

# Surrogate analysis for detecting nonlinear dynamics in normal vowels

Isao Tokuda<sup>a)</sup>

*Department of Computer Science and Systems Engineering, Muroran Institute of Technology, Muroran, Hokkaido 050-8585, Japan*

Takaya Miyano

*Department of Intelligent Machines and System Engineering, Hirosaki University, Hirosaki, Aomori 036-8561, Japan*

Kazuyuki Aihara

*Department of Mathematical Engineering and Information Physics, Faculty of Engineering, The University of Tokyo, Bunkyo-ku, Tokyo 113-8656, Japan and CREST, JST, Honmachi, Kawaguchi, Saitama 332-0012, Japan*

(Received 16 July 1999; revised 20 August 2001; accepted 28 August 2001)

Normal vowels are known to have irregularities in the pitch-to-pitch variation which is quite important for speech signals to be perceived as natural human sound. Such pitch-to-pitch variation of vowels is studied in the light of nonlinear dynamics. For the analysis, five normal vowels recorded from three male and two female subjects are exploited, where the vowel signals are shown to have normal levels of the pitch-to-pitch variation. First, by the false nearest-neighbor analysis, nonlinear dynamics of the vowels are shown to be well analyzed by using a relatively low-dimensional reconstructing dimension of  $4 \leq d \leq 7$ . Then, we further studied nonlinear dynamics of the vowels by spike-and-wave surrogate analysis. The results imply that there exists nonlinear dynamical correlation between one pitch-waveform pattern to another in the vowel signals. On the basis of the analysis results, applicability of the nonlinear prediction technique to vowel synthesis is discussed. © 2001 Acoustical Society of America. [DOI: 10.1121/1.1413749]

PACS numbers: 43.70.Gr, 43.25.Rq [AL]

## I. INTRODUCTION

In the studies of human speech, linear dynamical systems analysis, such as the power spectrum analysis and the linear predictive coding (LPC) model, is the most popular and standard methodology.<sup>1-5</sup> This is because acoustical characteristics of human speech are mainly due to the resonances of the vocal tract, which form the basic spectral structure of the speech signals.<sup>1</sup> In fact, linear dynamical systems analyses have been widely and successfully applied to speech analysis and synthesis. One example is the analysis of vowels which are known to be well characterized by their power spectral structures, especially by the locations of the several peak formant frequencies. Despite the successful applications of the linear systems analysis, human speech, strictly speaking, is a nonlinear dynamical phenomenon which involves nonlinear aerodynamic, biomechanical, physiological, and acoustic factors. In fact, a variety of vocal fold models are based on nonlinear modeling of the vocal fold physiology and nonlinear aerodynamics.<sup>6-9</sup> In speech synthesis, the nonlinear physiological models such as the two-mass model<sup>6</sup> and the glottal waveform models<sup>10-12</sup> are used for the excitation signals of LPC vocoders. Nonlinear dynamical information is also used implicitly in the standard speech coding schemes. For example, in the code-excited linear prediction (CELP) scheme,<sup>13</sup> a combination of codevectors from codebooks is used to model periodic and aperiodic

impulsive components of the excitation signals of LPC. These nonlinear techniques imply that some of the important qualities of speech are inherently characterized by nonlinear dynamics.

Despite the complicated vocal production mechanism, which is usually considered to be high-dimensional, the concept of dissipative nonlinear dynamics<sup>14</sup> implies a possibility that the complex vocal phenomena originate from deterministic nonlinear dynamics with only a small number of state variables. From this viewpoint, nonlinear dynamical system analysis<sup>15,16</sup> has been recently carried out for a variety of vocal phenomena.<sup>17-33</sup> For diagnosis of pathological voices, various nonlinear dynamics such as periodic, quasi-periodic, and chaotic dynamics have been analyzed<sup>17-21</sup> and in non-stationary infant cries possible bifurcation phenomena leading to chaos have been studied.<sup>22</sup> In fricative consonants chaotic dynamics has been discussed<sup>23</sup> and in normal phonation of vowels irregularity in pitch-to-pitch variation has been investigated in terms of low-dimensional nonlinear dynamics.<sup>24-33</sup>

Among these nonlinear speech studies, this article focuses on the nonlinear dynamics of vowels.<sup>24-33</sup> It has been known that in normal phonation of vowels cyclic changes in pitch amplitudes and pitch periods are observed.<sup>34</sup> By psychoacoustic experiments, it has been shown that this pitch-to-pitch variation is indispensable for speech signals to be perceived as natural human sound.<sup>35-38</sup> Since the naturalness of sound is an important factor for speech synthesis, the ir-

<sup>a)</sup>Electronic mail: tokuda@csse.muroran-it.ac.jp

regular property of the pitch-to-pitch variation of vowels is worthwhile investigating.

There have been several studies that considered the effect of pitch-to-pitch variation on the quality of synthesized vowels. It has been reported that the buzzerlike quality of the vowels synthesized by periodic excitation of LPC vocoders can be improved to some extent, if the standard deviations of the pitch amplitudes and pitch periods of the LPC excitation signals are optimized.<sup>39–44</sup> It has also been indicated that frequency characteristics of the sequences of the pitch periods and pitch amplitudes have strong influence on the voice quality and optimization of such frequency characteristics enhances the natural quality of the synthesized vowels.<sup>36–38</sup> If the original sequence of the pitch periods and pitch amplitudes obtained from real subjects are available, perceptually transparent speech can be synthesized by using standard speech coding schemes such as the code-excited linear prediction (CELP) scheme<sup>13</sup> and the multi-band excitation (MBE) scheme.<sup>45</sup> Such schemes, however, require a huge database or codebooks of pitch data for every voiced phonation of real speakers. They also provide no insight into the physiological mechanism that gives rise to the pitch-to-pitch variation of vowels, since they merely use a database of real pitch signals.

Compared to the conventional techniques, recently developed nonlinear prediction models for vowel synthesis are quite interesting. Townshend<sup>24</sup> and Banbrook *et al.*<sup>25</sup> used local linear function models, Sato *et al.*<sup>26</sup> and Tokuda *et al.*<sup>27</sup> used neural networks, Kubin<sup>28</sup> used polynomial function models, and Mann and McLaughlin<sup>29</sup> and Judd<sup>30</sup> used radial basis function models for the vowel synthesis. They reported that the irregular dynamical property of the pitch-to-pitch variation that contributes to natural vowel sounds is well reproduced by the nonlinear prediction models. Such nonlinear prediction models can provide speech synthesis techniques *possibly* simpler than the conventional ones in the sense that they are based on the function approximation techniques which do not require any huge database of pitch sequences. Although the conventional nonlinear prediction models that need optimization of many free parameters should be further refined for practical use, it is important to explore a new approach to vowel synthesis.

The studies of the nonlinear predictions imply that a dominant portion of the irregularity of vowels is due to low-dimensional possibly chaotic dynamics, because chaos is the only dynamics that deterministically gives rise to irregular behavior in nonlinear systems. In order to examine the plausibility of the nonlinear prediction models of vowels, it is important to study the irregular property in vowels from the viewpoint of nonlinear systems, especially deterministic chaos. In fact, there exist several studies that report chaotic dynamical properties in normal vowels. By fractal dimensional analysis with reliable dimension estimate technique,<sup>32</sup> noninteger fractal dimension lying between 1.0 and 3.0 was estimated for vowel signals. By the Wayland test,<sup>33</sup> deterministic nonlinearity was detected for normal vowels. Geometrical structure that resembles a typical chaotic orbit is observed by singular systems analysis of time-delay embedding of normal vowels.<sup>25,27</sup> By Lyapunov spectrum analysis, a

positive Lyapunov exponent<sup>27,31</sup> or weakly positive but close to zero Lyapunov exponent<sup>25</sup> was computed for normal vowels.

Despite these intensive studies, it is still difficult to confirm chaotic dynamics in normal vowels, because reliable estimation of nonlinear dynamical quantities from short-term speech data requires delicate numerical computation.<sup>46–48</sup> It should also be noted, on the other hand, that analysis of very-long-term data can suffer from nonstationarity. Moreover, we have to be very careful in analyzing and discussing low-dimensional chaos in real-world systems, since noisy data can sometimes mimic chaotic behavior.<sup>49–51</sup> Rapp *et al.*<sup>51</sup> demonstrated that the Grassberger–Procaccia algorithm falsely detects low-dimensional chaotic dynamics in artificial data generated by a simple filtering of a purely random number sequence. Since this kind of spurious result may often take place in laboratory experiments, nonlinear systems analysis combined with additional techniques such as surrogate data techniques is recommended.<sup>50,51</sup>

The present article does not directly prove chaotic dynamical properties in vowels. Instead, we investigate strength of nonlinearity in the irregular dynamics of the pitch-to-pitch variation of vowels. Our approach is based upon the method of surrogate data.<sup>52–54</sup> The surrogate data analysis is a kind of statistical hypothesis testing which is used to detect nonlinear dynamical structure in time series data observed from an unknown dynamical system. We test a null-hypothesis that

“There is no nonlinear dynamical correlation between one pitch waveform pattern to another.”

