Universal Communication Among Systems With Heterogeneous "Voices" and "Ears"

David N. Levin

***Abstract*--This paper describes representations of time-dependent signals that are invariant under any invertible signal distortion. Such a representation is 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. This technique makes it possible to establish universal communication among systems with a wide variety of "voices" and "ears". As in speech communication, the systems do not have to characterize the form of the signal distortion, which remains unknown. The technique's mathematical properties are illustrated by analytical examples and by applying it to acoustical waveforms of human speech. The method is also applied to the time-dependent spectra of synthetic speech-like signals. The results suggest that the new technique can create speech representations that are invariant across a wide range of speakers' voices and listeners' ears.**

***Index Terms*--sensor, channel, distortion, calibration, communication, speech recognition, pattern recognition, computer vision**

Paper presented at 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'Aquilla, Italy, August 6-12, 2001.

D. N. Levin's address: Dept. of Radiology, MC 2026, U. of Chicago, 5841 S. Maryland Ave., Chicago, IL, 60637, USA. Email: d-levin@uchicago.edu. Tel.: 773-702-6511. Web: http://www-radiology.uchicago.edu/faculty/Levin.html.

1. 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. Most communications 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. 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. 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 [1] in which subjects wore goggles creating severe geometric distortions of the visual field (e.g., 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 [2,3], 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 recently encountered signal levels. This dynamically 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 identical 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 [2, 3]. 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 dynamic rescaling technique can be used to faithfully communicate information in the presence of any invertible signal distortion, without resorting to traditional calibration procedures. 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 2, the method is derived, and then it is illustrated with analytic examples in order to demonstrate its mathematical characteristics. The technique is applied to the acoustic waveforms of human speech in Section 3. In Section 4, the method is applied to the parameterized spectra of synthetic speech-like signals in order to show its potential for creating speech representations that are invariant across a wide range of speakers' voices and listeners' ears. In Section 5, the implications of this work are discussed, particularly its application to speech recognition systems.

2. 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 $x \to x' = x'(x)$. 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 $x \to x'$ 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 $(y \to y')$ 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 $x \to x'$. 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 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, and, therefore, distortion-invariant information cannot be communicated then. This does not compromise the fidelity of communication, although its time efficiency is reduced. The inability to compute signal invariants at some time points means that the number of independent signal invariants (i.e., the number of time points at which $S(t)$ can be computed) may be less than the number of degrees of freedom in the raw signal from which the invariants were computed (i.e., the number of time points at which the signal $x(t)$ is transmitted). The particle analogy in Section 1 suggests that this is not surprising. Note that there are a number of linear relationships among the coordinates of the particles when they are expressed in the collection's intrinsic "center-of-mass" coordinate system. For example, their sum vanishes. Therefore, the number of independent invariants (i.e., the number of independent particle positions in the intrinsic coordinate system) is less than the number of degrees of freedom of the particle collection (i.e., the number of particle locations in an extrinsic coordinate system). This is because some of the collection's degrees of freedom were used to define the intrinsic coordinate system itself.

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 its distorted signal $x'(t)$, is the same as the rescaled signal $S(t)$, which the transmitter derives from its 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 two Sections.

3. EXPERIMENTS WITH ACOUSTIC WAVEFORMS OF HUMAN SPEECH

In this Section, the mathematical properties of dynamic rescaling are further illustrated by applying it to acoustic waveforms of human speech. An adult male American uttered English words with speed and loudness that were characteristic of normal conversation. These sounds were digitized with 16 bits of depth at a sample rate of 11.025 kHz. Figure 2a shows a 40 ms segment of digitized signal (*x(t)*), located at the midpoint of the 334 ms signal corresponding to the word "door". Figure 2b shows the "*s* representation" (i.e., the dynamically rescaled signal *S(t)*) that was derived from Fig. 2a by the method of Section 2. 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 $s = 50n$ 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. Figure 2e is the dynamically 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 2. It is interesting to note that this result is apparent when one listens to the sounds represented in Fig. 2. 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".

