PAPER

Rotational Properties of Vocal Tract Length Difference in Cepstral Space

Daisuke Saito[1], Nobuaki Minematsu[2], and Keikichi Hirose[2]

[1]Graduate School of Engineering, The University of Tokyo
[2]Graduate School of Information Science and Technology, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan
{dsk_saito,mine,hirose}@gavo.t.u-tokyo.ac.jp

Abstract

In this paper, we prove that the direction of cepstrum vectors strongly depends on vocal tract length and that this dependency is represented as rotation in a cepstrum space. In speech recognition studies, vocal tract length normalization (VTLN) techniques are widely used to cancel age- and gender-difference. In VTLN, a frequency warping is often carried out and it can be modeled as a linear transform in a cepstrum space; $\hat{\boldsymbol{c}}=\boldsymbol{A}\boldsymbol{c}$. In this study, the geometric properties of this transformation matrix $\boldsymbol{A}$ are made clear using n dimensional geometry and it is shown that the matrix can be approximated as rotation matrix. Further, for better approximation, a new method is proposed. Namely, using eigenvalues of $\boldsymbol{A}$, its quasi-rotational distortion is factorized into multiple true rotation operations and multiple magnification operations. This decomposition resolves the intrinsic ambiguity of the rotation angle based on the inner product, and it describes the detailed geometrical properties of the transformation caused by vocal tract length normalization. Experimental results using real and resynthesized speech samples demonstrate that the difference of cepstrum vectors extracted from different speakers is represented as rotation and magnification, and that the decomposition based on eigenvalues can capture it precisely.

Keywords: frequency warping, rotation matrix, vocal tract length, eigenvalue, rotational plane

1. Introduction

Speech includes rich information. However, the richness of information sometimes influences the accuracy of speech application. Information in speech can be divided into three kinds; linguistic, para-linguistic and non-linguistic information. Usually, a speech application focuses on only one of them. For example, speech recognition systems aim to extract linguistic information, and speaker recognition/verification systems are developed to extract non-linguistic information, i.e. speaker information. To focus on the desired information exclusively, generally, statistical approaches are often adopted to hide the other kinds of information. In every speech application, a feature vector is used to characterize the focused information well. However, this feature vector is modified easily due to the unfocused information. For example, an MFCC vector of /a/ is different between a male and a female.

In the case of speech recognition, speech acoustics vary due to differences in gender, age, microphone, room, lines, and a variety of factors. These factors strongly influence the accuracy of speech recognition. To deal with these variations, usually, thousands of speakers in different conditions are prepared to train acoustic models of the individual phonemes; called speaker-independent (SI) system. However, the recognition accuracy of SI systems is sometimes very low for certain individuals, or conditions. It means that the SI systems are not really SI.

To overcome the above mismatch problems caused by acoustical variations, feature normalization has been used in many systems. Feature normalization techniques can be divided into two approaches; one based on subtraction or taking differential and the other based on transformation. Cepstrum mean normalization (CMN) and the use of Δcepstrums correspond to the former, and vocal tract length normalization (VTLN) to the latter.

In CMN, the long-term average of the cepstrum is subtracted from each cepstrum frame [1]. This helps eliminate changes created not only by differences among individuals, but also by channel differences. Δcepstrums are calculated based on the first derivative of cepstral features, and they also have effects of cancelling some static mismatches.

VTLN techniques are widely used to cancel the difference of vocal tract length (VTL) [2]. In VTLN, the transformation matrix in a cepstrum space is estimated and used to transform the VTL of an input speaker to a predefined value. In this paper, we are interested in the transformation matrix, whose geometrical properties have not been well discussed. We mathematically and experimentally investigate how the transformation matrix influences cepstrum vectors and their Δs and $\Delta\Delta$s. If some tractable properties are found in the

geometrical aspect of the modification, it will be possible to exclude the unfocused information from speech features appropriately so that a collection of a huge amount of data may not be needed [3].