According to the null-hypothesis, we generate sets of spike-and-wave surrogate data and compute nonlinear dynamical statistics of the original and surrogate data. By observing whether there is any significant difference between estimates of the original and surrogate data, the null-hypothesis is tested.

To our knowledge, a comprehensive analysis of the vowel signals based on the above surrogate method has not been reported. In Refs. 32 and 55, Fourier transformed (FT) surrogate analysis was carried out for testing nonlinearity in normal vowels. The FT surrogate analysis that is to test a linear *Gaussian* property of vowels is not really interesting, because vowels are in general not considered to be generated from linear *Gaussian* processes in speech research. Instead, we examine nonlinear dynamical correlation between the pitch waveforms of vowels by the spike-and-wave surrogate analysis. By showing that there exists nonlinear dynamics in the pitch-to-pitch variation of vowels, we discuss the plausibility of modeling the vowels by nonlinear prediction techniques. Possible application of the nonlinear analysis results to the physiological modeling of vowels is also discussed.

The present article is organized as follows. In Sec. II, details of the vowel signals studied in this article are provided. Pitch-to-pitch variation observed in the vowel signals is also evaluated. In Sec. III, false nearest-neighbor analysis is carried out to study how many dimensions are necessary for nonlinear analysis of the vowels. In Sec. IV, nonlinear dynamics of the vowels are examined by spike-and-wave surrogate analysis. The final section is devoted to conclu-

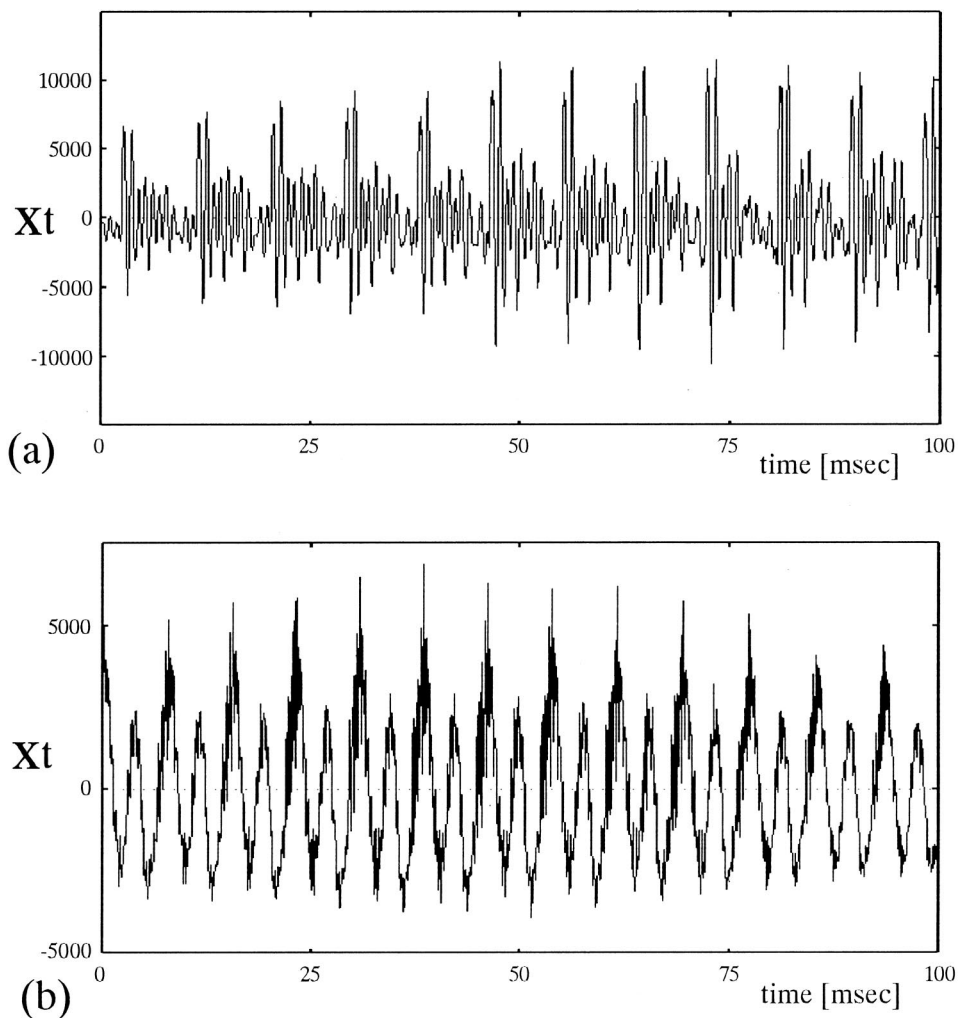


FIG. 1. (a) Speech signal  $\{x_t: t = 1, 2, \dots, 2048\}$  of vowel /a/ (subject: mau). (b) Speech signal of vowel /i/ (subject: mau).

sions of our experiments and discussions on possible application of nonlinear dynamics to speech synthesis.

## II. EXPERIMENTAL DATA

### A. Speech data

For our analysis, speech signals of five vowels /a/, /i/, /u/, /e/, and /o/ recorded from five subjects are exploited. Each vowel is spoken only once by each subject. We analyze five vowels so that we can consider dependency of nonlinear dynamical characteristics of vowels on the vocal tract shape. If the vocal tract shape gives rise to strong constriction at voiced phonation, we may expect that nonlinearity of the vocal fold dynamics is weakened by a filtering effect of the vocal tract. We study this effect for five vowels. The subject group is composed of three male speakers (mau, mms, mmy) and five female speakers (fsu, fyn) with no laryngeal pathology. The speech data are in the standard ATR (Advanced Telecommunications Research Institute International) database which is accessible at <http://www.ctr.atr.co.jp>. The speech signals are low-pass filtered with a cut-off frequency of 8 kHz and digitized with a sampling rate of 20 kHz and with 16-bit resolution. The initial transient phase and the final decay phase are removed from all data and the almost stationary part of the data is extracted. As examples, speech

signals denoted by  $\{x_t \in \mathbf{R}: t = 1, 2, \dots, N_{\text{data}}\}$  ( $N_{\text{data}} = 2048$ ) are drawn for two vowels /a/ and /i/ (subject: mau) in Figs. 1(a) and (b).

### B. Pitch-to-pitch variation

It is well known that cyclic changes in maximal peak amplitudes and pitch periods are observed in normal vowel signals.<sup>34</sup> Let us evaluate the level of this pitch-to-pitch variation in our speech data. First, maximum peak amplitudes and pitch periods are successively extracted from each vowel signal by using the peak-picking and zero-crossing method.<sup>56–60</sup> In our speech data, 15 to 35 pitch periods were extracted from each vowel signal. We call the sequences of the maximum peak amplitudes and the pitch periods amplitude sequence (AS) and period sequence (PS), respectively. Then, the standard deviation of the AS and PS is computed for each vowel signal. In order for normalization, the coefficient of variation (C.V.) is used as a measure for the size of fluctuations in AS and PS, where C.V. stands for the standard deviation of a sequence normalized by the mean.<sup>61</sup>

In Figs. 2(a) and (b), histograms of the mean and the standard deviation of pitch periods computed from our 25 vowel signals are respectively shown. The mean pitch period ranges from 3.5 to 9 ms and the standard deviation ranges from 0.06 to 0.48 ms. As is shown in Fig. 2(c), the C.V. of

