ORIGINAL ARTICLES

ANALYSIS OF NONLINEARITIES IN NONVERBAL VOICES

Toward Extraction of *Kansei* Information via Nonverbal Voices

Jun-ichi IMAI*, Hidenori TANAKA** and Tomomasa NAGASHIMA***

* Satellite Venture Business Laboratory, Muroran Institute of Technology, 27-1 Mizumoto-cho, Muroran, 050-8585, Japan

** Graduate School of Engineering, Muroran Institute of Technology, 27-1 Mizumoto-cho, Muroran, 050-8585, Japan

*** Faculty of Engineering, Muroran Institute of Technology, 27-1 Mizumoto-cho, Muroran, 050-8585, Japan

Abstract: In this paper, we study several kinds of nonverbal voices such as cries, screams and laughs. Nonverbal voices are the most primitive and typical emotional voices, and have much *Kansei* information. They are important from the viewpoint of communication between speakers and listeners. Especially, they are very important tools for communication for babies and children whose faculties of speech are not developed sufficiently. Therefore, it is a very significant subject to analyze nonverbal voices in order to investigate how *Kansei* information is expressed in them and how listeners recognize speakers' emotions through them. Recently, techniques for analyses of voices from the viewpoint of nonlinear deterministic dynamics have become a center of attraction. Some works suggest that the "nonlinearities" are strongly related to information in voices except linguistic information, especially *Kansei* information. It should be a challenging problem to tackle this suggestion. In order to approach this suggestion, it might be an important problem to extract *Kansei* information in nonverbal voices by using the nonlinear characteristics. In this paper, as a first step toward such a nonlinear analysis of nonverbal voices, we carry out a statistical testing to confirm the existence of nonlinearities in voices of cries, screams and laughs using the method of surrogate data. The largest Lyapunov exponent and prediction error by the locally linear method are used as discriminating statistics for testing. As a result, we confirmed the existence of nonlinearities in voices of cries and laughs statistically. Therefore, it was suggested that not only linear but also nonlinear methods are required for extracting much information from at least these two kinds of nonverbal voices. We also discuss functional interpretations of nonlinearities in nonverbal voices.

Keywords: *Kansei* information, Nonverbal voice, Testing for nonlinearity, Functional interpretation of nonlinearity

1. INTRODUCTION

Voice is the most natural means for our communication. The development of information technology such as speech synthesis and speech recognition has more and more increased the importance of voice even for interfaces between people and machines.

The conventional information technology has mainly focused on linguistic information in voices. However, we do not communicate each other using only linguistic information. In conversation between people, nonverbal means such as eyes, facial expressions, gestures and physical contacts also play important roles in natural and efficient communication. As for voices, we naturally transfer *Kansei* information such as our intensions, emotions and even information on physical conditions through our voices as well as linguistic information. Conversely, conventional techniques focusing on only linguistic information are quite insufficient for realizing the true *Kansei* communication being realized in human communication, which is more than mere mechanical exchange of linguistic information. Techniques for extracting *Kansei* information from voices are strongly required for the next stage of information technology.

Nonverbal voices, such as cries, screams and laughs, are the most primitive and typical emotional voices. Since nonverbal voices do not have any linguistic information, they have been intentionally excluded from the scopes of conventional studies on voices that have regarded linguistic information as of major importance. However, nonverbal voices have much *Kansei* information. Furthermore, they are also important from the viewpoint of communication between speakers and listeners. For instance, cries and screams can be regarded as emergency signals appealing to listeners for help, and laughs often invite listeners' laughs. Especially, nonverbal voices are very important tools for communication for babies and children whose faculties of speech are not developed sufficiently. Therefore, it is a very significant subject to analyze nonverbal voices in order to investigate how *Kansei* information is expressed in them and how listeners recognize speakers' emotions through them. This investigation must also provide suggestive clues for general emotional voices. In this paper, we study several kinds of these nonverbal voices such as cries, screams and laughs.

