2/13/02 Final 5.2.1

Representations of Sound That Are Insensitive to Spectral Filtering and Parameterization Procedures

David N. Levin
Department of Radiology
University of Chicago
Chicago, Illinois 60637

Accepted for publication in the *Journal of the Acoustical Society of America*

Send correspondence to:
David N. Levin
Department of Radiology, MC2026
University of Chicago
5841 S. Maryland Ave.
Chicago, IL 60637

Tel: 773-702-6511
Fax: 773-834-7610
Email: d-levin@uchicago.edu
Web: http://www-radiology.uchicago.edu/faculty/Levin.html

ABSTRACT

This paper describes representations of time-dependent signals that are invariant under any invertible signal distortion. Such a representation can be created by rescaling the signal in a non-linear dynamic manner that is determined by recently encountered signal levels. Information that is encoded in such representations will be faithfully communicated in the presence of severe signal distortions, which may originate in the transmitter, receiver, or the channel between them. As in speech communication, the receiver is "blind" and need not characterize the form of the signal distortion, which remains unknown. The method is applied to analytical examples, acoustic waveforms of human speech, and the short-term Fourier spectra of a bird song. The results suggest that the rescaled representation of a sound is insensitive to the way its spectra have been filtered and parameterized, as long as those processes do not obliterate the differences between the various spectra in the sound. Finally, the possible "speaker"-independence of these representations is explored in the context of a simple linear prediction model of vocal tracts with a single degree of freedom.

PACS numbers: 43.72.Ar

43.72.-p

43.72.Ne

43.60.Lq

I. INTRODUCTION

The fidelity of electronic communication is often degraded when the signal is distorted as it propagates through the transmitter, receiver, and the channel between them. Many telecommunications systems attempt to correct for these effects by periodically transmitting calibration data (e.g., test patterns) so that the receiver can characterize the distortion and then compensate for it by "unwarping" the signal (e.g., channel equalization). These techniques may be costly because they take the system "off-line" for brief periods or otherwise reduce its efficiency.

In contrast, humans perceive the information content of ordinary speech to be remarkably invariant, even though the signal may be transformed by significant alterations of the speaker's voice, the listener's auditory apparatus, and the channel between them (1, 2, 3). Yet there is no evidence that the speaker and listener exchange calibration data in order to characterize and compensate for these distortions. Evidently, the speech signal is redundant in the sense that listeners extract the same content from multiple acoustic signals that are transformed versions of one another. Human visual perception is also invariant when the raw signal is distorted by a variety of changes in observational conditions. This phenomenon is strikingly illustrated by experiments (4-8) in which subjects wore goggles creating severe geometric distortions of the visual field (e.g., inversion, reflection, and/or non-linear warping). Although the subjects initially perceived the distortion, their perceptions of the world returned to the pre-experimental baseline after several weeks of constant exposure to familiar stimuli seen through the goggles. Apparently, humans utilize recent sensory experiences to automatically "recalibrate" their perception of subsequent sensory data.

In earlier reports (9-11), the author showed how to design sensory devices that behave in this way. In such devices, the signal is represented by a non-linear function of its instantaneous level at each time, with the form of this scale function being determined by the signal levels in a chosen time interval (e.g., recently-encountered signal levels). This rescaled signal is invariant if the signal

levels at all times are invertibly transformed by the same distortion. This is because the transformation's effect on the signal level at any time is cancelled by its effect on the scale function at that time. This can be understood by considering the following analogy. The positions of particles in a plane can be described in terms of a "natural" coordinate system (or scale) that is rooted in the particle collection's intrinsic structure; i.e., the coordinate system that originates at the collection's center of "mass" and is oriented along its principal moments of "inertia". Each particle's position with respect to this intrinsic scale is invariant under rigid rotations and translations that change all particle coordinates in the extrinsic coordinate system. This is because each particle and the collection's intrinsic coordinate system are rotated and translated in the same manner, so that each particle's location with respect to that coordinate system is unchanged. Earlier papers showed how the signal levels recently detected by a sensory device may have an intrinsic structure that defines a non-linear coordinate system (or scale) on the manifold of possible signal levels (9-11). The "location" of the currently detected signal level with respect to this intrinsic coordinate system is invariant under any invertible transformation (linear or non-linear) of the entire signal time series. This is because the signal level at any time and the scale function at the same time point are transformed in a manner that leaves the rescaled signal level unchanged.

In this paper, we show how this self-referential rescaling technique can be used to faithfully communicate information in the presence of any invertible signal distortion, without resorting to the explicit calibration procedures that are often used in telecommunications. This is done by encoding the information in the above-described signal invariants, which are not affected by such distortions. The requirement of invertibility is relatively weak; it simply means that the distortion does not compromise the receiver's ability to distinguish between signal levels that are distinguished by the transmitter, and vice versa. In Section II, the method is derived, and then it is illustrated with analytic examples. The mathematical properties of the technique are also demonstrated by applying it to the acoustic waveforms of human speech in Section III. In Section IV, the method is applied to

the parameterized spectra of a bird song in order to show that the new sound representations are insensitive to the way the spectra have been filtered and parameterized. In Section V, a simple linear prediction model of vocal tracts with a single degree of freedom is used to explore the possible "speaker"-independence of the new sound representations. In Section VI the implications of this work are discussed, particularly the possible application to speech recognition systems.

II. THEORY

Let $x(t)$ be the time-dependent signal in the transmitter (e.g., the signal driving its antenna circuit), and let X be its value at time T. In this paragraph, we show how to rescale the signal level at this particular time point. The exact same procedure can be used to rescale the signal level at other times, thereby deriving a representation of the entire signal time series. Suppose that $x(t)$ passes through all of the signal levels in $[0, X]$ at one or more times during the interval $T - \Delta T \le t < T$. Here, ΔT is a parameter that can be chosen freely, although it influences the adaptivity and noise sensitivity of the method (see below). At each $y \in [0, X]$, define the value of the function $h(y)$ to be

$$h(y) = \left\langle \frac{dx}{dt} \right\rangle_y \qquad (1)$$

where the right side denotes the derivative averaged over those times in $T - \Delta T \le t < T$ when $x(t)$ passes through the value y. If $h(y)$ is non-vanishing for all $y \in [0, X]$, it can be used to compute the scale function $s(x)$ on this interval

$$s(x) = \int_0^x \frac{dy}{h(y)} \qquad (2)$$

The quantity $S = s(X)$ can be considered to represent the level of the transmitter signal X at time T, after it has been non-linearly rescaled by means of the function $s(x)$. Now, suppose that the signal in the receiver's detection circuit is related to the signal in the transmitter by the time-independent transformation $x \rightarrow x' = x'(x)$. The transformation $x'(x)$ could be the result of a time-independent distortion (linear or non-linear) that affects the signal as it propagates through the internal circuits of the transmitter and receiver, as well as through the channel between them. Furthermore, suppose that $x \rightarrow x'$ is invertible (i.e., $x'(x)$ is monotonic), and suppose that it preserves the null signal (i.e., $x'(0) = 0$). As mentioned earlier, the requirement of invertibility is relatively weak. It simply means that the distortion does not compromise the receiver's ability to distinguish between signal levels that are distinguished by the transmitter, and vice versa. The transformed signal $x'(t) = x'[x(t)]$ has the value $X' = x'(X)$ at $t=T$. During $T - \Delta T \le t < T$, $x'(t)$ passes through each of the values in $[0, X']$, because of our assumption that $x(t)$ attains all of the values in $[0, X]$ during that time interval. Therefore, for each $y' \in [0, X']$, the process in Eq.(1) can be applied to the transformed signal in order to define the function $h'(y')$ at time T

$$h'(y') = \left\langle \frac{dx'}{dt} \right\rangle_{y'} \qquad (1')$$

where the right side denotes the derivative averaged over those times in $T - \Delta T \le t < T$ when $x'(t)$ passes through the value y'. By substituting $x'(t) = x'[x(t)]$ in Eq.(1'), using the chain rule of differentiation, and noting that $x(t)$ passes through the value y when $x'(t)$ passes through the value $y' = x'(y)$, we find $h'(y') = \left. \frac{dx'}{dx} \right|_{y} h(y)$. The function $h'(y')$ is non-vanishing for $y' \in [0, X']$ because

the monotonicity of $x'(x)$ implies $dx'/dx \neq 0$. This means that the process in Eq.(2) can be used to compute a scale function $s'(x')$ on this interval

$$s'(x') = \int_0^{x'} \frac{dy'}{h'(y')} \qquad (2')$$

The quantity $S' = s'(X')$ represents the level of the receiver signal X' at time T, after it has been rescaled by means of a function $s'(x')$, which was derived from $x'(t)$ just as $s(x)$ was derived from $x(t)$. Because of our assumption that $x = 0$ transforms into $x' = 0$, a change of variables ($y \rightarrow y'$) in Eq.(2) implies $s'(x') = s(x)$ and, therefore, $S' = S$. This means that the rescaled value of a signal is invariant under the transformation $x \rightarrow x'$. In other words, the rescaled value S of the undistorted signal level at time T, computed from recently encountered undistorted signal levels, will be the same as the rescaled value S' of the distorted signal level at time T, computed from recently encountered distorted signal levels. Now, the above procedure can be followed in order to rescale the signal levels at times other than T. The resulting time series of rescaled signal levels $S(t)$, which the transmitter derives from the transmitted signal $x(t)$ in this way, will be identical to the time series of rescaled signal levels $S'(t)$, which the receiver derives from the received signal $x'(t)$. Thus, if the transmitter encodes information in the rescaled representation $S(t)$ of its signal, that information will be invariantly communicated to the receiver, even in the presence of invertible distortions of the propagating signal.

Notice that the above derivation assumes the prior knowledge that the unknown signal transformation satisfies $x'(0) = 0$. However, a weaker assumption will suffice: namely, prior knowledge of a single pair of "reference" amplitudes (a, a') that are transformed into one another by the unknown transformation: $x'(a) = a'$. In other words, the pair of reference amplitudes do not have to be zero, as was assumed above for simplicity. These reference amplitudes can be used to

define DC-offset signal amplitudes: $z = x - a$ and $z' = x' - a'$. Because x and x' are invertibly related and because the transformations $z(x)$ and $z'(x')$ are invertible, it follows that z and z' are invertibly related. Furthermore, the invertible transformation between z and z' satisfies $z'(0) = 0$ because the transformation $x'(x)$ satisfies $x'(a) = a'$. Therefore, the above-described methodology can be directly applied in order to derive invariant representations of $z(t)$ and $z'(t)$.