The above example suggests how dynamic rescaling might be used to enable universal communication among systems with a variety of "voices" and "ears". To see this, imagine that Figs. 2a and 2d are the signals in the detector circuits of two receivers, which are "listening" to the same transmission. The non-linear transformation that relates these raw signals (Fig. 2c) could be due to differences in the receivers' detector circuits (e.g., their gain curves), or it could be due to differences in the channels between the receivers and the transmitter, or it could be due to a combination of these mechanisms. As long as both receivers use dynamic rescaling to "decode" the detected signals, they will derive the same information content (i.e., the same function *S(t)*) from them. If one of the receivers is part of the system that originated the transmission (i.e., if this system is "listening" to its own transmission), then the information in the signal's *s* representation will be faithfully communicated to the other receiver, despite the fact that it has different "ears" than the transmitting system. Alternatively, imagine that Figs. 2a and 2d are the signals in a single receiver, when it detects the signals from two different transmitters. In this case, the non-linear transformation that relates these signals could be due to differences in the "voices" (i.e., the transmission characteristics) of the two transmitters. As long as the receiver "decodes" the detected signals by dynamically rescaling, it will derive the same information content (i.e., the same *S(t)*) from them. In other words, it will "perceive" the two transmitters to be broadcasting the same message in two different "voices". As mentioned above, the transmitters will derive the same information content as the receivers if they "listen" to their own transmissions and then dynamically rescale them. In this way, systems with heterogeneous "voices" and "ears" might be able to communicate accurately without using calibration procedures to characterize their transmission and reception characteristics. In Section 4, this proposition is demonstrated in the context of a

much more realistic model of differing voices and ears.

Some comments should be made about technical aspects of the example in Fig. 2. The dynamically rescaled signals in Figs 2b and 2e were computed by a minor variant of the method in Section 2. 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 2, this occurs at identical time points when dynamic 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 be computed at 92% of all time points.

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 ΔT = 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 2), 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 a change in the transmitter's "voice" or the receiver's "ears" or the channel between them. This adaptive behavior resembles that of the human subjects of the goggle experiments mentioned in Section 1.

Figure 4 illustrates the effect of noise on dynamic rescaling. Figure 4a was derived from Fig. 2d by adding white noise, chosen from a uniform distribution of amplitudes between –200 and +200. This causes a pronounced hiss to be superposed on the word "door" when the entire 334 ms sound exemplified by Fig. 4a is played. 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 2. 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.

4. EXPERIMENTS WITH SPECTRA OF SYNTHETIC SPEECH-LIKE SOUNDS

In the previous Section, dynamic rescaling was demonstrated by applying it to human speech waveforms that were related to one another by invertible transformations in the time domain. However, these transformations incorporated the effects of a relatively small range of speakers' "voices" and listeners' "ears". For example, signals related by such transformations did not mimic voices with a significant range of pitches. The signals produced by a much wider range of "voices" and detected by a much wider variety of "ears" can be created from a sound's short-term Fourier spectra, by distorting the spectral vectors with multidimensional non-linear transformations. In this Section, we demonstrate that time-dependent speech spectra, which are related by such invertible transformations, have identical dynamically rescaled representations. In other words, such a representation of a given utterance will be independent of the speakers and listeners, who produce and detect it. For computational simplicity, we consider synthetic speech-like signals that are generated by a "glottis" and "vocal tract" controlled by a single degree of freedom. These signals mimic the "1D speech" produced by multiple muscles whose motion is determined by the value of a single time-dependent parameter. The same approach can be applied to human speech signals, which are produced by a vocal apparatus with multiple degrees of freedom, by utilizing the multidimensional generalization [3] of the technique in Section 2.

The "1D speech" signals were generated by a standard linear prediction (LP) model [4]. 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 speech 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 glottal excitation. Each of these quantities was a function of a single parameter (g), which itself depended on time. These eight functions described the nature of the speaker's "voice", in the sense that they defined the manifold of all spectra that the speaker 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" of the speaker, in the sense that it determined how the speaker's vocal apparatus was configured at each time. In a musical analogy, the g-dependent functions of the LP model would describe the possible states of 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 speakers who produced "voiced" speech sounds that were driven by regular glottal impulses. However, it is straightforward to apply the same methods to "unvoiced" speech sounds that are driven by noise-like glottal excitation functions. The pitch of the first speaker's voice was taken to be constant and equal to 200 Hz.

The first listener's "ears" were described by his/her method of detecting and processing the time domain speech signals. The above-described signals were digitized at 10 kHz, and then short-term Fourier spectra were produced from the signals in a 10 ms Hamming window that was advanced in increments of 5 ms. Figure 5b shows the spectrogram that resulted from the signal generated by the first speaker's "voice", when it went through

the series of configurations described by the "gesture" function in Fig. 5a. The spectrum at each time point was parameterized by cepstral coefficients, which were generated by the discrete cosine transformation (DCT) of the log of the spectral magnitude, after the spectral magnitude had been averaged in equally spaced 600 Hz bins. The listener described in this paragraph (listener #1) was assumed to detect only the third, fourth, and fifth cepstral coefficients of each spectrum. The cepstral coefficients from each short-term spectrum defined a single point in this three-dimensional space. Each of these points fell on a 1D curve defined by the cepstral coefficients corresponding to all possible configurations of the speaker's vocal apparatus (i.e., all possible values of *g*). The precise shape of this curve depended on the nature of the speaker's voice (specified by the *g*-dependence of the speech model's poles and other parameters). Figure 5c shows the configuration of this curve for the voice of the speaker described in the previous paragraph. A convenient coordinate system (denoted by *x*) was established on this curve by projecting each of its points onto a connected array of chords that hugged the curve. The raw sensory signal for a specific utterance consisted of the temporal sequence of coordinates *x(t)* that were generated as the cepstrum traversed that curve. Because the spectrogram in Fig. 5b was generated by an oscillatory *g(t)*, previously generated spectra (and cepstra) were revisited from time to time, and the corresponding cepstral coefficients moved back and forth along the curve in Fig. 5c. The left side of Fig. 5d shows the oscillatory sensory signal *x(t)* that was generated in this way.