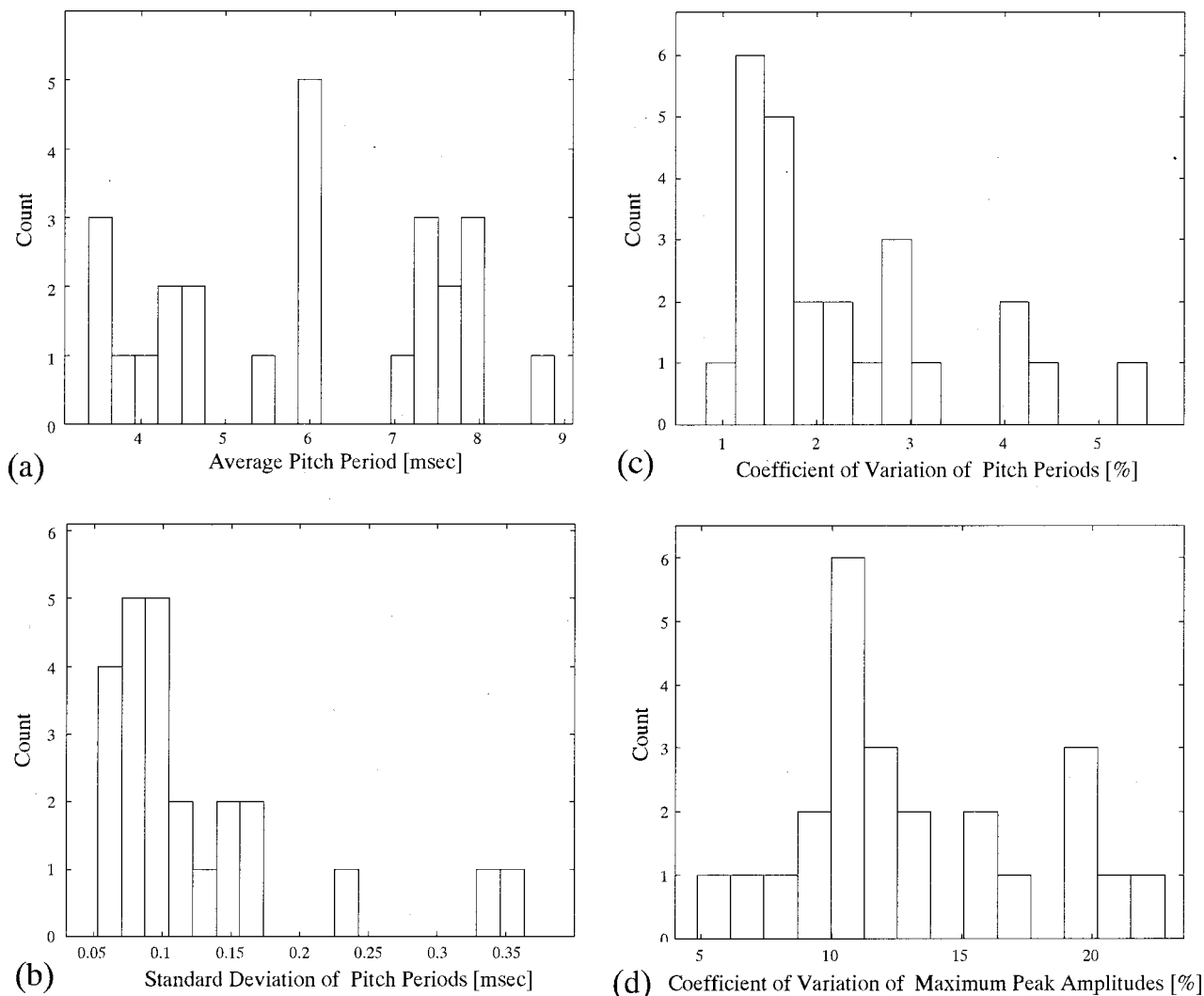


FIG. 2. (a) Distribution of the mean pitch period computed from 25 vowel signals. (b) Distribution of the standard deviation of pitch periods computed from 25 vowel signals. (c) Distribution of the coefficient of variation of pitch periods computed from 25 vowel signals. (d) Distribution of the coefficient of variation of maximum peak amplitudes computed from 25 vowels.

pitch periods ranges from 1% to 5.3% and its mode is located around 1.2%. The mode is within the normative range  $1.05 \pm 0.40\%$  which was reported for normal voiced sounds.<sup>62</sup>

Figure 2(d) shows a C.V. of maximum peak amplitudes computed from the 25 vowels. The C.V. of maximum peak amplitudes ranges from 5% to 29% and its mode is located around 10.6%. The mode is very close to the normative range  $6.68 \pm 3.03\%$  which was reported for normal voiced sounds.<sup>62</sup>

According to the evaluation of the cyclic changes in maximal peak amplitudes and pitch periods, we can observe a normal level of pitch-to-pitch variation in our speech data. In the following sections, we investigate irregular properties of this pitch-to-pitch variation from the view point of nonlinear dynamics.

### III. MINIMUM EMBEDDING DIMENSION OF VOWELS

In nonlinear dynamical systems analysis,<sup>15,16</sup> it is in general supposed that an observed time series with a single variable is generated by deterministic nonlinear dynamics with a

low-dimensional attractor. The first step for the nonlinear analysis of a single time series is to reconstruct a qualitatively similar dynamical trajectory to the original in a relatively low-dimensional delay-coordinate space as<sup>63,64</sup>

$$\mathbf{x}(t) = (x_t, x_{t-\tau}, \dots, x_{t-(d-1)\tau}), \quad (1)$$

where  $d$  and  $\tau$  stand for the reconstruction dimension and the time lag, respectively. Figures 3(a) and (b) show examples of two vowel signals /a/ and /i/ (subject: mau) reconstructed in the delay-coordinate space. As is discussed in Ref. 27, dynamical behavior that resembles a Shil'nikov-type chaos and a two-dimensional torus are recognized in the three-dimensional space of Figs. 3(a) and (b), respectively.

The result of Sauer *et al.*<sup>64</sup> states that when the original dynamical system that generates time series has a corresponding attractor with a box-counting dimension of  $d_A$ , a topologically equivalent attractor can be prevalently reconstructed in the delay-coordinate space when  $d > 2d_A$ . Although the mathematical result provides a sufficient topological condition for avoiding self-crossings of the trajectories in the delay-coordinate space, the natural ques-

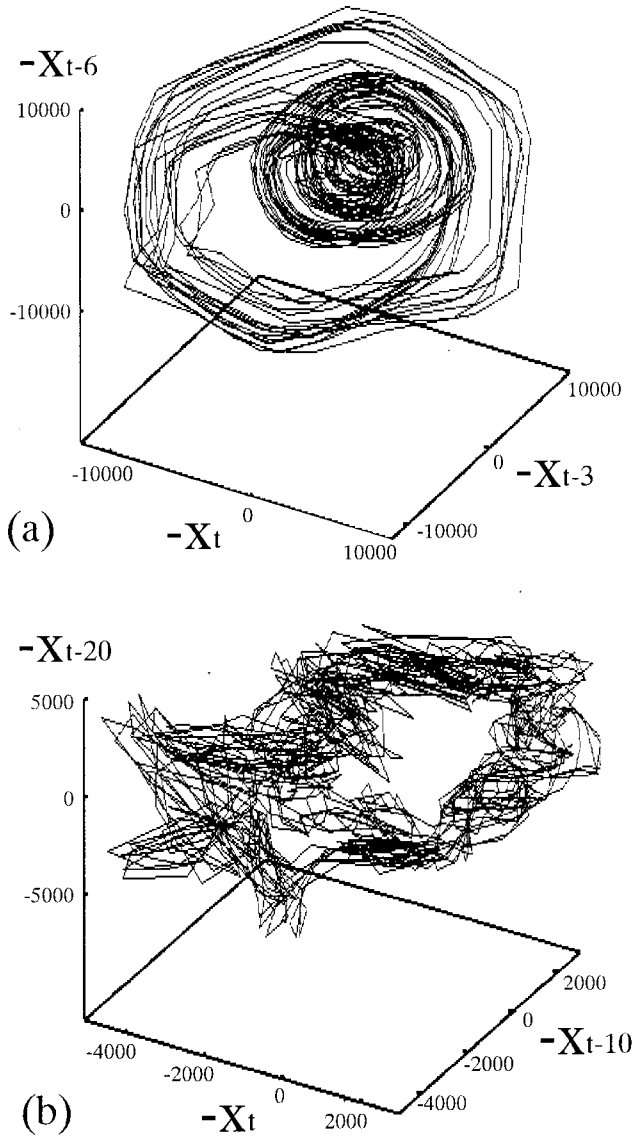


FIG. 3. (a) Reconstructed dynamics of the vowel /a/ of Fig. 1(a) in a three-dimensional delay-coordinate space  $(x_t, x_{t-3}, x_{t-6})$ . As is reported in Ref. 27, Shil'nikov-type dynamical structure can be recognized. (b) Reconstructed dynamics of the vowel /i/ of Fig. 1(b) in a three-dimensional delay-coordinate space  $(x_t, x_{t-10}, x_{t-20})$ . The dynamics resembles a quasi-periodic attractor.

tion is the following:

*Given a time series from an unknown dynamical system, how can the minimum embedding dimension  $d_E$  be determined for reconstructing the original dynamics?*

In order to determine the minimum embedding dimension  $d_E$ , let us analyze the speech data by the false nearest neighbor (FNN) method.<sup>65</sup> The FNN method provides a practical computational algorithm for estimating the minimum embedding dimension  $d_E$  of a time series data. Due to the simplicity of the algorithm and ease of its implementation, the FNN analysis has been widely applied to various real-world data.<sup>16</sup>