It is evident that the functional form of a signal's *s* representation depends on the choice of ΔT. However, the above analytic proof shows that untransformed and transformed signals will have the same *s* representation for any choice of ΔT, as long as the same value of ΔT is used to derive the *s* representation of each of these signals. In other words, the choice of ΔT influences the form of a signal's *s* representation, but it does not affect the invariance of the *s* representation for any given choice of ΔT. As demonstrated in Section III, the choice of ΔT also influences the manner in which the invariance of the *s* representation breaks down when the assumptions of the paper are violated (e.g., by time-dependence of the unknown signal transformation or by the presence of noise).

Notice that the forms of the scale functions $s(x)$ and $s'(x')$ (and of $h(y)$ and $h'(y')$) will usually be time-dependent because they are computed from the time course of previously encountered signals. At some times, both the receiver and transmitter may be unable to compute a rescaled signal level. This will happen if the scale function in Eq.(2) does not exist because the quantity $h(y)$ vanishes for some $y \in [0, X]$ or if the function $h(y)$ cannot even be computed at some values of *y* because these signal levels were not encountered recently. Because of the monotonicity of $x'(x)$, neither the transmitter nor the receiver can compute a signal invariant at such times. Therefore, if the information is exclusively encoded in signal invariants, no information is transmitted or received at these times. Note that this phenomenon does not result in loss of information or otherwise compromise the fidelity of communication, although it does reduce the time efficiency of the communication process.

It is useful to illustrate these results with a simple example. Suppose the transmitter signal $x(t)$ is a long periodic sequence of triangular shapes, like those in Fig. 1a. Let *a* and *b* be the slopes of the lines on the left and right sides, respectively, of each shape; Fig. 1a shows the special case: $a = 0.1$ and $b = -0.5$ (measured in inverse time units). If we choose ΔT to be an integral number of periods of $x(t)$, it is easy to see from Eqs.(1, 2) that the transmitter signal implies $h(y) = (a + b)/2$ and $S(t) = s[x(t)] = 2x(t)/(a + b)$ at each point in time. Figure 1b shows $S(t)$, which is the transmitted signal after it has been rescaled at each time point as dictated by its earlier time course. Now, suppose that the receiver detects a signal that is distorted by any of the following non-linear functions: $x'(x) = g_1 \ln(1 + g_2 x)$ where $g_2 > 0$. For example, if $g_1 = 0.5$ and $g_2 = 150$, the distorted signal in the receiver $x'(t)$ looks like Figure 1c. When Eq.(1') is used to compute $h'(y')$ from the received signal , the result is:

$$h'(y') = \frac{1}{2}(a + b) g_1 g_2 e^{-y'/g_1} \qquad (3)$$

at each point in time. Then, Eq.(2') shows that the rescaled version of the receiver signal is

$$S'(t) = s'[x'(t)] = \frac{2(e^{x'(t)/g_1} - 1)}{g_2(a + b)}, \qquad (4)$$

Substituting $x'(t) = x'[x(t)]$ into Eq.(4) shows that $S'(t) = S(t)$. In other words, the rescaled signal $S'(t)$, which the receiver derives from the distorted signal $x'(t)$, is the same as the rescaled signal $S(t)$, which the transmitter derives from the undistorted signal $x(t)$. This is because the effect of the invertible signal transformation on the signal level at any given time ($x(t) \rightarrow x'(t)$) is *cancelled* by its effect on the form of the scale function at that time ($s(x) \rightarrow s'(x')$). Notice that $s(x)$ and $s'(x')$ (as well as $h(y)$ and $h'(y')$) happen to be time-independent in this particular example, and this

implies that $x(t)$ and $x'(t)$ are rescaled in a time-independent fashion. This is because, in order to simplify the calculation, $x(t)$ was chosen to be periodic and ΔT was chosen to be an integral number of these periods. In the general case, the scale functions depend on time in a manner dictated by the earlier time course of the signal. However, the transmitter and receiver will still derive identical self-scaled signals (i.e., $S(t) = S'(t)$), as demonstrated by the proof at the beginning of this Section and as illustrated by the experimental examples in the next three Sections.

III. EXPERIMENTS WITH ACOUSTIC WAVEFORMS OF HUMAN SPEECH

In the context of speech recognition systems, the rescaling method should probably be applied to the parameterized short-term Fourier spectra of speech. As outlined in Section VI, this can be done by generalizing the techniques in Sections IV and V. However, in this Section, we simply apply rescaling to the acoustic waveforms of speech for the purpose of further illustrating the mathematical behavior of the method. The acoustic signals were generated by an adult male American who uttered English words with speed and loudness that were characteristic of normal conversation. These sounds were digitized with 16 bits of precision at a sample rate of 11.025 kHz. Figure 2a shows a 40 ms segment of digitized data ($x(t)$), located at the midpoint of the 334 ms signal corresponding to the word "door". The spectrograms in this Figure and in the rest of this Section were produced by computing the short-term Fourier transform of the acoustic waveform in 256-sample Hamming windows that were centered at 64-sample intervals. Figure 2b shows the "*s* representation" (i.e., the rescaled signal *S(t)*) that was derived from Fig. 2a by the method of Section II. The value of *S* was determined at each time point by a scale function *s(x),* which was derived from the previous 10 ms of signal (i.e., ΔT = 10 ms). These scale functions are shown by the horizontal lines in Fig. 2a, which denote values of x corresponding to $s = \pm 50n$ for *n=1, 2, … .* Figure 2d shows the signal that was derived from Fig. 2a by means of the non-linear transformation

($x'(x)$) shown in Fig. 2c. For example, this could represent the effect of a non-linear microphone or amplifier. Figure 2e is the rescaled signal that was derived from Fig. 2d with the parameter ΔT chosen to be 10 ms. Although there are significant differences between the "raw" signals in Figs 2a and 2d, their *s* representations (Figs. 2b and 2e) are almost identical, except for a few small discrepancies that can be attributed to the discrete methods used to compute derivatives. Thus, the *s* representation was invariant under a non-linear signal distortion, as expected from the derivation in Section II. It is interesting to note that this result is apparent when one listens to the sounds represented in Fig. 2 (see the sounds posted at the Web site in reference 12). Although all four signals in Fig. 2 sound like the word "door", there is a clear difference between the sounds of the two raw signals, and there is no perceptible difference between the sounds of their rescaled representations. In general, the rescaled signals sound like the word "door", uttered by a voice degraded by slight "static".

Some comments should be made about technical aspects of the example in Fig. 2. The rescaled signals in Figs 2b and 2e were computed by a minor variant of the method in Section II. Specifically, we assumed that all signal distortions were *monotonically positive*, and we restricted the contributions to Eq. (1) and Eq.(1') to those time points at which the signal had a *positive* time derivative as it passed through the values y and y', respectively. The rescaled signal is still invariant because monotonically positive transformations do not change the sign of the signal's time derivative, and, therefore, the functions $h(y)$ and $h'(y')$ were still constructed from time derivatives at identical collections of time points. At each time point, we attempted to compute the rescaled signal from the signal time derivatives encountered during the most recent 10 ms ($\Delta T = 10$ ms). At some times, the signal could not be rescaled because the signal level at that time was not attained during the previous 10 ms, and, therefore, there were no contributions to the right side of Eq.(1) for some values of y. For example, this happened at $t \sim 163$, 174, and 185 ms in Fig. 2. At such times, a signal invariant could not be computed, and communication of distortion-invariant information

was not possible. As mentioned in Section II, this occurs at identical time points when rescaling is applied to the "undistorted" signal (e.g., Fig. 2a) and to any distorted version of it (e.g., Fig. 2d). This means that the *s* representations of all of these signals are non-existent at identical time points and that at all other times they exist and have the same values. Therefore, this phenomenon does not corrupt the invariance of the signal's *s* representation, although it does reduce its information content. In this experiment, the *s* representation could not be computed at 8% of all time points. The rescaled signal was set equal to zero at those times.

Figure 3 shows what happened when the nature of the distortion changed abruptly. The signal in Fig. 3b was derived by applying the non-linear transformation in Fig. 2c to the first half (i.e., the first 167 ms) of the signal excerpted in Fig. 2a and by applying the non-linear transformation in Fig. 3a to the second half of that signal. Figure 3c shows the *s* representation derived by dynamically rescaling Fig. 3b with $\Delta T = 10$. Comparison of the latter to Fig. 2b shows that the *s* representation was invariant except during the time period $167\,ms \le t \le 177\,ms$. These discrepancies can be understood in the following way. During this time interval, the rescaled signal in Fig. 3c was derived from a mixed collection of signal levels, some of which were transformed as in Fig. 2c and some of which were transformed as in Fig. 3a. This violates the proof of invariance (Section II), which assumed the time-independence of the transformation between the "undistorted" and "distorted" signals. Notice the transitory nature of this corruption of the *s* representation. The rescaled signals in Figs. 2b and 3c became identical again, once sufficient time (ΔT) elapsed for the distortion to become constant over the time interval utilized by the rescaling procedure. In other words, the dynamic rescaling process was able to adapt to the new form of the distortion and thereby "recover" from the disturbance. Therefore, if communicating systems are encoding information in the signal's *s* representation, faithful communication will be reestablished ΔT time units after the onset of a change in the transmitter or the receiver or the channel between them. This adaptive behavior resembles that of the human subjects of the goggle experiments mentioned in Section I.

Figure 4 illustrates the effect of noise on dynamic rescaling. Figure 4a was derived from Fig. 2d by adding white noise so that the signal-to-noise ratio was equal to 15 dB. This causes a pronounced hiss to be superposed on the word "door" when the entire 334 ms sound exemplified by Fig. 4a is played (see the sounds posted at the Web site in reference 12). Figure 4b is the *s* representation, derived by dynamically rescaling Fig. 4a with ΔT=10 ms. Comparison of Figs. 4b, 2e, and 2b shows that the noise has caused some degradation of the invariance of the *s* representation. This is expected because additive noise ruins the invertibility of the transformations relating Figs. 4a, 2d, and 2a, thereby violating the proof of the invariance of *S* in Section II. The noise sensitivity of the *s* representation can be decreased by increasing ΔT, because this increases the number of contributions to the right side of Eq.1, which tends to "average out" the effects of noise. However, such an increase in ΔT means that more time is required for the dynamic rescaling process to adapt to a sudden change in distortion.

IV. EXPERIMENTS WITH THE SPECTRA OF A BIRD SONG

In this Section, we consider a bird song whose short-term Fourier spectra oscillated along a curve in the multidimensional space of spectral parameters. At any given time, the sound's spectrum was characterized by $x(t)$, a function that gave its position along this curve. We demonstrate that this function has a self-referential representation $S(t)$, which is unaffected by passing the sound through a wide variety of filters. We also demonstrate the insensitivity of this self-referential representation to the method of parameterizing the sound's spectra.