The ears of a second listener were modeled in the following manner. The second listener was assumed to compute the short-term Fourier spectra of the time domain signal, as described above. However, instead of calculating the cepstrum of each spectrum, the second listener was assumed to compute the DCT of its magnitude (*not* its log magnitude), after it (the spectral magnitude) had been averaged in equally spaced 600 Hz bins. This listener detected only the second, third, and sixth of these DCT coefficients. The voice of the above-described speaker was characterized by the 1D curve (Fig. 6a) swept out by the spectral DCT coefficients for all possible sounds that could be generated by the vocal apparatus (i.e., all possible values of *g*). As before, a convenient coordinate system (*x'*) was established on this curve by projecting each of its points onto a connected array of chords that hugged the curve. The left side of Fig. 6b is the raw sensory signal $x'(t)$ that was induced in listener #2 by the sound in Fig. 5b. Note that the *x* and *x'* coordinate systems could bear any relationships to the curves in Figs 5c and 6a, respectively, and need not have any definite or known relationship to one another, except that $x = 0$ and $x' = 0$ must correspond to the same sound (e.g., the same value of *g*). In this example, this condition was satisfied by defining the *x* and *x'* coordinate systems so that $x = x' = 0$ corresponded to the first short-term spectrum in the utterance in Fig. 5b. Alternatively, this could be arranged by having both listeners hear any single sound produced by speaker #1 and agree to originate their coordinate systems at the corresponding point on the speaker's "voice" curve; this 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. Finally, the raw sensory signal in each listener, $x(t)$ and $x'(t)$, was processed by dynamic rescaling with $\Delta T = 500\ ms$. The results are shown on the right sides of Figs. 5d and 6b, respectively. Notice the similarity between these *s* representations despite the differences between the raw sensory signals, $x(t)$ and $x'(t)$, from which they were created. This means that the two listeners created the same dynamically rescaled representation of the utterance, despite the dramatic differences in their "ear" mechanisms (Figs. 5c and 6a). The dynamically rescaled representations were the same because the raw sensory signals, $x(t)$ and $x'(t)$, were related to one another by an invertible transformation that preserved the null amplitude. This was true because each listener was sensitive to the spectral changes produced by all changes in *g*, and, therefore, each raw sensory signal was

invertibly related to *g(t)*. Furthermore, for the same reason, *any other* gesture function $\tilde{g}(t)$ that is invertibly related to the function in Fig. 5a will generate an utterance with the dynamically rescaled representation in the right panel of Fig. 5d. In other words, the utterances that are produced by these "different" gesture functions will be internally represented as the same message uttered in two *different tones of voice*. Finally, notice that the dynamically rescaled representation of $g_1(t) \equiv g(t) - g(0)$ is identical to the dynamically rescaled representations of $x(t)$ and $x'(t)$. This is expected because $g_1(t)$ is invertibly related to each of these sensory signals in a way that transforms $g_1 = 0$ into $x = x' = 0$. This means that the speaker creates identical internal representations of both the transmitted signal and the "motor" signal that controls the configuration of the vocal apparatus.

The voice of a second speaker was modeled by choosing different *g*-dependent functions for the 8 quantities in the LP model of the vocal apparatus. Specifically, the glottal pitch was set equal to 125 Hz, and the poles of the vocal tract transfer function were chosen to be significantly different functions of *g* than for the first voice. Figure 7a shows the spectrogram produced by this second "voice" when it made the "articulatory gesture" in Fig. 5a. Figure 7b is the curve in DCT coefficient space induced in listener #2 by this voice, when it produced all possible spectra (i.e., spectra corresponding to all possible values of *g*). Figures 6a and 7b show that the listener #2 characterized the first and second voices by dramatically different curves in DCT coefficient space. The left side of Fig. 7c depicts the raw sensory signal $x'(t)$ induced in listener #2 by the utterance in Fig. 7a. As before, the origin of the *x'* coordinate system along the curve in Fig. 7b was chosen to correspond to the first sound spectrum emitted by the speaker. Finally, the right side of Fig. 7c is the dynamically rescaled representation of this raw sensory signal. Notice that there is no significant difference between the dynamically rescaled representations in Figs. 7c and 6b, despite the fact that they were derived from the utterances of different voices and corresponded to raw sensory signals from different spectrograms (Figs. 7a and 5b). This is because these raw sensory signals are related by an invertible transformation. Such a transformation exists because each raw sensory signal is invertibly related to the same gesture function (i.e., *g(t)* in Fig. 5a).

