Low-order ARMA approximation using a perceptually-based criterion, for sound synthesis *

Rémi Mignot and 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). Because the aim of this work is the audio synthesis, the final judge of the approximation is the human listening, then a perceptually based criterion is defined and minimized. We propose the use of the iterative Gauss-Newton method for the solving of this non-linear and non-convex problem. This algorithm is adapted to guarantee the causality/stability of the obtained filter, and eventually its minimum phase property. The benefit of this new method is illustrated, and its robustness is discussed.

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 a physical or signal-based meaning.

It is known that in a general case a spectral envelope has a sparser representation with an ARMA model than a purely AR 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.: the Prony's method [2], the Shanks's method [3], the Iterative Prefiltering [4], the Durbin method [5] or the Modified Yule-Walker Equations [6] (or cf. e.g. [7] 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 [8, 9, 10] a warped frequency scale is used to fit the Bark scale, cf. [11, 12], and a warped AR filter is obtained. Unfortunately, first we have shown in [**?**] that for low-orders, the warped modeling is not satisfying in a perceptual sense, probably because the optimization criterion is not fully perceptually based, and second because 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.

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 input frequency response is obtained by a spectral envelope estimation of an original sound. It can be done by the DAP method of [13] or the True Envelope of [14, 15] for example. Note that it is also possible to use a post-processing, MTE, MTELPC [16] or PCF [17], which provides a "quasi-perceptual" pre-smoothing. This point is not detailed here.

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 algorithm is given in Sec. 4. Section 5 gives one practical example, and presents an evaluation procedure dealing with the robustness. Finally, section 6 concludes this paper and gives some perspectives.

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

This unpublished document provides temporarily technical details about the ARMA approximation used in [1]. It should be published soon, with some additional materials.

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

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

where 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_s the sampling rate, Q and P are the orders of the numerator B and the denominator A respectively. The gain g and 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 and the model $\widetilde{H}$:

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

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. [18, 19], and $L_s = s(L_p) = 2^{(L_p-40)/10}$ is the conversion to the sone scale, cf. e.g. [20].

Frequency scale Second, the measure $M(df)$ takes into account of the frequency resolution of the ear, 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 [21], $m(f) = 2595\log_{10}(1 + f/700)$.

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:

$$\mathcal{C}_2 = \sum_{n=1}^{N} \frac{\left(|H_n| - |\widetilde{H}_n|\right)^2 \sigma'(H_n, f_n)^2}{\sigma(H_n, f_n)^2}\, m'(f_n), \quad (3)$$

where the frequencies f_n uniformly sample the range $[0, F_s/2]$ and $H_n = H(f_n)$. Note that in this work σ and its derivative are computed using the analytical expression of [22].

If the phase of the target response H is known, we can replace $(|H_n| - |\widetilde{H}_n|)^2$ by $|H_n - \widetilde{H}_n|^2$. 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. [23].

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

$$\mathcal{C} = \sum_{n=1}^{N} \left|H_n - \widetilde{H}_n\right|^2 W_n^2 \qquad (5)$$

$$\text{with} \quad W_n = \frac{\underline{\sigma}'(H_n, f_n)}{\underline{\sigma}(H_n, f_n)} \sqrt{m'(f_n)}. \qquad (6)$$

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

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. [24], calculated from the target response H. Nevertheless, this strategy seems hazardous because H and $\widetilde{H}$ are not "concrete" spectra, but "abstract" spectral envelopes.

4 Optimization algorithm

In this section an algorithm is proposed to minimize the cost function $\mathcal{C}$ of (5) with the ARMA model of (1). Unfortunately, the optimization is not trivial because the error is non-linear with the coefficients a_p of the denominator A and not convex.

We propose in this section an adaptation of the iterative Gauss-Newton algorithm, cf. e.g. [25], 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.

4.1 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:

$$\mu^{k+1} = \mu^k - \Omega_{\mathcal{C}}^{-1}(\mu^k)\nabla_{\mathcal{C}}(\mu^k), \tag{7}$$

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_n, cf. e.g. [25]. With

$$e_n := (H_n - \widetilde{H}_n)W_n, \tag{8}$$

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

Defining the Jacobian matrix of E by $J_e(\mu)$, such that $J_e[i,n] = \partial e_n/\partial\mu_i$, the gradient vector is now written $\nabla_{\mathcal{C}}(\mu) = 2J_e(\mu)^H E(\mu)$, and the approximated Hessian matrix is given by $\Omega_{\mathcal{C}}(\mu) = 2J_e(\mu)^H J_e(\mu)$. Consequently