The FNN algorithm determines the minimum embedding dimension  $d_E$  by focusing on a topological change in the reconstructed dynamics in delay-coordinate space. Suppose that a time series  $\{x_t\}$  is reconstructed in delay-coordinate space by Eq. (1) with the reconstruction dimen-

sion  $d$ . For each data point  $\mathbf{x}(t)$ , denote its  $r$ th nearest neighbor by  $\mathbf{x}(t_r)$ . Then the square of the Euclidean distance between  $\mathbf{x}(t)$  and  $\mathbf{x}(t_r)$  is given by

$$R_d^2(t, r) = \|\mathbf{x}(t) - \mathbf{x}(t_r)\|^2 = \sum_{k=0}^{d-1} [x_{t-k\tau} - x_{t_r-k\tau}]^2. \quad (2)$$

Let us see a change in the distance  $R_d$  when the reconstruction dimension is increased as  $d \rightarrow d+1$ . The addition of the new  $(d+1)$ -th coordinate increases the distance between  $\mathbf{x}(t)$  and  $\mathbf{x}(t_r)$  by

$$R_{d+1}^2(t, r) = R_d^2(t, r) + [x_{t-d\tau} - x_{t_r-d\tau}]^2. \quad (3)$$

If the increase in the distance from  $R_d(t)$  to  $R_{d+1}(t)$  is significantly large as

$$\left[ \frac{R_{d+1}^2(t, r) - R_d^2(t, r)}{R_d^2(t, r)} \right]^{1/2} > R_{\text{tol}} \quad (R_{\text{tol}}: \text{threshold value}), \quad (4)$$

then  $\mathbf{x}(t_r)$  can be considered as a “false” nearest neighbor to  $\mathbf{x}(t)$  caused possibly by the self-crossing of orbit in the  $d$ -dimensional reconstruction space. Hence the condition (4) provides a first criterion for false nearest neighbors.

There is a second criterion for false nearest neighbors. Since we deal with time series with finite data points, the trajectory distribution can be sparse in the reconstruction space and some nearest neighbors to  $\mathbf{x}(t)$  might not be so close, i.e.,  $R_d(t) \approx R_A$  ( $R_A$ : an attractor size). If such distant nearest neighbors are “false” nearest neighbors, addition of a new  $(d+1)$ -th coordinate may stretch their distances by the attractor size and will result in  $R_{d+1}(t) \approx 2R_A$ . Hence, for such distant neighbors, the second criterion for false neighbors is given by

$$\frac{R_{d+1}(t)}{R_A} > 2, \quad (5)$$

where the attractor size  $R_A$  can be computed as

$$R_A^2 = \frac{1}{N_{\text{data}} - (d-1)\tau} \sum_{t=1+(d-1)\tau}^{N_{\text{data}}} \|\mathbf{x}(t) - \bar{\mathbf{x}}\|^2, \quad (6)$$

$$\bar{\mathbf{x}} = \frac{1}{N_{\text{data}} - (d-1)\tau} \sum_{t=1+(d-1)\tau}^{N_{\text{data}}} \mathbf{x}(t). \quad (7)$$

The “false” nearest neighbor is finally defined as the nearest neighbor that satisfies either of the first criterion (4) or the second criterion (5).

Figures 4(a) and (b) show results of the FNN analysis applied to two subject speakers mau and mms, where percentages of the false nearest neighbors of two vowels /a/ and /i/ are drawn simultaneously. The time lag is selected as  $\tau=3$  so that the window length of the delay-coordinates  $w=(d-1)\tau$  is set to be nearly equal to the first zero-crossing point of the auto-correlation function when vowel /a/ (mau) is reconstructed in three-dimensional space. The threshold value is set as  $R_{\text{tol}}=10$  and the reconstruction dimension is varied from  $d=1$  to  $d=7$ . In this analysis, “true” or “false” of only the first nearest neighbor is considered, i.e.,  $r=1$ .

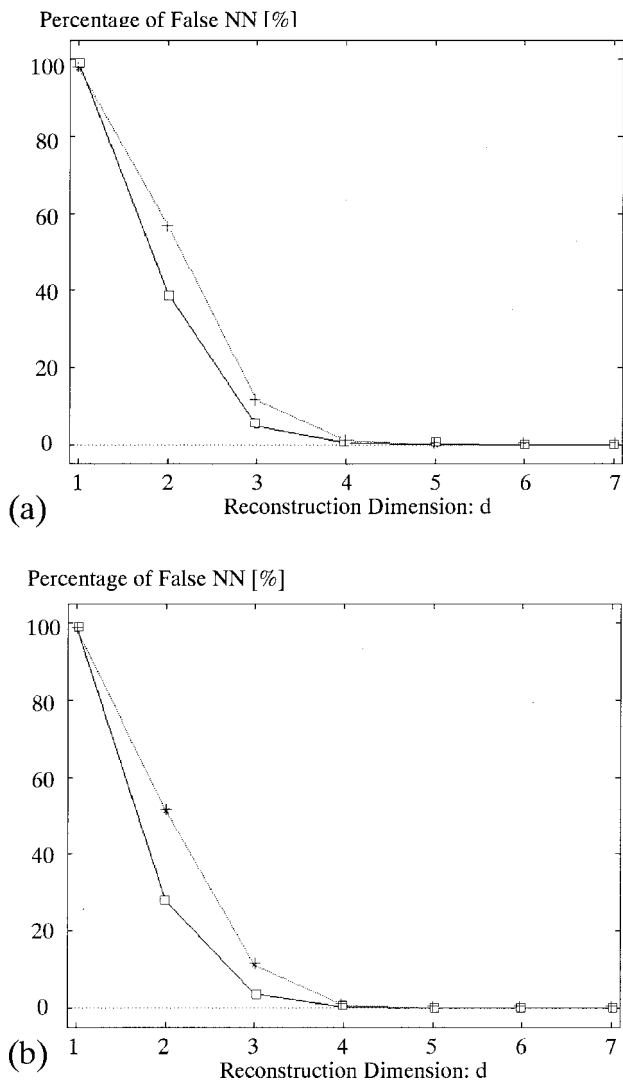


FIG. 4. Results of the FNN analysis applied to two subjects, (a) mau and (b) mms. In each figure, percentages of the false nearest neighbors of two vowels /a/ (solid line with squares) and /i/ (dotted line with crosses) are drawn simultaneously.

As the reconstruction dimension is increased from  $d = 1$ , we see that the percentage of false nearest neighbors is decreased and becomes almost zero for the reconstruction dimension higher than  $d=4$  for the two vowels. It is discussed in Ref. 65 that convergence to zero false-nearest-neighbor cannot be obtained in a noisy random data, since random data have practically infinite degrees of freedom. This implies that the speech signal of the two vowels can be characterized by relatively low-dimensional dynamics as  $d \leq 7$  and the minimum embedding dimension would be  $d_E = 4$ . For five vowels (/a/, /i/, /u/, /e/, /o/) and for five subjects (mau, mms, mmy, fsu, fyn), similar results have been obtained. Hence, the results of the FNN analysis do not seem to depend upon either the vowels or the subjects.

There are preceding studies of FNN analysis of vowels. Behrman<sup>21</sup> reported that six to eight (sometimes less) reconstruction dimensions are required for nonlinear analysis of normal vowels and Judd<sup>30</sup> reported that four dimensions are necessary to unfold the topological structure of a normal vowel. The present results basically agree with their results.

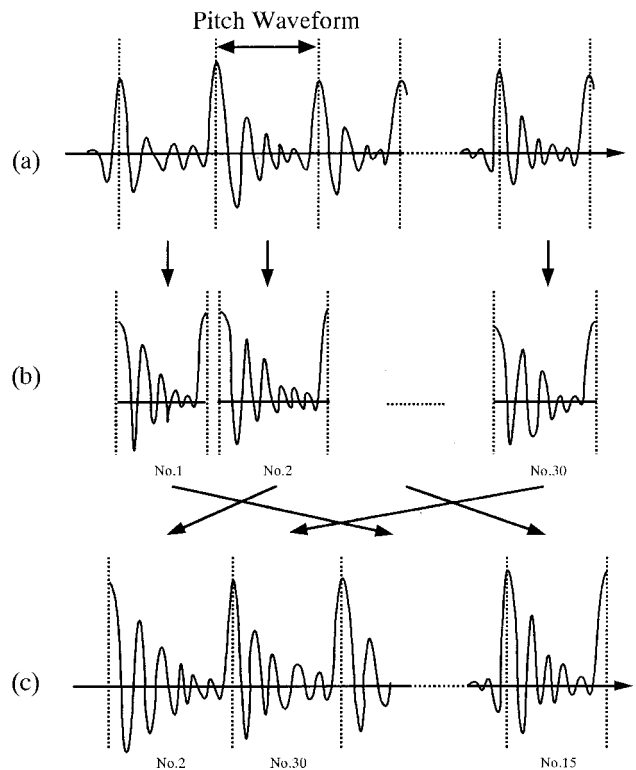


FIG. 5. Method of creating spike-and-wave surrogate data. (a), (b) Divide the original speech signal into pitch-waveform patterns by cutting the signal at maximal peak amplitudes. (c) Shuffle the pitch-waveforms and reconnect them with each other in random order.