This paper proves that the distortion caused by VTL difference can be approximated to rotational properties in a cepstral space. To reveal it, the definition of rotation matrix and the transformation matrix in VTLN are compared, and their similarity is focused on. In addition, this quasi-rotational properties are factorized into multiple true rotational operations and multiple magnification modifications, by diagonalizing the transformation matrix with eigenvalues. In order to verify these properties of the transformation caused by VTL difference, experiments using resynthesized speech samples using STRAIGHT [4] and real speech samples are carried out. We also conduct some other experimental investigations on correlations between the new features, rotation and magnification, and speaker information. Although the discussion in this paper is mainly based on the properties of the transformation matrix $\boldsymbol{A}$ in VTLN, revealed properties may be useful for representing some speaker information.

The paper is organized as follows. In Section 2, representation of difference in vocal tract length using a frequency warping and its linear modeling is described. In Section 3, we prove that the transformation matrix for VTLN is approximated as rotation matrix. In Section 4, we describe the decomposition of the rotational properties based on eigenvalue analysis. Section 5 presents experimental results using resynthesized and real speech samples. Finally, we summarize this paper in Section 6.

2. Difference in VTL and its effects

2.1 Frequency warping

In VTLN, the distortion caused by VTL differences is often modeled by a warping function in a spectrum space. Here, we adopt a first order all-pass transform function, which is formulated as

$$\hat{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, \;\; z = e^{j\omega}, \;\; \hat{z} = e^{j\hat{\omega}}, \tag{1}$$

where α is a warping parameter and $|\alpha| < 1$; ω and $\hat{\omega}$ are frequencies before and after transformation, respectively. In the case of $\alpha < 0$, formants are transformed to be lower and the VTL longer. $\alpha > 0$ realizes the opposite effect. Figure 1 shows a few examples of warping functions.

2.2 Linear modeling of frequency warping

We now describe a frequency warping by a linear transformation. Emori [6] converted a frequency warping of Equation 1 to a linear transformation in a cepstrum space. If power coefficients (c_0 and $\hat{c}_0$) are excluded, Equation 1 can be expressed as

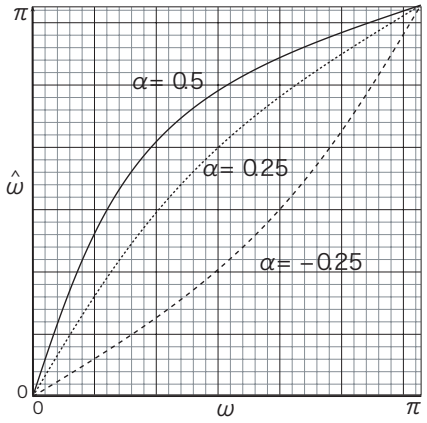


Figure 1: Examples of frequency warping functions for different values of α. $\alpha < 0$ transforms VTL longer and $\alpha > 0$ does VTL shorter. These functions are also used for implementation of mel-frequency warping [5].

$$\hat{\boldsymbol{c}} = \boldsymbol{A}\,\boldsymbol{c}, \tag{2}$$

where

$$\hat{\boldsymbol{c}} = [\hat{c}_1\ \hat{c}_2\ \hat{c}_3\ \hat{c}_4 \cdots]^\top,$$
$$\boldsymbol{A} = \begin{pmatrix} 1-\alpha^2 & 2\alpha-2\alpha^3 & \cdots & \cdots \\ -\alpha+\alpha^3 & 1-4\alpha^2+3\alpha^4 & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix},$$
$$\boldsymbol{c} = [c_1\ c_2\ c_3\ c_4 \cdots]^\top. \tag{3}$$

$(\cdot)^\top$ denotes the transpose of a vector or a matrix. From Pitz *et al.* [7], the element a_{ij} of $\boldsymbol{A}$ can be written using α as

$$a_{ij} = \frac{1}{(j-1)!} \sum_{m=m_0}^{j} \binom{j}{m} \frac{(m+i-1)!}{(m+i-j)!} (-\alpha)^{(m+i-j)} \alpha^m, \tag{4}$$

where m_0=max$(0, j-i)$ and

$$\binom{j}{m} = \begin{cases} {}_jC_m & (j \geq m) \\ 0 & (j < m). \end{cases} \tag{5}$$

3. Rotation in a cepstrum space

3.1 Rotation in a two dimensional cepstrum space

In this section, we discuss the properties of matrix $\boldsymbol{A}$ in Equation 3 geometrically. To facilitate the discussion, at first, we focus only on the first and second dimensions of the cepstrum space. Then, the discussion will be expanded into n dimensions.