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

Nevertheless, with (7), the algorithm may diverge in some cases. Then, it is usual to introduce a relaxation factor $\lambda_k \leq 1$, and the algorithm becomes

$$\mu^{k+1} = \mu^k - \lambda_k \Omega_{\mathcal{C}}^{-1}(\mu^k)\nabla_{\mathcal{C}}(\mu^k). \tag{9}$$

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 Optimization of the ARMA model

The model is given by $\widetilde{H}(Z) = gB(Z)/A(Z)$, cf. (1), 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. 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 $Z_n = \mathrm{e}^{j2\pi f_n/F_s}$ and $\mu = [\gamma, a_1, \dots a_P, b_1, \dots b_Q]^T$, the Jacobian matrix $J_e(\mu)$ is given by

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

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. 4.1 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.

4.3 Initial filter

The initial filter of the iterative Gauss-Newton algorithm can be obtained by any ARMA approximation. Nevertheless, it must be causal/stable and must be a minimum phase system if desired. We tested the well-known methods of Prony [2], Shanks [3] and the Iterative Prefiltering [4]. Unfortunately, the two first do not guarantee the minimum phase property, even if the target H is a minimum phase system, and the last one does not guarantee the causality/stability. To solve this problem, the simple method is to retest the method with lower orders until the obtained filter verifies the condition.

Otherwise, if the minimum phase property is desired, we can apply the following iterative method: first the denominator $A^{(0)}$ is initialized using the covariance method, cf. e.g. [7], with the initial gain g_0 as the predictive gain, cf. e.g. [26]. Then the following steps are iteratively computed:

- Deconvolution and Inverse autocorrelation: $r_a = \text{iDFT}\{1/|HA^{(k-1)}|^2\}$.
- Q-order LPC estimate: $b^{(k)} = \text{LPC}(r_a, Q)$.
- Inverse filtering and Autocorrelation: $r_b = \text{iDFT}\{|H/B^{(k)}|^2\}$.
- P-order LPC estimate: $a^{(k)} = \text{LPC}(r_b, P)$.
- Predictive gain: $g_k = \sqrt{\langle a^{(k)}, r_b \rangle}$.

Thanks to the properties of Linear Prediction (Yule method) the obtained filter is guaranteed to be causal/stable and with the minimum phase property.

5 Results and discussions

5.1 Illustrations

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 [17]. Then, from the magnitude of the obtained frequency response, the phase has been recovered assuming a minimum phase system, cf. e.g. [23]. Finally, the PLF is computed using an ARMA(8,8) model. In this example, the initial filter is obtained by the method previously presented in Sec. 4.3. 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 technics (cf. [9, 10]), but especially to those frequencies where the target response H is above the auditory threshold. Note that the initial filter H_0 does not fit H 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 fits H when it is audible, and it strongly smooths it when it is imperceptible.

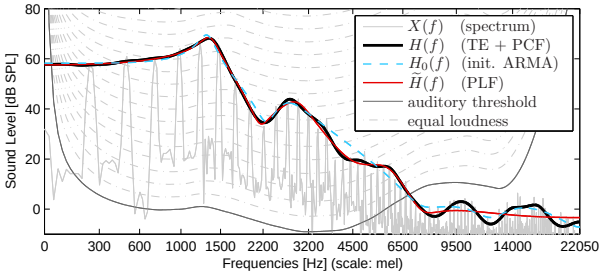


Figure 1: Illustration of the Perceptual Linear Filter (PLF). The orders of the ARMA model are $Q = 8$ and $P = 8$.

5.2 Robustness

In practice, we observe a weak robustness of the method. Indeed, with some initial filters H_0, the nearest local minimum of $\mathcal{C}$ may be outside the constraint domain (causality/stability or minimum phase), even if H_0 verifies the conditions. Then the algorithm is blocked at the limit of the domain and the criterion cannot significantly decrease.

Since this phenomenon strongly depends on the initial filter, we compared the robustness of the

PLF using 4 different initial methods: the Prony method (PRONY), the Shanks method (SHANKS), the Iterative Prefiltering of Steiglitz and McBride (STMCB) and the iterative linear prediction proposed in Sec. 4.3 (ILP). The target H is computed using the True Envelope of a speech frame.

Figure 2 displays the PLF error, value of $\mathcal{C}$ after the last iteration, as a function of the orders Q and P, and the initial method (cf. above). As a result, since the use of the ILP method, as the initial ARMA method, provides the lowest PLF error in most of the cases, in comparison of the other methods, we conclude that it is more robust.

