is given in the frequency domain, but considering the Hermitian symmetry of H, $\widetilde{H}$, and W, and using the Parseval theorem, we can write it in the discrete time domain to avoid complex numbers. Whereas the computation of h_n, the time response of H, does not cause any issue, the direct inverse Fourier transform of W

makes a non-causal response because W is real. Nevertheless, $\mathcal{C}'$ is invariant by adding a phase to W, then to avoid time aliasing, we define w_n as the minimum phase solution of W, cf. e.g. [24].

With $y := (h * w)/A'$ and $x := w/A'$, where the symbol $*$ denotes the convolution product and $./A'$ denotes the prefiltering by the AR filter $1/A'$, the secondary criterion $\mathcal{C}'$ is written

$$\mathcal{C}' = \frac{1}{2}\sum_{n=0}^{N-1}\Big(y_n + \sum_{p=1}^{P} a_p y_{n-p} - \sum_{q=0}^{Q} b_q x_{n-q}\Big)^2, \qquad (8)$$

Note that, even if the computations of w and $(h * w)$ may be quite expensive, they are done only once before the first iteration.

Then, for $n \in [1, N]$, $p \in [1, P]$ and $q \in [1, Q+1]$, and with the matrix transpose $.^T$, we define the column vectors Y and μ such that $Y_n = y_{n-1}$ and $\mu = [a_1, \dots a_P, b_0, b_1, \dots b_Q]^T$, and we define the block matrix $\Phi = [-\Phi_y, \Phi_x]$, with the Toeplitz matrices $\Phi_y[n, p] = y_{n-1-p}$ and $\Phi_x[n, q] = x_{n-q}$. Note that considering causal signals, $y_n = 0$ and $x_n = 0$ for $n < 0$.

Consequently, the matrix form of the secondary criterion is: $\mathcal{C}' = \frac{1}{2}(Y - \Phi\mu)^T(Y - \Phi\mu)$, and if Φ is full rank, the optimal solution in the LMS sense is given by solving the linear problem $(\Phi^T\Phi)\mu = (\Phi^T Y)$, which can be written, cf. e.g. [26],

$$\mu = \left(\Phi^T\Phi\right)^{-1}\Phi^T Y = \Phi^\dagger Y. \qquad (9)$$

As it is implicitly mentioned in [3], at the first iteration, we simply choose $A' = 1$. Note that without weight W, at the first iteration $x_n = \delta_n$, the Dirac distribution, and the first estimated B and A are the solutions of Prony's method.

4.1.3. Properties

Remark that the positions of the roots of A and B are not ensured to be inside the unit circle, which means that the causality/stability and the minimum phase property cannot be controled. Even if this problem occurs rarely if the target H checks these properties, it may be overcome by testing the desired properties at every iteration, using the Jury criterion for example [28], and by recomputing the LMS solution with lower orders, P and Q. This strategy usually leads to good properties, but with eventually a worse $\mathcal{C}'$.

As mentioned in [3], the convergence of this iterative procedure is not guaranteed. Nevertheless, we observed in every experiment an efficient decrease in the criterion $\mathcal{C}$ and we observed the convergence of the coefficients of A. Unfortunately, first, some conditioning problems usually appear after some iterations, when $\Phi^T\Phi$ is numerically singular, and second, even if $\mathcal{C}'$ get closer to $\mathcal{C}$, the partial derivatives of $\mathcal{C}'$ are different from those of $\mathcal{C}$, which explains why this algorithm usually does not converge to a local minimum in the sense of $\mathcal{C}$.

In [3], a second iterative procedure, the respective Mode 2, has been proposed to improve the estimation of the first one. This point is not detailed here, we refer the interested readers to [3]. In favorable cases, this new mode converges to the closest local minimum, but again, the convergence is not guaranteed, and may diverge if its initial value is far from a local minimum. Moreover, in our experiments, some conditioning problems may still appear. Finally, as with the Mode 1, the causality/stability, and the minimum phase property, of the obtained filter cannot be clearly guaranteed.

In the next section, we proposed another Mode 2 which is based on the Gauss-Newton algorithm. First, this method has a better convergence, second, the conditionning is efficiently improved, and third, this approach can guarantee the causality/stability and the minimum phase property of the estimated filter.