5. 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 produced by the transmitter, receiver, or the channel between them. This form of the signal is found by rescaling the signal at each time, in a manner that is determined by its recent time course. The dynamically 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 be used to establish universal communication among systems with heterogeneous "voices" and "ears", differing 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 need not explicitly characterize the distortion, which remains unknown; 3) if the nature of the distortion changes, faithful communication resumes after a period of adaptation.

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 dynamic 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 dynamic 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 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 dynamic rescaling process to adapt to changes in signal distortion.

The family of all signals $x(t)$ that dynamically 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 dynamic rescaling is applied N times in succession, where $N \geq 2$. Successive applications of dynamic rescaling may eventually create a function that is not changed by further applications of the procedure (i.e., the serial dynamic rescaling process may reach a fixed "point"). For example, it is easy to show that, if the dynamic scale of a signal is time-independent (i.e., if $h(y)$ and $s(x)$ are time-independent), it will dynamically rescale to such a fixed point. Such a signal is loosely analogous to music, in the sense that musical compositions are also based on a time-independent scale (e.g., 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 [5]. The signals from the system's detectors would be dynamically 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 position and/or orientation of the microphone), 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 4, dynamic rescaling was demonstrated by applying it to synthetic speech signals produced by a variety of "voices" and detected by a variety of "ears." These experiments showed that the utterance of any one speaker produced the same dynamically rescaled representations in listeners with different ears (Figs. 5 and 6). Likewise, identical dynamically rescaled representations were induced in any one listener by the utterances of two speakers, who sought to transmit the same message (Figs. 6 and 7). The listener-independence and speaker-independence of the dynamically rescaled representations is quite general, even though it was demonstrated in the context of a specific family of voice and ear models. To see this, recall from Section 2 that any two functions have the same dynamically rescaled representations as long as they are related by an invertible transformation. Assume that each listener is sensitive to the differences between any two configurations of a speaker's vocal apparatus. It follows that those configurations are invertibly related to the listener's raw sensory states when the vocal apparatus is used to create sounds. Therefore, if the speaker's utterance is heard by two different listeners with this sensitivity, their raw sensory states will be invertibly related to one another and, consequently, have identical dynamically rescaled representations. Similarly, consider a single listener

who sensitively hears the utterances of two different speakers. Assume that there is an invertible transformation between the two speakers' vocal configurations when they utter the same message. In other words, assume that their vocal configurations are invertibly related because one speaker mimics the other in a consistent fashion or because both speakers "read" from the same "text" in a consistent manner. Then, the raw sensory signals induced in the listener by the two speakers will also be invertibly related. This is because these sensory signals are invertibly related to vocal configurations, which are themselves invertibly related. It follows that the listener will construct an identical dynamically rescaled representation of each speaker's utterance. As mentioned in Section 4, because the raw sensory signals are invertibly related to the vocal apparatus configurations producing those signals, the "motor" signal $g(t)$ controlling an utterance will have the same dynamically rescaled representation as the utterance itself. This is consistent with the "motor" theory of speech perception [6].

Although the experiments in Section 4 were performed with 1D-speech signals, it is straightforward to generalize the methodology to signals produced by models with multiple degrees of freedom. For example, consider the spectra generated by a vocal apparatus with two degrees of freedom. Each spectrum will correspond to a point on a 2D subspace (i.e., a sheet-like surface) in the space of spectral parameters (e.g., cepstral coefficients), and each utterance will be characterized by a trajectory on this 2D surface. In reference 3, the author demonstrated several techniques for dynamically rescaling such signals with two (or more) degrees of freedom. It may be computationally practical to apply this technique to human speech that is generated by a vocal apparatus with a relatively small number of degrees of freedom. For the reasons cited in the previous paragraph, such a device would generate the same internal (dynamically rescaled) representation of any given utterance by a wide variety of speakers. Therefore, a speech recognition device with such a "front end" may not need extensive retraining when the speaker's voice or certain other conditions are changed. Furthermore, the adaptive nature of dynamic rescaling might enable it to account for coarticulation [4] 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 dynamically 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

REFERENCES

[1] R. Held and R. Whitman, *Perception: Mechanisms and Models*, W. H. Freeman, San Francisco, CA, 1972.

[2] D. N. Levin, "Time-dependent signal representations that are independent of sensor calibration", *J. Acoust. Soc. Am.*, Vol. 108, November 2000, p. 2575.

[3] D. N. Levin, "Stimulus representations that are invariant under invertible transformations of sensor data", *Proc. SPIE Internat. Sympos. Med. Imag.*, San Diego, CA, 17-22 February 2001.