Recently, analyses of voices from the viewpoint of nonlinear deterministic dynamics have become a center of at-

traction [1–4]. These analyses have revealed that there exist nonlinearities in vowels. Some works have pointed out that these nonlinearities play important roles in naturalness of voices [5–7]. Furthermore, there has been an attempt to predict speaker’s fatigue and drowsiness using nonlinear characteristics in his or her voice [8]. These examples suggest that the “nonlinearities” are strongly related to information in voices except linguistic information, especially *Kansei* information. It should be a challenging problem to tackle this suggestion. In order to approach this suggestion, it might be an important problem to extract *Kansei* information in nonverbal voices by using the nonlinear characteristics in them.

In this paper, as a first step toward such a nonlinear analysis of nonverbal voices, we carry out a statistical testing to confirm the existence of nonlinearities in voices of cries, screams and laughs using the method of surrogate data [9]. The largest Lyapunov exponent and prediction error by the locally linear method [10] are used as discriminating statistics for testing. By this testing, we suggest that not only linear but also nonlinear methods are required for extracting much information from nonverbal voices. We also discuss functional interpretations of nonlinearities in nonverbal voices.

2. MATERIALS

Time series data cut out from nonverbal voices (4 cries, 4 screams and 4 laughs uttered by babies and children) recorded in “The Nightingale Voice Box: Just Kids and Babies” (Nightingale Music Productions, 2001) are analyzed. Descriptions of voice data are shown in Table 1. It is considered that babies and children depend on nonverbal voices much more strongly in communication than adults with sufficient faculties of speech. Since it is expected that *Kansei* information in nonverbal voices uttered by babies and children are expressed more naturally and obediently, we analyze nonverbal voices of babies and children in this paper.

The sampling frequency is 44.1 [kHz]. Each time series data for analysis is cut out from a stationary utterance within the voice data. The length of time series N is 21,870 points (about 0.50 [s]) for cries and screams, and 10,368 points (about 0.24 [s]) for laughs. Since voices of laughs generally consist of repetition of relatively short utterances, the length of time series for laughs is set to be shorter than others.

Waveforms and spectrograms of cry #1, scream #1 and

Table 1: Descriptions of voice data.

	#	Description
Cry	1	Baby, 1 year old, angry
	2	Baby, 1 year old, whining, sobbing
	3	Boy, sobbing
	4	Girl, sobbing
Scream	1	Child, shriek, very high-pitched
	2	Boy, low-pitched, yell
	3	Child, shriek, high-pitched, scared
	4	Child, shriek, high-pitched, classic horror
Laugh	1	Girl, 10 years old, having fun
	2	Girl, 2 years old, high, delighted, giggle
	3	Boy, 10 years old, evil, cackle
	4	Baby, 1 year old, happy

laugh #3 are shown in Figure 1 as typical examples. For each data, the top figure denotes the waveform of whole voice data. Second one is an indicator of the position of time series data cut out for analysis. Lower two figures denote the waveform of time series and its spectrogram respectively. Each spectrogram is computed using a 256-point fast Fourier transform.

The spectrogram of scream #1 in Figure 1 (b) shows clear stripes, which imply a harmonic structure associated with fundamental frequencies. On the other hand, the spectrograms of cry #1 and laugh #3 shown in Figure 1 (a) and (c) are more obscure than that of scream #1. All other data also have same tendencies.

3. METHODS

3.1 Testing for nonlinearity

We carry out a statistical testing to confirm the existence of nonlinearities in nonverbal voices using the method of surrogate data [9].

In our analysis, a following null hypothesis H_0 on nonverbal voice data is formulated.

H_0 : “The observed time series data has been generated by a linear Gaussian stochastic process under a possibly nonlinear static transform.”

Under this hypothesis, surrogate data are generated using the algorithm called iterative amplitude adjusted Fourier transform (IAAFT) [11]. Surrogate data generated by IAAFT algorithm exactly preserve the original histogram and approximately preserve the original power spectrum. As described below, 39 surrogates are used for each voice

Figure 1: Examples of waveforms and spectrograms.

data in this paper. We use the “TISEAN” software package [12, 13] for generating surrogate data, which implements IAAFT algorithm.