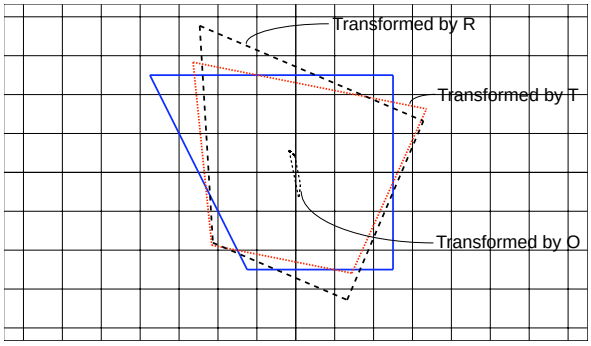


Figure 2: Effects of transformations of $\boldsymbol{T}$, $\boldsymbol{R}$, and $\boldsymbol{O}$ for $\alpha = 0.2$. A trapezoid is rotated clockwise after transformation by $\boldsymbol{T}$, and $\boldsymbol{O}$ has a very small influence.

In the two dimensional space, Equation 2 is

$$\begin{pmatrix} \hat{c}_1 \\ \hat{c}_2 \end{pmatrix} = \begin{pmatrix} 1-\alpha^2 & 2\alpha-2\alpha^3 \\ -\alpha+\alpha^3 & 1-4\alpha^2+3\alpha^4 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}. \quad (6)$$

We call the transformation matrix in Equation 6 $\boldsymbol{T}$, and $\boldsymbol{T}$ can be decomposed into

$$\boldsymbol{T} = \boldsymbol{R} + \boldsymbol{O}, \quad (7)$$

where

$$\boldsymbol{R} = \begin{pmatrix} 1-2\alpha^2 & 2\alpha(1-\frac{1}{2}\alpha^2) \\ -2\alpha(1-\frac{1}{2}\alpha^2) & 1-2\alpha^2 \end{pmatrix}, \quad (8)$$

$$\boldsymbol{O} = \begin{pmatrix} \alpha^2 & -\alpha^3 \\ -\alpha & -2\alpha^2+3\alpha^4 \end{pmatrix}. \quad (9)$$

$\boldsymbol{R}$ can be viewed as a rotation matrix in a two dimensional space by well-known approximation that $(1+t)^k \simeq 1+kt$, i.e.

$$\boldsymbol{R} \simeq \begin{pmatrix} 1-2\alpha^2 & 2\alpha\sqrt{1-\alpha^2} \\ -2\alpha\sqrt{1-\alpha^2} & 1-2\alpha^2 \end{pmatrix} \quad (10)$$

$$= \begin{pmatrix} \cos 2\theta & \sin 2\theta \\ -\sin 2\theta & \cos 2\theta \end{pmatrix} (\alpha = \sin\theta). \quad (11)$$

$\boldsymbol{R}$ is a rotation matrix and it rotates clockwise any vector by 2θ around the original point.

On the other hand, we can say that $\boldsymbol{O}$ has a very small influence on transformation by $\boldsymbol{T}$ because $|\alpha| < 1$ and three elements of $\boldsymbol{O}$ are composed of α^n where $n \geq 2$. Hence, transformation in a two dimensional space by $\boldsymbol{T}$ nearly equals transformation by matrix $\boldsymbol{R}$, i.e. rotation. Figure 2 shows how a trapezoid in a two dimensional space is transformed by $\boldsymbol{T}$, $\boldsymbol{R}$ and $\boldsymbol{O}$. Three large trapezoids drawn by solid, dotted, and dashed lines are the ones before and after transformation by $\boldsymbol{T}$ and $\boldsymbol{R}$ with $\alpha = 0.2$. A small quadrilateral around the origin is the one transformed by $\boldsymbol{O}$. It is clearly shown that a trapezoid is rotated clockwise after transformation by $\boldsymbol{T}$ and this rotation is reasonably similar to that of transformation by $\boldsymbol{R}$. $\boldsymbol{O}$ has a very small influence, where all the points in a space are compressed around the origin because $\boldsymbol{O}$ is close to a zero matrix.

