Invariance in vowel systems

Masatoshi Funabashi[a)]
Fundamental Research Laboratory, Sony Computer Science Laboratories, Inc., Tokyo 141-0022, Japan

(Received 11 July 2014; revised 27 March 2015; accepted 9 April 2015)

This study applies information geometry of normal distribution to model Japanese vowels on the basis of the first and second formants. The distribution of Kullback-Leibler (KL) divergence and its decomposed components were investigated to reveal the statistical invariance in the vowel system. The results suggest that although significant variability exists in individual KL divergence distributions, the population distribution tends to converge into a specific log-normal distribution. This distribution can be considered as an invariant distribution for the standard-Japanese speaking population. Furthermore, it was revealed that the mean and variance components of KL divergence are linearly related in the population distribution. The significance of these invariant features is discussed. © *2015 Acoustical Society of America*. [http://dx.doi.org/10.1121/1.4919360]

[JFL] Pages: 2892–2900

I. INTRODUCTION

The sound analysis of vowel systems has been mainly focused on the time-series spectra, especially the parameters called formant. The first and second formants are known to give a distinctive feature to most vowels.[1]

Many linguists have investigated the difference or variation in vowel formants with respect to the speaker's physiological and social profiles such as their age, gender, occupation, community, etc.[2–6]

On the other hand, there have recently been studies on a more universalist approach with the use of various phonetic features in frequency space. The structural order of vowel composition enables us to integrate the notions of language as physical and cognitive systems, and some studies even refer to the foundation of a language faculty from its structural invariance.[7–9]

In any case, the frequency space representing vowel formants or cepstrum vectors is taken *a priori* as the only mathematical space in which to perform the analysis. However, the coordinate transformation conserving the sufficient statistics can be considered as a change of the observation method, and there is a particular kind of coordinate accessible to a strong statistical theory called information geometry.[10] Information geometry treats geometrical structure of probability distributions, based on the Fisher information matrix as Riemannian metric and the dual-flat connection that conforms to the nature of statistical theory. Information geometry is applied in various domains and increasingly showing its validity both in theory and application (for example, Ref. 11). Although applications of information geometry exist in speech recognition, the physical law preserved in the entity of the vowel system has been little investigated.[12,13]

In this article, we apply an information geometrical formalization to the five Japanese vowels formants and analyze them from a universalist point of view to find their invariant characteristics.

[a)]Electronic mail: masa_funabashi@csl.sony.co.jp

II. MODELING OF VOWEL SYSTEM WITH INFORMATION GEOMETRY

A. Sampling of vowel formants

The five conventional vowels of standard Japanese (defined as /a/, /e/, /i/, /o/, and /u/ in the Hepburn system) were recorded being pronounced with a monotonic accent and analyzed with the use of Praat software.[14] The first and second formants of each vowel were extracted for 500 steps with a 0.01 s lapse and 0.025 s window length for the short time Fourier transform. The data were obtained from 26 male and 29 female Japanese volunteers living in or around Tokyo, ages ranging between 20 and 69 years, and whose parents are all Japanese. The distributions of the five vowels in the first and second formants space can each be approximated with two-dimensional Gaussian distribution, as shown in Fig. 1.

Gaussian distribution belongs to the exponential family on which basic results of information geometry is valid. Representation of Gaussian distribution on the statistical manifold with Fisher information metric can incorporate both mean and variance parameters into the geometrical distance between distributions. For instance, a larger difference of mean values leads to a longer distance between distributions, while a larger variance will result in less distance, which represents the noise effect.[15] This property can be extended to the multi-variate Gaussian distribution in general, and it is a geometrically desirable property to investigate the discrepancy in vowel systems with information theoretical measures.

B. Model description

Here, we derive the expression of a two-dimensional Gaussian distribution as an exponential family with dual-affine coordinates $\theta = (\theta_1, \ldots, \theta_5)$ and $\eta = (\eta_1, \ldots, \eta_5)$. The Gaussian distribution $p(\mathbf{x})$ with two-dimensional continuous variables $\mathbf{x} = (x_1, x_2)^T$ can be defined as follows:

$$p(\mathbf{x}) = \frac{1}{2\pi\sqrt{|S|}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^T S^{-1}(\mathbf{x}-\mu)\right\}, \quad (1)$$

 0001-4966/2015/137(5)/2892/9/$30.00

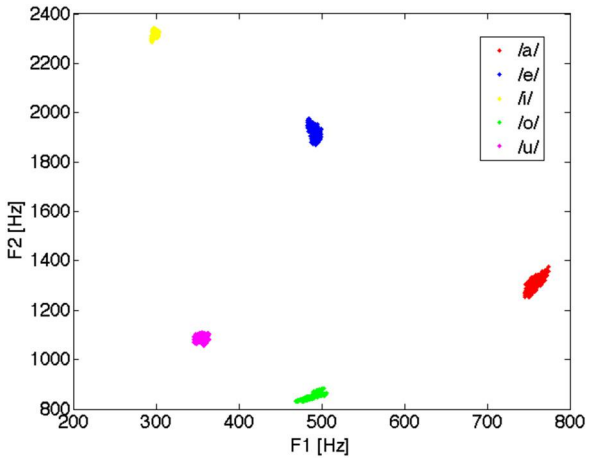

FIG. 1. (Color online) Example of $F1$–$F2$ distribution of five Japanese vowels. Horizontal axis: first formant frequency. Vertical axis: second formant frequency. Data of only one person is depicted.

where $\mu = (\mu_1, \mu_2)$ is the mean value vector, and

$$S = E[(\mathbf{x} - \mu)(\mathbf{x} - \mu)^T] = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$$

is the variance-covariance matrix of the variable. The superscript T means the transpose of vectors. Note that $\sigma_{12} = \sigma_{21}$ always holds. We then need five different functions of the parameters $\mu_1, \mu_2, \sigma_{11}, \sigma_{12} = \sigma_{21}, \sigma_{22}$ as coordinates to specify a distribution.

We first have the expression of Eq. (1) as an exponential family of distribution by the following variable and parameter transformations:

$$p(\mathbf{x}) = \exp \sum_{i=1}^{5} \theta_i F_i(\mathbf{x}) - \Psi(\theta), \tag{2}$$

where $F_1(\mathbf{x}) = x_1$, $F_2(\mathbf{x}) = x_2$, $F_3(\mathbf{x}) = x_1^2$, $F_4(\mathbf{x}) = x_2^2$, $F_5(\mathbf{x}) = x_1 x_2$, $\theta_1 = A(-2\sigma_{22}\mu_1 + 2\sigma_{12}\mu_2)$, $\theta_2 = A(2\sigma_{12}\mu_1 - 2\sigma_{11}\mu_2)$, $\theta_3 = A\sigma_{22}$, $\theta_4 = A\sigma_{11}$, $\theta_5 = A(-2\sigma_{12})$, the potential $\Psi(\theta) = \log(2\pi\sqrt{|S|}) - A(\sigma_{22}\mu_1^2 - 2\sigma_{12}\mu_1\mu_2 + \sigma_{11}\mu_2^2)$, and $A = -(1/2|S|)$.