The null hypothesis H_0 is tested as follows. Let t_0 be a discriminating statistic estimated for original data. Let t_k ($k = 1, 2, \dots, B$) be those for surrogate data, which are sorted in descending order, that is,

$$t_1 \geq t_2 \geq \dots \geq t_B, \quad (1)$$

where B denotes the number of surrogates. In two-sided test, H_0 can be rejected at significance level α if

$$t_0 > t_{(B+1)\alpha/2} \quad (2)$$

or

$$t_0 < t_{(B+1)(1-\alpha/2)}. \quad (3)$$

We set $B = 39$ and $\alpha = 0.05$ in this paper. Therefore, H_0 can be rejected at significance level 0.05 if t_0 is greater than t_k for all k , or less than t_k for all k .

3.2 Nonlinear discriminating statistics for testing

In this paper, the largest Lyapunov exponent (LLE) and nonlinear prediction error (NPE) are used as discriminating statistics for nonlinearity testing. We estimate these statistics using the same “TISEAN” software package.

(i) Largest Lyapunov exponent

The largest Lyapunov exponent (LLE) is a crucial index to identify orbital instability of deterministic dynamics, that is, a measure of sensitive dependence on initial conditions. A positive LLE indicates the existence of chaos in time series data.

For estimating the Lyapunov exponents, reconstruction of attractor [14] is needed. That is, the scalar time series (i.e. the original or surrogate data) $x(t)$ ($t = 1, 2, \dots, N$) is transformed into a series of m -dimensional vector $\mathbf{x}(t)$ ($t = 1, 2, \dots, N - (m - 1)\tau$) as follows.

$$\mathbf{x}(t) = (x(t), x(t - \tau), \dots, x(t - (m - 1)\tau)), \quad (4)$$

where τ denotes a constant time lag parameter. m is called an embedding dimension.

The Lyapunov exponents are estimated by investigating temporal change of distances between a point $\mathbf{x}(t)$ and its neighborhood points on this reconstructed attractor. We analyze the Lyapunov exponents based on the algorithm by Sano and Sawada [15]. LLE is the largest value of estimated Lyapunov exponents.

In our analysis, the time lag τ is set to be an appropriate value derived by considering high order correlations [16, 17] for each data individually. The embedding dimension m is a parameter.

(ii) Nonlinear prediction error

The nonlinear prediction error (NPE) by the locally linear method [10], which is one of nonlinear prediction methods, is also used for nonlinearity testing.

In this method, reconstruction of attractor is needed as well as the estimation of LLE. Then a relation between $x(t + p)$ and $x(t)$ is approximated by a following linear equation.

$$x(t + p) \approx a_{p,t} \cdot x(t) + b_{p,t} \quad (5)$$

The coefficients $a_{p,t}$ and $b_{p,t}$ in Equation (5) are determined by the method of least squares minimizing a following error function.

$$Error = \sum_{x(i) \in \mathcal{U}_t} \{x(i + p) - (a_{p,t} \cdot x(i) + b_{p,t})\}^2, \quad (6)$$

where $\mathcal{U}_t$ is the neighborhood of $x(t)$, excluding $x(t)$ itself.

This technique is called the locally linear method because the “local” dynamics of data is approximated by “linear” models (i.e. Equation (5)). Note that “locally linear” implies essentially a nonlinear model, since the prediction model depends on the location of $x(t)$ on the reconstructed attractor.

The accuracy of the p -steps ahead prediction $\hat{x}(t + p)$ ($= a_t \cdot x(t) + b_t$) is evaluated by the following relative prediction error (i.e. nonlinear prediction error).

$$NPE = \frac{\sqrt{\sum_{t=1}^P (\hat{x}(t) - x(t))^2}}{\sqrt{\sum_{t=1}^P (x(t) - \bar{x})^2}}, \quad (7)$$

where $\bar{x} = (\sum_{t=1}^P x(t))/P$, and $P (= N - (m - 1)\tau - p)$ denotes the length of predicted time series. This error NPE is used as the discriminating statistic for testing.