#### IV. SURROGATE ANALYSIS

The FNN analysis of the previous section has shown that characteristic dynamics of the vowels can be reconstructed in a relatively low-dimensional delay-coordinate space of  $4 \leq d \leq 7$ . On the basis of the FNN analysis, let us further examine the nonlinear dynamical structure of the vowels by the method of surrogate data.<sup>52-54</sup>

The surrogate data analysis is a kind of statistical hypothesis testing which is to test a null-hypothesis  $H_0$  that the speech signal is generated from a particular class of non-deterministic dynamical process. In accordance with the null-hypothesis  $H_0$ , sets of artificial time series, called *surrogate data*, which preserve some of the statistical properties of the original speech signal are created by a surrogate algorithm. Then a discriminating statistic  $T$  is computed for the original and the surrogate data. If the original discriminating statistic  $T_{\text{origin}}$  is significantly different from that of the surrogate data, the null-hypothesis  $H_0$  can be rejected. The surrogate data have the property of “constrained realization,”<sup>53</sup> which is to randomize the original data by strictly preserving some of the original statistical properties. It is empirically known that the surrogate analysis is effective for statistical hypothesis testing when a *nonlinear* discriminating statistic  $T$ , whose distribution function is not well known, is utilized.

Among a variety of surrogate data analyses, spike-and-wave surrogate analysis is carried out in this study. The spike-and-wave surrogate analysis has been introduced by Theiler<sup>54</sup> for the analysis of epileptic EEG signals. The epileptic EEG signals are characterized by repeated occurrence of spike-and-wave patterns, where the variation of spike-and-

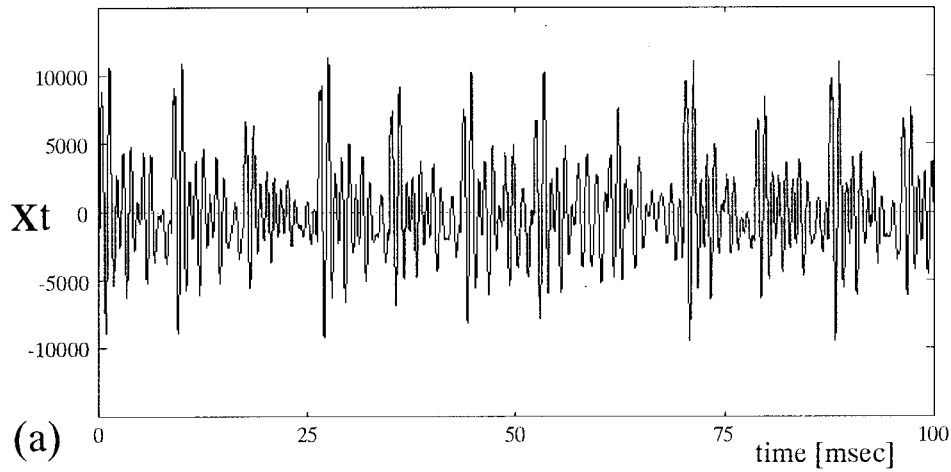
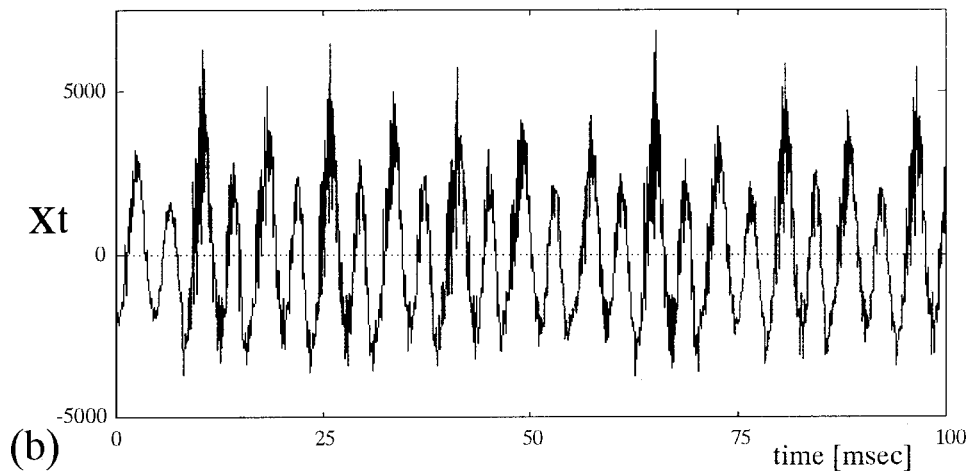


FIG. 6. (a) Spike-and-wave surrogate data made from the vowel signal /a/ of Fig. 1(a). (b) Spike-and-wave surrogate data made from the vowel signal /i/ of Fig. 1(b).



wave patterns had been considered as noisy components of limit cycle dynamics. By the surrogate analysis that examines nonlinear dynamical correlation between the spike-and-wave patterns, nonlinear dynamics that underlies the irregularity of the spike-and-wave patterns was detected in the epileptic EEG signals. In a similar manner, we study nonlinear dynamical correlation between pitch-waveform patterns in the vowel signals.

In the spike-and-wave surrogate analysis, we consider the following null-hypothesis:

$H_0$ : “There is no nonlinear dynamical correlation from one pitch-waveform pattern to another.”

### A. Surrogate data

In the surrogate analysis, it is a necessary condition that the subject data are stationary. Stationarity means that statistical characteristics of time series do not change in time. Since speech production is inherently a nonstationary dynamical process, we have to be careful when applying the surrogate analysis to speech. Even in a single phonation of a vowel, it is known that vocal tract configuration slightly changes in time. In order to apply the surrogate analysis to stationary parts of the data, relatively short-term vowel signals ( $\approx 100$  ms) consisting of about 15 to 35 pitch waveforms are extracted.

Of course, there exists a drawback of using such short-term data for the surrogate analysis. Especially for comput-

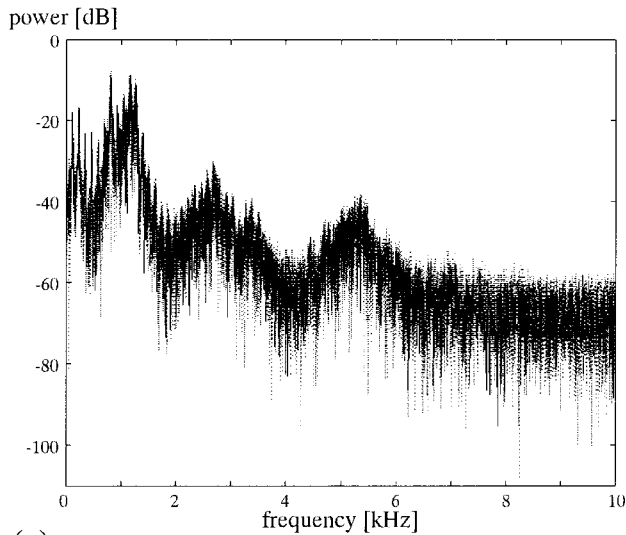
ing a discriminating statistic, reliable estimation of the nonlinear dynamical quantity from short-term data is quite a difficult task.<sup>46–48</sup> In this sense, there is a limitation of analyzing speech signals by the surrogate method which needs to reliably estimate nonlinear dynamical quantities from short-term stationary data.

The spike-and-wave surrogate data can be generated as follows (see Fig. 5).

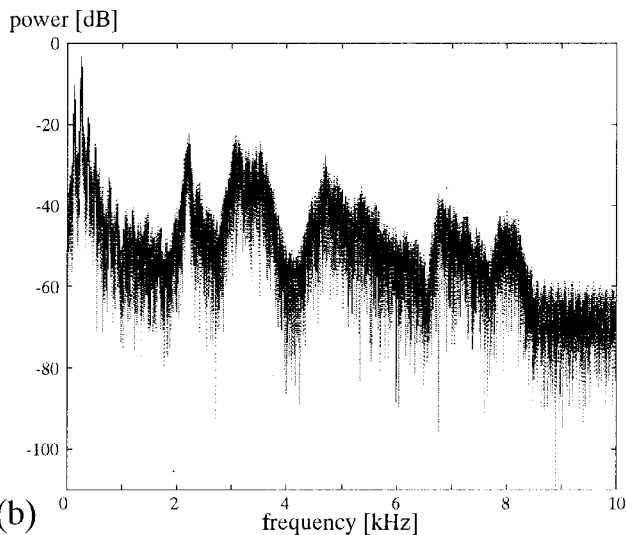
- (1) Divide the original speech signal into pitch-waveforms by cutting the signal at the maximal peak amplitudes.
- (2) Shuffle the pitch-waveforms and reconnect them with each other in random order.