We then define the new coordinate $\eta = (\eta_1, \ldots, \eta_5)$ that is dual to $\theta = (\theta_1, \ldots, \theta_5)$ and its corresponding potential $\Phi(\eta)$ as follows:

$$\eta_i = E[F_i(\mathbf{x})], \tag{3}$$

$$\Phi(\eta) = \sum_{i=1}^{5} \theta_i \eta_i - \Psi(\theta), \tag{4}$$

which gives in accordance with the formant mean and variance parameters

$$\begin{aligned} &\eta_1 = \mu_1, \quad \eta_2 = \mu_2, \quad \eta_3 = \sigma_{11} + \mu_1^2, \\ &\eta_4 = \sigma_{22} + \mu_2^2, \quad \eta_5 = \sigma_{12} + \mu_1\mu_2. \end{aligned} \tag{5}$$

Hence θ and η are dual affine coordinates. By introducing θ and η, the set of all $p(\mathbf{x})$ forms a dual-flat space with respect to the following Fisher information matrix (g_{ij}), (g^{ij}) as the Riemannian metric, and the e- and m-connection coefficients $\Gamma^{(1)}_{jik}$, $\Gamma^{(-1)}_{ijk}$, respectively,[10]

$$g_{ij} = \frac{\partial}{\partial\theta_i}\frac{\partial}{\partial\theta_j}\Psi(\theta), \tag{6}$$

$$g^{ij} = \frac{\partial}{\partial\eta_i}\frac{\partial}{\partial\eta_j}\Phi(\eta), \tag{7}$$

$$\Gamma^{(\alpha)}_{jik} = [ji;k] - \frac{\alpha}{2}T_{ijk}\,(\alpha = \pm 1), \tag{8}$$

where

$$[ji;k] = \frac{1}{2}\left(\frac{\partial}{\partial\theta_i}g_{jk} + \frac{\partial}{\partial\theta_j}g_{ik} - \frac{\partial}{\partial\theta_k}g_{ij}\right), \tag{9}$$

$$T_{jik} = E\left[\frac{\partial}{\partial\theta_i}\log p(\mathbf{x})\frac{\partial}{\partial\theta_j}\log p(\mathbf{x})\frac{\partial}{\partial\theta_k}\log p(\mathbf{x})\right]. \tag{10}$$

Fisher information is invariant under any invertible non-linear transformation of the variables. Therefore, using Fisher information metric to define a statistical manifold refers to the invariant structure that does not depend on the way of observation. One can freely choose the observation method as long as it conserves sufficient statistics. The definition of the dual affine connection $\Gamma^{(\pm 1)}_{jik}$ is essential in information geometry to define geodesics and the orthogonality between dual affine coordinates θ and η. Although standard Fisher distance is defined with Levi-Civita connection ($\alpha = 0$), significant connections in information geometry can only be found with $\alpha = \pm 1$ that conform to the flat projections on the dual coordinates.

C. Decomposition of Kullback-Leibler divergence to the first- and the second-order statistics

According to the definition of the Riemannian metric, information geometry provides the following theorem:

The coordinates $\theta_2 = (\theta_3, \theta_4, \theta_5)$ are orthogonal to the coordinates $\eta_1 = (\eta_1, \eta_2)$.

Therefore, we can compose the mixed orthogonal coordinates ζ as

$$\zeta = (\eta_1; \theta_2) = (\eta_1, \eta_2; \theta_3, \theta_4, \theta_5). \tag{11}$$

Since all parameters $\mu_1, \mu_2, \sigma_{11}, \sigma_{12} = \sigma_{21}, \sigma_{22}$ are included in ζ, the mixed coordinates are sufficient to specify a probability distribution.

We use the Kullback-Leibler (KL) divergence to measure the discrepancy between two vowels $v_1, v_2 \in \{/a/, /e/, /i/, /o/, /u/\}$. By denoting the probability distribution of the vowels v_1, v_2 as $p_{v_1}(\mathbf{x})$ and $p_{v_2}(\mathbf{x})$, respectively, the KL divergence $D[p_{v_1} : p_{v_2}]$ is defined as follows:

$$D[p_{v_1} : p_{v_2}] = \int\int p_{v_1}(\mathbf{x}) \log \frac{p_{v_1}(\mathbf{x})}{p_{v_2}(\mathbf{x})} dx_1 dx_2. \tag{12}$$

The KL divergence between continuous distributions is invariant under any continuous transformation of the parameters. Therefore, it is an invariant measure to quantify the discrepancy that does not depend either on the way of

observation or the function of perception, in case those are assumed to be a continuous transformation in general. In a different way, we can also calculate $D[p_{v_1} : p_{v_2}]$ with $\theta^{v_1} = (\theta_1^{v_1}, \ldots, \theta_5^{v_1})$ as the θ-coordinates of v_1, $\eta^{v_2} = (\eta_1^{v_2}, \ldots, \eta_5^{v_2})$ as the η-coordinates of v_2, and their corresponding potentials $\Psi^{v_1}(\theta^{v_1})$, $\Phi^{v_2}(\eta^{v_2})$,

$$D[p_{v_1} : p_{v_2}] = -\sum_{i=1}^{5} \theta_i^{v_2} \eta_i^{v_1} + \Psi^{v_2}(\theta^{v_2}) + \Phi^{v_1}(\eta^{v_1}). \quad (13)$$

Using the orthogonality between θ- and η-coordinates, we obtain the following decomposition of $D[p_{v_1} : p_{v_2}]$:

$$D[p_{v_1} : p_{v_2}] = D[p_{v_1} : p_{v_1 v_2}] + D[p_{v_1 v_2} : p_{v_2}], \quad (14)$$

where $p_{v_1 v_2}$ are given by $\zeta^{v_1 v_2} = (\eta_1^{v_1}; \theta_2^{v_2}) = (\eta_1^{v_1}, \eta_2^{v_1}; \theta_3^{v_2}, \theta_4^{v_2}, \theta_5^{v_2})$.

Since the coordinates $\theta_2^{v_2}$ include only the variance and covariance parameters, the term $D[p_{v_1} : p_{v_1 v_2}]$ represents the discrepancy in the second-order statistics of p_{v_2} from p_{v_1}, fixing the mean values μ^{v_1} as specified by p_{v_1}. Then, $D[p_{v_1 v_2} : p_{v_2}]$ represents the residual discrepancy purely in the mean values. This means that we are able not only to evaluate the discrepancy between the vowels but also to decompose its dependence into different orders of statistics. An intuitive explanation of this theorem is shown in Fig. 2 using one-dimensional Gaussian distributions.