4.2. Mode 2: Non-linear optimization

We propose in this section an adaptation of the iterative Gauss-Newton algorithm, cf. e.g. [26], with constraints for the causality/stability of the filter, and eventually its minimum phase property. Compared to the standard gradient descent, its convergence is usually faster, and it avoids the successive 1D optimizations along the direction of maximal descent.

4.2.1. Gauss-Newton algorithm

Newton's algorithm is based on a second-order limited development of the criterion. Starting from an initial parametrization of the model, the parameters are iteratively updated by the optimum solution of the quadratic form given by the limited development around the previous parameters. If the cost function is quadratic, the algorithm converges in one step, and if it is not quadratic but sufficiently regular, it naturally converges to the nearest local minimum in some iterations.

With μ^k the column vector collecting the current parameters of the model, the following parameters are given by:

$$\mu^{k+1} = \mu^k - \Omega_{\mathcal{C}}^{-1}(\mu^k)\nabla_{\mathcal{C}}(\mu^k), \qquad (10)$$

with $\nabla_{\mathcal{C}}(\mu)$ the gradient vector and $\Omega_{\mathcal{C}}(\mu)$ the Hessian matrix: $\nabla_{\mathcal{C}}[i] = \partial\mathcal{C}/\partial\mu_i$ and $\Omega_{\mathcal{C}}[i, j] = \partial^2\mathcal{C}/\partial\mu_i\partial\mu_j$.

The Gauss-Newton algorithm differs from the previous one by the approximation of the Hessian matrix. This approximation facilitates the computation and is justified by the fact that the criterion is the squared sum of the magnitude of the error e_m, cf. e.g. [26]. With

$$e_m := (H_m - \widetilde{H}_m)W_m, \qquad (11)$$

the criterion is written $\mathcal{C} = E^H E$, where E is the column vector of the error e_m and $.^H$ is the Hermitian transpose.

Now, defining $J_e(\mu)$ as the Jacobian matrix of E, such that $J_e[m, i] = \partial e_m/\partial\mu_i$, the gradient vector of $\mathcal{C}$ becomes $\nabla_{\mathcal{C}}(\mu) = 2J_e(\mu)^H E(\mu)$, and the approximated Hessian matrix is written $\Omega_{\mathcal{C}}(\mu) = 2J_e(\mu)^H J_e(\mu)$. Consequently

$$\Omega_{\mathcal{C}}^{-1}\nabla_{\mathcal{C}} = \left(J_e^H J_e\right)^{-1} J_e^H E = J_e^\dagger E. \qquad (12)$$

Nevertheless, with (10), the algorithm may diverge in some cases. Then, it is usual to introduce a relaxation factor $\lambda_k \leq 1$, and the algorithm becomes

$$\mu^{k+1} = \mu^k - \lambda_k\Omega_{\mathcal{C}}^{-1}(\mu^k)\nabla_{\mathcal{C}}(\mu^k). \qquad (13)$$

A simple strategy for the choice of λ_k is to successively reduce its value until $\mathcal{C}(\mu^{k+1}) < \mathcal{C}(\mu^k)$. Note that if the Hessian matrix is positive-definite, there always exists a $\lambda_k > 0$ providing a decreasing criterion. Here, we first test $\lambda = 1$ to accelerate the convergence, and we divide it by 2 if $\mathcal{C}$ does not decrease.

4.2.2. Optimization of the ARMA model

Starting from the standard ARMA modeling of (1), to improve the conditioning we introduce a gain g and we force $b_0 = 1$, without loss of generality. The model is then given by $\widetilde{H}(z) = gB(z)/A(z)$, and the parameters to identify are the gain g and the coefficients a_p and b_q of the polynomials $A(z)$ and $B(z)$ respectively, with $p \in [1, P]$ and $q \in [1, Q]$. Moreover, to avoid the singular case in $g = 0$, we do the change of variable $g = \zeta\,\mathrm{e}^\gamma$,

where ζ is the sign of the initial gain, and we optimize γ on $\mathbb{R}$ instead of g.