Either the embedding dimension m or the prediction step p is a parameter in our analysis. We set $p = 1$ when m is a parameter, and $m = 5$ when p is a parameter conversely.

4. RESULTS

4.1 Cries

The results of estimating nonlinear discriminating statistics for the voice of cry #1 are shown in Figure 2 as a typical example.

(a) Reconstructed attractor.

(b) Largest Lyapunov exponent.

(c) Nonlinear prediction error versus embedding dimension ($p = 1$).

(d) Nonlinear prediction error versus number of prediction steps ($m = 5$).

Figure 2: Results for voice of cry #1.

Table 2: Largest Lyapunov exponents for nonverbal voices ($m = 5$).

#	Cry	Scream	Laugh
1	4.759×10^{-2}	—	4.672×10^{-2}
2	5.067×10^{-2}	—	2.149×10^{-2}
3	3.425×10^{-2}	—	3.706×10^{-2}
4	2.463×10^{-3}	2.052×10^{-3}	2.309×10^{-2}
Mean	3.374×10^{-2}	—	3.209×10^{-2}
Variance	4.857×10^{-4}	—	1.440×10^{-4}

The reconstructed attractor for $m = 3$ is shown in Figure 2 (a). This figure implies the existence of some nonlinear structure in this voice.

Figure 2 (b), (c) and (d) show estimated LLE and NPE (versus the embedding dimension m and the number of prediction steps p) respectively. In these figures, discriminating statistics for original data are less than those for corresponding surrogate data. Therefore, the null hypothesis H_0 formulated in Section 3.1 can be rejected at significance level $\alpha = 0.05$. Similarly, in our analysis, H_0 can also be rejected at significance level 0.05 for all other cries.

LLE (at $m = 5$) for each data is shown in Table 2. For all data of cries, LLE is small but positive. These results suggest that there exist nonlinearities in voices of cries.

4.2 Screams

The results for the voice of scream #1 are shown in Figure 3 as a typical example. The reconstructed attractor (for $m = 3$) shown in Figure 3 (a) seems concentric circles. In Figure 3 (b), (c) and (d), each line for original data is almost hidden by the lines for corresponding surrogate data. Therefore, the null hypothesis H_0 cannot be rejected at significance level 0.05. The results of other screams are similar. In our analysis, H_0 can be rejected at significance level 0.05 only for scream #4.

LLE (at $m = 5$) only for scream #4 is shown in Table 2. LLE for scream #4 takes the smallest value in our analysis.

4.3 Laughs

The results for the voice of laugh #3 are shown in Figure 4. Almost all statistics for original data are less than those for corresponding surrogate data in Figure 4. Therefore, the null hypothesis H_0 can be rejected at significance level 0.05 for such statistics. However, H_0 cannot be rejected for LLE at $m = 2$ and 3, although it can be rejected

(a) Reconstructed attractor.

(b) Largest Lyapunov exponent.

(c) Nonlinear prediction error versus embedding dimension ($p = 1$).

(d) Nonlinear prediction error versus number of prediction steps ($m = 5$).

Figure 3: Results for voice of scream #1.

(a) Reconstructed attractor.

(b) Largest Lyapunov exponent.

(c) Nonlinear prediction error versus embedding dimension ($p = 1$).

(d) Nonlinear prediction error versus number of prediction steps ($m = 5$).

Figure 4: Results for voice of laugh #3.

at higher embedding dimension. It is considered because $m = 2$ and 3 are too low dimensions for embedding this voice data. (The minimal sufficient embedding dimension determined by the false nearest neighbor method [18] is $m = 4$ for laugh #3.) H_0 is rejected for all other laughs.

LLE (at $m = 5$) for each data is shown in Table 2. LLE is positive for all data of laughs being similar to cries.