The unknown η- and θ-coordinates of $p_{v_1 v_2}$ can be derived analytically from $\zeta^{v_1 v_2}$. This means each term of Eq. (14) is possible to calculate directly from the formants data, without calling for a numerical method. The above theoretical formalization plays an essential basis to assure the accuracy of calculation. Other numerical pitfalls exist in a zero parameter problem, such as the case $|S| = 0$ in the denominator, which was confirmed not to be the case with the used dataset.

For simplicity, we hereinafter call the logarithm of the first term the variance component and the logarithm of the second term the mean value component of KL divergence.

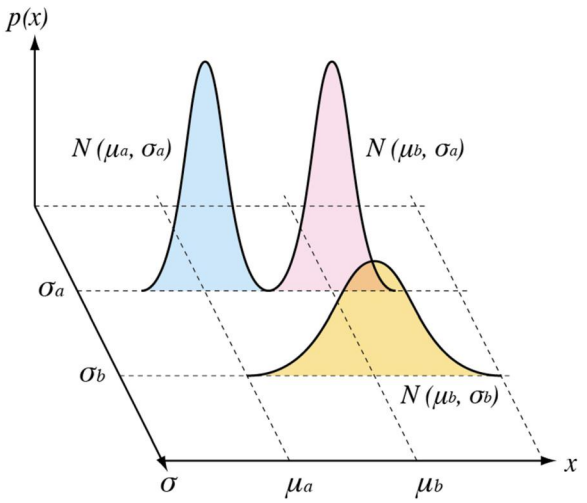

FIG. 2. (Color online) Intuitive explanation of the Pythagorean theorem. One-dimensional Gaussian distributions $N(\mu, \sigma)$ are taken as an example. The KL divergence between $N(\mu_a, \sigma_a)$ and $N(\mu_b, \sigma_b)$ can be decomposed into the mean and variance discrepancy elements: $D[N(\mu_a, \sigma_a) : N(\mu_b, \sigma_b)] = D[(\mu_a, \sigma)_a : N(\mu_b, \sigma_a)] + D[(\mu_b, \sigma_a) : N(\mu_b, \sigma_b)]$.

We also define the logarithmic variance/mean component ratio α as $\alpha = \log D[p_{v_1} : p_{v_1 v_2}] / \log D[p_{v_1 v_2} : p_{v_2}]$ for further correlational analysis. Considering the perceptual difference in these parameters, this decomposition can find phonetical meaning to study the equilibrium of the vowel system (see Sec. III). Hence, in KL divergence, the temporal fluctuation of the formants are also taken into account, while in conventional cepstrum analysis, for instance, the definition of distance between vowels is instantaneous and does not consider time-averaged higher-order statistics. In other words, the Euclidian distance between cepstrum vectors is the distance between stochastic variables, not between their probability distributions.

III. RESULT AND DISCUSSION

A. Distribution of KL divergence between vowels

We calculated the KL divergence between each set of vowels for each person and obtained a histogram summing up the sample population. This can be considered as the inter-individual distribution of Japanese vowels. The log-normal distribution well fits both the male and female populations, as shown in Figs. 3 and 4. The fitting of the histogram was performed by estimating the mean value and the unbiased variance.

We also calculated the KL divergence between each set of vowels of one male and obtained a histogram. The calculation was performed from 50 samples of the same individual. This can be considered as the intra-individual distribution of Japanese vowels. The individual histogram also well follows the log-normal distribution, as shown in Fig. 5.

Next, we consider the relationship between the population and the individual distribution. Assuming that these distributions follow the log-normal distribution, the estimated probability density of KL divergence is depicted in Fig. 6.

We first performed the F-test to reveal whether these distributions significantly vary or not. The two-sided significance level was set to a conventional value, 0.25. The results are listed in Table I. The individual distribution has significantly different variance from that of the male and female population distributions. On the other hand, the male and female population distributions do not significantly differ.

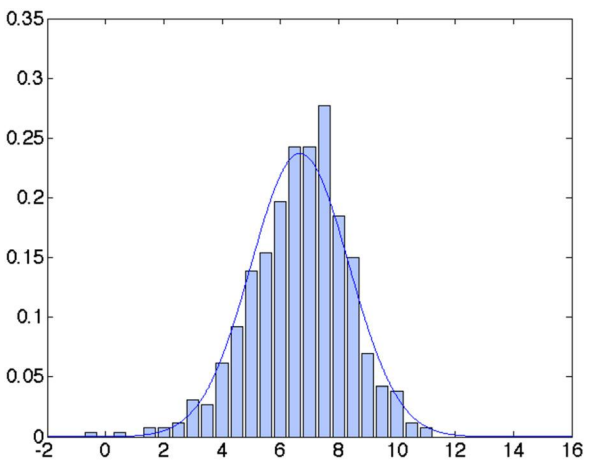

FIG. 3. (Color online) Distribution of KL divergence between vowels in population of 26 males. Horizontal axis: logarithm of KL divergence. Vertical axis: probability density.

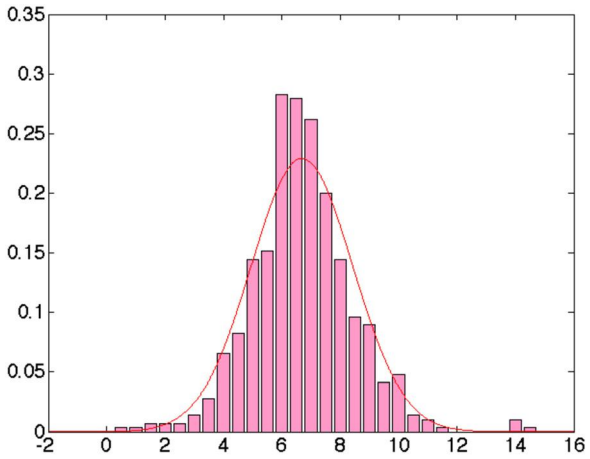

FIG. 4. (Color online) Distribution of KL divergence between vowels in population of 29 females. Horizontal axis: logarithm of KL divergence. Vertical axis: probability density.