With $z_m = \mathrm{e}^{j2\pi f_m/F_s}$ and $\mu = [\gamma, a_1, \dots a_P, b_1, \dots b_Q]^T$, the Jacobian matrix $J_e(\mu)$ is given by

$$\begin{cases} \dfrac{\partial e_m}{\partial \gamma} = -\zeta \dfrac{B(z_m)}{A(z_m)} \dfrac{\partial\, \mathrm{e}^{\gamma}}{\partial \gamma} W_m & = -\widetilde{H}(z_m) W_m, \\ \dfrac{\partial e_m}{\partial a_p} = g \dfrac{B(z_m)}{A(z_m)^2} \dfrac{\partial A(z_m)}{\partial a_p} W_m & = z_m^{-p} \dfrac{\widetilde{H}(z_m)}{A(z_m)} W_m, \\ \dfrac{\partial e_m}{\partial b_q} = \dfrac{-g}{A(z_m)} \dfrac{\partial B(z_m)}{\partial b_q} W_m & = z_m^{-q} \dfrac{-g}{A(z_m)} W_m. \end{cases}$$

Remark that in (2) and (5), we only have considered unilateral spectra for $f \in [0, F_s/2]$. Then the solution given by (13) could lead to complex coefficients g, a_p and b_q, with no consideration of the range $[F_s/2, F_s]$. Instead of summing the error on the full range $[0, F_s]$, we prove the equivalence of the following update equation

$$\mu^{k+1} = \mu^k - \lambda_k \mathrm{Re}\{\Omega_{\mathcal{C}}(\mu^k)\}^{-1} \mathrm{Re}\{\nabla_{\mathcal{C}}(\mu^k)\}. \qquad (14)$$

where $\mathrm{Re}\{.\}$ is the real part operator, and where $\Omega_{\mathcal{C}}$ and $\nabla_{\mathcal{C}}$ are still computed on the frequency range $[0, F_s/2]$. This equation can be fastly computed by splitting the real parts and the imaginary parts of J_e and E, cf. (12).

Concerning the causality/stability of the obtained filter, and eventually its minimum phase property, it is necessary to add this constraint in the algorithm. Remind that an ARMA filter is a minimum phase system if and only if both poles and zeros are strictly inside the unit circle. To guarantee the desired property at every iteration, we adapt the choice of the relaxation factor λ_k as it is done in Sec. 4.2.1 for the convergence. To study the location of the roots of the polynomials A and B, we use the Jury stability criterion. Note that at some iterations, at the point μ^k, the local properties of $\mathcal{C}$ may attract the algorithm outside the constraint domain, even if the nearest local minimum is inside. Then, choosing a small λ_k allows to stay inside the domain, and most of the time, from the new position μ^{k+1} the algorithm naturally reconverges to the minimum.

4.2.3. Summary of the algorithm

To summarize the complete algorithm, first the Mode 1 iterations, weighted Iterative Prefiltering, are computed using the initialization $A' = 1$. As mentioned above, even if the Mode 1 usually converges, the primary criterion $\mathcal{C}$ may not be strictly decreasing, which means that with a finite number of iterations, the last result may not be the best one in the sense of $\mathcal{C}$. Then, to initialize the Mode 2, the Gauss-Newton algorithm, among all successive results of the Mode 1, we retain this one which minimizes $\mathcal{C}$. Finally, since the convergence of the Gauss-Newton is well-defined, we can use standard stop criteria. Here the algorithm is stopped when the maximal number of iterations is attained, or when the relative difference of two consecutive criteria is smaller than a threshold defined in [%].

5. EXPERIMENTATIONS

5.1. Illustration

The proposed method, which we call the *Perceptual Linear Filter* (PLF), is illustrated in Fig. 1. Using an Oboe tone, B3 ($\sim$ 247 Hz), first the spectral envelope has been estimated with the True Envelope (TE) of [14] and has been slightly smoothed using the PCF approach of [18]. Then, from the magnitude of the obtained frequency response, the phase has been recovered assuming a minimum phase system, cf. e.g. [24]. Finally, the PLF is computed by the algorithm presented in Sec. 4, Mode 1 and 2, using an ARMA(8,8) model, and is compared to Prony's method with the same orders. All frequency responses are displayed in the dB SPL scale, together with the auditory threshold.