We considered a 260 ms sound of a warbler that consisted of one long "chirp" and two short "chirps" (Canary: The Cornell Bioacoustics Workstation, Version 1.2, Cornell Laboratory of Ornithology, Ithaca, N.Y.). This sound was digitized at 22.3 kHz with 16 bits of precision and subjected to a short-term Fourier transform, after it had been "windowed" with a Hamming function

in 128-sample frames, which were centered at 32 sample intervals. The time domain waveform and spectrogram of the sound are shown in Figure 5a. The magnitude of each spectrum was subjected to a discrete cosine transformation (DCT), after it had been smoothed by averaging over bins of 800 Hz width. Thus, each spectrum was represented by a single point in the 64-dimensional space of DCT coefficients. Figure 5b displays these points after they were projected into the three-dimensional subspace of the coefficients with indices i = 1, 4, and 7. It is evident from this Figure (as well as from the points' positions in other 3D subspaces) that all of the sound's spectra lie close to an arc-like curve in the space of all DCT coefficients. During the first chirp, the spectrum moved from a position at the densely populated end of the curve to its sparsely populated distal segment and then followed the reverse trajectory back to its initial position. During each of the second and third chirps, the spectrum made oscillatory movements along the curve's proximal segment, before coming to rest near its initial position. We considered a simulated sound detector that "sensed" the nine spectral DCT coefficients with indices i= 1, 4, 7, 10, 12, 13, 15, 16, 18. The detector's sensor state $x(t)$ consisted of the spectrum's position along the above-described curve in this 9-dimensional subspace. There are many dimensional reduction procedures that can be used to assign each spectral point to position on this one-dimensional manifold. In this example, this was done by defining a piecewise linear trajectory that hugged the curvilinear collection of spectral points in the 9-dimensional subspace (Fig. 5b), and each spectrum was assigned the position (x) of the nearest point on this collection of chords. The x coordinate system along these chords was defined so that its origin was at the position of the first spectrum in the sound (i.e., it was defined so that $x(0) = 0$). The left panel of Fig. 5c shows the spectrum's position as a function of time, which constitutes the detector sensor state elicited by the sound in Fig. 5a. The right panel of Fig. 5c shows the function $S(t)$ obtained by using Eqs. (1, 2) to rescale the left panel, with the choice ΔT = 300 ms. Notice that the rescaling process has segmented the sound into the three chirps, separated by time intervals during which invariants could not be computed. As proved in Section II, the same rescaled function

$S(t)$ would have been computed if we had used *any* x coordinate system along the sound's trajectory, as long as it was invertibly related to the above-described scale and originated at the same point on the curve. For instance, we would have obtained the same rescaled sensor state if we had projected the spectral points onto *any* piece-wise continuous collection of chords that was sufficiently close to the arc-like collection of spectral points.

We attenuated the high frequency components of the sound in Fig. 5 by convolving it with a Hamming function corresponding to a filter passing frequencies in the range 1.5 ± 5.6 kHz. Comparison of Figs. 6a and 5a shows the suppression of the high frequencies in this sound, which had a noticeably lower "pitch" than the one in Fig. 5a (see the sounds posted at the Web site in reference 12). As before, the DCT coefficients of the sound's spectra defined points that tended to cluster along an arc-like line in 64-dimensional DCT space. Figure 6b shows these points after they were projected onto the subspace defined by the DCT coefficients with indices i = 1, 4, and 7. The left panel of Fig. 6c shows each spectrum's position along the corresponding curve through the nine-dimensional subspace of DCT coefficients "sensed" by the simulated detector (i.e., coefficients with indices i= 1, 4, 7, 10, 12, 13, 15, 16, 18). This function represents the detector's sensor state $x(t)$ as it "listened" to the filtered sound in Fig. 6a. The right panel of Fig. 6c shows the function $S(t)$ that was obtained by rescaling the left panel with the choice $\Delta T = 300$ ms. Figures 5c and 6c demonstrate that the original and filtered sounds had similar rescaled representations, even though they produced significantly different sensor states in the simulated detector. This is due to the fact that the unrescaled sensor states were related by an invertible transformation that preserved null values. Such a transformation existed because two conditions were satisfied: 1) an invertible transformation related each unfiltered spectrum's position on the one-dimensional manifold of the sound's spectra to the corresponding filtered spectrum's position on the one-dimensional manifold of spectra in the filtered sound; 2) for both the unfiltered and filtered sounds, the position of each spectrum on the one-dimensional manifold of the sound's spectra was invertibly related to its

coordinate along the sound's curve in the subspace of "sensed" DCT coefficients. The first condition was satisfied because the filtering operation did not obliterate all of the differences between any pair of spectra in the sound; i.e., it did not map different spectra in the original sound onto the same filtered spectrum. Specifically, because the width of the filter's impulse response was less than the width of the window used to create short-term spectra, at every time point each component of the filtered spectrum was equal to the corresponding component of the unfiltered spectrum, multiplied by the corresponding spectral component of the filter's impulse response. Because the filter's spectrum was non-zero for most important frequencies, this transformation did not map two different spectra in the unfiltered sound onto a single spectrum of the filtered sound. The second condition means that the "sensed" DCT coefficients were sensitive to the differences among the sound's spectra. This was true because the spectra followed a trajectory that was never simultaneously orthogonal to the axes of all of the sensed coefficients. Because the above two conditions were satisfied in this example, there was an invertible transformation between the sensor states derived from unfiltered and filtered spectra at identical times. Furthermore, this mapping preserved the null value. This is because the origins of the x coordinate systems for the two sounds (unfiltered and filtered) were both defined to be at the positions of the first spectrum in each sound, and the mapping transforms these positions into one another. In other words, in this example, two detectors, which sensed a sound through unfiltered and filtered channels, used the first detected spectrum to define a common origin for their sensor state scales. This procedure is analogous to having a choir leader play a pitch pipe in order to establish a common origin of the musical scale among the singers.

Figure 7a shows the sound produced by attenuating the low frequency components of the sound in Fig. 5a. This sound was created by convolving the sound in Fig. 5a with a Hamming function corresponding to a filter passing frequencies in the range 6.0 ± 2.8 kHz. Comparison of Figs. 7a and 5a shows the suppression of the low frequencies of this sound, which had a high-pitched quality, noticeably different from the sound in Fig. 5a (see the sounds posted at the Web site in

reference 12). Figure 7b shows three DCT coefficients (indices i = 1, 4, and 7) of the sound's spectra, which followed a curvilinear trajectory in the full DCT coefficient space, as well as in the depicted three-dimensional subspace. The left panel of Fig. 7c shows the detector's sensor state $x(t)$ as it "listened" to the filtered sound in Fig. 7a; namely, the Figure shows each spectrum's position along the sound's curve through the subspace of the nine "sensed" DCT coefficients (indices i= 1, 4, 7, 10, 12, 13, 15, 16, 18). The right panel of Fig. 7c shows the function $S(t)$ that was obtained by rescaling the left panel with the choice ΔT = 300 ms. It is evident from Figs. 5c and 7c that similar rescaled representations were derived from the original and filtered sounds, even though they produced noticeably different unrescaled sensor states. As before, this is expected because the sensor state time series produced by the two sounds were related by an invertible transformation. Notice that this transformation was non-linear even though the spectral coefficients of the two sounds were related by a linear filtering operation. This is because the transformation between sensor states was a composite of mappings that related the positions of spectra in non-linear subspaces (i.e., arc-like curves) in the spaces of spectral coefficients.

Next consider what happened when the sound in Fig. 5a was sensed by a simulated detector with different "sensors". This detector "sensed" eight DCT coefficients (with indices i= 5, 6, 8, 9, 11, 14, 17, 19) that were completely different than those recorded by the previously described simulated detector. The sound's spectra were represented by points in the eight-dimensional space of these coefficients, and these points were on (or nearly on) a curve. This is illustrated by Fig. 8a, which depicts the projection of these points on the 3D subspace of DCT coefficients with indices i= 5, 8, 11. The left panel of Fig. 8b shows the time course of this detector's sensor data as it "heard" the sound in Fig. 5a, and the right panel shows the rescaled sensor state representation. Comparison of Figs. 5c and 8b shows that the two simulated detectors produced nearly identical rescaled sensor state representations, despite the fact that they were "equipped" with different "sensors". This is expected because the sensor states of the two detectors were related by an invertible mapping. To

see this, recall that each detector was sensitive to the differences among the sound's spectra. Therefore, in each detector, the sensor state induced by a spectrum was invertibly related to the position of that spectrum within the one-dimensional collection of the sound's spectra in the space of all spectral parameters.

One can argue that rescaled representations are insensitive to the way sound spectra are filtered or parameterized because they reflect an "inner" property of the motion of the warbler's vocal tract, rather than reflecting the "outer" signal propagation and detection processes. Because the spectra of the warbler's song were confined to a one-dimensional trajectory in the space of all possible spectra, it is likely that the warbler's vocal tract moved through a one-dimensional manifold in the space of all possible configurations (e.g., its vocal tract muscles were continuously reconfigured in a coordinated fashion that depended on just one time-dependent parameter). If the "sensed" DCT coefficients were sensitive to each of these motions, there must be an invertible mapping between the detector's sensor state (x) and the "inner" parameter that describes the configuration of the vocal tract. It follows that the time courses of the sensor state and this parameter must have identical rescaled representations; i.e., the sound's rescaled representation describes an intrinsic property of the vocal tract's motion that is independent of the way it is observed. The next Section explicitly demonstrates this phenomenon in the context of a family of simulated vocal tracts with one degree of freedom.

V. EXPERIMENTS WITH SPECTRA OF SYNTHETIC SPEECH-LIKE SOUNDS

In this Section, we consider a family of simple simulated vocal tracts that are configured by varying a single parameter, and, in the context of these models, we demonstrate the "speaker"-independence of the rescaled sound representation. Specifically, we consider a pair of sounds that

are generated by two different simulated vocal tracts controlled by the same time-dependent parameter, and we show that these sounds have identical rescaled representations.