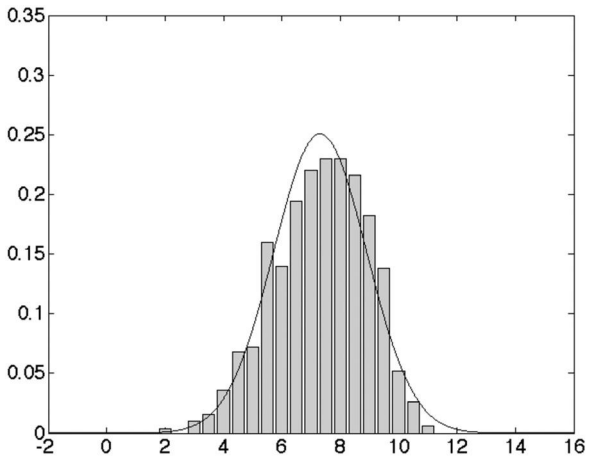

FIG. 5. Distribution of KL divergence between vowels in 50 samples from 1 male (provisionally called individual B). Horizontal axis: logarithm of KL divergence. Vertical axis: probability density.

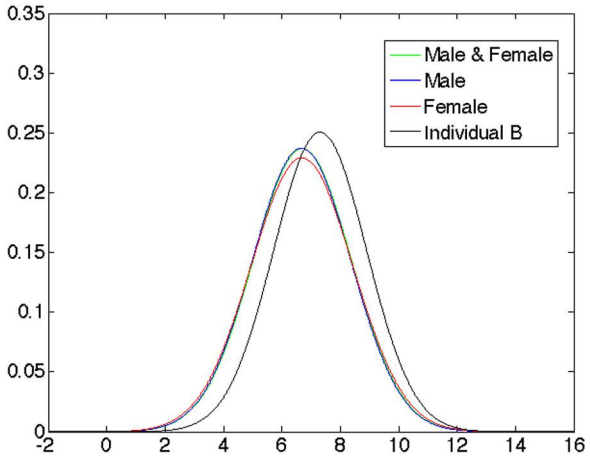

FIG. 6. (Color online) Estimated probability density function of KL divergence between vowels. The estimation was performed by calculating the mean value and the unbiased variance of each distribution. Horizontal axis: logarithm of KL divergence. Vertical axis: probability density.

TABLE I. *F*-test between population and individual distribution of KL divergence. The *F*-test is designed to test if two population variances are equal. The *F*-values, their left and right critical values of the significance level 0.25, and the truth value of the null hypothesis are listed for each combination of the distributions. I, M, and F denote the individual B, the male population, and the female population, respectively.

	I vs M and F	I vs M	I vs F	M vs F
F-Value	0.8935	0.8955	0.8357	0.9332
Left critical value	0.9312	0.9166	0.9192	0.9061
Right critical value	1.0736	1.0932	1.0896	1.1032
Null hypothesis	False	False	False	True

On the basis of the results of the *F*-test, we performed a *t*-test to investigate whether the mean values of these distributions significantly differ or not. The two-sided significance level was set to 0.05. The results are listed in Table II. Again, the individual distribution has a mean value significantly different from that of the male and female population distributions, while there is no significant difference between males and females.

These results imply that the individual distributions have statistically significant fluctuation. However, since the population distribution is equivalent to the mixture of individual distributions, the fact that the population statistics converge to a particular distribution gives a collective limit to the individual fluctuation. This relationship is not trivial when considering the fact that the mixture of different normal distributions is not restricted to another normal distribution. Mathematically, it can even approximate any continuous and differentiable function. The convergence of population distribution is proof of a collective order in individual variability. This reveals a hierarchical structure of the KL divergence distributions between the individual and population. The individual distributions are distributed in accordance with the population distribution, and despite the restrictions inside population statistics, each one is still able to express its proper variation. Therefore, the population distribution can be considered as an invariant distribution for the standard-Japanese speaking population.

The coincidence of male and female distributions strongly supports the notion of invariance above the superficial phonetic variation. The gender difference is often studied to reveal distinctive features between the two groups,[16–21] though the converged distribution implies the existence of mutual order regardless of gender profile.

Considering the acquisition process of vowel sounds, this structure may reflect the learning process, because the individual statistics are collectively bounded by the

TABLE II. *t*-Test between population and individual distribution of KL divergence. Depending on the results of the *F*-test, the first three columns are the results of the Welch's *t*-test, while the last column is that of the student's *t*-test. The *p*-values and the truth value of the null hypothesis with respect to the significance level 0.05 are listed for each combination of the distributions. I, M, and F denote the individual B, the male population, and the female population, respectively.

	I vs M and F	I vs M	I vs F	M vs F
p-Value	3.0731×10^{-17}	1.9563×10^{-12}	2.6790×10^{-12}	0.8965
Null hypothesis	False	False	False	True

population one. In this sense, the population distribution also makes part of the structural invariance of the vowel system, which links the individual perception to the collective definition of the Japanese vowel system.[7]

B. Relationship between the mean value component and the variance component of KL divergence

It has been revealed that the perception of each vowel largely depends on the first and second formant frequencies.[1,2] In this sense, the mean value of each vowel's formant is essential to recognize which vowel it is. In fact, the five vowels in most of our experimental data form distinctive clusters in $F1$–$F2$ space (Fig. 7), as reported widely in vowel systems (for example, Ref. 13).

On the other hand, the variances represent the fluctuation range of the formants, and it gives an additional phonetic feature. The difference in the formants' variances can be recognized as a part of the differences in so-called voice quality. Indeed, certain fluctuation of the formants is considered to relate to the naturalness of vowel sounds.[4] The frequency variance is also associated with the naturalness of tone timbre.[5] Furthermore, distributions of vowels in variance parameter space show certain localization of each vowel, which may give another distinctive feature (Fig. 8).

In our setting, the differences between vowel distributions depend on these parameters, the mean values and the variance-covariance matrix, that encode qualitatively different perceptual information. Since the KL divergence between two vowels gives mathematical discrepancy between the two distributions depending on the parameters that affect our perception, it is natural to consider that it also reflects our cognitive distinctiveness between these vowels. The decomposition of the KL divergence into the mean value and the variance component enables us to investigate whether there exists a balance related to the distinctiveness between them, when comparing two vowels.

The relationship between the mean value and the variance/mean component ratio α are plotted with linear

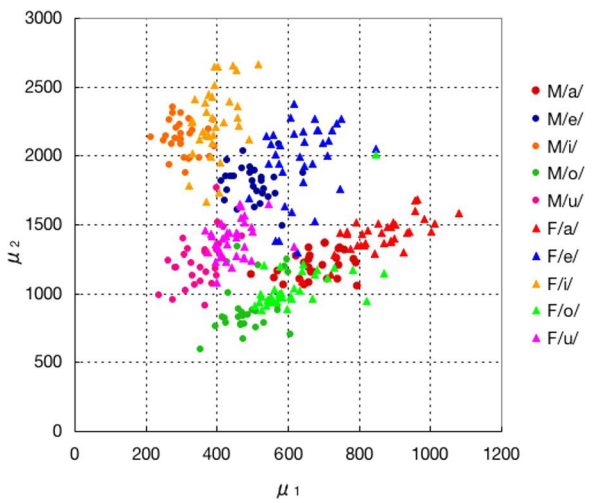

FIG. 7. (Color online) Population distribution of five vowels' $F1$ and $F2$ mean values. M and F denote male and female, respectively. Each vowel forms cluster in its characteristic frequency region.