By this surrogate shuffling, both histogram and pitch-waveform patterns of the original vowel signal are exactly preserved. For the statistical test, 39 sets of spike-and-wave surrogate data are created for each vowel. Figures 6(a) and (b) show spike-and-wave surrogate signals made from the original speech signals of Figs. 1(a) and (b), respectively. We see that the pitch-waveform structures of the original speech are preserved in the surrogate data.

In Figs. 7(a) and (b), the power spectra of the vowel signals /a/ and /i/ (subject: mau) are compared with those of their surrogate signals. In each figure, the bold line indicates the power spectrum of the original data, while the dotted lines indicate the power spectra of 39 sets of surrogate data. The power spectrum of the original data is covered with



(a)



(b)

FIG. 7. (a) Power spectra of the original speech signal of Fig. 1(a) (bold line), and its spike-and-wave surrogate signals (dotted line). (b) Power spectra of the original speech signal of Fig. 1(b) (bold line), and its spike-and-wave surrogate signals (dotted line).

those of the surrogate data and hence the original data structure cannot be distinguished from the surrogate data. This implies that linear dynamical quantities such as the power spectrum cannot detect a difference between the original vowel and its spike-and-wave surrogates.

Ifukube *et al.*<sup>35</sup> studied the effect of pitch waveform fluctuations on the perception of natural vowels. They created a surrogate data from a vowel /a/ in a similar manner with the spike-and-wave method and carried out a listening test. Their psychoacoustic test reported that the surrogate data shuffling instantly destroyed natural perception of the vowel. This implies that the human auditory system is capable of perceiving the naturalness of human sound in terms of the irregular dynamical structure of vowels. Presently, there is no way to quantify naturalness in terms of the irregular property of the vowels. If a nonlinear dynamical quantity can differentiate the original vowel signal from the surrogate data, such a nonlinear quantity might be a candidate for characterizing the naturalness of the vowels.

## B. Wayland translation error

As a discriminating statistic  $T$  of the surrogate analysis, the Wayland translation error<sup>66</sup> is exploited.

The Wayland algorithm assumes that a time series  $\{x_t\}$  is generated from a continuous nonlinear dynamical system and the reconstructed trajectory in the delay-coordinate space  $\{\mathbf{x}(t): t=1+(d-1)\tau, \dots, N_{\text{data}}\}$  is described by a continuous mapping  $f: R^d \rightarrow R^d$  as  $\mathbf{x}(t+1)=f(\mathbf{x}(t))$ . Since  $f$  is continuous, “nearby” data points, e.g.,  $\mathbf{x}(t)$  and  $\mathbf{x}(s)$ , are transformed to nearby states in  $T$ -step future,  $\mathbf{x}(t+T)$  and  $\mathbf{x}(s+T)$ , in the delay coordinate space. With respect to the assumption of continuity in the reconstructed dynamics, the Wayland translation error  $e_{\text{trans}}$  can be calculated as follows.

For a fixed data point  $\mathbf{x}(t_0)$ , called a *translation center*, find its  $k$ -nearest neighbors  $\mathbf{x}(t_1), \dots, \mathbf{x}(t_k)$ . Then, with respect to a *translation horizon*  $T$ , the translation vectors  $\{\mathbf{v}_j = \mathbf{x}(t_j+T) - \mathbf{x}(t_j): j=0, \dots, k\}$  are computed. If the neighboring points  $\mathbf{x}(t_1), \dots, \mathbf{x}(t_k)$  are transformed to neighboring points  $\mathbf{x}(t_1+T), \dots, \mathbf{x}(t_k+T)$  in  $T$ -step future states, the translation vectors  $\{\mathbf{v}_j\}$  are expected to point in similar directions. With respect to the diversity of the translation vectors, the translation error is calculated as

$$e_{\text{trans}} = \frac{1}{k+1} \sum_{j=0}^k \frac{\|\mathbf{v}_j - \bar{\mathbf{v}}\|^2}{\|\bar{\mathbf{v}}\|^2} \left( \bar{\mathbf{v}} = \frac{1}{k+1} \sum_{j=0}^k \mathbf{v}_j \right). \quad (8)$$

Figures 8(a) and (b) show the translation errors computed for speech signals of vowels /a/ and /i/ (subject: mau) and 39 sets of their surrogates. In each figure, error curves are drawn for the original speech data (solid line with circles) and for the surrogate data (solid lines with no circles). In order to reduce the statistical error for estimating the translation error of each time series, 20 sets of 301 translation centers  $\mathbf{x}(t_0)$  are randomly chosen and the median of each set of translation errors is calculated. Then the average over the 20 medians is estimated as the translation error  $e_{\text{trans}}$ . The reconstruction dimension is varied as  $d=2, \dots, 15$  and other parameters are set to  $\tau=10$ ,  $k=4$ , and  $T=50$ .

According to the numerical studies of the Wayland algorithm,<sup>66</sup> *Gaussian* white noise gives rise to a translation error of  $e_{\text{trans}} \approx 1$  independently of  $d$ . Colored noise, on the other hand, exhibits a translation error which monotonically decreases to  $\sim 0.5$  with increase in  $d$  due to the sustained autocorrelation.<sup>67</sup>

In Figs. 8(a) and (b), the original speech data give rise to translation error much less than 0.5, namely,  $e_{\text{trans}} \ll 0.5$ , with the minimum at  $d \approx 11$  in case of (a) and  $d \approx 8$  in case of (b). This implies that the vowel signals are described by neither *Gaussian* noise nor colored noise and that the qualitative dynamics of the vowel is well reconstructed with the dimension of  $d \leq 11$ . Moreover, the original data exhibit translation error curve which is distinctively lower than those of the 39 sets of the spike-and-wave surrogate data. In fact, for all five vowels (/a/, /i/, /u/, /e/, /o/) of the five subjects (mau, mms,

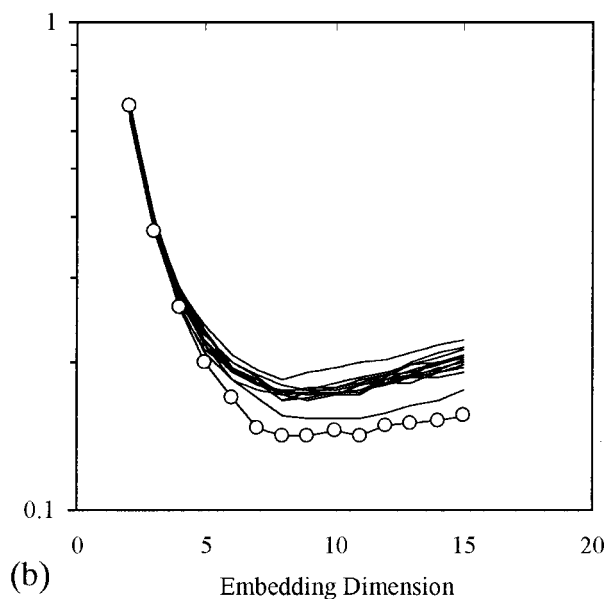
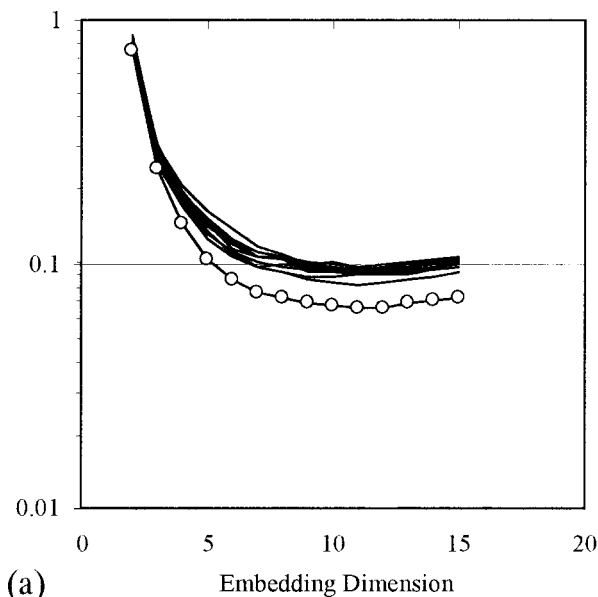


FIG. 8. (a) The translation error curve  $e_{\text{trans}}(d=2, \dots, 15)$  of the original vowel /a/ (subject: mau) (solid line with circles) and 39 sets of its spike-and-wave surrogates (solid lines with no circle). (b) The translation error curve of the original vowel /i/ (subject: mau) (solid line with circles) and 39 sets of its spike-and-wave surrogates (solid lines with no circle).

mmy, fsu, fyn), the translation errors of the original data are lower than those of the surrogate data. This means that for all vowel signals the spike-and-wave surrogate hypothesis is rejected with a level of  $\alpha=0.05$  ( $=\frac{2}{40}$ ). This is in general a strong rejection level for a statistical test and the results are independent of the vowels and the subjects. Therefore, we may conclude that there is nonlinear dynamical correlation between the pitch-waveforms of the vowel signals and such nonlinear dynamics has been destroyed by the spike-and-wave shuffling.

