Seeking maximum linearity of transfer functions

Filipi N. Silva, Cesar H. Comin, and Luciano da F. Costa
São Carlos Institute of Physics, University of São Paulo, P.O. Box 369, São Carlos, SP, Brazil

(Received 24 September 2016; accepted 15 November 2016; published online 8 December 2016)

Linearity is an important and frequently sought property in electronics and instrumentation. Here, we report a method capable of, given a transfer function (theoretical or derived from some real system), identifying the respective most linear region of operation with a fixed width. This methodology, which is based on least squares regression and systematic consideration of all possible regions, has been illustrated with respect to both an analytical (sigmoid transfer function) and a simple situation involving experimental data of a low-power, one-stage class A transistor current amplifier. Such an approach, which has been addressed in terms of transfer functions derived from experimentally obtained characteristic surface, also yielded contributions such as the estimation of local constants of the device, as opposed to typically considered average values. The reported method and results pave the way to several further applications in other types of devices and systems, intelligent control operation, and other areas such as identifying regions of power law behavior. *Published by AIP Publishing.* [http://dx.doi.org/10.1063/1.4969058]

I. INTRODUCTION

Several situations in applied sciences involve transforming a signal from an input to an output domain. This includes measuring any physical property through a sensor, conditioning a signal through a filter[1] or amplifier,[2] and transducing an electrical signal into some action (e.g., a force). Any of these situations can be conveniently summarized in terms of a systems approach such as shown in Fig. 1(a), where the transforming system T receives an input signal $x(t)$ and outputs a signal $y(t)$. The effect of the transformation can be clearly characterized in terms of transfer function of the system, illustrated in Fig. 1(b). Though the *transfer function* can be applied to other situations (e.g., frequency transforms), here we use this term to quantify the action of a system (represented by a function) on the amplitude of an input signal to produce a respective output.[3]

Oftentimes, a linear mapping is desired between input and output, which ensures no modification, distortion, or delay to the signal other than eventual scaling, value shifting, or delay (linear phase). Unfortunately, the linearity of real-world transfer functions is never perfect, being limited in several aspects, such as by noise and distortions. Yet, some of the regions of the transfer function are closer to being linear, and it becomes important to devise methods capable of selecting the best region for operation of the system. Three main problems can be considered: (i) a maximum deviation from linearity E_{max} is imposed on the sought region of a given length L of the transfer function; (ii) given L (along the input domain), find the region that minimizes the deviation from linearity; and (iii) given a maximum deviation from linearity E_{max}, search for the longest region in the given transfer function. In the former situation, the application requires a maximum acceptable distortion; in (ii), the objective is to select the best region of operation for a given application. Observe that criterion (ii) is a particular case of (i) as it optimizes the error for the same required L. In the present work, we concentrate on criterion (ii), which is often found in practice, in the sense that L is pre-specified (e.g., in sensors and amplifiers applications, the desired output extension is often a design constraint). Such a methodology can be useful for best exploring the intrinsic capabilities of any sensor, amplifier, or transducer, in the sense that maximum linearity operation can therefore be achieved for a given L, as illustrated in Fig. 1(c). Frequently, this region of interest is associated or defined by an operation (or quiescent) point Q, such as in Fig. 1(c), which corresponds to the operation of the system under the absence of signal (which defines the null level). In most cases, the linear region should extend symmetrically along both sides from Q, in order to allow the maximum linearity.

Experimentally, the continuous transfer function of a system, sensor, or transducer is never available and needs to be sampled in terms of a sequence of points S. The devised procedure (to be explained in detail in Section II) to find the most linear region for a given S and L performs minimization of the least mean square residues for several candidate regions. The suggested procedure is evaluated in terms of the sigmoid function, which represents the transfer functions typically found in electronic systems.[4]

To illustrate the practical usefulness of the introduced methodology with respect to a real-world situation, we apply it to the problem of determining the best operating points of a simple one-stage class A current amplifier configuration[5] based on a single generic low signal transistor. We observe that we do not address a complete, operational, amplifier circuit, but only one of the most basic configurations with minimal circuitry, so that the analysis becomes more directly related to the device than to circuit setting. Yet, the choice of a class A configuration as a case example in this work is justified because this type of circuit is often appreciated by its linearity and simplicity, though typically at the expense of increased power consumption.[6,7] Also, we observe that the current approach is limited to resistive loads.

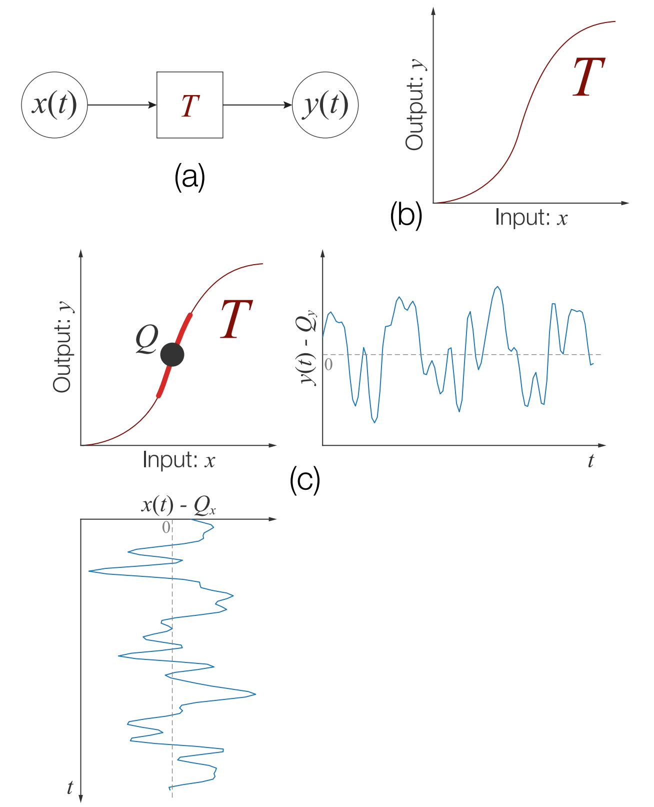

FIG. 1. (a) Illustration of an input signal $x(t)$ being transformed by a system T into an output signal $y(t)$. (b) The transfer function specifying the system. (c) It is often desired to have the operation point $Q = (Q_x, Q_y)$ of the system in the center of the most linear region of the transfer curve, so that the shifted input signal $x(t) - Q_x$ is transformed into the output signal $y(t) - Q_y$ with little distortion.

The paper is organized as follows. Section II presents the in-detail description of the proposed methodology to obtain the most linear regions of a transfer curve. In Section III, the methodology is illustrated and validated with respect to a Sigmoid function. In Section IV, we illustrate the application of the methodology to a real device (a generic, small signal bipolar junction transistor (BJT)). We start from the experimentally obtained characteristic surface and then quantify the linearity of several possible circuit configurations.