5. DISCUSSIONS

As mentioned in the previous section, the null hypothesis H_0 can be rejected and LLE takes a positive value for all data of cries and laughs. These results suggest that there exist nonlinearities in voices of cries and laughs, and that not only linear but also nonlinear methods are required for analysis, modeling, and engineering applications of these voices. It should be noticed that we could not necessarily have obtained the confirmation of “determinism” in these time series, even though the null hypothesis H_0 based on a linear stochastic process has been rejected. An analysis using the method for recognizing determinism [19, 20] will be necessary for this issue.

On the other hand, H_0 cannot be rejected for many data of screams. Therefore, we cannot claim the existence of nonlinearities in voices of screams from these results, and a further investigation will be required.

It can be seen from Table 2 that the mean value of LLE for cries is relatively larger than that for laughs. Furthermore, the variance for cries is also larger than that for laughs. It can be considered that this tendency is a reflection of some specific properties of these two categories. It is expected that these specific properties of nonverbal voices can be grasped more exactly by using several nonlinear statistics together, and a further investigation will be required to confirm it. Moreover, a study on modeling the mechanism of generating these tendencies of nonlinear characteristics among nonverbal voices will also be required as a significant topic of *Kansei Engineering*.

In the remainder of this section, we discuss functional interpretations of the nonlinearities in nonverbal voices. There are three possible interpretations as follows.

(i) Naturalness of voice

It is considered that the nonlinearities in nonverbal voices relate to naturalness of voices as well as those in general voices. Generally, natural voices give more intimate impressions to listeners than artificial voices. Furthermore, we suppose that naturalness in nonverbal voices

probably plays an important role also in listeners' recognition of emotions of speakers. In fact, screams that were not confirmed the existence of nonlinearities in our analysis sound like artificial buzzers or emergency signals rather than emotional voices. A further investigation will be required to support this supposition.

(ii) Individual variation

We can see from Table 2 that the largest value of LLE is more than double of the smallest one even in the category of laughs, which has the relatively small variance. It can be considered that the individual variation is one of the main factors causing this variation of the degree of nonlinearity. These variations probably enable listeners to identify the individual nonverbal voice, such as their own baby's cry for example.

(iii) Expression of *Kansei* information

It can be considered that the variation of the degree of nonlinearity also depends on slight differences of speakers' mental states affected by his or her craving as well as the individual variation. For example, similar cries uttered by same baby often imply different declaration of intention, such as "I am hungry," "painful" or "sleepy." It is sufficiently expected that measures of the degree of nonlinearity are important clues for discrimination between these slight differences of voices within the same category of emotions such as laughs or cries.

It is expected that we can efficiently extract these *Kansei* information in detail from their voices using the techniques based on both of conventional frequency analysis and nonlinear deterministic dynamics complementarily. These techniques will also be useful for synthesis of natural emotional voices.

6. CONCLUSIONS

In this paper, we have carried out the statistical testing to confirm the existence of nonlinearities in voices of 4 cries, 4 screams and 4 laughs. As a result, we have confirmed the existence of nonlinearities in voices of cries and laughs statistically. Therefore, it has been suggested that not only linear but also nonlinear methods are required for extracting much information from at least these two kinds of nonverbal voices. We have also investigated differences in the degree of nonlinearity in these nonverbal voices and discussed functional interpretations of the nonlinearities.

As a future work, a large-scale analysis of nonverbal voices should be addressed. In this paper, we have analyzed a small set of voice data in order to investigate each data in detail. Hereafter, the investigation of a larger number of nonverbal voice data will be required. Furthermore, we have not made a clear distinction between babies' voices and children's ones in this paper. However, it is probable that the degree of importance of nonverbal voices in communication is different between for babies with no faculties of speech and for children, and that characteristics of their nonverbal voices are also different. A further analysis for this problem will be necessary. In addition, subjects indicated in the section 5 will also be important.

Moreover, in this paper, only stationary utterance parts within the voice data have been analyzed. Transitions between qualitatively different dynamics, or bifurcations, within the same voice data have not been considered. The investigation of bifurcations within nonverbal voices will also be required.