As a general trend, we observe that the PLF method focuses the approximation at the lower frequencies, as the standard warping technique (cf. [8, 9]), but especially to those frequencies where the target response $H(f)$ is above the auditory threshold. We can observe that Prony's estimate $\widetilde{H}_1(f)$ does not fit $H(f)$ around 4.5 kHz, but it fits the last formant after 14 kHz which is imperceptible in principle. On the contrary, thanks to the perceptual weighting $W(f)$, cf. (6), the PLF filter $\widetilde{H}_2(f)$ fits $H(f)$ when it is audible, and it strongly smooths it when it is imperceptible.

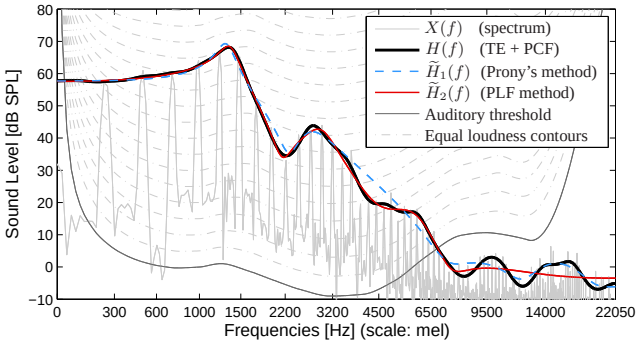

Figure 1: Illustration of the Perceptual Linear Filter (PLF). The orders of the ARMA model are $Q = 8$ and $P = 8$.

5.2. Perceptual evaluation

This section proposes a perceptual evaluation by comparing the PLF method and other methods, using periodic signals imitating instrument sounds. We prefer to perform automatic and objective perceptual tests in order to have an exhaustive evaluation; with many orders, fundamental frequencies, and instruments. A listening test would have required too much time to be done in practice.

First, we define the perceptual measure of the approximation error following some concepts of the PEAQ method, cf. [29, 30]. Then we describe the procedure of the objective tests, and finally the results are presented. Note that to use a neutral evaluation, which does not favor the PLF method, we have to choose an error measure which is as different as possible from the criterion $\mathcal{C}$.

5.2.1. Perceptual Mean Square Error

Let G_m^{ref} and G_m^{test} be the magnitudes in [Pa] of the m-th harmonic of the reference and the test sounds, with the frequencies $f_m = mF_0$ in [Hz] with F_0 the fundamental frequency. Because the approximation is evaluated here, G_m^{ref} and G_m^{test} sample the responses of the target $H(f)$ and the estimate $\widetilde{H}(f)$, at the frequencies f_m.

First, the effect of the middle-ear response is taken into account by multiplying the magnitudes by $\Gamma(f) = 10^{\gamma(f)/20}$ where

$$\gamma(f) = -3.6\left(\frac{f}{1000}\right)^{-0.8} - 0.001\left(\frac{f}{1000}\right)^{4} + 6.5\,\mathrm{e}^{-0.6\left(\frac{f}{1000}-3.3\right)^{2}} \quad (15)$$

The function $\gamma(f)$, which is similar to the middle-ear modeling of [29], is actually the inversion of the auditory threshold modeling given in [31]. Hence, the auditory threshold just corresponds to $\Gamma(f_m)G_m = p_0$ with $p_0 = 2\times10^{-5}$ Pa the reference sound level.

Then, to imitate the auditory system's critical bands, the power of the corrected harmonics, $(\Gamma(f_m)G_m)^2$, is summed by processing a filter bank as done with the PEAQ or the MFCC computation, cf. e.g. [32]. We use here a triangular window with a single overlapping, and 100 filters uniformly spaced in the Bark scale of [10].

Finally, with L_k the outputs of the filter bank, the measure of the perceptual error ε is given by

$$\varepsilon = \frac{1}{K}\left(\sum_{k=1}^{K}\frac{\left(L_k^{\text{ref}} - L_k^{\text{test}}\right)^2}{(L_k^{\text{ref}} + p_0)(L_k^{\text{test}} + p_0)}\right)^{\frac{1}{2}} \quad (16)$$

Here, the auditory threshold is implicitly taken into account because of p_0 which imitates the presence of an inner-ear noise, as with the PEAQ method. Moreover, a relative difference is used here in order to take account for the logarithmic sensitivity of the ear to the sound level. Note that this choice is similar to this one of Sec. 3, but it does not favor the PLF method because all the other methods also minimize a relative error in frequency, as the LPC, cf. [5].