II. METHODOLOGY

In the following, we consider a given sequence of points $S = ((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n))$, describing the relationship between variables x and y. An example of such a sequence is shown in Figure 2. Although we consider S to be a generic sequence, it can have different meanings, such as data sampled from a known continuous function or from an experiment. A contiguous subsequence $S_{k,q}$ of S is defined as the sequence of $m = q - k + 1$ points in S having index i in the range $[k, q]$.[8]

As mentioned in Sec. I, the linearity of the transfer curve of a system (e.g., sensor, filter, and amplifier) should be optimal in the expected operation range L of the system. Therefore, we only consider subsequences $S_{k,q}$ having a size $W_{k,q} = x_q - x_k$ which is as close as possible to the desired target range L. This is done by selecting subsequences contained in S that obey the following conditions:

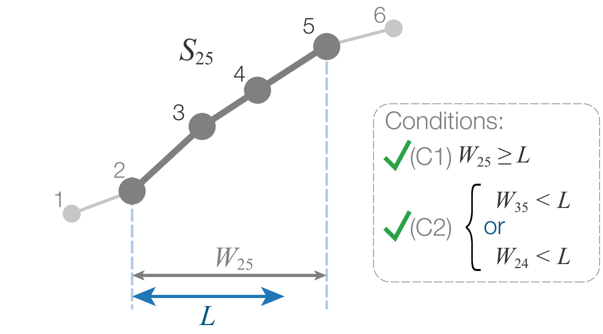

FIG. 2. Example of subsequence having a valid target size L. Given the original sequence of six points, the highlighted subsequence $S_{2,5}$ has size $W_{2,5} \geq L$, which obeys condition (C1). When removing point 2 or 5, the size of the subsequence becomes smaller than L, which is in agreement with condition (C2).

$$\textbf{C1}: \; W_{k,q} \geq L \text{ and } \textbf{C2}: \begin{cases} W_{k+1,q} < L \\ \text{or} \\ W_{k,q-1} < L \end{cases} .$$

These conditions are illustrated in Figure 2. The subsequence $S_{2,5}$ shown in the figure follows both conditions because its size is larger than L (condition 1) and, after removing one of its endpoints, its size becomes smaller than L (condition 2). Subsequences that follow these two criteria are considered valid for linearity quantification.

In order to assess how linear a given subsequence is, we need to quantify the deviation, E, of such a subsequence from a straight line. This deviation can have different definitions. One traditional approach is to calculate the sum of the squared distances, in the y coordinate, between the points and a candidate straight line adjusted to the data.[9] The process of finding the straight line that minimizes the sum of squared distances is known as linear least squares regression,[10,9] and the respective error of the linear regression can be used to quantify the linearity of the points in a candidate subsequence. This error is given by

$$E_{k,q} = \sqrt{\frac{1}{m} \sum_{i=k}^{q} (y_i - \alpha x_i - \beta)^2}, \tag{1}$$

where α and β are, respectively, the slope and the y intercept value of the best-fitting linear function.

The proposed methodology consists in applying the linear least squares regression to all subsequences of S following conditions (C1) and (C2). A simple, but not optimal, approach for such a task is to explore all existing subsequences in the investigated sequence of points S. This can be done by varying both k and q, such that $1 \leq k < q \leq |S|$, and checking if the resulting range $[k, q]$ follows the aforementioned conditions. In addition, all subsequences must have at least 3 data points for the analysis, since 2 points always define a linear subsequence. We note that the process can be optimized by preemptively discarding ranges containing subsequences that were already considered valid for linearity quantification. A linear least squares regression is then applied to the points

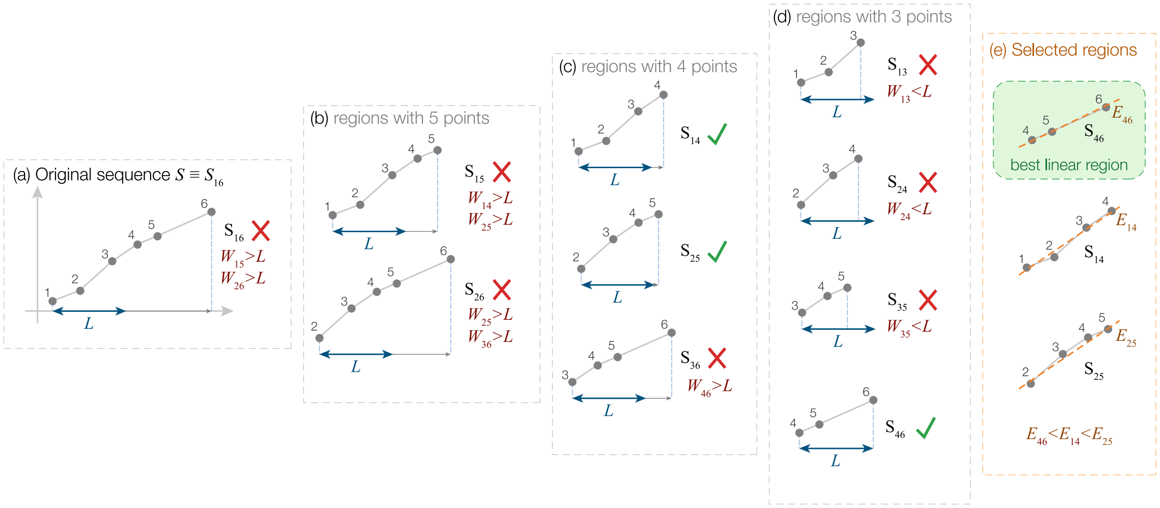

FIG. 3. Example of application of the methodology. The original sequence S, containing six points, is shown in (a). All subsequences of S with at least 3 points are considered for the initial selection. Sequences with 5, 4, and 3 points are presented in, respectively, (b), (c), and (d). The target range L is indicated below each subsequence. Check marks indicate subsequences that comply with conditions (C1) and (C2), while discarded subsequences are marked with an **X**. The selected subsequences are shown in (e), where the most linear subsequence, i.e., the one having the lowest residue $E_{k,q}$, is highlighted.

belonging to each valid subsequence $S_{k,q}$. Next, the respective error, $E_{k,q}$, of each regression is calculated. Finally, the subsequence associated with the lowest error defines the most linear extent of S. Figure 3 illustrates the application of the methodology to a small sequence of points. In the figure, all possible subsequences (ten in total) that can be applied to the sequence of six points are shown. A check mark is used to indicate subsequences that follow the aforementioned conditions.

Algorithm 1 summarizes the process of finding the best linear region in a sequence of points S for a given L. The function **bestLinearFitError**($S_{k,q}$) calculates the residue obtained when applying the least squares method to the subsequence $S_{k,q}$.