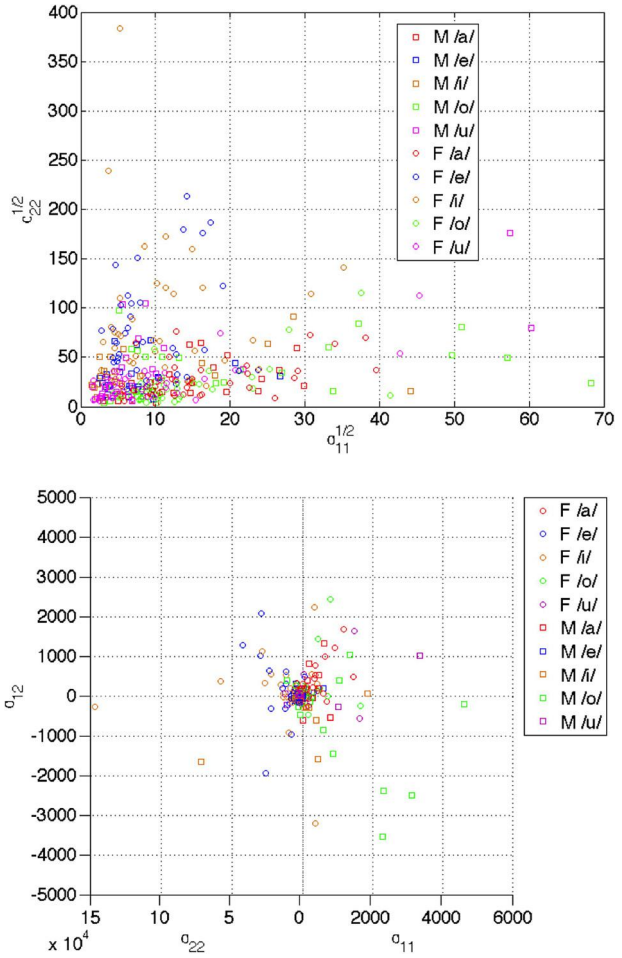

FIG. 8. (Color online) Population distribution of five vowels' $F1$ and $F2$ variances. M and F denote male and female, respectively. The top figure is the distribution on $\sqrt{\sigma_{11}} - \sqrt{\sigma_{22}}$ plane. The bottom figure is the side view of three-dimensional plot. /a/, /o/, and /u/ are localized in relatively low $F2$ variance region compared with $F1$ variance, while /e/ and /i/ are in relatively low $F1$ variance region. The covariance axis does not seem to give a distinctive feature among vowels.

regression in Fig. 9. The correlation coefficients between the mean value component, the variance component, and α are listed in Table III. In each distribution, the variance component is correlated to the mean value component to a certain extent. However, if we compare the mean component to α, less correlation exists. Specifically, the sum of the male and female population shows little correlation. This fact implies that α is the invariant ratio between the mean value and the variance component in population distribution. Indeed, α is symmetrically distributed with a sharp peak and fits well with normal distribution, as depicted in Fig. 10. The sharp symmetric peak supports the invariance of α in population distribution.

More precisely, the variance component consists of two elements: one is linearly proportional to the mean value component, and the other is not linear. The linear element is canceled out when measuring the correlation with α. The latter certainly exists in individual distribution as shown in Fig. 11, though it is consistent in any distribution in which the correlations between the mean component and α show only weak correlation. Therefore, we deduce that the α is an

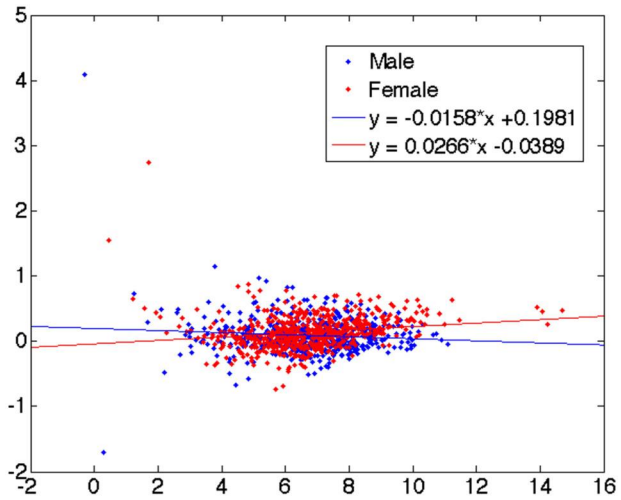

FIG. 9. (Color online) Relationship between mean value component vs variance/mean component ratio α of KL divergence between vowels in populations of 26 males and 29 females. Horizontal axis: mean value component. Vertical axis: variance/mean component ratio α. The lines are the linear regressions.

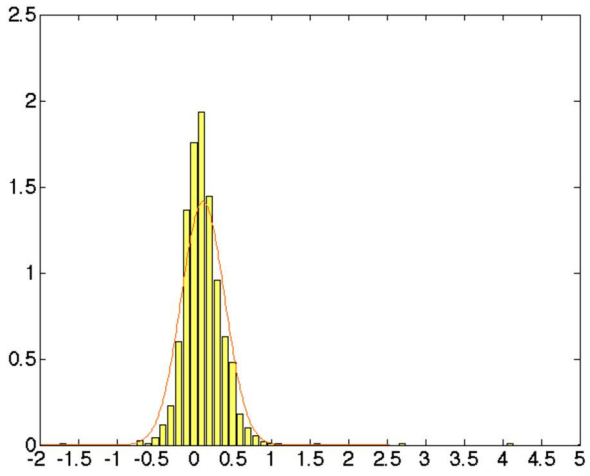

FIG. 10. (Color online) Distribution of variance/mean component ratio α in populations of 26 males and 29 females. Horizontal axis: value of α. Vertical axis: probability density.

invariant feature in population distribution, while it accepts a certain fluctuation in each individual that cancels each other out at the population level.

C. Vowel combination-wise distribution of KL divergence

We further investigate the content of log-normal distribution between vowels by decomposing it into vowel combination-wise distribution. Figures 12 and 13 show the distribution of KL divergence for different combinations of vowels. The combination-wise distributions also show the tendency to fit log-normal distribution with various mean and variance, but they contain certain fluctuation. This may be due to the combinatorial decrease of the sample number in combination-wise distribution, since the combination /a/-/e/, for example, is only 1/20 of all possible combinations between the five vowels. The fact that the fluctuation reduces by taking larger combinations such as /a/-/e//i//o//u/ also supports this notion. The distribution between five vowels of the same sample number order also show large fluctuation, as shown in Fig. 14. Although more accurate forms of the vowel combination-wise distributions can only be verified by augmenting the sample number, the circumstantial evidence suggests a hierarchical structure between the individual and combination-wise distributions similar to that of the population and individual ones: The vowel combination-wise distributions seem to follow different log-normal distributions, under global constraints of the individual distribution.