Each sound was generated by a standard linear prediction (LP) model (13). In other words, the signals' short-term Fourier spectra were equal to the product of an "all pole" transfer function and a "glottal" excitation function. The transfer function had six poles, two real and four complex (forming two complex conjugate pairs). The resulting sound spectra depended on the values of eight real quantities, six that described the positions of the poles and two that described the pitch and amplitude ("gain") of the excitation. Each of these quantities was a function of a single parameter (g), which itself depended on time. These eight functions described the physical nature of the simulated vocal tract, in the sense that they defined the manifold of all spectra that it could produce as g ranged over all of its possible values. The actual sound produced at any given time was determined by these eight functions, together with the value of $g(t)$. The latter function defined the "articulatory gesture" associated with the sound, in the sense that it determined how the simulated vocal apparatus was configured at each time. In a musical analogy, the g-dependent functions of the LP model would describe the possible spectra produced by a musical instrument played with one finger, and the function $g(t)$ would describe the motions of the musician's finger as it configures the instrument during a particular tune. In these examples, we considered "voiced" sounds that were driven by regular excitation functions. However, it is straightforward to apply the same methods to "unvoiced" sounds that are driven by noise-like excitation functions. In this experiment, we considered two simulated vocal tracts that were described by LP models, whose transfer functions were markedly different functions of g. The two vocal tracts were driven by regular excitation functions with pitches of 200 Hz and 125 Hz, respectively.

Figures 9b and 10a show the spectrograms of sounds that were generated when each of these simulated vocal tracts was controlled by the "gesture" function in Fig. 9a. These spectrograms were created by digitizing each sound at 10 kHz and then performing short-term Fourier transformations

in 10 ms Hamming windows that were advanced in increments of 5 ms. Although the sounds are audibly different, there is a noticeable similarity in the underlying pattern of variation (see the sounds posted at the Web site in reference 12).

Each sound was sensed by a detector that was simulated as follows. The spectrum at each time point was parameterized by the discrete cosine transformation (DCT) of its magnitude after it had been averaged in equally spaced 600 Hz bins. The simulated detector was assumed to sense the three coefficients with indices i = 2, 3, 6, and each sound spectrum defined a single point in the corresponding 3D space. These points fell on a curve generated by the DCT coefficients of the spectra produced by all possible configurations of the simulated vocal tract (i.e., all possible values of g). Figures 9c and 10b show the different configurations of this curve for the above-described vocal tract models. The precise shape of each curve depended on the nature of the modeled vocal tract (i.e., on the nature of the g-dependence of the model's poles and other parameters). As in Section IV, the detector's sensor state consisted of the position of the detected spectrum as it moved along the curve traversed by the sound. This position was measured in a coordinate system (denoted by x) that was established on each curve by projecting each of its points onto a connected array of chords that hugged the curve. As in Section IV, each x coordinate system was defined so that its origin coincided with the position of the first spectrum in the sound. The left panels of Figures 9d and 10c show the sensory signals produced by the sounds from the first and second simulated vocal tracts (Figures 9b and 10a), and the right panels show their corresponding rescaled representations (ΔT = 500 ms). Notice that the rescaled representations are nearly identical despite the differences between the simulated vocal tracts, sounds, and sensor states that produced them. This is because the two sensor signals were related by an invertible transformation that preserved the null state. This follows from the fact that the quantity g parameterizes the curves in Figs. 9c and 10b, and, therefore, there is an invertible transformation between $g(t)$ and the sensor signal $x(t)$ for each sound. For the same reason, the rescaled representation of $g_1(t) \equiv g(t) - g(0)$ is identical to the rescaled

representations of $x(t)$ for each sound. This means that the rescaled representation of each sound is an "inner" property of the vocal tract motion that produced it (i.e., a property that is independent how the sound propagates or is detected).

VI. DISCUSSION

This paper describes a non-linear signal processing technique for identifying the “part” of a signal that is invariant under any invertible signal distortion. This form of the signal is found by rescaling the signal at each time, in a manner that is determined by its time course in a chosen time interval. The rescaled signal (called its *s* representation) is unchanged if the original signal time series is subjected to any time-independent invertible transformation. Therefore, if a transmitter encodes information in this representation, it will be faithfully communicated to the receiver despite severe distortions of the propagating signal. This technique can also be used to establish communication among systems with heterogeneous transmitters and receivers that differ by unknown invertible mappings. Such a communication system resembles speech in the sense that: 1) the same information is carried by signals that are related to one another by a wide variety of distortions; 2) the transmitter and receiver can be "blind" to the nature of the distortion, which remains unknown; 3) if the distortion changes, faithful communication resumes after a period of adaptation.

The method in this paper differs significantly from techniques for multidimensional scaling or dimensional reduction (14-16). In each of these methods, it is necessary to impose an *ad hoc* measure of "distance" between each pair of neighboring data points (17) or, at least, to rank the distances between pairs of neighboring points (18, 19). In each case, the defined distances or rankings are not invariant under general, non-linear coordinate transformations. Therefore, the scale values assigned to each data point are also not transformation-independent, unlike the rescaled representations described in this paper. However, it should also be mentioned that multidimensional scaling methods are applicable to data that do not form a time series, unlike the technique in this paper.

Signals have the same *s* representation, as long as they are related to one another by *time-invariant* invertible transformations. Signals that are related by a time-dependent distortion may have different *s* representations immediately after each change in the nature of the distortion (e.g., Fig. 3). However, the rescaling process eventually adapts to the new form of the distortion, and the invariance of the signal's *s* representation is re-established. The length of this period of adaptation is ΔT, the user-defined parameter that determines the length of the signal history that is used to derive the scale at each time point. Decreasing ΔT can reduce the duration of this transient corruption of the *s* representation. However, this strategy will tend to reduce the *number* of signal time derivatives contributing to the computation of the signal scale and thereby increase the noise sensitivity of the scaling process. Conversely, the noise sensitivity of the *s* representation can be limited by increasing ΔT, at the cost of increasing the time required for the rescaling process to adapt to changes in signal distortion.

The family of all signals $x(t)$ that rescale to a given function $s(t)$ can be considered to form an equivalence class. If such a class includes a given signal, it also includes all invertible transformations of that signal. Signals can be assigned to even larger equivalence classes of all signals that lead to the same result when rescaling is applied N times in succession, where $N \geq 2$. For example, suppose that two signals do not have the same representation after one application of rescaling, but the same function is produced by two applications of rescaling. Then, the two signals can be considered to be equivalent at a deeper level, in the sense that the "inner" forms of their "inner" forms are the same. Such a relationship is analogous to the relationship between two spoken sentences or phrases, which have different meanings at a superficial level but have the same internal structure at a deeper level. Notice that successive applications of rescaling may result in more and more time points at which an intrinsic scale cannot be computed (e.g., because $h(y)$ cannot be computed or because it vanishes for some values of y). In other words, as in Fig. 5c, the serial rescaling process may segment the signal into information-bearing fragments that are separated by

time intervals containing no invariants (analogous to the words or phrases of spoken language). Successive applications of rescaling may eventually create a function that is not changed by further applications of the procedure (i.e., the serial rescaling process may reach a fixed "point"). For example, it is easy to show that, if the natural scale of a signal is time-independent (i.e., if $h(y)$ and $s(x)$ are time-independent), it will rescale to such a fixed point. Such a fixed point will also be reached if a signal is rescaled by means of a scale that has a fixed and time-independent form dictated by convention (e.g., if the frequency content of a simple melody is rescaled in terms of the equally tempered scale of Western music).

This technology may provide a useful "front end" for intelligent sensory devices, such as computer vision and speech recognition systems (20). The signals from the system's detectors would be rescaled before they are passed to the system's pattern recognition module for higher level analysis. The *s* representation of a stimulus is invariant under changes in observational conditions that cause invertible transformations of the states of the system's detectors. Such changes may include: 1) alterations of the internal characteristics of the device's detector (e.g., alterations of a microphone's gain characteristics), 2) changes in the observational environment that is external to the sensory device and the stimuli (e.g., changes in the acoustic channel), 3) modifications of the presentation of the stimuli themselves (e.g., modifications of the speaker's voice). Unlike conventional sensory systems, a device with this type of "representation engine" need not be periodically recalibrated with test stimuli, and its pattern recognition software need not be retrained when conditions change. This is advantageous because calibration procedures may be logistically impractical in some situations (remote, unsupervised devices), and, in any event, they reduce the device's duty cycle by taking it "off-line".

In Section IV, rescaling was demonstrated by applying it to a bird song whose spectra were filtered and parameterization in a variety of ways. The results suggested that the rescaled representation of a sound is insensitive to the exact nature of the filtering and parameterization

operations, as long as these procedures do not obscure the differences between the various spectra contained in the sound. In Section V, sounds produced by different vocal tract models had the same rescaled representations, as long as there was an invertible mapping between the underlying parameters controlling the vocal tract configurations. One can argue that rescaled representations will exhibit this kind of invariance for even more general classes of vocal tracts and detectors than those exemplified in Sections IV and V. To see this, assume that a detector of interest is sensitive to the differences between the spectra generated by any two configurations of a simulated vocal apparatus. It follows that those configurations are invertibly related to the detector's sensory states when the vocal tract model is used to create sounds. Therefore, if sound is sensed by two different detectors with this sensitivity, their sensory states will be invertibly related to one another and, consequently, have identical rescaled representations. Similarly, consider a single detector that sensitively records the sounds from two different vocal tract models. Assume that there is an invertible transformation between the configurations of the two vocal tracts when they generate comparable sounds. This might happen because one vocal tract mimics the other in a consistent fashion (in analogy with two speakers "reading" from the same "text" in a consistent manner). Then, the sensory signals induced in the detector by the two vocal tracts will be invertibly related. This is because these sensory signals are invertibly related to vocal tract configurations, which are themselves invertibly related. It follows that the detector will construct identical rescaled representations of the sounds from the two vocal tracts. Thus, rescaled sound representations should be "speaker"-independent and detector-independent for a wide class of speaker and detector models. Furthermore, as mentioned in Section V, because the vocal apparatus configurations producing a sound are invertibly related to the sensory signals induced by the sound, the underlying vocal tract parameter ($g(t)$) controlling that configuration will have the same rescaled representation as the sound itself. If this parameter's time course is regarded as a record of the articulatory gesture of the vocal tract and the rescaled sensor state representation is regarded as the "percept" induced by the

sound, this is consistent with the “motor” theory of speech perception (21). From this point of view, the rescaled representation of a sound represents an "inner" property of the vocal tract's motion, and that is why it is independent of "outer" features of the detection process (such as the nature of the detectors and the acoustic channel).