ALGORITHM 1. Algorithm to determine the best linear region of a sequence S for given L.

```
input : A sequence of points
        $S = ((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n))$.
input : Minimum region length $L$.
output : A tuple $[k_{\text{best}}, q_{\text{best}}]$ corresponding to the best
         fitted subsequence and its residue $E_{\text{best}}$
windows ← [ ]; $E_{\text{best}} \leftarrow \infty$; $k_{\text{best}} \leftarrow \emptyset$; $q_{\text{best}} \leftarrow \emptyset$
for $k \leftarrow 1 \rightarrow |S|$ do
    for $q \leftarrow (k+2) \rightarrow |S|$ do
        if $[k, q]$ contains no ranges of windows then
            if $x_q - x_k \geq L$
            and $(x_{q-1} - x_k < L$ or $x_q - x_{k+1} < L)$ then
                append $[k, q]$ to windows;
                $E \leftarrow$ bestLinearFitError($S_{k,q}$);
                if $E < E_{\text{best}}$ then
                    $E_{\text{best}} = E$;
                    $k_{\text{best}} = k$;
                    $q_{\text{best}} = q$;
                end
            end
        end
    end
end
```

III. LINEARITY ON ARTIFICIAL DATA

In order to illustrate the potential of the methodology to quantify linearity, in this section we present the application of the methodology to a sigmoid function. For such a task, we considered the logistic function, given by

$$f(x) = \frac{1}{1 + e^{-x}}. \tag{2}$$

This function was chosen because it has a clear linear region around $x = 0$, while the non-linearity of the function increases with $|x|$, until reaching saturation. This behavior is indicated in Figure 4, where we plot the logistic function and its respective curvature.[11] Note that we considered the interval $[-3, 3]$ for the function domain. The plot shows that at $x = 0$ the curvature is zero, meaning that the function is locally linear at this point. The curvature increases when going away from $x = 0$, until it starts to decrease again since the logistic function tends to a constant value for $|x| \rightarrow \infty$, due to saturation. Therefore, $x = 0$ should represent the optimal operation point of a logistic transfer function as far as linearity is concerned.

In order to verify the robustness of the methodology for identifying linear regions, we added different levels of noise to the function $f(x)$. Since $f(x)$ has its values defined in the interval $[0, 1]$, the noise level is represented as a fraction r of this interval, or equivalently, as a percentage $100r$ of the function range. Given a noise level r, we define a new function

$$g_r(x) = f(x) + \zeta(x), \tag{3}$$

where ζ is a random variable having a uniform distribution in the interval $[-r/2, r/2]$. In such a case, the region near the origin should be considered the most linear by the methodology.

We tested the methodology for different noise levels r and distinct values for the minimum range L. The results are shown in Figure 5. Each row of plots corresponds to a distinct noise level, while each column corresponds to a different L. The largest linear region of each considered case is indicated

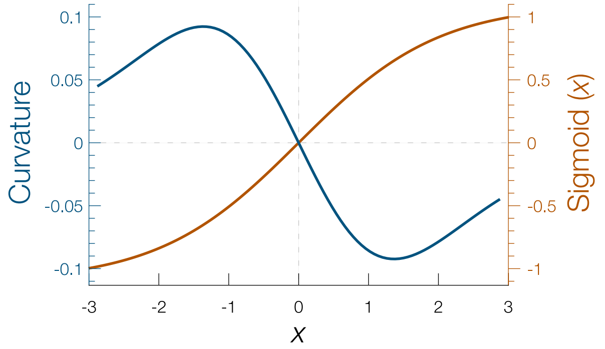

FIG. 4. The logistic function and its respective local curvature.

in red. The results show that the methodology identifies the region near the origin as being the most linear, as expected by the properties of the logistic function, as well as by a visual inspection of the function shape. We observe a small variation on the central position of the most linear region when L is comparable to the noise level added to the function. Therefore, the results indicate that the methodology is robust against random perturbations on the analyzed function. As could be expected, the error E tends to increase with the intensity of added noise. However, it is interesting to observe that the intensity of noise tends to have a greater influence on the value of the error in a more linear part of the curve than in a less-linear portion. This is a consequence of the fact that a small perturbation in a more linear, and consequently symmetric, part of the curve has a greater *relative* effect on the overall symmetry than the same level of perturbation applied to a region already containing larger scales of non-linearity.

In order to generalize the results obtained when applying the methodology to the logistic function, we considered distinct realizations of the noise ζ added to the function $f(x)$ and calculated the optimal operation point for each realization. Then, the respective standard deviation of the calculated positions was estimated, for different values of L. The results are shown in Figure 6. Each curve in the plot is relative to a distinct noise level r, as indicated. The plot shows that the position of the most linear region can have large changes depending on the noise level and the parameter L. Still, the position always tends to 0 for large L, showing that a proper choice of the minimum range is important for the methodology.

IV. CASE EXAMPLE: CLASS A ONE-STAGE TRANSISTOR AMPLIFIER

Given their ability to change the amplitude of electronic signals, amplifiers are part of many electronic systems. In particular, audio amplifiers play a critical role in transforming the low power audio signals generated by the source (e.g., CD player and DAD) into audible sound. In a high fidelity (hifi) system, the amplifier should only uniformly affect the amplitude of the input signal, which requires a nearly linear transfer function covering the respective operation