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

[5] D. N. Levin, patents pending.

[6] A. M. Liberman, F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy, "Perception of the speech code", *Psych. Rev.*, Vol. 74, 431-461, 1967.

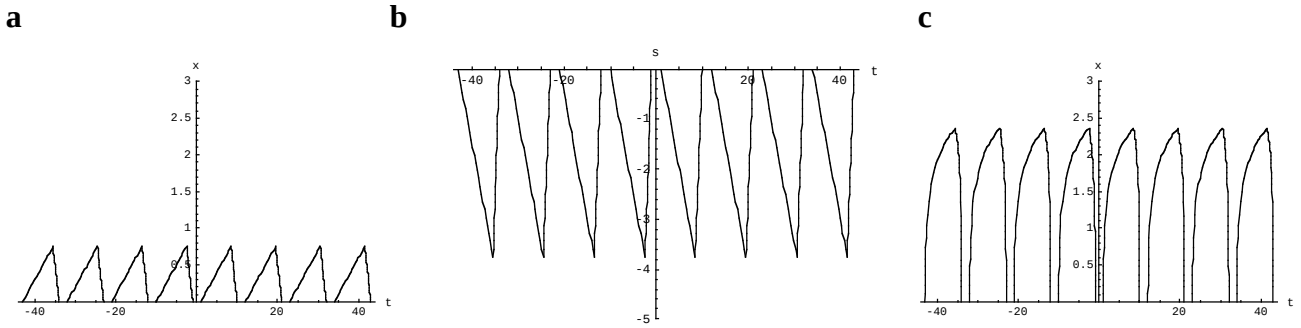


Figure 1. a) The transmitter signal $x(t)$ describing the transmission of a long succession of identical pulses that are uniformly spaced in time. b) The signal representation $S(t)$ that results from applying the dynamic rescaling method in Section 2 either to the signal in panel *a* or to the distorted version of that signal in panel *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 $g_2 = 150$.