Although the experiments in Sections IV and V were performed with sounds depending on a single underlying degree of freedom, it is straightforward to generalize the methodology to signals produced by vocal tracts with multiple degrees of freedom. For example, consider the spectra generated by a vocal apparatus with two degrees of freedom. The set of all possible spectra from that apparatus define a 2D subspace (i.e., a sheet-like surface) in the space of spectral parameters (e.g., DCT coefficients), and the shape and location of that subspace characterize the range of possible sounds, which that vocal apparatus can produce. Any particular utterance will be characterized by a trajectory on this 2D surface, although the form of the unrescaled trajectory will generally be dependent on the nature of the detectors, channel, and speaker. In reference 10, the author demonstrated several techniques for rescaling such signals with two (or more) degrees of freedom. It may be computationally practical to apply these techniques to human speech, because there is evidence that the human vocal tract has a relatively small number of degrees of freedom (e.g., 3-5; see references 22-23). As in the previous paragraph, it can be argued that the same rescaled representation would be generated from any given utterance produced by a wide variety of speakers and propagated through a wide variety of channels and detectors. Therefore, a speech recognition device with such a “front end” may not need extensive retraining when the speaker’s voice, the acoustic channel, and/or the detectors are changed. One can also speculate that the adaptive nature of the rescaling process might enable it to account for coarticulation (13) during human speech. Recall that the manner in which each sound (i.e., each parameterized spectrum) is rescaled depends on the nature of recently encountered sounds. It could also depend on the nature of sounds to be encountered in the near future, if the interval ΔT is defined to include times *after* the

sound to be rescaled. In other words, the rescaled representation of each sound spectrum depends on its acoustic *context* (defined by the endpoints of ΔT), similar to the contextual dependence of speech perception that is the hallmark of the coarticulation phenomenon. Finally, the foregoing considerations make it tempting to speculate that the human brain itself decodes speech signals by constructing some type of dynamically rescaled version of speech spectra. This could account in part for the ease of speech communication involving a variety of speakers, listeners, and acoustic environments. Of course, the considerations in this paragraph are purely theoretical and must be tested by application to actual human speech spectra.

VI. REFERENCES

1. K. M. Ponting (Ed.). *Computational Models of Speech Pattern Processing*. Berlin: Springer, 1999.
2. L. C. Nygaard and D. B. Pisoni. Talker-specific learning in speech perception. *Perception and Psychophysics*, **60**, 355-376 (1998).
3. D. B. Pisoni. Some thoughts on "normalization" in speech perception. In: Johnson, K. and Mullennix, J. W. (Eds.), *Talker Variability in Speech Processing*, San Diego: Academic Press, 1997, pp. 9-32.
4. G. M. Stratton. Some preliminary experiments on vision without inversion of the retinal image. *The Psychological Review*, **3**, 611-617 (1896).
5. G. M. Stratton. Vision without inversion of the retinal image. *The Psychological Review*, **4**, 341-360 (1897).
6. G. M. Stratton. Vision without inversion of the retinal image (concluded). *The Psychological Review*, **4**, 463-481 (1897).

7. J. J. Gibson. Adaptation, after-effect, and contrast in the perception of curved lines. *Journal of Experimental Psychology*, **16**, 1-31 (1933).

8. R. Held and R. Whitman. *Perception: Mechanisms and Models*. San Francisco: W. H. Freeman, 1972.

9. D. N. Levin. Time-dependent signal representations that are independent of sensor calibration. *J. Acoust. Soc. Am.*, **108**, 2575 (2000).

10. D. N. Levin. Stimulus representations that are invariant under invertible transformations of sensor data. *Proceedings of the Society of Photoelectronic Instrumentation Engineers*, **4322**, 1677-1688 (2001). This paper can be downloaded from http://www.geocities.com/dlevin2001/reprint1.html.

11. D. N. Levin. Universal communication among systems with heterogeneous "voices" and "ears". *Proceedings of the International Conference on Advances in Infrastructure for Electronic Business, Science, and Education on the Internet*, Scuola Superiore G. Reiss Romoli S.p.A., L'Aquila, Italy, August 6-12, 2001. This paper can be downloaded from http://www.ssgrr.it/en/ssgrr2001/index.htm or from http://www.geocities.com/dlevin2001/reprint2.html.

12. The sounds corresponding to each figure in Sections III-V can be played or downloaded at http://www.geocities.com/dlevin2001/preprint1.html.

13. L. Rabiner and B-H. Juang. *Fundamentals of Speech Recognition*. Englewood Cliffs, N.J.: Prentice Hall, 1993.

14. R. N. Shepard. The analysis of proximities: multidimensional scaling with an unknown distance function. I. *Psychometrika*, **27**, 125-140 and II. *Psychometrika*, **27**, 219-246 (1962).

15. J. D. Carroll and P. Arabie. Multidimensional scaling. *Annual Reviews of Psychology*, **31**, 607-649 (1980).

16. T. Cox and M. Cox. *Multidimensional Scaling*. London: Chapman & Hall, 1994.

17. S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, **290**, 2323-2326 (2000).

18. E. W. Holman. Completely nonmetric multidimensional scaling. *Journal of Mathematical Psychology*, **18**, 39-51 (1978).

19. J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, **290**, 2319-2323 (2000).

20. D. N. Levin. Patents pending.

21. A. M. Liberman, F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy. Perception of the speech code. *Psych. Rev.*, **74**, 431-461 (1967).

22. N. Tishby. A dynamical systems approach to speech processing. *Proceedings, 1990 International Conference on Acoustics, Speech, and Signal Processing*, pp. 365-368 (1990).

23. B. Townshend. Nonlinear prediction of speech signals. In: Casdagli, M. and Eubank, S. (Eds.), *Nonlinear Modeling and Forecasting. Proceedings of the Workshop on Nonlinear Modeling and Forecasting, Sante Fe, N. M, 1990.* New York: Addison-Wesley, 1992.

Figure Captions

Figure 1. a) The transmitter signal $x(t)$ describing the transmission of a long succession of identical pulses that are uniformly spaced in time. b) The signal representation $S(t)$ that results from applying the rescaling method in Section II either to the signal in *a* or to the distorted version of that signal in *c*. c) The receiver signal obtained by subjecting the transmitter signal in *a* to the distortion: $x'(x) = g_1 \ln(1 + g_2 x)$ where $g_1 = 0.5$ and $g_2 = 150$.

Figure 2. a) Upper panel: The digitized acoustic signal of the word "door", uttered by a male speaker of American English. A 40 ms segment in the middle of the 334 ms signal is shown, with time given in ms. The horizontal lines show signal amplitudes that have dynamically rescaled values equal to $s = \pm 50\,n$ for $n=1, 2, \ldots$. Lower panel: The spectrogram of the signal excerpted in the upper panel. b) Upper panel: the signal $S(t)$ (in units of μs) obtained by rescaling the signal in the upper part of *a*, with the parameter ΔT=10 ms. Lower panel: the spectrogram of the rescaled signal excerpted in the upper panel. c) The non-linear function $x'(x)$ that was used to transform the signal in *a* into the one in *d*. d) Upper panel: a distorted version of the signal in *a*, obtained by applying the non-linear transformation in *c* to its acoustic waveform. Lower panel: the spectrogram of the signal excerpted in the upper panel. e) Upper panel: the signal obtained by rescaling the acoustic waveform in *d* with the parameter ΔT=10 ms. Lower panel: the spectrogram of the rescaled signal excerpted in the upper panel.

Figure 3. This figure shows the effects of an abrupt change in the distortion of the signal. a) A non-linear signal distortion. b) Upper panel: the signal obtained by applying the transformation in Fig. 2c to the first half (167 ms) of the acoustic waveform excerpted in Fig. 2a and by applying the distortion in panel *a* to the second half of that waveform. Lower panel: the spectrogram of the distorted acoustic waveform excerpted in the upper panel. c) Upper panel: the signal obtained by rescaling the acoustic waveform in panel *b*, using the parameter ΔT=10 ms. Lower panel: the spectrogram of the rescaled signal excerpted in the upper panel.

Figure 4. The effect of noise on the rescaling process. a) Upper panel: the signal derived from the signal in Fig. 2d by adding white noise so that the signal-to-noise ratio is 15 dB. Lower panel: the spectrogram of the entire time course of the signal excerpted in the upper panel. b) Upper panel: the signal obtained by rescaling the signal in panel *a* with ΔT=10 ms. Lower panel: the spectrogram of the entire time course of the signal excerpted in the upper panel.

Figure 5. a) The spectrogram (upper) and acoustic waveform (lower) of a warbler's song. b) The positions of the spectra in *a* in the space of the three spectral DCT coefficients with indices *i*=1, 4, 7. The solid lines depict a piecewise linear curve that is close to the nearly one-dimensional collection of points. c) Left panel: the sensor state of a detector that measured the position of each sound spectrum along a curve in the space defined by the nine spectral DCT coefficients with indices *i*=1, 4, 7, 10, 12, 13, 15, 16, 18. Right panel: the rescaled representation of the sensor data on the left.

Figure 6. The effect of filtering out high frequency components of the sound in Fig. 5a. a) The spectrogram (upper) and acoustic waveform (lower) of the warbler's song in Fig. 5a, after it was passed through a Hamming filter (1.5 ± 5.6 kHz). b) The positions of the spectra in *a* in the space of the spectral DCT coefficients with indices *i*=1, 4, 7. c) Left panel: the sensor state of a detector that measured the position of each sound spectrum along a curve in the space defined by the nine spectral DCT coefficients with indices *i*=1, 4, 7, 10, 12, 13, 15, 16, 18. Right: the rescaled representation of the sensor data on the left.

Figure 7. The effect of filtering out low frequency components of the sound in Fig. 5a. a) The spectrogram (upper) and acoustic waveform (lower) of the warbler's song in Fig. 5a, after it was passed through a Hamming filter (6.0 ± 2.8 kHz). b) The positions of the spectra in *a* in the space of the spectral DCT coefficients with indices *i*=1, 4, 7. c) Left panel: the sensor state of a detector that measured the position of each sound spectrum along a curve in the space defined by the nine spectral DCT coefficients with indices *i*=1, 4, 7, 10, 12, 13, 15, 16, 18. Right: the rescaled representation of the sensor data on the left.

Figure 8. The effect of using a different simulated detector to sense the sound in Fig. 5a. a) The positions of the spectra in Fig. 5a in the space of the spectral DCT coefficients with indices *i*=5, 8, 11. b) Left panel: the sensor state of a detector that measured the position of each sound spectrum along a curve in the space defined by the eight spectral DCT coefficients with indices *i*= 5, 6, 8, 9, 11, 14, 17, 19. Right: the rescaled representation of the sensor data on the left.

Figure 9. Simulated vocal tract #1. a) The time course of the parameter *g* that described how each simulated vocal apparatus was configured as it generated a sound. Time is in seconds. b) The spectrogram of the sound produced by the first vocal tract while it was controlled by the "articulatory gesture" shown in *a*. Time is in ms. c) The curve swept out by the second, third, and sixth DCT coefficients of the spectra produced by the first simulated vocal tract when it passed through all of its possible configurations (i.e., when the parameter *g* passed through all of its possible values). d) Left panel: the sensor state of a detector that measured the position of each sound spectrum in *b* along the curve in *c*. Right: the rescaled representation of the sensor data on the left.