Even if the error measure ε and the criterion $\mathcal{C}$ are based on similar concepts, they are different. This fact allows unbiased results, which does not favor the PLF method.

5.2.2. Experimental procedure

For every half-tone between 220 and 440 Hz, the spectrum envelope of a frame is estimated using the True Envelope of [14]. This frame is chosen around the middle of the sustain part. Then, an accurate AR modeling is done using the TELPC method of [33]. This high-order modeling of the spectral envelope gives the target response $H(f)$.

All tested ARMA methods are computed for the obtained frequency responses H of all half-tones. They provide an ARMA(q,q) approximation in the linear frequency scale. The tested methods are the following:

- **Prony:** The well-known Prony method of [1].
- **StMcB:** The Iterative Prefiltering of Steiglitz and McBride, cf. [3], Mode 1 and 2.
- **WLP*:** The warped LPC modeling, cf. [7, 9], for which the warping factor λ^* is this one which optimally fits the Bark scale, cf. [11]. For $F_s = 44.1$ kHz, $\lambda^* = 0.7564$.
- **WLP.6:** The warped LPC modeling with $\lambda = 0.6$.
- **PLF1:** The Mode 1 of the proposed PLF method, cf. Sec. 4.1.
- **PLF1&2:** The proposed PLF method, Mode 1 and Mode 2 of Secs. 4.1 and 4.2.

Remark that a warped AR(q) filter can be converted in principle into a linear ARMA(q,q) filter. Because the purpose of this paper is the low-cost simulation, we prefer to compare the methods with equal simulation complexity, even if the warped methods have less degrees of freedom.

To cancel the effect of the fundamental frequency, the results of the perceptual tests are printed as a function of the adimensional order $\alpha = q/n_h$, where $n_h = 0.5F_s/F_0$ is the number of harmonics between 0 and $F_s/2$. We tested the adimensional orders $\alpha \in \{0.1, 0.2, 0.3\}$. Table 1 summarizes the used orders q for the lower and the higher fundamental frequencies, 220 and 440 Hz.

	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$
$F_0 = 220$ Hz	10	20	30
$F_0 = 440$ Hz	5	10	15

Table 1: Value of the order q as a function of the adimensional order α and the fundamental frequency F_0, for $F_s = 44.1$ kHz.

For all target $H(f)$ and all approximations $\widetilde{H}(f)$, we derive five harmonic spectra which uniformly sample the associated responses. The fundamental frequencies F_k are chosen on a range of two half-tones around the original fundamental frequency F_0, which means $F_k = F_0\,2^{\frac{k}{2\times12}}$, with $-2 \le k \le 2$. This procedure allows to have a refined evaluation of the response approximation.

Finally, every test spectrum, which samples the approximation $\widetilde{H}(f)$, is compared with its associated reference spectrum, which samples $H(f)$. The perceptual measure of the distance is detailed in Sec. 5.2.1.

5.2.3. Results

The results of the objective evaluation are printed in Fig. 2. The original musical sounds come from the sound database of [34]. Since for all the 13 estimations (half-tones between 220 and 440 Hz), 5 discrete spectra have been synthesized and compared, the mean and the standard deviation of Fig. 2 are computed using 65 computations of the perceptual distance, separately for each method, each tested instrument, and each order α. Here the tested instruments are: clarinet, horn, trumpet and violin; we also tested other sustained instruments, such as: trombone, cello, saxophone, flute, and similar results are obtained.

As a general trend, we observe that the proposed PLF method is among the best methods in all cases, whereas the other methods fail at least once. In consequence, even if the PLF method is not clearly the best method in all cases, it is significantly the more robust. Moreover, comparing with the PLF Mode 1 alone, we observe a slight improvement due to the Mode 2 as expected.