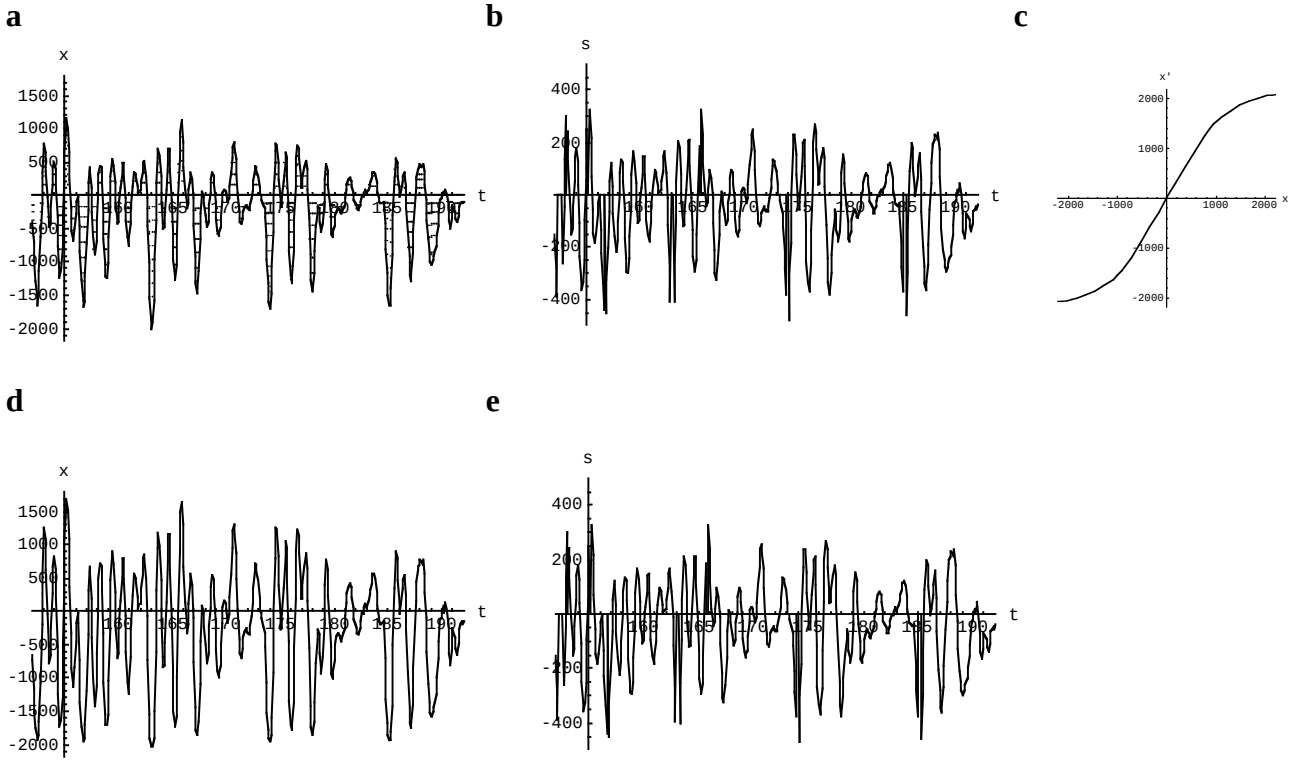


Figure 2. a) A signal obtained by digitizing the acoustic signal of the word "door", uttered by a male speaker of American English. A 40 ms segment in the middle of the 334 ms signal is shown, with time given in ms. The horizontal lines show signal amplitudes that have dynamically rescaled values equal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

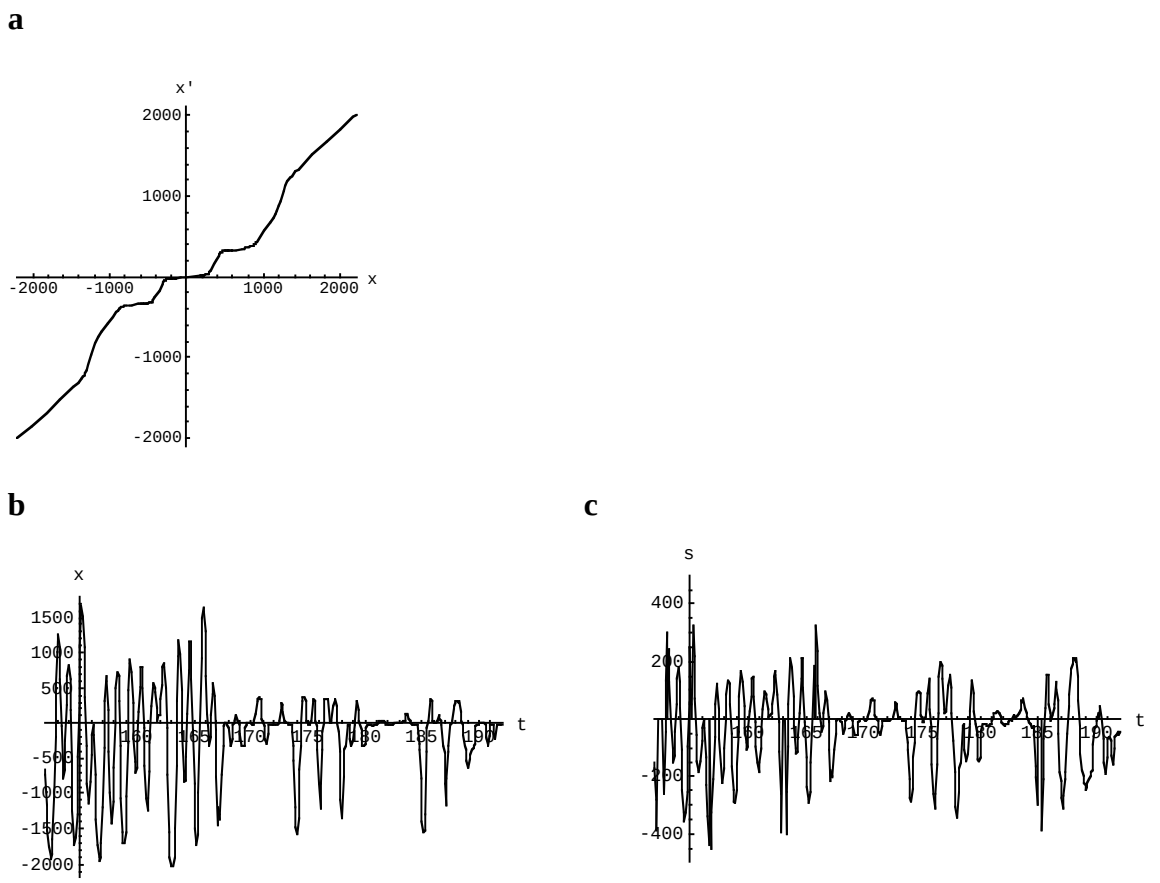


Figure 3. This figure shows the effects of an abrupt change in the distortion of the signal. a) A non-linear signal distortion. b) The signal obtained by applying the transformation in Fig. 2c to the first half (167 ms) of the signal excerpted in Fig. 2a and by applying the distortion in panel *a* to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