In several studies, the pitch-to-pitch variation of vowels is considered merely as noisy components of limit cycles.<sup>17,20,23</sup> If the pitch-to-pitch variation was generated from a purely stochastic random noise added to periodic cycles, statistical characteristics of the vowel signals may not

have been changed by the surrogate shuffling so significantly. The present results therefore indicate that the pitch-to-pitch variation of vowels cannot be simply regarded as random noise added to limit cycles.

We note that the method of generating the spike-and-wave surrogate data is based on the extraction of pitch waveforms and their reconnection in a randomized order. By this shuffling, discontinuity can occur at the reconnected points of the surrogate data. This could have an influence on the numerical results of the surrogate analysis. In order to avoid such a problem, an algorithm can be improved by applying, e.g., a smoothing filter to the reconnection points. We consider, however, that this may not change our main results, since our results show strong rejection levels for all vowel signals.

## V. CONCLUSIONS AND DISCUSSIONS

The dynamical structure and characteristics of normal vowels have been investigated by nonlinear systems analysis. By the false nearest-neighbor analysis, the minimum embedding dimension required for nonlinear analysis of vowels was estimated to be  $d_E=4$ . The analysis also revealed that the nonlinear dynamics of the vowels is well reconstructed and analyzed in relatively low-dimensional delay-coordinate spaces with  $4 \leq d \leq 10$ . Then, nonlinear dynamics of the vowels were further studied by the spike-and-wave surrogate analysis which exploits Wayland translation error as the nonlinear discriminating statistic. On the basis of the surrogate analysis, we have shown a possibility that there exists nonlinear dynamical correlation between one pitch-waveform pattern to another in the vowel signals.

Let us consider the present results in the light of the LPC modeling of vowels. In speech synthesis, vowels are usually synthesized by LPC models excited by impulse trains. In the sense that the intervals of the spike trains correspond to the pitch periods of the vowel signals, the present surrogate test implies that the spike intervals for the LPC model should not be mutually independent, but they may have nonlinear dynamical correlation. In conventional speech coding techniques such as the CELP scheme, the impulse excitation signals are selected from codebooks of periodic and aperiodic spike signals. Such schemes are not always efficient in the sense that they require a huge database of pitch sequences. If the intervals of the spike trains have a nonlinear dynamical correlation, such spike trains can be modeled by nonlinear prediction models. The nonlinear prediction models can provide speech synthesis techniques *possibly* simpler than the conventional ones in the sense that they are based on function approximation techniques which do not require any pitch database. It is our future work to examine plausibility of modeling LPC excitation signals by nonlinear prediction models that reproduce irregular properties of vowels with natural human sound quality.

We finally note that one of the most important applications of nonlinear analysis of vowels is to aid the development of nonlinear models for voice production, which can clarify the physiological mechanism that gives rise to the pitch-to-pitch variation of vowels. It is, however, still difficult to develop such physiological models from the present

study, since the present nonlinear analysis is not directly related to voice physiological models. There are several works that attempted to relate nonlinear analysis results to voice physiological models. For example, Herzel *et al.* interpreted nonsynchronous chaotic dynamics of pathological voice in terms of nonsymmetric vocal fold vibrations.<sup>68,69</sup> A similar approach to interpret the pitch-to-pitch variation of vowels will be explored in a future study.

## ACKNOWLEDGMENTS

This study was presented at the Second International Conference on Voice Physiology and Biomechanics in Berlin in March 1999. The authors express their special gratitude to Dr. Anders Löfqvist and two anonymous reviewers for their careful and thoughtful comments on the original manuscript. They also thank Professor J. Kurths, Professor H. Herzel, and Professor A. I. Mees and Dr. K. Judd, Dr. T. Ikeguchi, Dr. S. McLaughlin, and Dr. N. Aoki for stimulating discussions and valuable comments on the present work. This research was partially supported by Grant-in-Aid for Scientific Research (No. 10750259) of Japanese ministry of education, science, sports and culture and by Special Coordination Funds for Promoting Science and Technology (SCF).

<sup>1</sup>G. Fant, *Acoustic Theory of Speech Production* (Mouton, Gravenhage, 1960).  
<sup>2</sup>B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Am.* **50**, 637–655 (1971).  
<sup>3</sup>J. D. Markel and A. H. Gray, *Linear Prediction of Speech* (Springer-Verlag, Berlin, 1976).  
<sup>4</sup>D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Am.* **87**, 820–857 (1990).  
<sup>5</sup>D. G. Childers and C. K. Lee, "Vocal quality factors: Analysis, synthesis, and perception," *J. Acoust. Soc. Am.* **90**, 2394–2410 (1991).  
<sup>6</sup>K. Ishizaka and J. L. Flanagan, "Synthesis of voiced sounds from a two-mass model of the vocal cords," *Bell Syst. Tech. J.* **51**(6), 1233–1268 (1972).  
<sup>7</sup>I. R. Titze and D. T. Talkin, "A theoretical study of the effects of various laryngeal configurations on the acoustics of phonation," *J. Acoust. Soc. Am.* **66**, 60–74 (1979).  
<sup>8</sup>R. C. Scherer, I. R. Titze, and J. F. Curtis, "Pressure-flow relationships in two models of the larynx having rectangular glottal shapes," *J. Acoust. Soc. Am.* **73**, 668–676 (1983).  
<sup>9</sup>N. Miki, "Recent progress of the acoustic theory of speech production process," *J. Acoust. Soc. Jpn.* **48**(1), 15–19 (1992).  
<sup>10</sup>A. E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," *J. Acoust. Soc. Am.* **49**, 583–590 (1971).  
<sup>11</sup>G. Fant and Q. Lin, "Frequency domain interpretation and derivation of glottal flow parameters," *Speech Trans. Lab. Q. Prog. Stat. Rep.* **2**(3), 1–21 (1988).  
<sup>12</sup>H. Fujisaki and M. Ljungqvist, "Proposal and evaluation of models for glottal source waveform," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing* (1986), pp. 1605–1608.  
<sup>13</sup>M. R. Schroeder and B. S. Atal, "Code-excited linear prediction (CELP): High-quality speech at very low bit rates," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Tampa, FL (1985), pp. 937–940.  
<sup>14</sup>G. Nicolis and I. Prigogine, *Self-Organization in Nonequilibrium Systems* (Wiley, New York, 1977).  
<sup>15</sup>W. Lauterborn and U. Parlitz, "Methods of chaos physics and their application to acoustics," *J. Acoust. Soc. Am.* **84**, 1975–1993 (1988).  
<sup>16</sup>H. D. I. Abarbanel, R. Brown, J. J. Sidorowich, and L. S. Tsimring, "The analysis of observed chaotic data in physical systems," *Rev. Mod. Phys.* **65**, 1331–1392 (1993).  
<sup>17</sup>I. R. Titze, R. J. Baken, and H. Herzel, "Evidence of chaos in vocal fold