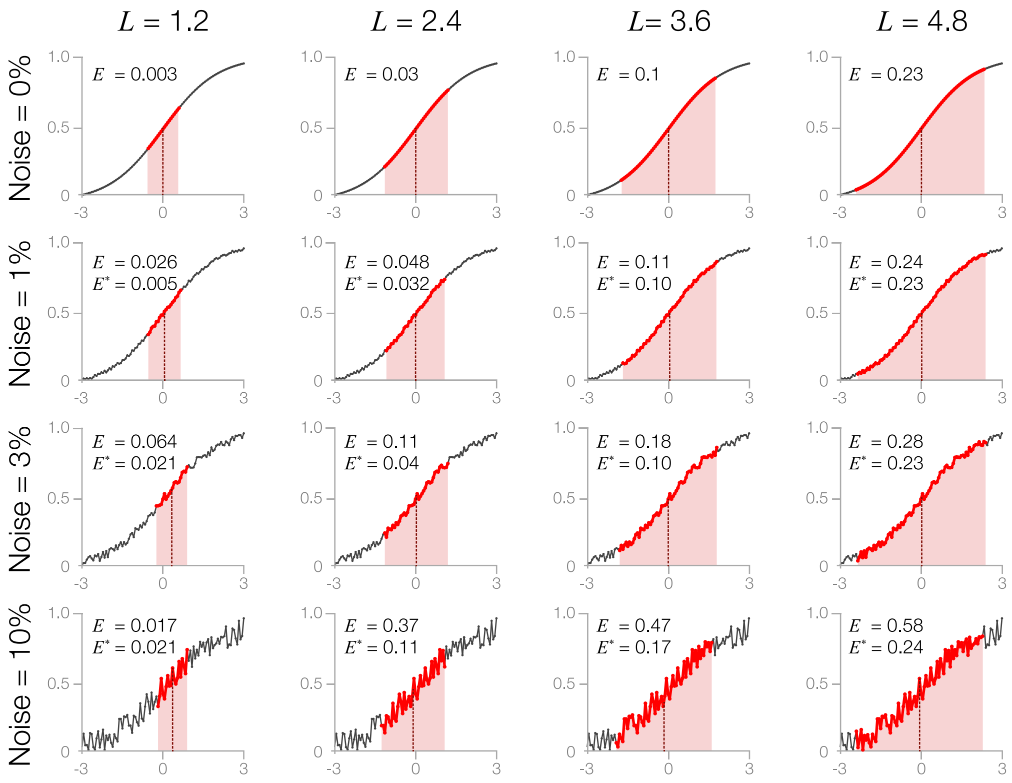

FIG. 5. Identification of the most linear region of the logistic function. Each row of plots corresponds to the logistic function having different noise levels, while each column contains the results for a distinct minimum window size L. Regions marked in red represent the most linear interval found by the method. The values of the linearity deviation E for the best selected regions are also given respectively in the plots, as well as the values E^*, indicating the errors in the absence of noise.

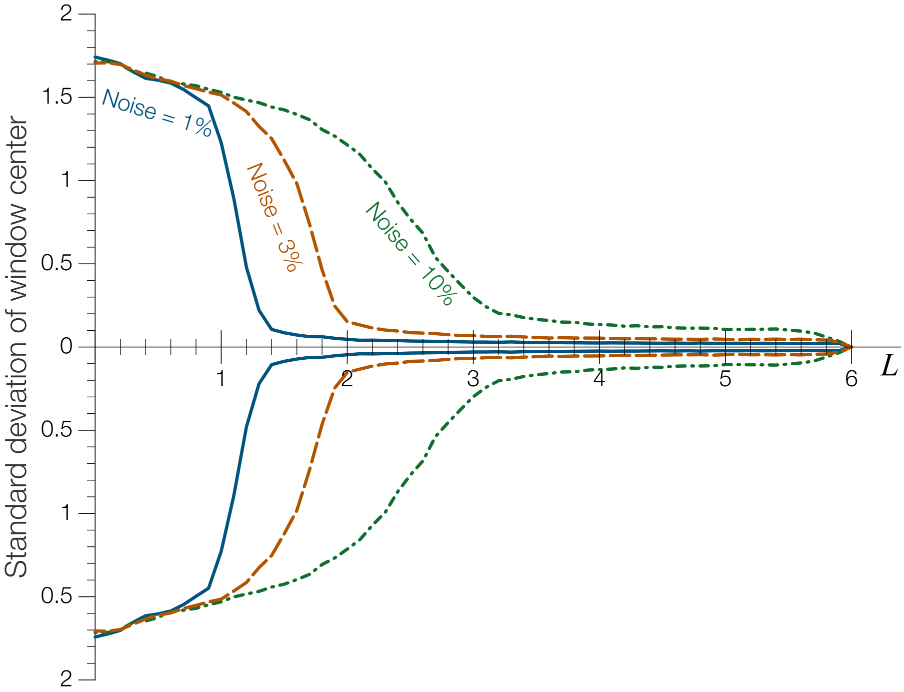

FIG. 6. Standard deviation of the most linear region of the noisy logistic function, as a function of the minimum window size L. Each line represents a different noise level r added to the data.

region. Typically, several stages are required in order to accomplish the desired amplification, which demands special care in achieving good linearity levels at each stage. Here, we consider analog audio amplifiers, particularly those in the class A, which is characterized by 100% of the signal being used.[6,12] In addition, we use a low signal BJT (bipolar junction transistor)[13] as the amplification device. Additional information on transistor amplifiers are provided in Appendix A.

We apply the method proposed in Section II to the problem of choosing the operation point of a one stage class A amplifier in order to maximize linearity, given a desired input range. For generality's sake, we are not restricted to finding the best configuration along a load line, instead we consider many putative load lines derived from the characteristic surface defining the device operation. In other words, given the device characteristics, the range of operation, and type of circuit, the reported methodology is capable of identifying the best operation point. First, experimental data are obtained and interpolated as the characteristic surface, in order to allow accurate estimation of the partial derivatives required for modeling the transfer function. We show that different

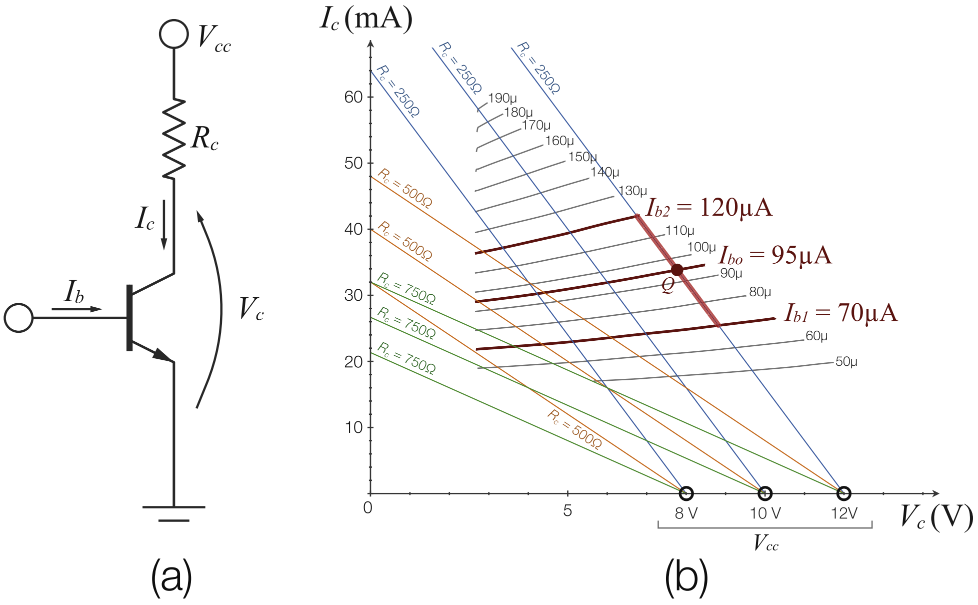

FIG. 7. (a) Circuit of a one-stage class A amplifier with resistive load. (b) Load lines defined by distinct values of V_{cc} and R_c, shown in the $V_c \times I_c$ plane. The regions near the saturation and cutoff were excluded.