Additionally, the behavior of the warped methods has been already observed in [12] using a listening test. With the optimal warping factor $\lambda^* = 0.7564$, in the sense of [11], for the lower orders the results are sometimes worse than the results of $\lambda = 0.6$. This phenomenon has been explained by analyzing the frequency responses. Indeed, with a strong warping, the high frequencies are compressed around $F_s/2$, and the natural slope of the spectrum becomes stronger in the warped frequency scale. As a result, because of the properties of the LPC, cf. [5], the frequency response at high frequencies is overestimated. One solution is then to reduce the value of λ.

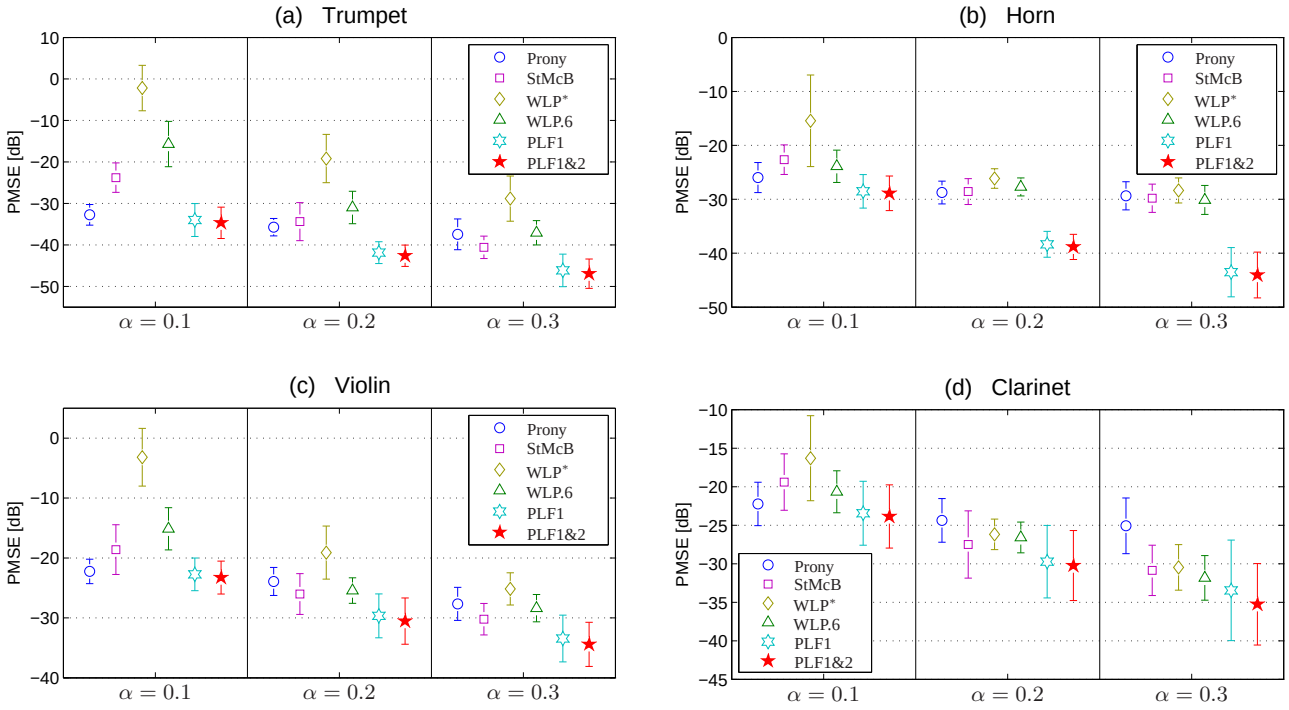

Figure 2: Results of the perceptual evaluation of six ARMA approximation methods, for four instruments and three adimensional orders α. The mean of the perceptual square error (PMSE) is displayed in the decibel scale together with the standard deviation. The tested methods are: Prony's method (Prony), the Iterative Prefiltering Modes 1 and 2 of Steiglitz and McBride (StMcB), the optimal warped LPC (WLP*), the warped LPC with $\lambda = 0.6$ (WLP.6), the PLF method Mode 1 (PLF1) and the PLF method Mode 1 and 2 (PLF1&2).