REFERENCES

1. H. Herzel, D. Berry, I. R. Titze and M. Saleh; Analysis of Vocal Disorders with Method from Nonlinear Dynamics, *Journal of Speech and Hearing Research*; 37, 1008–1019 (1994).
2. S. S. Narayanan and A. A. Alwan; A Nonlinear Dynamical Systems Analysis of Fricative Consonants, *The Journal of the Acoustical Society of America*; 97 (4), 2511–2524 (1995).
3. M. Banbrook, S. McLaughlin and I. Mann; Speech Characterization and Synthesis by Nonlinear Methods, *IEEE Transactions on Speech and Audio Processing*; 7 (1), 1–17 (1999).
4. A. Behrman; Global and Local Dimensions of Vocal Dynamics, *The Journal of the Acoustical Society of America*; 105 (1), 432–443 (1999).
5. D. H. Klatt and L. C. Klatt; Analysis, Synthesis, and Perception of Voice Quality Variations among Female and Male Talkers, *The Journal of the Acoustical Society of America*; 87 (2), 820–857 (1990).
6. A. V. McCree and T. P. Barnwell III; A Mixed Excitation LPC Vocoder Model for Low Bit Rate Speech Coding, *IEEE Transactions on Speech and Audio Processing*; 3 (4), 242–250, (1995).
7. N. Aoki and T. Ifukube; Two $1/f$ Fluctuations in Sustained Phonation and Their Roles on Naturalness

- of Synthetic Voice, Proceedings of IEEE International Conference on Electronics, Circuits and Systems, 96TH8229 (1), 311–314 (1996).
8. K. Shiomi and S. Hirose; Fatigue and Drowsiness Predictor for Pilots and Air Traffic Controllers, The 45th Annual Air Traffic Control Association in their Conference Proceedings (2000).
9. 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).
10. J. D. Farmer and J. J. Sidorowich; Predicting Chaotic Time Series, *Physical Review Letters*; 59 (8), 845–848 (1987).
11. T. Schreiber and A. Schmitz; Improved Surrogate Data for Nonlinearity Tests, *Physical Review Letters*; 77 (4), 635–638 (1996).
12. R. Hegger, H. Kantz and T. Schreiber; Practical Implementation of Nonlinear Time Series Methods: The TISEAN Package, *Chaos*; 9 (2), 413–435 (1999).
13. T. Schreiber and A. Schmitz; Surrogate Time Series, *Physica D*; 142, 346–382 (2000).
14. F. Takens; Detecting Strange Attractors in Turbulence. In; D. A. Rand and L. S. Young (Eds.) *Dynamical Systems and Turbulence*, Lecture Notes in Mathematics; 898, Berlin, Springer-Verlag, 366–381, (1981).
15. M. Sano and Y. Sawada; Measurement of the Lyapunov Spectrum from a Chaotic Time Series, *Physical Review Letters*; 55 (10), 1082–1085 (1985).
16. A. M. Albano, A. Passamante and Mary Eileen Farrell; Using Higher-Order Correlations to Define an Embedding Window, *Physica D*; 54, 85–97 (1991).
17. T. Ikeguchi, M. Takakuwa and K. Aihara; Reconstructing Attractors with Higher Order Correlation, *Proceedings of International Symposium on Nonlinear Theory and Its Applications*, 2, 1149–1152 (1997).
18. M. B. Kennel, R. Brown and H. D. I. Abarbanel; Determining Embedding Dimension for Phase-Space Reconstruction Using a Geometrical Construction, *Physical Review A*; 45 (6), 3403–3411 (1992).
19. R. Wayland, D. Bromley, D. Pickett and A. Passamante; Recognizing Determinism in a Time Series, *Physical Review Letters*; 70 (5), 580–582 (1993).
20. Q. Yao and H. Tong; A Bootstrap Detection for Operational Determinism, *Physica D*; 115, 49–55 (1998).