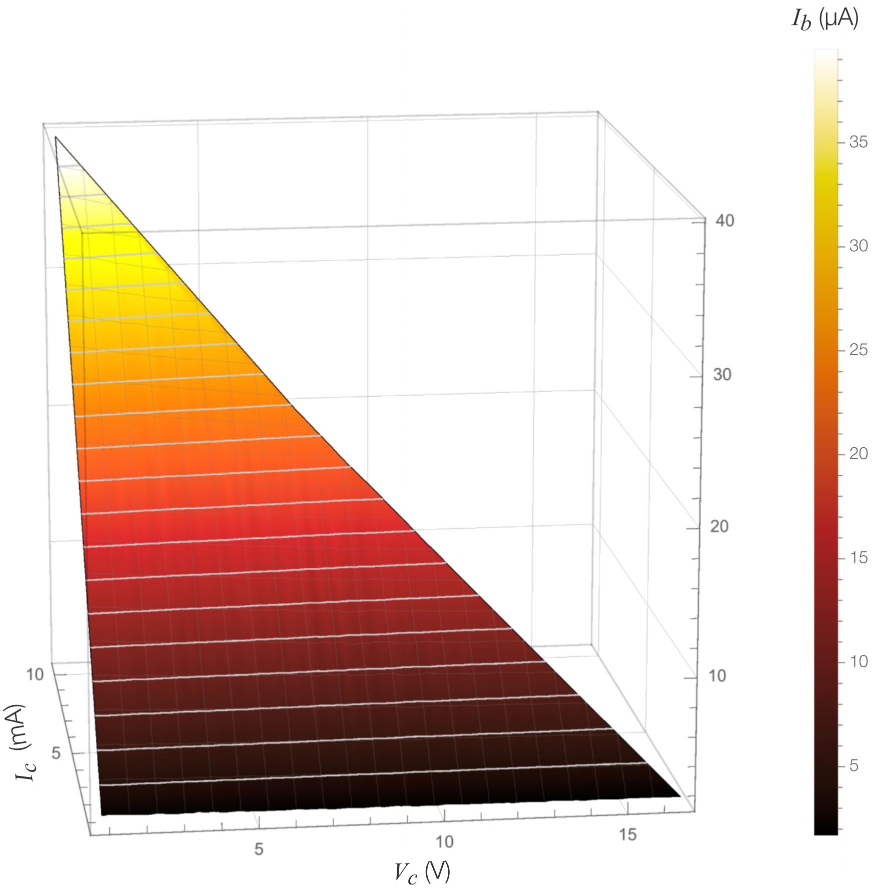

FIG. 8. Interpolated surface obtained for the device properties I_c, V_c, and I_b. The isolines of I_b are shown in grey.

operation points lead to varying compromises between output resistance, current gain, and linearity.

The basic circuit for the one-stage class A amplifier is shown in Figure 7(a). In order to simplify the analysis, we consider a purely resistive load. In the figure, I_b and I_c are, respectively, the input and output currents of the transistor and V_c the collector voltage.[14] The considered circuit has two parameters, the main power supply (V_{cc}) and the resistance

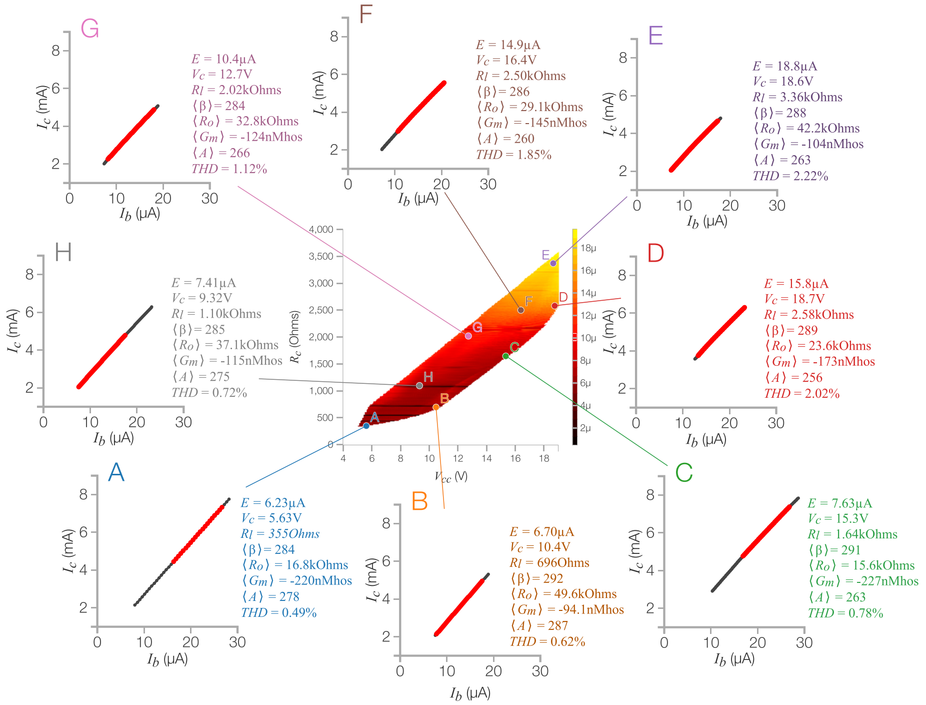

FIG. 9. Heat map of the linearity error for distinct values of V_{cc} and R_c. The relationship between the input current, I_b, and output current, I_c, is shown for the chosen load lines.

TABLE I. Throughout the discussion we consider some particularly interesting load lines for the circuit. The table shows the relevant properties of such lines. These properties, and their respective physical units, correspond to the circuit voltage (V_{cc}, in Volts), circuit resistance (R_c, in Ohms), linearity error (E, in μ A), average circuit amplification (A), average current gain (β), average transconductance (G_m, in nMhos), average output resistance (R_o, in k Ohms), and total harmonic distortion (THD, in percentage).

Label	V_{cc}	R_c	E	$\langle A \rangle$	$\langle \beta \rangle$	$\langle G_m \rangle$	$\langle R_o \rangle$ (k)	THD (%)
A	5.63	355	6.23	278	284	−220	16.8	0.49
B	10.4	696	6.70	287	292	−94.1	49.6	0.62
C	15.3	1.64k	7.63	263	291	−227	15.6	0.78
D	18.7	2.58k	15.8	256	289	−173	23.6	2.02
E	18.6	3.36k	18.8	263	288	−104	42.2	2.22
F	16.4	2.50k	14.9	260	286	−145	29.1	1.85
G	12.7	2.02k	10.4	266	284	−124	32.8	1.12
H	9.32	1.10k	7.41	275	285	−115	37.1	0.72

(R_c). These two parameters define a load line for the transistor, which restricts the relationship between I_c and I_b to a line in the $V_c \times I_c$ plane. Examples of load lines are shown in Figure 7(b). Also shown in Figure 7(b), in particular for the rightmost load line, is a specific configuration of operation point defined by $I_{bo} = 95\ \mu$A as well as a region of operation extending between $I_{b1} = 70\ \mu$A and $I_{b2} = 120\ \mu$A. Observe that the operation point is defined by the intersection between the load line and the isoline I_{bo}. A relevant property of the circuit is the total current gain, A, for a given region of operation defined as

$$A = \frac{dI_c}{dI_b}. \tag{4}$$

This property describes the actual current amplification imposed by the circuit for given BJT constants and circuit parameters. Please refer to Appendices A and B for more information on this property. Typically, the aim of a hifi amplifier is to provide a linear relationship between I_b and I_c for a selected load line.