TABLE III. Correlation coefficients between the logarithms of the mean value component, the variance component, and the variance/mean component ratio α of KL divergence. For simplicity, the components are referred to as Mean, Variance, and α. The results of the population and individual distributions are listed. I, M, and F denote the individual B, the male population, and the female population, respectively.

	M and F	M	F	I
Mean vs variance	0.3242	0.1050	0.4923	0.3986
Mean vs α	0.0444	−0.0890	0.1749	0.2365

D. The origin of log-normal distribution in vowel system: Weber-Fechner law in η coordinates

We investigate the origin of the observed log-normal distribution by analyzing the variation of vowels distribution in η coordinates.

Independent multiplicative processes are known to converge to a log-normal distribution[22,23] The distributions of five vowels in η coordinates do not show correlation. If the variance of each vowel in η coordinates is a multiplicative noise, the KL divergence between any combination of five vowels naturally converges to a log-normal distribution, because its definition is based on the linear combination of η coordinates [see Eq. (13)].

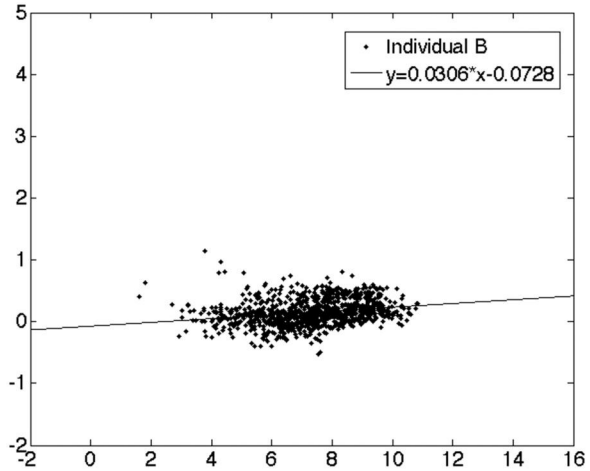

FIG. 11. Relationship between mean value component vs variance/mean component ratio α of KL divergence between vowels for one male (provisionally called individual B). Horizontal axis: mean value component. Vertical axis: variance/mean component ratio α. The line is the linear regression.

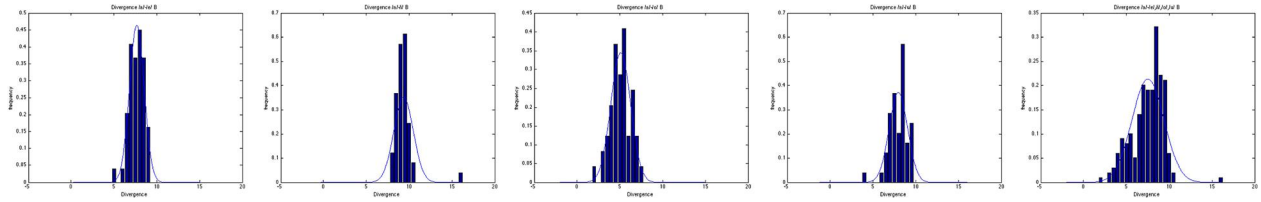

FIG. 12. (Color online) Vowel combination-wise distribution of KL divergence of individual B. From left to right: distribution of /a/-/e/, /a/-/i/, /a/-/o/, /a/-/u/, and /a/-/e//i//o//u/.

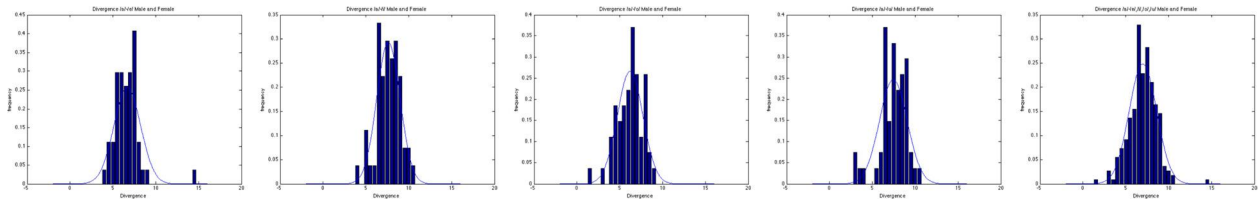

FIG. 13. (Color online) Vowel combination-wise distribution of KL divergence of males and females. From left to right: distribution of /a/-/e/, /a/-/i/, /a/-/o/, /a/-/u/, and /a/-/e//i//o//u/.

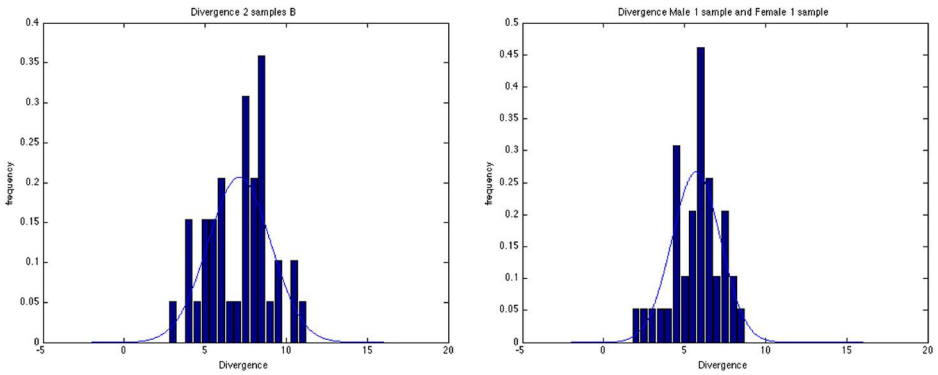

FIG. 14. (Color online) Distribution of KL divergence between five vowels with reduced sample number. Left: two samples of individual B (40 combinations of KL divergence). Right: one sample each from both male and female populations (40 combinations of KL divergence).