Figure 10. Simulated vocal tract #2. a) The spectrogram of the sound produced by the second simulated vocal tract while it was controlled by the "articulatory gesture" shown in Fig. 9a. Time is in ms. b) The curve swept out by the second, third, and sixth DCT coefficients of the spectra produced by the second simulated vocal tract when it passed through all of its possible configurations (i.e., when the parameter *g* passed through all of its possible values). c) Left panel: the sensor state of a detector that measured the position of each sound spectrum in *a* along the curve in *b*. Right: the rescaled representation of the sensor data on the left.

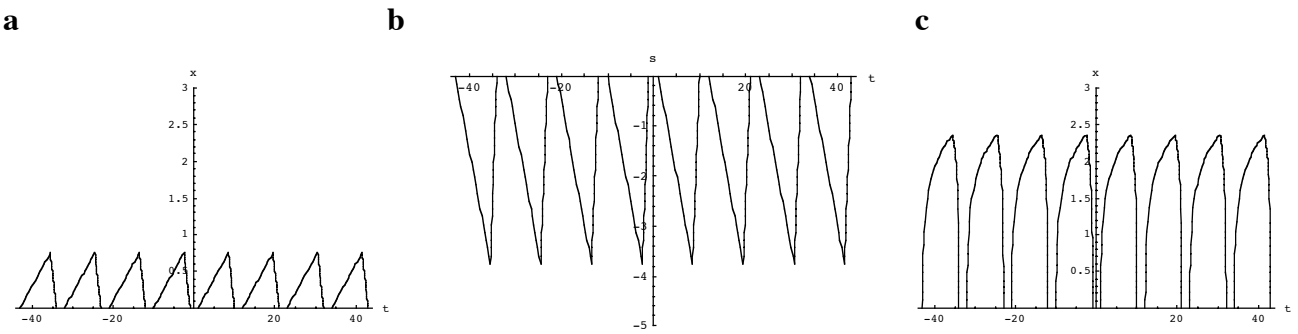


Figure 1. a) The transmitter signal $x(t)$ describing the transmission of a long succession of identical pulses that are uniformly spaced in time. b) The signal representation $S(t)$ that results from applying the rescaling method in Section II either to the signal in *a* or to the distorted version of that signal in *c*. c) The receiver signal obtained by subjecting the transmitter signal in *a* to the distortion: $x'(x) = g_1 \ln(1 + g_2 x)$ where $g_1 = 0.5$ and $g_2 = 150$.

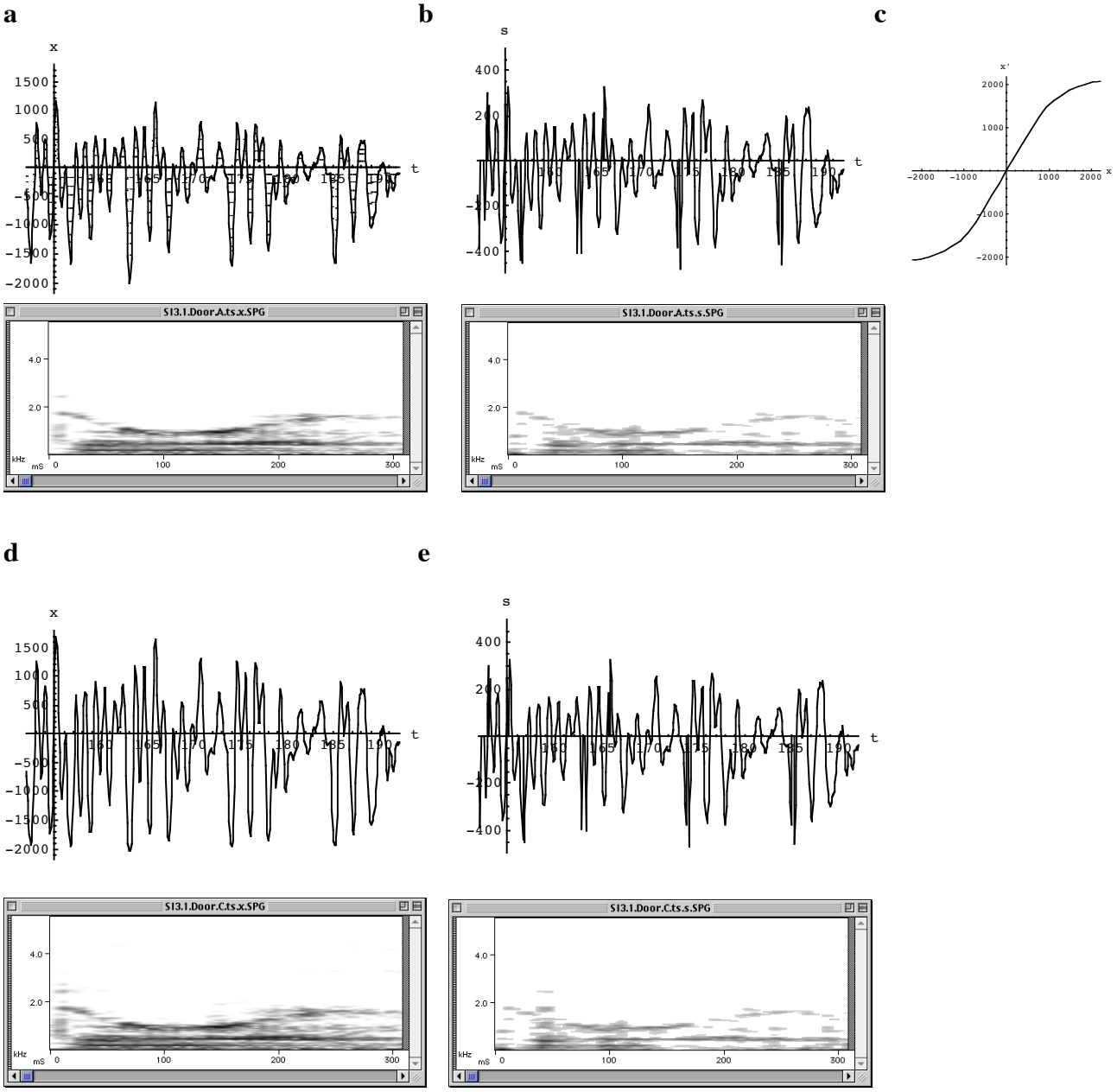


Figure 2. a) Upper panel: The digitized acoustic signal of the word "door", uttered by a male speaker of American English. A 40 ms segment in the middle of the 334 ms signal is shown, with time given in ms. The horizontal lines show signal amplitudes that have dynamically rescaled values equal to $s = \pm 50\,n$ for $n=1, 2, \ldots$. Lower panel: The spectrogram of the signal excerpted in the upper panel. b) Upper panel: the signal $S(t)$ (in units of μs) obtained by rescaling the signal in the upper part of *a*, with the parameter ΔT=10 ms. Lower panel: the spectrogram of the rescaled signal excerpted in the upper panel. c) The non-linear function $x'(x)$ that was used to transform the signal in *a* into the one in *d*. d) Upper panel: a distorted version of the signal in *a*, obtained by applying the non-linear transformation in *c* to its acoustic waveform. Lower panel: the spectrogram of the signal excerpted in the upper panel. e) Upper panel: the signal obtained by rescaling the acoustic waveform in *d* with the parameter ΔT=10 ms. Lower panel: the spectrogram of the rescaled signal excerpted in the upper panel.

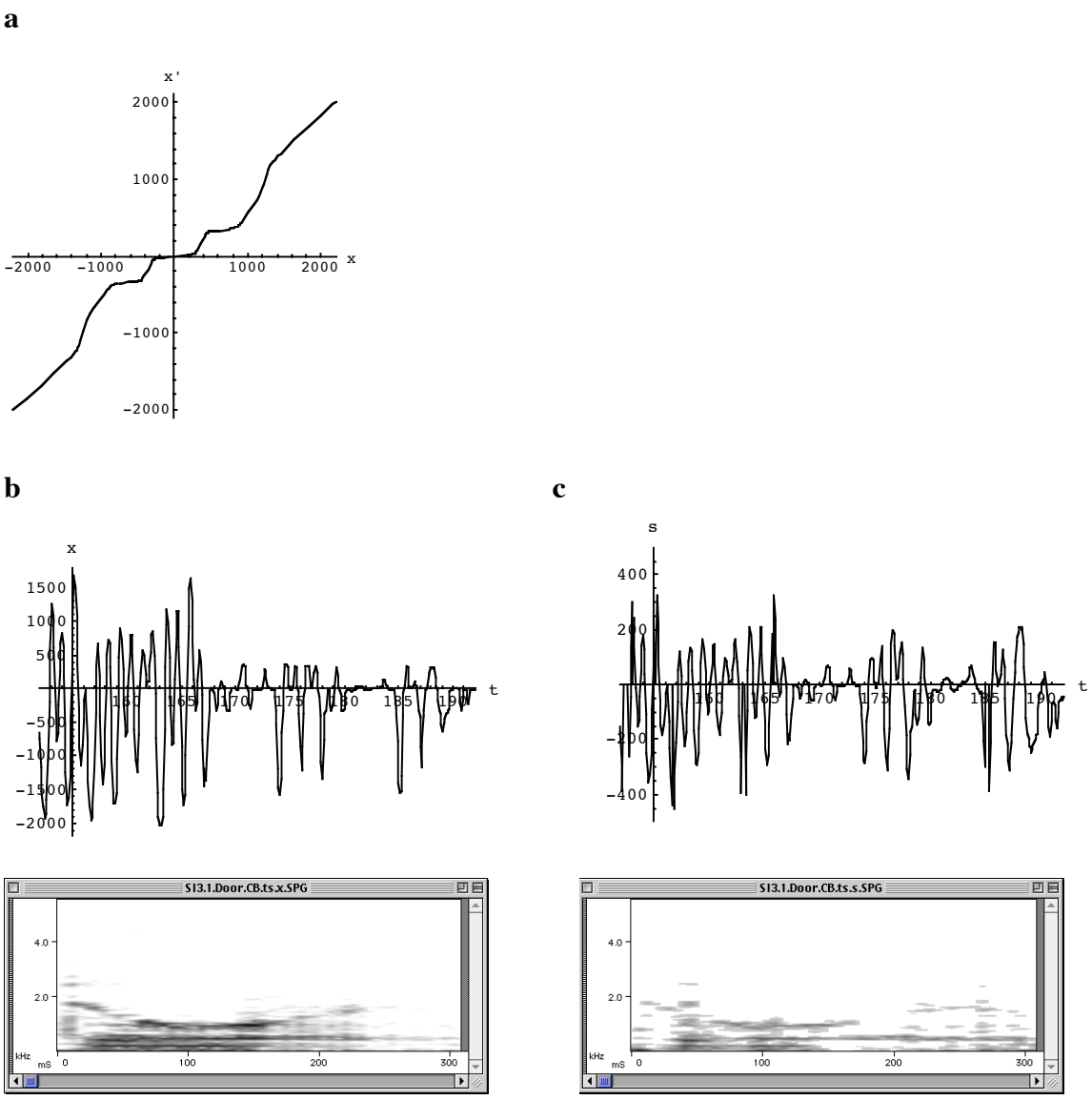


Figure 3. This figure shows the effects of an abrupt change in the distortion of the signal. a) A non-linear signal distortion. b) Upper panel: the signal obtained by applying the transformation in Fig. 2c to the first half (167 ms) of the acoustic waveform excerpted in Fig. 2a and by applying the distortion in panel *a* to the second half of that waveform. Lower panel: the spectrogram of the distorted acoustic waveform excerpted in the upper panel. c) Upper panel: the signal obtained by rescaling the acoustic waveform in panel *b*, using the parameter ΔT=10 ms. Lower panel: the spectrogram of the rescaled signal excerpted in the upper panel.