In order to obtain the $S(V_c, I_c, I_b)$ surface associating variables V_c, I_c, and V_c, we experimentally sampled the $I_c(V_c)$ curves along load lines with fixed R_c for a sequence of V_{cc} values. Next, we employed a triangle-based interpolation method[15] over the scattered data points, resulting in the surface $S(V_c, I_c, I_b)$, which is shown in Figure 8.

As mentioned above, each pair of circuit parameters (V_{cc}, R_c) implies a load line that defines the operation of the circuit. The systematic variation of parameters V_{cc} and R_c allows a thorough analysis of the circuit properties at distinct operation conditions. These parameters are bounded by the adopted values of the transistor constants, shown in Figure 11 of Appendix A. By considering all these allowed values of V_{cc} and R_c, we can define an operation domain $\mathcal{S}$ for the circuit. The considered load lines are specified by sampling this domain with 500 points of resolution for each of the circuit parameters. The methodology presented in Section II was applied to each considered load line, given a target input range of $L = 10\ \mu$A. The resulting linearity error, E, over $\mathcal{S}$ is shown in Figure 9. It is clear that the error increases steadily upwards along the vertical. The most linear regions are to be found precisely for low values of R_c and V_{cc}. In Figure 9 we also show the transfer curves defined by a few chosen load lines. These load lines were chosen as they were found to provide a good representation of the circuit properties inside domain $\mathcal{S}$, since the linearity shows smooth variation along $\mathcal{S}$. The selected operation range L of each transfer curve is indicated in red. The first four columns of Table I present the values of V_{cc}, R_c, and E for each of the load lines, specified by labels.

Besides the requirement that a proper load line should provide a highly linear relationship between I_c and I_b, other properties of the circuit are often also sought. For instance, one may seek a large amplification and/or transconductance, the latter being typically useful to minimize the influence of reactive loads. In Figures 12(a) and 12(b) of Appendix A, we show the averages of, respectively, the amplification, $\langle A \rangle$, and transconductance, $\langle G_m \rangle$, obtained for the load lines. The averages were calculated along the respective operation range found by the linearity methodology for each load line. The values of the average amplification, current gain, transconductance, and output resistance for some load lines are indicated in Table I. The total harmonic distortion (THD), defined in Appendix A, was also applied to the chosen load lines indicated in Figure 9, and the obtained values are shown in Table I. This table can be used as a reference to many distinct amplifier properties associated with the linearity error found by the presented methodology. For instance, even though load lines A and B define highly linear transfer functions (given their low linearity deviation E), load line A has a much smaller output resistance, which is often a desired property for amplifiers.

V. CONCLUSIONS

Linear operation has been of paramount importance in most theoretical and applied areas, as a consequence of its ability to preserve the properties of signals, avoiding distortions, and other unwanted effects. Yet, relatively few approaches have been proposed in order to objectively quantify the linearity of a given region of operation in a sensor, device, or transducer. The moving least squares method[16,17] bears similarities with our methodology in the sense that the least squares regression is applied to different parts of the point sequence. Nevertheless, the method is used for interpolating or treating missing points in the data, and not for finding an optimal region of operation. In the present work, we developed a methodology capable of, given a transfer function, finding its respective operation interval allowing maximum linearity. The reported approach is based on least squares regression, but also incorporates the constraint given by the extent of the desired region of operation. In addition, all possible intervals are considered by scanning a window along the domain of the transfer function.

The methodology has been characterized with respect to the analytical situation involving sigmoid transfer functions in the presence of varying levels of noise and also for real-world data related to the properties of a low power, one-stage class A transistor amplifier operating with resistive load. In the former case, we verified that the method was capable of identifying the optimal region, centered at the origin of the coordinate axis of the sigmoid function, where the curvature is known to be smallest. The application to the amplifier incorporates several interesting results, such as the determination of the surface of the transistor operation (i.e., $S(V_c, I_c, I_b)$) by using interpolation, which allowed the detailed estimation of the transistor constants along a domain in the $V_c \times I_c$ space by using partial derivatives, and the estimation of the linearity error in terms of amplification and output resistance. A complex structure was found to underlie the characteristic surface of the adopted small-signal transistor.

It should be observed that the results obtained for the amplifier are specific to the considered configuration, device, and parameters and cannot be directly extended to other situations. In addition, it should be reminded that the proposed methodology for selecting linear regions was conceived with resistive loads in mind and is, in principle, restricted to that case. Other situations, e.g., involving reactive loads, imply the transfer function to have hysteresis and, consequently, to have its behavior split into two or more parts. Such situations could be eventually approached by applying the reported methodology in a piecewise fashion, which constitutes a possible future development. It would also be possible to extend the methodology to *n*-port representations of systems,[3] involving multiple inputs and outputs, in which case the problem would become to find linear patches in surfaces or hypersurfaces.

The reported methodology and results provide several additional possibilities for future investigation. For instance, it would be interesting to apply the method to optimize the operation of sensors and transducers, as well as of amplifiers involving other configurations and devices (e.g., class AB, vacuum tubes, and integrated circuits). Other linearity criteria could be used, for instance, THD. The complex structure of the characteristic surface obtained for the small signal transistor also motivates further investigation, including other models of transistors and devices. Another interesting situation to be addressed is the amplifiers involving reactive loads. It would also be interesting to develop intelligent control systems using the proposed linearity optimization approach in order to dynamically and interactively set up the best operation points in such devices and systems. In addition, it should be also observed that, though presented here in the context of electronics and instrumentation, the proposed methodology can be directly used to tackle many important problems in other areas, such as identifying linear regions underlying power-law relationships in logarithmically related measurements (e.g., scale free complex networks[18]).

ACKNOWLEDGMENTS

F. N. Silva acknowledges FAPESP (Grant No. 15/08003-4). C. H. Comin thanks FAPESP (Grant No. 15/18942-8) for the financial support. L. da F. Costa thanks CNPq (Grant No. 307333/2013-2) and NAP-PRP-USP for their support. This work has been supported also by FAPESP Grant No. 11/50761-2.