6. CONCLUSION

In this paper, a novel ARMA approximation for audio signals is presented. It is based on a perceptually meaningful criterion which takes into account the sensitivity of the ear to the frequencies and to sound level via the loudness conversion. The solving algorithm is split into 2 consecutive modes: the first one is a weighted version of the *Iterative Prefiltering* of [3], and the second one is an adaptation of the Gauss-Newton algorithm.

Let's remark that the perceptually-based criterion and the proposed algorithm are two independent contributions of this paper. First the proposed criterion may be optimized using another model or method, second the proposed algorithm can be used with a different frequency weighting. Moreover, even without weighting, for the reasons mentioned earlier (convergence, stability control and conditioning), the proposed Mode 2, is preferable compared to the original Mode 2 of [3].

As illustrated in Fig. 1, this method efficiently focuses the criterion where the original frequency response is audible, and provides less accurate fitting where it is inaudible but with coherent results. A perceptual evaluation is given in Sec. 5.2. Even if the proposed approach does not lead to outstanding results, we notice its stronger robustness. Whereas the other methods may fail in some cases, the PLF method always provides one of the best results.

As a possible improvement of the proposed method, we envisage to apply it with a warped ARMA modeling, cf. [11]. For example, we can notice that a warped ARMA(q,q) filter can be directly converted to an equivalent linear ARMA(q,q) filter. The benefit is to better adapt the model to the criterion $\mathcal{C}$, together with the same number of degrees of freedom, and the same simulation cost. Unfortunately, with a high warping factor λ or high order q, some numerical problems usually occur. In this case, even if the equivalent linear ARMA filter is stable in theory, the finite precision of the floating numbers makes the filter numerically unstable. For example, with $q = 15$, these problems might appear if $\lambda > 0.4$ with the single-precision floating-point.

Unfortunately, this approach may not be suitable in the case of a frame-by-frame analysis-synthesis framework. Indeed, we usually observe strong discontinuities between the estimated spectra of two consecutive frames, which leads to annoying effects. Note that it is also the case in many other ARMA approximation methods. Nevertheless, in the case of the synthesis of a quasi-static spectral envelope, which is under interest in the context of our work, cf. e.g. [35], the proposed PLF method is fully satisfying.

7. REFERENCES

[1] G. Baron 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'alkool, à différentes températures," *J. de l'Ecole Polytechnique (Paris)*, vol. 1, no. 2, pp. 24–76, 1795.

[2] J.L. Shanks, "Recursion filters for digital processing," *Geophysics*, vol. 32, no. 1, pp. 33–51, 1967.

[3] K. Steiglitz and L. McBride, "A technique for the identification of linear systems," *IEEE Trans. Automatic Control*, vol. 10, no. 4, pp. 461–464, 1965.

[4] J. Durbin, "The fitting of time-series models," *Revue de l'Institut International de Statistique*, pp. 233–244, 1960.

[5] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, April 1975.

[6] M.H. Hayes, *Statistical Digital Signal Processing and Modeling*, Wiley, 1999, 624 pages.

[7] A. Härmä, "Linear predictive coding with modified filter structures," *IEEE Trans. Speech and Audio Process.*, vol. 9, no. 8, pp. 769–777, 2001.

[8] A. Härmä, M. Karjalainen, L. Savioja, V. Välimäki, U.K. Laine, and J. Huopaniemi, "Frequency-warped signal processing for audio applications," *J. Audio Eng. Soc.*, vol. 48, no. 11, pp. 1011–1031, 2000.

[9] H.W. Strube, "Linear prediction on a warped frequency scale," *J. Acoust. Soc. Amer.*, vol. 68, no. 4, pp. 1071–1076, 1980.

[10] E. Zwicker and E. Terhardt, "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency," *J. Acoust. Soc. Amer.*, vol. 68, pp. 1523, 1980.

[11] J.O. Smith and J.S. Abel, "Bark and ERB bilinear transforms," *IEEE Trans. Speech and Audio Process.*, vol. 7, no. 6, pp. 697–708, 1999.

[12] R. Mignot, H.-M. Lehtonen, and V. Välimäki, "Warped low-order modeling of musical tones," in *Proc. SMC / SMAC*, Stockholm, Sweden, August 2013.