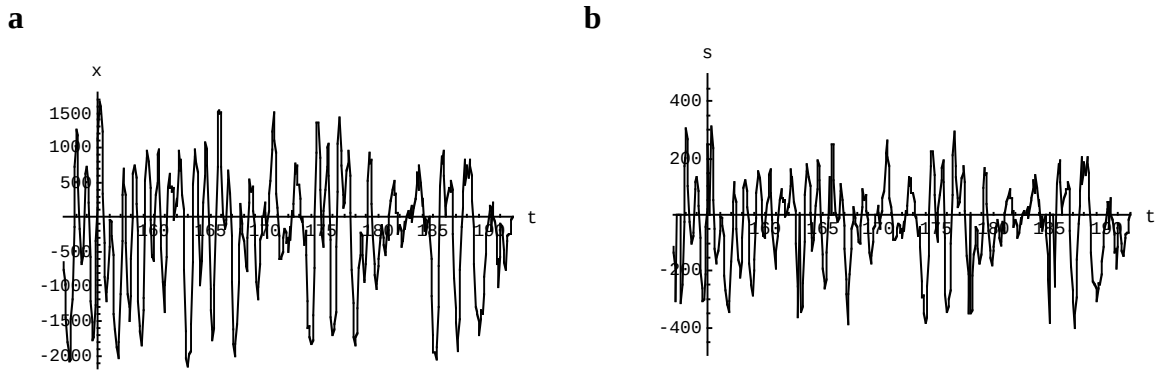
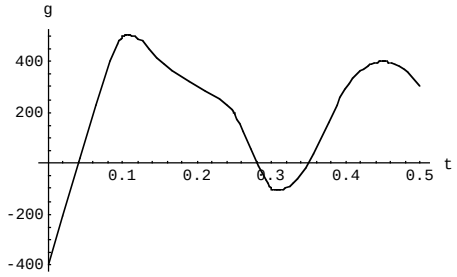
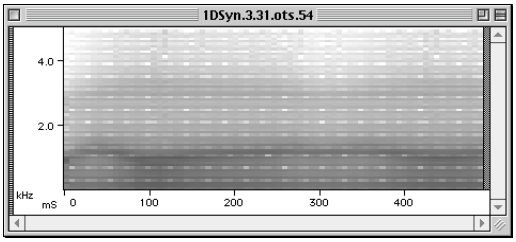


Figure 4. The effect of noise on the dynamic rescaling process. a) A signal derived from the signal in Fig. 2d by adding white noise with amplitudes randomly chosen from a uniform distribution between –200 and +200. b) The signal obtained by dynamically rescaling the signal in panel *a* with ΔT=10 ms.

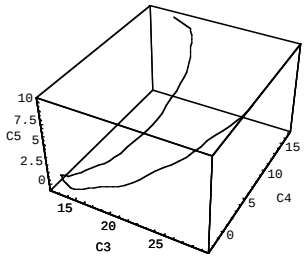
a



b



c



d

Figure 5. Speaker #1 and listener #1. a) The time course of the parameter *g*, which describes the state of speaker #1's vocal apparatus, during a particular utterance. Time is in seconds. b) The spectrogram of the sound produced by speaker #1 during the utterance described by *g(t)* in panel *a*. Time is in ms. c) The curve swept out by the third, fourth, and fifth cepstral coefficients of the spectra produced by speaker #1's 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: the raw sensory signal induced in listener #1 when speaker #1 uttered the sound produced by the sequence of vocal apparatus configurations in panel *a*. Here, *x* denotes the instantaneous position of the sound spectrum's cepstral coefficients with respect to a convenient coordinate system along the curve in panel *c*. Time is in seconds. Right: the dynamically rescaled representation of the raw sensory signal on the left.

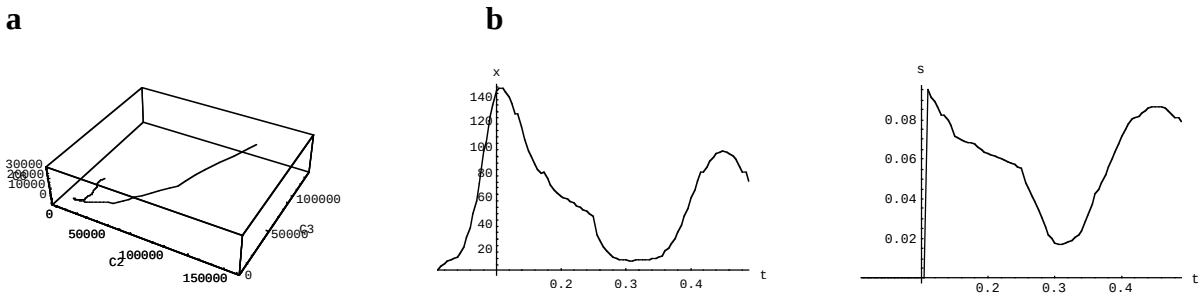


Figure 6. Speaker #1 and listener #2. a) The curve swept out by the second, third, and sixth DCT coefficients of the spectra produced by speaker #1's vocal tract, when it passed through all of its possible configurations. b) Left: the raw sensory signal induced in listener #2 when speaker #1 uttered the sound produced by the sequence of vocal apparatus configurations in Fig. 5a. Here, *x'* denotes the instantaneous position of the sound spectrum's DCT coefficients with respect to a convenient coordinate system along the curve in panel *a*. Time is in seconds. Right: the dynamically rescaled representation of the raw sensory signal on the left.