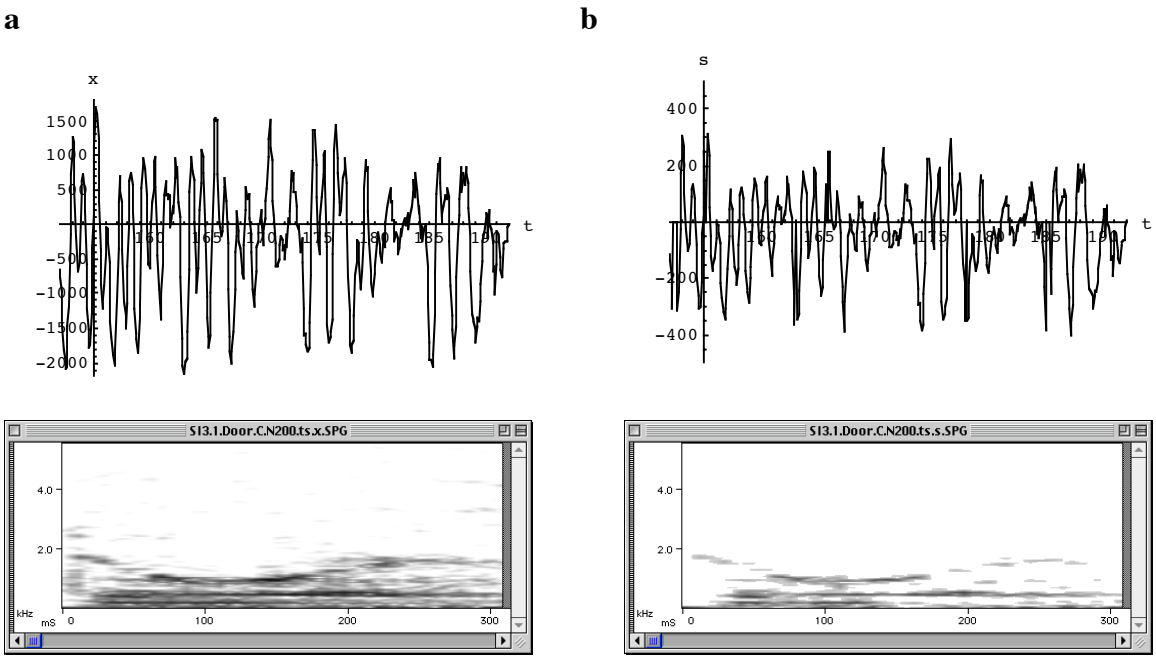
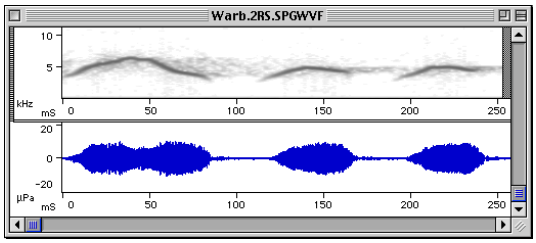
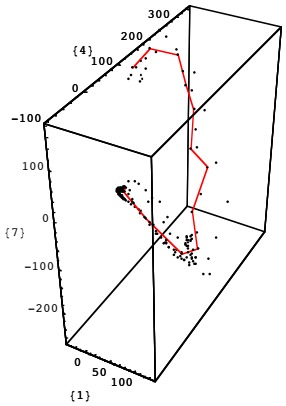


Figure 4. The effect of noise on the rescaling process. a) Upper panel: the signal derived from the signal in Fig. 2d by adding white noise so that the signal-to-noise ratio is 15 dB. Lower panel: the spectrogram of the entire time course of the signal excerpted in the upper panel. b) Upper panel: the signal obtained by rescaling the signal in panel *a* with ΔT=10 ms. Lower panel: the spectrogram of the entire time course of the signal excerpted in the upper panel.

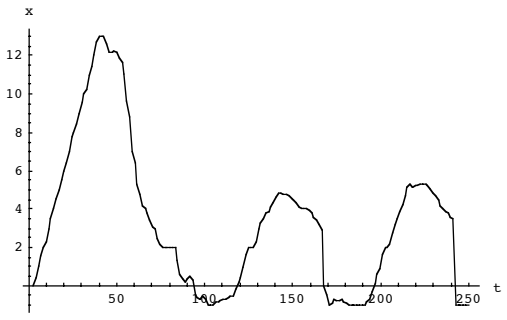
a



b



c



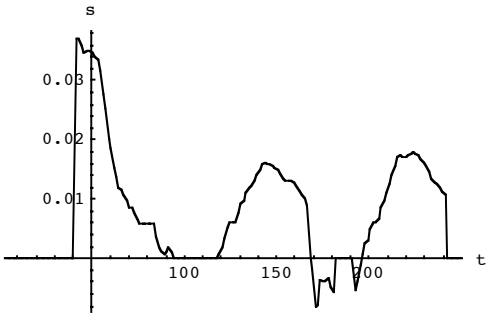
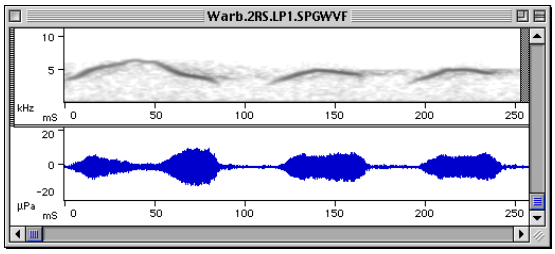
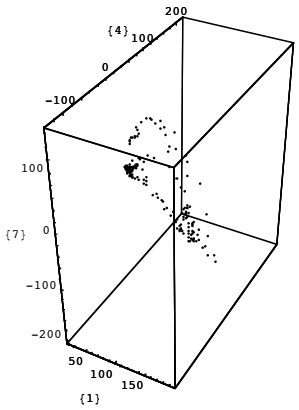


Figure 5. a) The spectrogram (upper) and acoustic waveform (lower) of a warbler's song. b) The positions of the spectra in *a* in the space of the three spectral DCT coefficients with indices *i*=1, 4, 7. The solid lines depict a piecewise linear curve that is close to the nearly one-dimensional collection of points. c) Left panel: the sensor state of a detector that measured the position of each sound spectrum along a curve in the space defined by the nine spectral DCT coefficients with indices *i*=1, 4, 7, 10, 12, 13, 15, 16, 18. Right panel: the rescaled representation of the sensor data on the left.

a



b



c

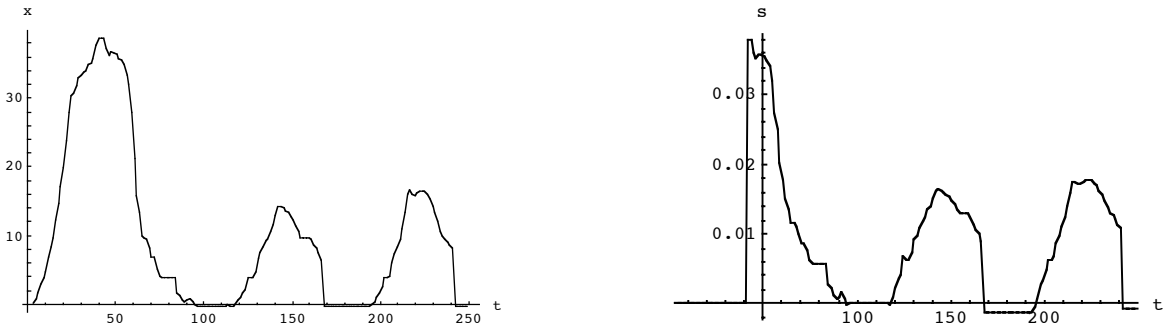
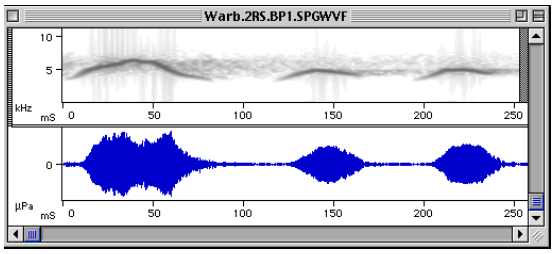
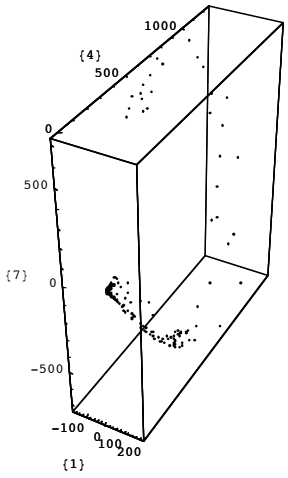


Figure 6. The effect of filtering out high frequency components of the sound in Fig. 5a. a) The spectrogram (upper) and acoustic waveform (lower) of the warbler's song in Fig. 5a, after it was passed through a Hamming filter (1.5 ± 5.6 kHz). b) The positions of the spectra in *a* in the space of the spectral DCT coefficients with indices i=1, 4, 7. c) Left panel: the sensor state of a detector that measured the position of each sound spectrum along a curve in the space defined by the nine spectral DCT coefficients with indices i=1, 4, 7, 10, 12, 13, 15, 16, 18. Right: the rescaled representation of the sensor data on the left.