[13] A. El-Jaroudi and J. Makhoul, "Discrete all-pole modeling," *IEEE Trans. Signal Process.*, vol. 39, no. 2, pp. 411–423, 1991.

[14] A. Röbel and X. Rodet, "Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation," in *Proc. Int. Conf. on Digital Audio Effects (DAFx'05)*, Madrid, Spain, September 2005, pp. 30–35.

[15] S. Imai and Y. Abe, "Spectral envelope extraction by improved cepstral method," *Electronics and Communication*, vol. 62-A, no. 4, pp. 10–17, 1979, in Japanese.

[16] R. Mignot and V. Välimäki, "True discrete cepstrum: an accurate and smooth spectral envelope estimation for music processing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'14)*, Florence, Italy, May 2014, pp. 7515–7519.

[17] F. Villavicencio, A. Röbel, and X. Rodet, "Extending efficient spectral envelope modeling to Mel-frequency based representation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'08)*, Las Vegas, Nevada, USA, March-April 2008, pp. 1625–1628.

[18] R. Mignot and V. Välimäki, "Perceptual cepstral filters for speech and music processing," in *Proc. IEEE Workshop on Applicat. of Signal Process. to Audio and Acoust. (WASPAA'13)*, New Paltz, NY, USA, October 2013.

[19] D.W. Robinson and R.S. Dadson, "A re-determination of the equal-loudness relations for pure tones," *British Journal of Applied Physics*, vol. 7, no. 5, pp. 166, 1956.

[20] B.C.J. Moore, B.R. Glasberg, and T. Baer, "A model for the prediction of thresholds, loudness, and partial loudness," *J. Audio Eng. Soc.*, vol. 45, no. 4, pp. 224–240, 1997.

[21] S.S. Stevens, "A scale for the measurement of a psychological magnitude: loudness," *Psychological Review*, vol. 43, no. 5, pp. 405, 1936.

[22] D. O'Shaughnessy, *Speech Communication: Human and Machine*, Addison-Wesley, 1987.

[23] "ISO 226 (2003). Contours, Acoustics–Normal Equal-Loudness-Level," *International Organization for Standardization*, Geneva, Switzerland.

[24] A.V. Oppenheim and R.W. Schafer, *Discrete-Time Signal Processing*, Prentice Hall, 3rd edition, 2009, 1120 pages.

[25] S.A. Gelfand, *Hearing: An Introduction to Psychological and Physiological Acoustics*, Taylor & Francis, 4th edition, 2004, 512 pages.

[26] E. Walter and L. Pronzato, "Identification of parametric models," *Communications and Control Engineering*, 1997, 413 pages.

[27] K. Steiglitz, "On the simultaneous estimation of poles and zeros in speech analysis," *IEEE Trans. Acoust., Speech and Signal Process.*, vol. 25, no. 3, pp. 229–234, 1977.

[28] E.I. Jury, *Inners and Stability of Dynamic Systems*, Wiley-Interscience, New York, 1974, 328 pages.

[29] ITU-R Recommendation BS.1387-1, "Method for Objective Measurements of Perceived Audio Quality," November 2001.

[30] T. Thiede, W.C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J.G. Beerends, and C. Colomes, "PEAQ-The ITU standard for objective measurement of perceived audio quality," *J. Audio Eng. Soc.*, vol. 48, no. 1/2, pp. 3–29, 2000.

[31] E. Terhardt, "Calculating virtual pitch," *Hearing research*, vol. 1, no. 2, pp. 155–182, 1979.

[32] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoustics, Speech and Signal Process.*, vol. 28, no. 4, pp. 357–366, 1980.

[33] F. Villavicencio, A. Röbel, and X. Rodet, "Improving LPC spectral envelope extraction of voiced speech by True-Envelope estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'06)*, Toulouse, France, May 2006, pp. I869–I872.

[34] University of IOWA, Electronic Music Studios, "Musical Instrument Samples," http://theremin.music.uiowa.edu/MIS.html.

[35] R. Mignot and V. Välimäki, "Extended subtractive synthesis of harmonic musical tones," in *136th Audio Engineering Society Convention (AES136)*, Berlin, Germany, April 2014, number 9038.