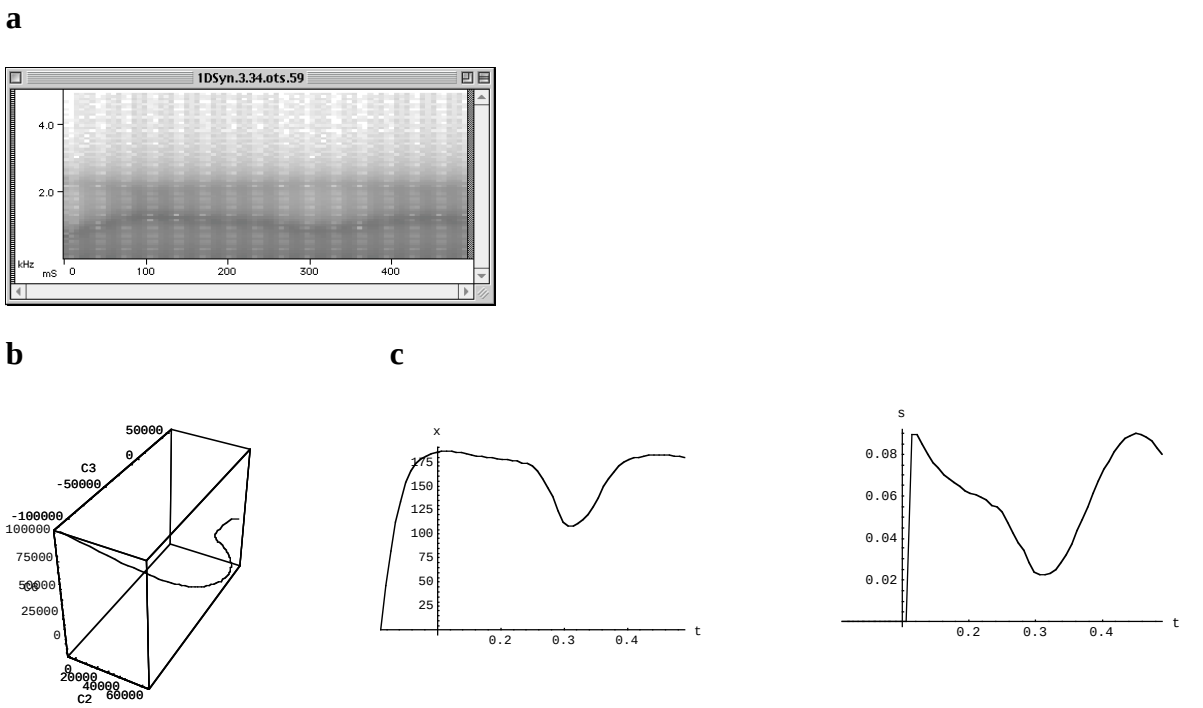


Figure 7. Speaker #2 and listener #2. a) The spectrogram produced when speaker #2 uttered the sound described by the "gesture" function *g(t)* in Fig. 5a. Time is in ms. b) The curve swept out by the second, third, and sixth DCT coefficients of the spectra produced by speaker #2's vocal tract when it passed through all of its possible configurations (i.e., when the parameter *g* passed through all of its possible values). c) Left: the raw sensory signal produced in listener #2 when speaker #2 uttered the sound produced by the sequence of vocal apparatus configurations in Fig. 5a. Here, *x'* denotes the instantaneous position of the spectrum's DCT coefficients with respect to a convenient coordinate system along the curve in panel *b*. Time is in seconds. Right: the dynamically rescaled representation of the raw sensory signal on the left.

David N. Levin received his Ph.D. in theoretical physics from Harvard University in 1970 and did research in quantum field theory until 1977, when he entered medical school at the University of Chicago. He joined the faculty after receiving an M.D. and completing radiology residency at the University. During 1987-1999, he was Director of Clinical MRI at the University. He is currently Professor in the Department of Radiology and co-directs the University's Brain Research Imaging Center. This facility is equipped with a 3 T MRI scanner that is dedicated to brain research with functional MRI and MR spectroscopy. His past research interests have included multimodality 3D brain imaging, computer-assisted neurosurgery, and image segmentation. His current research is focused on new methodology for mapping the brain with functional MRI, novel techniques for using prior knowledge to increase the speed of MR image acquisition, and applications of non-linear signal processing to the design of intelligent sensory systems.