APPENDIX A: ADDITIONAL AMPLIFIER CHARACTERISTICS

In the main text we show an application of the linearity methodology for identifying the most linear operating region of an amplifier. However, seeking for an operating region solely based on linearity can lead to undesired properties for the amplifier. Here we present additional properties that can influence an amplifier operation and show how they vary according to the parameters optimized by the linearity.

1. Amplifier characteristics—Definitions

The schematics of an NPN bipolar junction transistor (BJT) is shown in Figure 10(a). Mathematically, the transistor operation can be described in terms of the state variables $I_c(I_b, V_c)$, $I_b(I_c, V_c)$, and $V_c(I_b, I_c)$, where I_b and I_c are, respectively, the input and output currents of the transistor and V_c, the collector voltage.[14] Therefore, a given transistor has a well-defined surface in the $I_c \times I_b \times V_c$ space, defined by the relationship between these three properties. It is a

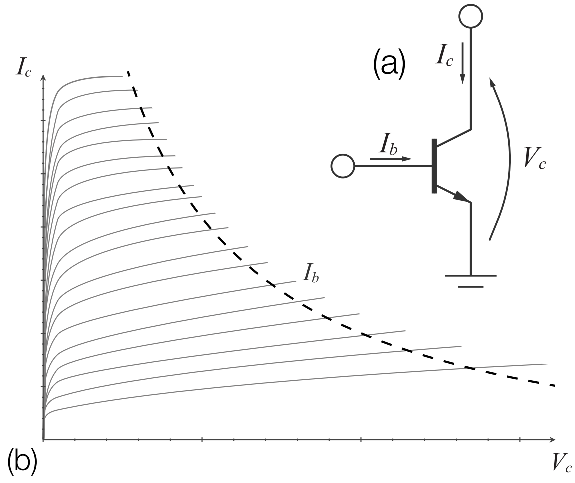

FIG. 10. A generic NPN BJT (a) and the characterization of its properties in terms of isolines in the $V_c \times I_c$ space (b). The maximum dissipation power is shown by the dashed curve.

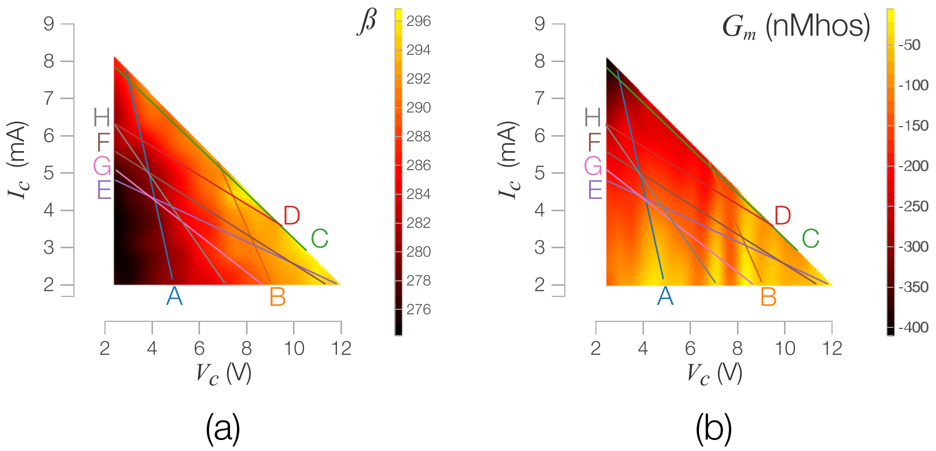

FIG. 11. Transistor constants calculated from the interpolated values shown in Figure 8. The constants, and the respective equations defining them, are (a) current gain (Equation (A1)) and (b) transconductance (Equation (A2)). Some load lines considered throughout the discussion are shown in each figure.

common practice to visualize such a surface as isolines in a 2D $V_c \times I_c$ space. An example of such visualization is shown in Figure 10(b).

The $S(I_c, I_b, V_c)$ surface properties of a transistor are specified by a set of the so-called constants, referred as *current gain* (β), *transconductance* (G_m), and *output resistance* (R_o). These constants can be defined in terms of the partial derivatives of the transistor state variables, that is,

$$\beta = \frac{\partial I_c}{\partial I_b}, \tag{A1}$$

$$G_m = \frac{\partial I_b}{\partial V_c}, \tag{A2}$$

$$R_o = \frac{\partial V_c}{\partial I_c}. \tag{A3}$$

R_o, however, can be calculated in terms of β and G_m as $R_o = -\frac{1}{G_m \beta}$.

The parameter β expresses the current gain, i.e., how much the collector current can be modified by the base current. Typically, β should be large so as to promote amplification. The transconductance G_m has an analogue interpretation, but regarding the collector voltage with respect to the base current. The combined consideration of these two parameters underlies the power amplification that can be achieved by using the device. In addition, a high value of this constant is useful to minimize undesired effects from the reactive components in the circuit. The output resistance influences the transfer of power to the load. Another important parameter of an amplifier is its total current gain A, which is given by Equation (4). The relationship between A and the transistor constants β and R_o is presented in Appendix B.

2. Amplifier characteristics—Experimental values

Typically, transistor amplifiers incorporate a high degree of feedback, which reduces the effect of wide variability of real-world device constants such as β.[12,19] However, in the present work we consider a relatively less common circuit, devoid of feedback, so as to provide a more diversified operation and linearity behavior as the circuit parameters are varied, therefore allowing a better validation of the proposed linearity method.

As described in Section IV, values of I_c, I_b, and V_c were experimentally obtained for many distinct combinations of circuit parameters R_c and V_{cc}. This allowed the calculation of the characteristic surface $S(V_c, I_c, I_b)$ of the transistor, shown in Figure 8 of the main text. The obtained surface is smooth enough to allow differentiation. From this surface we estimated the constants β and G_m of the transistor. The

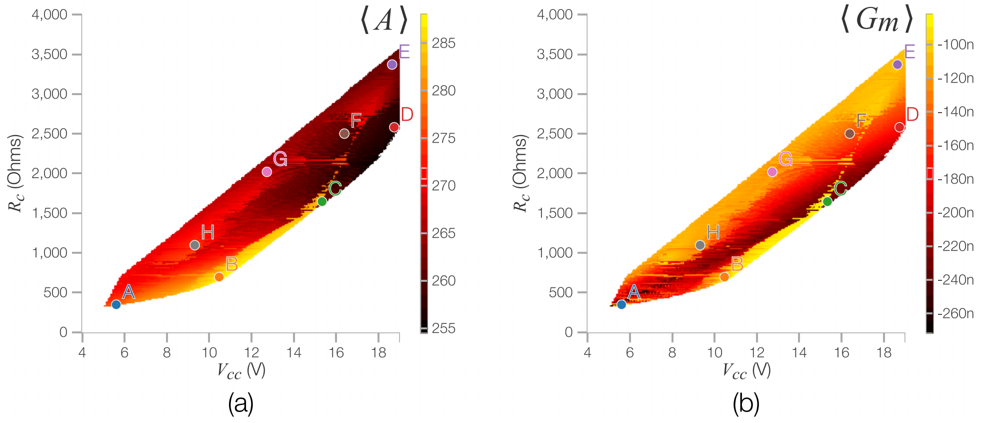

FIG. 12. Circuit and transistor properties calculated for distinct values of V_{cc} and R_c. (a) Average circuit amplification and (b) transconductance. We also show in each plot the position of the load lines indicated in Figure 11.