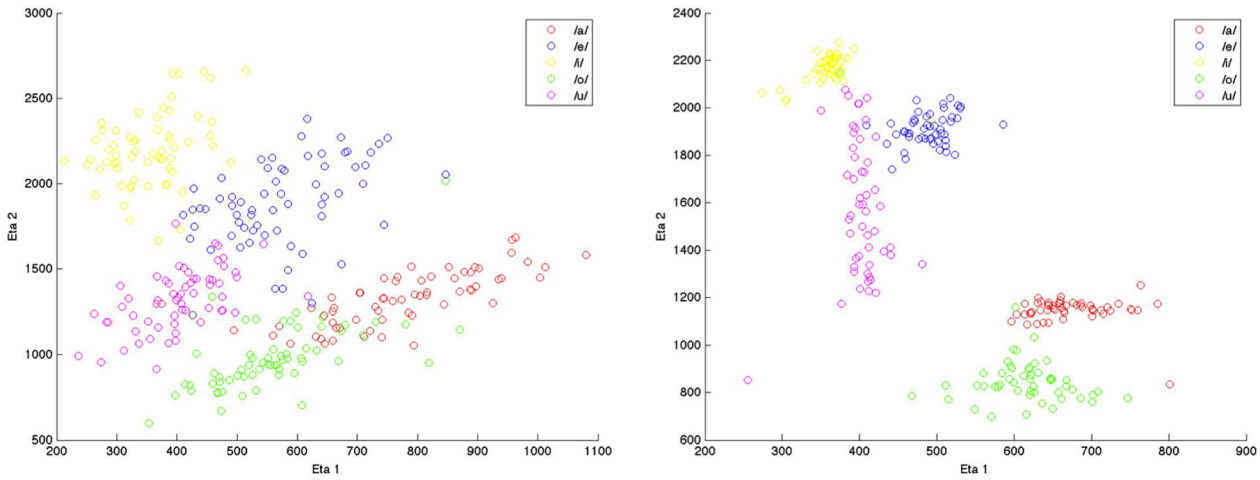

FIG. 15. (Color online) Distribution of five Japanese vowels in the $\eta_1 - \eta_2$ plane. Left: Male and female. Right: Individual B.

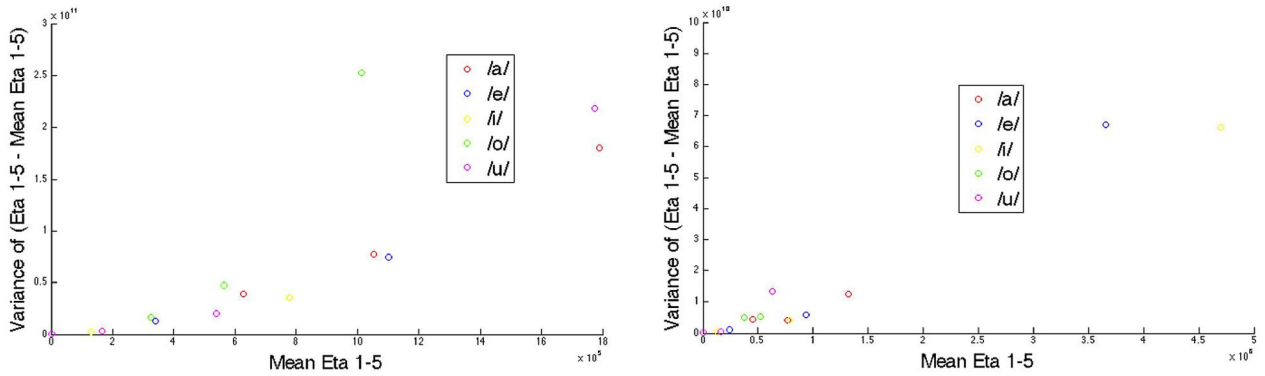

FIG. 16. (Color online) Mean vs additive variance of five Japanese vowel distributions in η coordinates. Left: Male and female. Right: Individual B. The proportional increase of the additive variance implies multiplicative noise.

Figure 15 shows the distributions of five vowels projected in the $\eta_1 - \eta_2$ plane. To verify the existence of multiplicative noise, we calculate the following two kinds of variances. The first is the simple variance of each vowel in η coordinates, namely, the additive variance. The second is also the variance of each vowel in η coordinates, but divided by their mean value, namely, the multiplicative variance. If the variance of the vowels is multiplicative, the additive variance increases proportionally with respect to the mean value in η coordinates, while the multiplicative variance remains constant.

Figures 16 and 17 show the results that support the existence of multiplicative variation. The multiplicative tendency exists both in population and individual distributions in all dimensions of η coordinates. Therefore, the distribution on η coordinates implies that the origin of the observed log-normal distributions is grounded to its multiplicative variation. Note that the η_1 and η_2 coordinates correspond to $F1$ and $F2$ values, respectively.

The multiplicative variation in frequency space has a perceptual meaning known as Weber-Fechner law: For notes spaced equally apart to the human ear, the frequencies are related by a multiplicative factor. Humans hear pitch in a logarithmic or geometric ratio-based fashion. For this reason, musical scales are always based on geometric relationships. In the case of the vowel system, this relationship would support the distinctiveness of each vowel. The observed log-normal distribution can be considered as the result of repulsive localization of each vowel according to the constant perceptual distance between them.

The multiplicative factor is also interesting when we assume a constant degree of accuracy in voice control. If the control precision of vocal tract is constant, the vocalization is naturally associated with multiplicative noise with respect to the produced frequency.

E. Relationship to ecological linguistics: Invariants of Gibson expanded

We started the analysis from the universalist point of view, seeking the common structure of the vowel system that would support our perception of harmonized resonance in human language. In visual perception, Gibson insisted that geometrical invariants in optical flow are the foundation of perceptual significance.[24] The observed invariant relationships clearly relate to the principle of Gibson's invariants, but outreach simple pictures such as formants localization[18] and invariant quantity under affine transformation.[7] The observed invariance has a hierarchical structure between individuals and the population, different orders of statistics, and possibly the single pair and the whole combination of the vowels. One way to explain such a complex structure was introduced in reference to the Weber-Fechner law, but the convergence of the population distribution and the relationship between different orders of statistics remain untouched. These relationships may become a phonetic expansion of Gibson's invariants, which reflect the complexity of phonetic perception deeply linked to