a



b



c

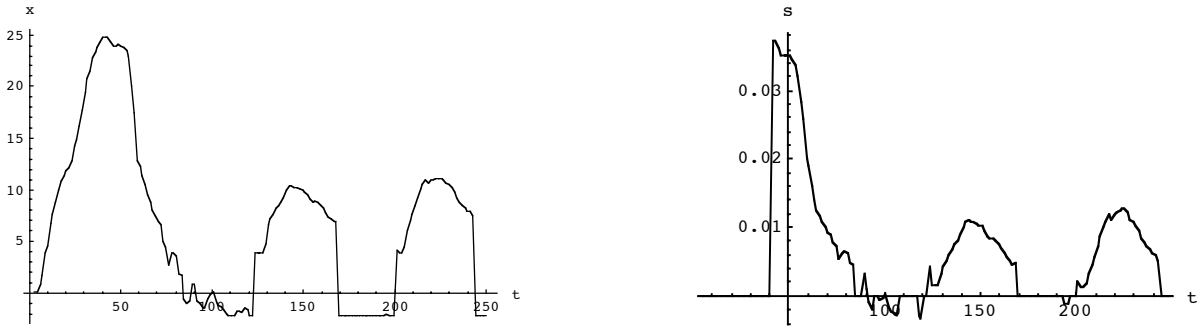
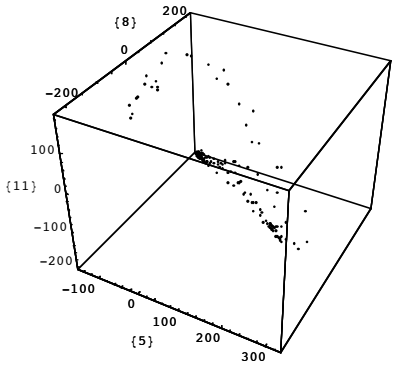
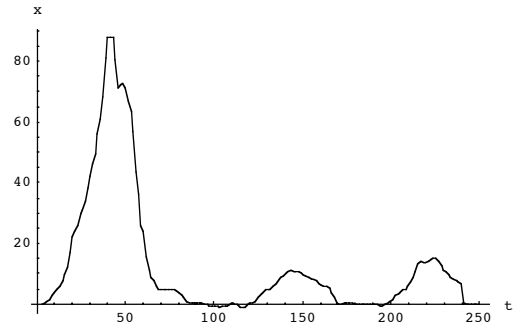


Figure 7. The effect of filtering out low frequency components of the sound in Fig. 5a. a) The spectrogram (upper) and acoustic waveform (lower) of the warbler's song in Fig. 5a, after it was passed through a Hamming filter (6.0 ± 2.8 kHz). b) The positions of the spectra in *a* in the space of the spectral DCT coefficients with indices *i*=1, 4, 7. c) Left panel: the sensor state of a detector that measured the position of each sound spectrum along a curve in the space defined by the nine spectral DCT coefficients with indices *i*=1, 4, 7, 10, 12, 13, 15, 16, 18. Right: the rescaled representation of the sensor data on the left.

a



b



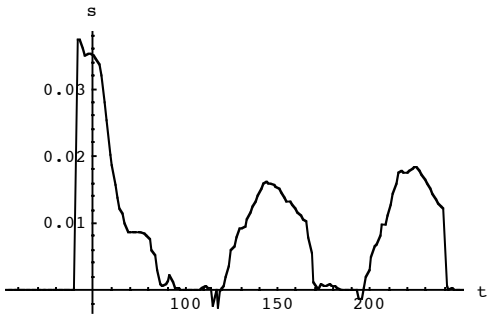


Figure 8. The effect of using a different simulated detector to sense the sound in Fig. 5a. a) The positions of the spectra in Fig. 5a in the space of the spectral DCT coefficients with indices *i*=5, 8, 11. b) Left panel: the sensor state of a detector that measured the position of each sound spectrum along a curve in the space defined by the eight spectral DCT coefficients with indices *i*= 5, 6, 8, 9, 11, 14, 17, 19. Right: the rescaled representation of the sensor data on the left.

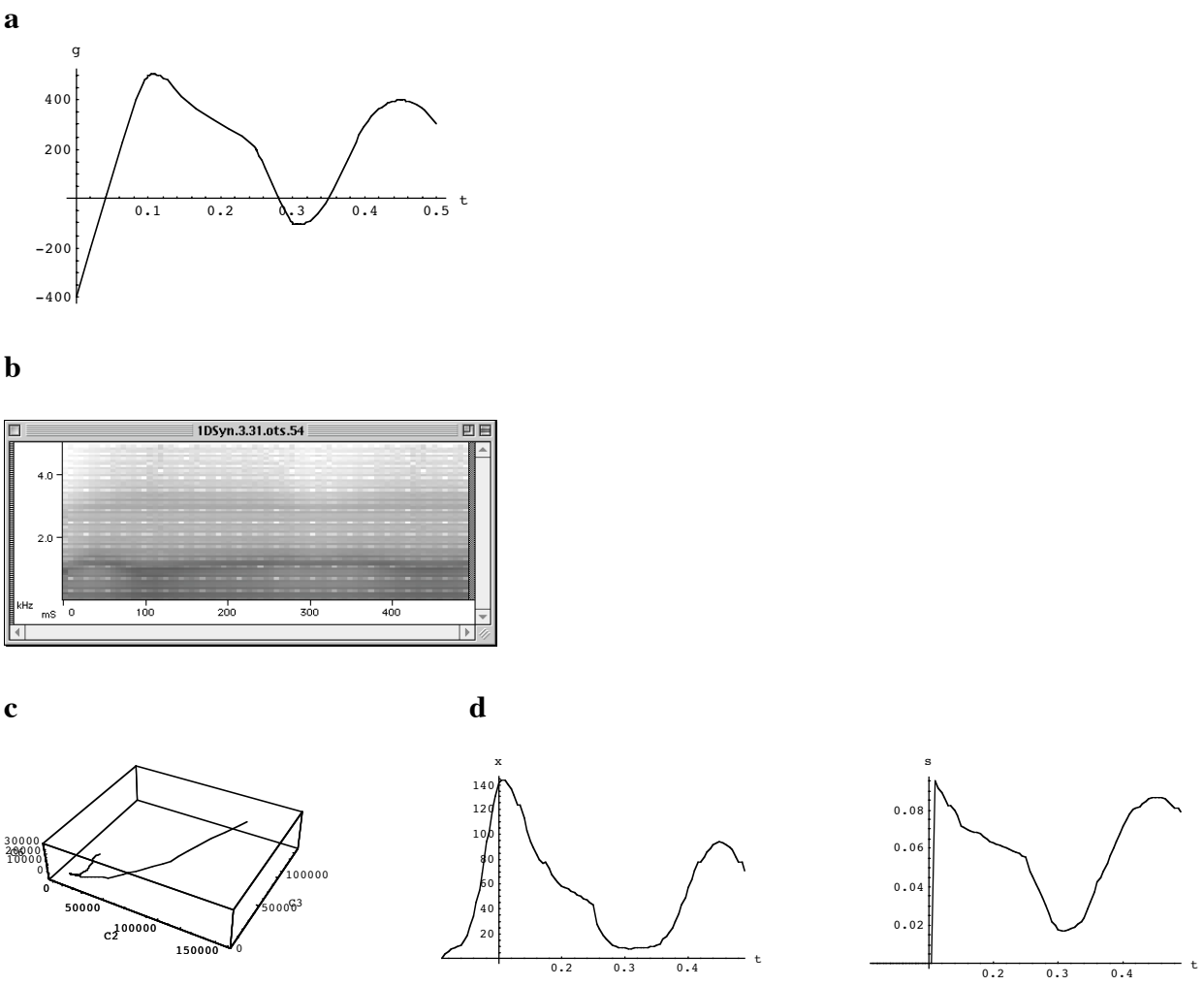
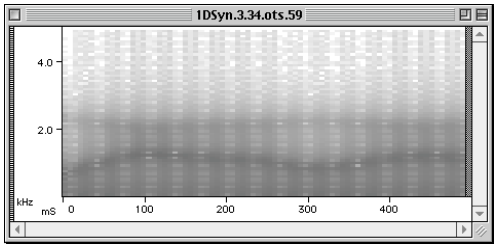


Figure 9. Simulated vocal tract #1. a) The time course of the parameter *g* that described how each simulated vocal apparatus was configured as it generated a sound. Time is in seconds. b) The spectrogram of the sound produced by the first vocal tract while it was controlled by the "articulatory gesture" shown in *a*. Time is in ms. c) The curve swept out by the second, third, and sixth DCT coefficients of the spectra produced by the first simulated vocal tract when it passed through all of its possible configurations (i.e., when the parameter *g* passed through all of its possible values). d) Left panel: the sensor state of a detector that measured the position of each sound spectrum in *b* along the curve in *c*. Right: the rescaled representation of the sensor data on the left.

a



b **c**

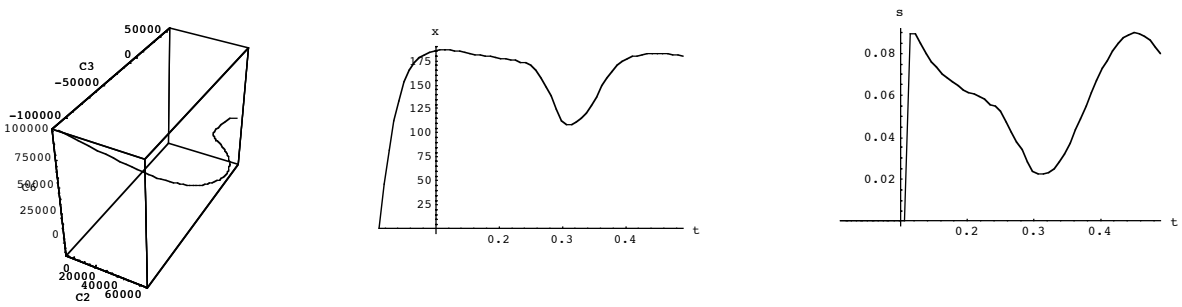


Figure 10. Simulated vocal tract #2. a) The spectrogram of the sound produced by the second simulated vocal tract while it was controlled by the "articulatory gesture" shown in Fig. 9a. Time is in ms. b) The curve swept out by the second, third, and sixth DCT coefficients of the spectra produced by the second simulated vocal tract when it passed through all of its possible configurations (i.e., when the parameter *g* passed through all of its possible values). c) Left panel: the sensor state of a detector that measured the position of each sound spectrum in *a* along the curve in *b*. Right: the rescaled representation of the sensor data on the left.