results, shown in Figure 11, provide a much more informative characterization of these two properties than the minimum and maximum values typically given in transistor data sheets. The region shown in this figure, which corresponds to the circuit configurations covered by the experimental procedure and interpolation, is henceforth called *polyhedron*. The β values vary from 242 to 437, with average 360 and standard deviation 30, reaching its highest values at the right lower region of the polyhedron in Figure 11(a). The obtained G_m values, depicted in Figure 11(b), range from −400 nMhos to −10 nMhos, peaking at the upper corner of the polyhedron. The surfaces obtained for these transistor constants present some lump-like irregularities, which are in agreement with the variation of beta suggested by the changing slopes of experimental isolines sometimes found in the literature (e.g., Ref. 12).

In the main text, we present the optimal linearity for different values of circuit parameters R_c and V_{cc}, as indicated in Figure 9. It is also interesting to verify how other amplifier properties vary for the same circuit parameters. This can be done using the following procedure. For each load line of the circuit, defined by a pair of parameters (R_c, V_{cc}), an optimal operation range is found using the linearity methodology. Then, the average of a given amplifier property is calculated for this optimal range. As a consequence, a surface can be defined, associating the amplifier property with the considered values of circuit parameters R_c and V_{cc}. In Figures 12(a) and 12(b) we show, respectively, the average of total amplification and transconductance for distinct values of R_c and V_{cc}. This figure, which is discussed in Section IV, can be compared with Figure 9, containing the linearity deviation of the amplifier for the same values of R_c and V_{cc}.

3. Total harmonic distortion

A traditional way to study the linearity of an amplifier is by estimating its total harmonic distortion (THD).[7] For a given frequency f, this measurement can be obtained by using a pure sinusoidal function with frequency f as input, identifying new harmonic components in the output (a perfectly linear amplifier would produce no such components), and taking the ratio between the magnitudes of these spurious harmonics (V_{2f}, V_{3f}, etc.) and of the fundamental (V_f). More formally, the THD can be calculated as

$$THD(f) = \frac{\sqrt{V_{2f}^2 + V_{3f}^2 + V_{4f}^2 + \cdots}}{V_f}. \tag{A4}$$

Because the load is purely resistive, the same THD will be attained irrespectively of the input frequency f. Therefore, we considered a sinusoidal function with $f = 1$ kHz. In the main text, we use THD to provide an additional characterization of the linearity besides the error of the least squares regression.

APPENDIX B: DERIVATION OF THE TOTAL CURRENT GAIN

The total current gain of the circuit used for the experiments (shown in Figure 7) is given by Equation (4). Since I_c, the collector current, is a function of V_c, the collector voltage, and I_b, the base current, we can also write A in terms of the partial derivatives of I_c, that is,

$$A = \frac{dI_c}{dI_b} = \left(\frac{\partial I_c}{\partial V_c}\frac{dV_c}{dI_b} + \frac{\partial I_c}{\partial I_b}\frac{dI_b}{dI_b}\right). \tag{B1}$$

Replacing the partial derivatives by the transistor properties indicated in Equations (A1) and (A3), we obtain

$$A = \frac{dI_c}{dI_b} = \frac{1}{R_o}\frac{dV_c}{dI_b} + \beta. \tag{B2}$$

Since V_c and I_c are related through the circuit parameters according to

$$V_{cc} = R_c I_c + V_c, \tag{B3}$$

the total derivative dV_c/dI_b can be rewritten in terms of A and R_c as

$$\frac{dV_c}{dI_b} = -R_c\frac{dI_c}{dI_b} = -R_c A. \tag{B4}$$

Therefore,

$$A = -\frac{R_c}{R_o}A + \beta \Longrightarrow A = \frac{R_o\beta}{R_o + R_c}. \tag{B5}$$

[1] L. R. Rabiner and B. Gold, *Theory and Application of Digital Signal Processing* (Prentice-Hall, Inc., 1975), Vol. 1.

[2] M. H. Rashid, *Power Electronics Handbook: Devices, Circuits and Applications* (Academic Press, 2010).

[3] W.-K. Chen, *Active Network Analysis: Feedback Amplifier Theory* (World Scientific, 2016), Vol. 15.

[4] H. S. Lee and M. Tomizuka, "Robust motion controller design for high-accuracy positioning systems," IEEE Trans. Ind. Electron. **43**, 48–55 (1996).

[5] RCA, *RCA Receiving Tube Manual RC-20* (RCA Corporation, 1960).

[6] D. Self, *Audio Power Amplifier Design*, 6th ed. (Focal Press, 2013).

[7] B. Cordell, *Designing Audio Power Amplifiers* (McGraw-Hill, 2011).

[8] Specifically, the subsequence is given by $S_{k,q} = ((x_k, y_k), (x_{k+1}, y_{k+1}), \ldots, (x_{q=k+m-1}, y_{q=k+m-1}))$.

[9] P. R. Bevington and D. K. Robinson, *Data Reduction and Error Analysis* (McGraw-Hill, 2003).

[10] N. R. Draper, H. Smith, and E. Pownell, *Applied Regression Analysis* (Wiley, New York, 1966), Vol. 3.

[11] L. da F. Costa and R. M. Cesar, Jr., *Shape Classification and Analysis: Theory and Practice* (CRC Press, Inc., 2009).

[12] R. F. Shea, *Transistor Audio Amplifiers* (Wiley, 1955).

[13] G. Parker, *Introductory Semiconductor Device Physics* (CRC Press, 2004).

[14] H. J. Zimmerman and S. J. Mason, *Electronic Circuit Theory* (Wiley, 1959).

[15] R. Renka and A. Cline, "A triangle-based C^1 interpolation method," Rocky Mt. J. Math. **14**, 223 (1984).

[16] P. Lancaster and K. Salkauskas, "Surfaces generated by moving least squares methods," Math. Comput. **37**, 141–158 (1981).

[17] D. Levin, "The approximation power of moving least-squares," Math. Comput. Am. Math. Soc. **67**, 1517–1531 (1998).

[18] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," Science **286**, 509–512 (1999).

[19] G. Palumbo and S. Pennisi, *Feedback Amplifiers: Theory and Design* (Springer Science+Business Media, 2007).