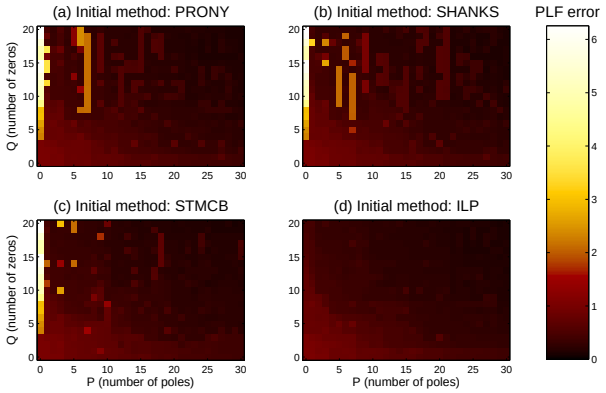


Figure 2: Illustration of the robustness.

Although Fig. 2d encourages the use of the ILP method, its robustness is not satisfying in few cases. A possible solution is to test the PLF with several orders $q_0 \leq Q$ and $p_0 \leq P$, and to use the obtained filters as initialization of a new PLF computation with the orders Q and P, with a 0-padding of the polynomials. Then, the case which gives the lower PLF error is conserved. In practice this approach provides a significantly better robustness, but in a frame-by-frame analysis-synthesis framework, the poles and zeros may move suddenly which may imply some undesirable artifacts during the synthesis. However, the target response is efficiently fitted at every frame as wanted, and this solution is satisfactory in the case of the synthesis of a quasi-static spectral envelope.

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 of the sensitivity of the ear to the frequencies and to sound level via the loudness conversion. As illustrated, 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.

Unfortunately, this approach may not be robust enough in the case of a frame-by-frame analysis-synthesis framework as discussed. Note that it is also the case in many other ARMA approximation methods. Then, in a next improvement, it is necessary to derive a low-cost algorithm which provides a frame-by-frame sound synthesis without artifacts.

Nevertheless, in the case of the synthesis of a quasi-static spectral envelope, which is under interest in the context of our work, the solution proposed in Sec. 5.2 is fully satisfying. For example, in [1], this method is used in order to efficiently approximate the "representative timbre" of a given note signal.

References

[1] R. Mignot and V. Välimäki. Extended subtractive synthesis of harmonic musical tones. In *136th Audio Engineering Society Convention*, Berlin, Germany, 2014.

[2] 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, a différentes temperatures. *J. de l'Ecole Polytechnique (Paris)*, 1(2):24–76, 1795.

[3] J.L. Shanks. Recursion filters for digital processing. *Geophysics*, 32(1):33–51, 1967.

[4] K. Steiglitz and L. McBride. A technique for the identification of linear systems. *IEEE Trans. Automatic Control*, 10(4):461–464, 1965.

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

[6] B. Friedlander and B. Porat. The modified Yule-Walker method of ARMA spectral estimation. *IEEE Trans. Aerospace and Electronic Systems*, AES-20(2):158–173, 1984.

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

[8] A. Härmä. Linear predictive coding with modified filter structures. *IEEE Trans. Speech and Audio Process.*, 9(8):769–777, 2001.

[9] 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.*, 48(11):1011–1031, 2000.

[10] H.W. H.W. Strube. Linear prediction on a warped frequency scale. *J. Acoust. Soc. Amer.*, 68(4):1071–1076, 1980.

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

[12] J.O. Smith and J.S. Abel. Bark and ERB bilinear transforms. *IEEE Trans. Speech and Audio Process.*, 7(6):697–708, 1999.

[13] A. El-Jaroudi and J. Makhoul. Discrete all-pole modeling. *IEEE Trans. Signal Process.*, 39(2):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)*, pages 30–35, Madrid, Spain, 2005.

[15] S. Imai and Y. Abe. Spectral envelope extraction by improved cepstral method. *Electronics and Communication*, 62-A(4):10–17, 1979. in Japanese.

[16] 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)*, pages 1625–1628, Las Vegas, Nevada, USA, 2008.

[17] 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, 2013.

[18] D.W. Robinson and R.S. Dadson. A re-determination of the equal-loudness relations for pure tones. *British Journal of Applied Physics*, 7(5):166, 1956.

[19] 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.*, 45(4):224–240, 1997.

[20] S.S. Stevens. A scale for the measurement of a psychological magnitude: loudness. *Psychological Review*, 43(5):405, 1936.

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

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

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

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

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

[26] J. Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580, 1975.