vibration," in *Vocal Fold Physiology*, edited by I. R. Titze (Singular, San Diego, 1993), pp. 143–188.  
<sup>18</sup>H. Herzel, D. Berry, I. R. Titze, and M. Saleh, "Analysis of vocal disorders with method from nonlinear dynamics," *J. Speech Hear. Res.* **37**, 1008–1019 (1994).  
<sup>19</sup>M. Tigges, P. Mergel, H. Herzel, T. Wittenberg, and U. Eysholdt, "Observation and modelling of glottal biphonation," *Acustica* **83**, 707–714 (1997).  
<sup>20</sup>A. Behrman and R. J. Baken, "Correlation dimension of electroglottographic data from healthy and pathologic subjects," *J. Acoust. Soc. Am.* **102**, 2371–2379 (1997).  
<sup>21</sup>A. Behrman, "Global and local dimensions of vocal dynamics," *J. Acoust. Soc. Am.* **105**, 432–443 (1999).  
<sup>22</sup>W. Mende, H. Herzel, and I. R. Titze, "Bifurcations and chaos in newborn cries," *Phys. Lett. A* **145**, 418–424 (1990).  
<sup>23</sup>S. S. Narayanan and A. A. Alwan, "A nonlinear dynamical systems analysis of fricative consonants," *J. Acoust. Soc. Am.* **97**, 2511–2524 (1995).  
<sup>24</sup>B. Townshend, "Nonlinear prediction of speech signals," in *Nonlinear Modeling and Forecasting*, edited by M. Casdagli and S. Eubank, SFI Studies in Sciences of Complexity (Addison-Wesley, Reading, MA, 1992), pp. 433–453.  
<sup>25</sup>M. Banbrook, S. McLaughlin, and I. Mann, "Speech characterization and synthesis by nonlinear methods," *IEEE Trans. Speech Audio Process.* **7**(1), 1–17 (1999).  
<sup>26</sup>M. Sato, K. Joe, and T. Hirahara, "APOLONN brings us to the real world," *Proc. Int. Joint Conf. Neural Networks* **1**, 581–587 (1990).  
<sup>27</sup>I. Tokuda, R. Tokunaga, and K. Aihara, "A simple geometrical structure underlying speech signals of the Japanese vowel /a/," *Int. J. Bifurcation Chaos Appl. Sci. Eng.* **6**(1), 149–160 (1996).  
<sup>28</sup>G. Kubin, "Nonlinear processing of speech," in *Speech Coding and Synthesis*, edited by W. B. Kleijin and K. K. Paliwal (Elsevier Science, Amsterdam, 1995), pp. 557–610.  
<sup>29</sup>I. Mann and S. McLaughlin, "Nonlinear dynamical modelling for speech synthesis using radial basis functions," preprint (2000).  
<sup>30</sup>K. Judd, "Nonlinear modelling: Keep it simple, vary the embedding, and make sure the dynamics are right," presented at Newton Institute Workshop on Nonlinear Dynamics and Statistics, Cambridge, 21–25 September 1998.  
<sup>31</sup>A. Kumar and S. K. Mullick, "Nonlinear dynamical analysis of speech," *J. Acoust. Soc. Am.* **100**, 615–629 (1996).  
<sup>32</sup>T. Ikeguchi and K. Aihara, "Estimating correlation dimensions of biological time series with a reliable method," *J. Int. Fuzzy Sys.* **5**(1), 33–52 (1997).  
<sup>33</sup>T. Miyano, "Are Japanese vowels chaotic?," *Proc. 4th Int. Conf. Soft Computing* (1996), Vol. 2, pp. 634–637.  
<sup>34</sup>L. Dolanský and P. Tjernlund, "On certain irregularities of voiced speech waveforms," *IEEE Trans. Audio Electroacoust.* **16**(1), 51–56 (1968).  
<sup>35</sup>T. Ifukube, M. Hashiba, and J. Matsushima, "A role of 'waveform fluctuation' on the naturality of vowels," *J. Acoust. Soc. Jpn.* **47**(12), 903–910 (1991).  
<sup>36</sup>T. Kobayashi and H. Sekine, "The role of fluctuations in fundamental period for natural speech synthesis," *J. Acoust. Soc. Jpn.* **47**(8), 539–544 (1991).  
<sup>37</sup>O. Komuro and H. Kasuya, "Characteristic of fundamental period variation and its modeling," *J. Acoust. Soc. Jpn.* **47**(12), 928–934 (1991).  
<sup>38</sup>N. Aoki and T. Ifukube, "Analysis and perception of spectral 1/f characteristics of amplitude and period fluctuations in normal sustained vowels," *J. Acoust. Soc. Am.* **106**, 423–433 (1999).  
<sup>39</sup>R. W. Wendahl, "Laryngeal analog synthesis of harsh voice quality," *Folia Phoniatri.* **15**, 241–250 (1963).  
<sup>40</sup>R. W. Wendahl, "Laryngeal analog synthesis of jitter and shimmer auditory parameters of harshness," *Folia Phoniatri.* **18**, 98–108 (1966).  
<sup>41</sup>S. Hiki, K. Sugawara, and J. Oizumi, "On the rapid fluctuation of voice pitch," *J. Acoust. Soc. Jpn.* **22**, 290–291 (1966).  
<sup>42</sup>R. F. Coleman, "Effect of median frequency levels upon the roughness of jittered stimuli," *J. Speech Hear. Res.* **12**, 330–336 (1969).  
<sup>43</sup>R. F. Coleman, "Effect of waveform changes upon roughness perception," *Folia Phoniatri.* **23**, 314–322 (1971).  
<sup>44</sup>A. J. Rozsypal and B. F. Miller, "Perception of jitter and shimmer in synthetic vowels," *J. Phonetics* **7**, 343–355 (1979).  
<sup>45</sup>D. R. Griffin and J. E. Lim, "Multiband excitation vocoder," *IEEE Trans. Acoust., Speech, Signal Process.* **36**(8), 1223–1235 (1988).  
<sup>46</sup>J. Theiler, "Spurious dimension from correlation algorithms applied to limited time-series data," *Phys. Rev. A* **34**(3), 2427–2432 (1986).

- <sup>47</sup>L. A. Smith, "Intrinsic limits of on dimension calculations," *Phys. Lett. A* **133**(6), 283–288 (1988).
- <sup>48</sup>D. Ruelle, "Deterministic chaos: Science and fiction," *Proc. R. Soc. London, Ser. A* **427**, 244–248 (1990).
- <sup>49</sup>D. Ruelle, "Where can one hope to profitably apply the ideas of chaos?" *Phys. Today* **July**, 24–30 (1994).
- <sup>50</sup>P. E. Rapp, "Chaos in the neurosciences: cautionary tales from the frontier," *Biologist* **40**(2), 89–94 (1993).
- <sup>51</sup>P. E. Rapp, A. M. Albano, T. I. Schmah, and L. A. Farwell, "Filtered noise can mimic low-dimensional chaotic attractors," *Phys. Rev. E* **47**, 2289–2297 (1993).
- <sup>52</sup>J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, and J. D. Farmer, "Testing for nonlinearity in time series: the method of surrogate data," *Physica D* **58**, 77–94 (1992).
- <sup>53</sup>J. Theiler and D. Prichard, "Constrained-realization Monte-Carlo method for hypothesis testing," *Physica D* **94**, 221–235 (1996).
- <sup>54</sup>J. Theiler, "On the evidence for low-dimensional chaos in an epileptic electroencephalogram," *Phys. Lett. A* **196**, 335–341 (1995).
- <sup>55</sup>T. Miyano, A. Nagami, I. Tokuda, and K. Aihara, "Detecting nonlinear determinism in voiced sounds of Japanese vowel /a/," *Int. J. Bifurcation Chaos Appl. Sci. Eng.* **10**(8), 1973–1979 (2000).
- <sup>56</sup>H. Hollien, J. Michel, and E. T. Doherty, "A method for analyzing vocal jitter in sustained phonation," *J. Phonetics* **1**, 85–91 (1973).
- <sup>57</sup>Y. Horii, "Some statistical characteristics of voice fundamental frequency," *J. Speech Hear. Res.* **18**, 192–201 (1975).
- <sup>58</sup>Y. Horii, "Fundamental frequency perturbation observed in sustained phonation," *J. Speech Hear. Res.* **22**, 5–19 (1979).
- <sup>59</sup>I. R. Titze, Y. Horii, and R. C. Scherer, "Some technical considerations in voice perturbation measurements," *J. Speech Hear. Res.* **30**, 252–260 (1987).
- <sup>60</sup>I. R. Titze and H. Liang, "Comparison of *F0* extraction methods for high-precision voice perturbation measurements," *J. Speech Hear. Res.* **36**, 1120–1133 (1993).
- <sup>61</sup>N. B. Pinto and I. R. Titze, "Unification of perturbation measures in speech signals," *J. Acoust. Soc. Am.* **87**, 1278–1289 (1990).
- <sup>62</sup>R. C. Scherer, V. J. Vail, and C. G. Guo, "Required number of tokens to determine representative voice perturbation values," *J. Speech Hear. Res.* **38**, 1260–1269 (1995).
- <sup>63</sup>F. Takens, "Detecting strange attractors in turbulence," in *Lecture Notes in Math* (Springer, Berlin, 1981), Vol. 898, pp. 366–381.
- <sup>64</sup>T. Sauer, J. A. York, and M. Casdagli, "Embedology," *J. Stat. Phys.* **65**(3), 579–616 (1991).
- <sup>65</sup>M. B. Kennel, R. Brown, and H. D. I. Abarbanel, "Determining embedding dimension for phase-space reconstruction using a geometric construction," *Phys. Rev. A* **45**(6), 3403–3411 (1992).
- <sup>66</sup>R. Wayland, D. Bromely, D. Pickett, and A. Passamante, "Recognizing determinism in a time series," *Phys. Rev. Lett.* **70**(5), 580–582 (1993).
- <sup>67</sup>T. Miyano, "Time series analysis of complex dynamical behavior contaminated with observational noise," *Int. J. Bifurcation Chaos Appl. Sci. Eng.* **6**, 2031–2045 (1996).
- <sup>68</sup>P. Mergel and H. Herzel, "Modeling biphonation," *Speech Commun.* **22**, 141–154 (1997).
- <sup>69</sup>I. Steinecke and H. Herzel, "Bifurcations in an asymmetric vocal-fold model," *J. Acoust. Soc. Am.* **97**, 1874–1884 (1995).