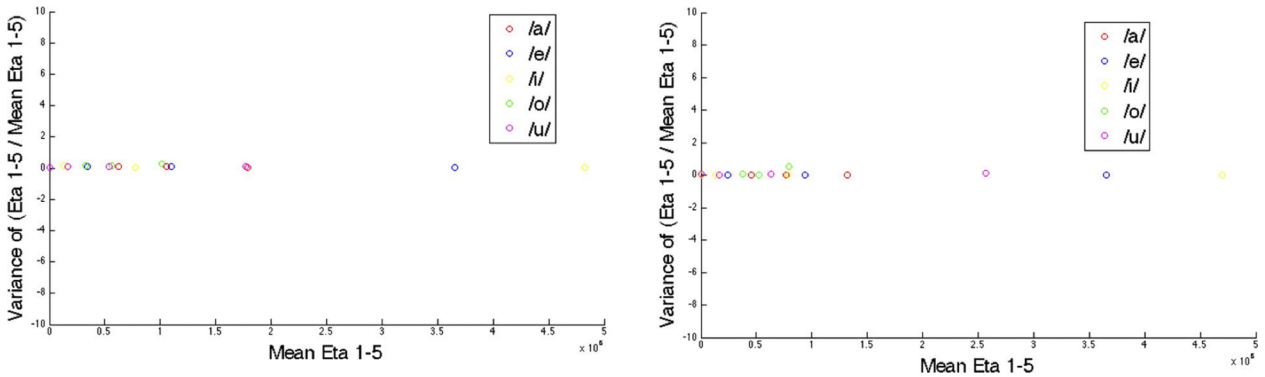

FIG. 17. (Color online) Mean vs multiplicative variance of five Japanese vowel distributions in η coordinates. Left: Male and female. Right: Individual B. The invariance of the multiplicative variance implies multiplicative noise.

our language faculty. A parallel study on the acquisition of the vowel system in children would be necessary to clarify the development of these features.[19,20]

It is of further interest how such invariance affects our perception of vowels, especially when they are perturbed. A synthetic approach to the vowel system controlling the discovered invariant relationships would be fruitful to further address these questions.

IV. CONCLUSION

We investigated the distributions of KL divergence between five Japanese vowels and insisted that the population log-normal distribution and the variance/mean component ratio α are invariant features. The hierarchical relationship between the population distribution, individual distribution, and vowel combination-wise distributions are also investigated. The origin of log-normal distribution is shown to be based on the multiplicative variation in formant frequency.

It is of further interest whether such invariance can also be observed in other vowel systems or by simply increasing the heterogeneity of the linguistic profile in a sample Japanese population. As well as statistical invariance, the geometrical composition of each vowel in the dual coordinate space remains to be analyzed. Information geometry already provides strong theoretical and numerical framework to analyze geometric composition related to the invariance of distributions such as centroids, which is compatible to multivariate Gaussian distribution.[25] Further developmental and multilingual comparative studies will be needed to relate the discovered invariance of vowel system to our cognitive mechanism.

ACKNOWLEDGMENTS

The author thanks all of the Hippo Family Club members for bringing many insights through their daily activities. Akira Tamaki, Akiko Okada, Eiichiro Osawa, Hiroyasu Kawakami, Chizuru Funabashi, Makie Murao, and Yukiko Sakakibara have especially supported the data analysis. This study was partially supported by the long-term study abroad support program of the University of Tokyo and by the French government (Promotion Simone de Beauvoir).

[1]S. Furui, *Digital Speech Processing, Synthesis, and Recognition* (Marcel Dekker, Inc., New York, 1989), p. 390.
[2]T. Chiba and M. Kajiyama, *Vowel: Its Nature and Structure* (Phonetic Society of Japan, Tokyo, Japan, 1958), p. 236.
[3]R. E. Turner and R. D. Patterson, "An analysis of the size information in classical formant data: Peterson and Barney (1952) Revisited," J. Acoust. Soc. Jpn. **33**, 585–589 (2003).
[4]D. F. 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).
[5]M. Saitou and T. Tsumura, "Relation between timbre and depth of frequency fluctuation—Comparison between listening condition through headphone and free field," J. Acoust. Soc. Jpn. **90**(5), 3–10 (1990).
[6]W. Labov, "La transmission des changements linguistiques" ("The transmission of linguistic changes"), Languages **26**(108), 16–33 (1992) (in French).
[7]N. Minematsu, T. Nishimura, K. Nishinari, and K. Sakuraba, "Theorem of the invariant structure and its derivation of speech Gestalt," in *Speech Recognition and Intrinsic Variation (SRIV2006)*, Toulouse, France (May 20, 2006), pp. 47–52.
[8]S. Takano and S. Nakamura, "Multilingual environment and natural acquisition of language," AIP Conf. Proc. **519**, 785–796 (2000).
[9]B. de Boer, *The Origins of Vowel Systems* (Oxford University Press, Oxford, UK, 2001), 184 p.
[10]S. Amari and H. Nagaoka, "Method of information geometry," in *Translations of Mathematical Monograph*, Vol. 191 (Oxford University Press, Oxford, UK, 2000), 206 pp.
[11]G. Verdoolaege (editor), Entropy Special Issue: "Information geometry," *Entropy*, MDPI AG, Basel, Switzerland (2014), http://www.mdpi.com/journal/entropy/special_issues/information-geometry (Last viewed March 31, 2015).
[12]M. K. Sonmez, "Information geometry of topology preserving adaptation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP* (2000), Vol. 6, pp. 3743–3746.
[13]A. Gunawardana, "The information geometry of EM variants for speech and image processing," Ph.D. dissertation, The Johns Hopkins University, Baltimore, MD, 2001.
[14]http://www.fon.hum.uva.nl/praat/ (Last viewed July 8, 2014).
[15]S. I. R. Costa, S. A. Santos, and J. E. Strapasson, "Fisher information distance: A geometrical reading," Discrete Appl. Math., available online http://dx.doi.org/10.1016/j.dam.2014.10.004.
[16]A. P. Simpson, "Gender-specific articulatory-acoustic relations in vowel sequences," J. Phonetics **30**(3), 417–435 (2002).
[17]A. P. Simpson, "Dynamic consequences of differences in male and female vocal tract dimensions," J. Acoust. Soc. Am. **109**(5), 2153–2164 (2001).
[18]L. Menard, J.-L. Schwarz, and J. Aubin, "Invariance and variability in the production of height feature in French vowels," Speech Commun. **50**(1), 14–28 (2008).
[19]U. G. Goldstein, "An articulatory model for the vocal tracts of growing children," Ph.D. dissertation, Massachusetts Institute of Technology (1980).
[20]S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," J. Acoust. Soc. Am. **105**(3), 1455–1468 (1999).
[21]Y. Samuelsson, "Gender effects on phonetic variation and speaking styles. A literature study," GSLT Speech Technology Term Paper, pp. 1–8 (2006).
[22]*Lognormal Distributions: Theory and Applications*, edited by E. L. Crow and K. Shimizu (CRC Press, Boca Raton, FL, 1987), 387 pp.
[23]M. Mitzenmacher, "A brief history of generative models for power law and lognormal distributions," Internet Math. **1**(2), 226–251 (2004).
[24]J. J. Gibson, *The Ecological Approach to Visual Perception* (Houghton Mifflin Harcourt, Boston, MA, 1979), 350 pp.
[25]F. Nielsen and R. Nock, "Sided and symmetrized Bregman centroids," IEEE Trans. Inf. Theory **55**, 2882–2904 (2009).