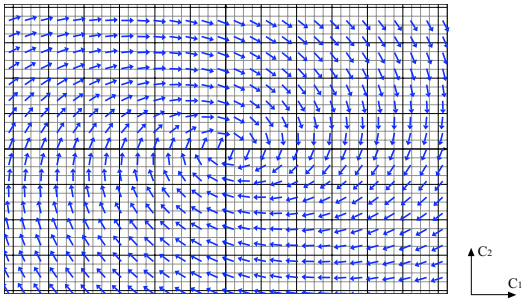


Figure 3: Vector filed given by Equation 12 for $\alpha = 0.2$.

Figure 3 shows the properties of $\boldsymbol{T}$ graphically from another viewpoint, which is a vector field given by vector-valued function;

$$\boldsymbol{y} = (\boldsymbol{T} - \boldsymbol{I})\boldsymbol{c} = \hat{\boldsymbol{c}} - \boldsymbol{c}, \quad (12)$$

where $\boldsymbol{I}$ is a two-dimensional identity matrix. $\boldsymbol{y}$ represents the influence at each point caused by transformation $\boldsymbol{T}$ because matrix $(\boldsymbol{T} - \boldsymbol{I})$ means the difference between before and after the transformation. From Figure 3, the vector field given by Equation 12 looks like a vortex. It means that $\boldsymbol{T}$ has a strong function of rotation.

3.2 Rotation in an n dimensional cepstrum space

In an n dimensional space, it is not so easy to extract the rotation properties from a given transformation matrix as in the case of a two dimensional space. Then, in this section, on the basis of the general definition of n dimensional rotation matrix, the geometrical properties of $\boldsymbol{A}$ are examined. Rotation matrix $\boldsymbol{R}$ is generally defined as

$$\boldsymbol{R}^\top \boldsymbol{R} = \boldsymbol{R}\boldsymbol{R}^\top = \boldsymbol{I} \quad (13)$$

$$\det \boldsymbol{R} = 1. \quad (14)$$

If it is assumed that $|\alpha| \ll 1$, $\boldsymbol{A}$ can be approximated as

$$\boldsymbol{A}_n = \begin{pmatrix} 1 & 2\alpha & 0 & \cdots & \cdots \\ -\alpha & 1 & 3\alpha & 0 & \cdots \\ 0 & -2\alpha & 1 & 4\alpha & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}, \quad (15)$$

in which α^ns with $n \geq 2$ are ignored [6]. The elements a_{ij} of $\boldsymbol{A}_n$ are

$$a_{ij} = \begin{cases} 1 & (i = j) \\ \mathrm{sgn}(j-i) * j\alpha & (|i-j| = 1) \\ 0 & (\text{otherwise}), \end{cases} \quad (16)$$

where $\mathrm{sgn}(j-i)$ returns $+1$ if $j-i>0$, or -1 if $j-i<0$. Now we will prove that both of $\boldsymbol{A}_n^\top \boldsymbol{A}_n$ and $\boldsymbol{A}_n \boldsymbol{A}_n^\top$ are close to $\boldsymbol{I}$.

$$\boldsymbol{A}_n^\top \boldsymbol{A}_n = \begin{pmatrix} 1+\alpha^2 & \alpha & -3\alpha^2 & 0 & \cdots \\ \alpha & 1+8\alpha^2 & \alpha & -8\alpha^2 & \cdots \\ -3\alpha^2 & \alpha & 1+18\alpha^2 & \alpha & \cdots \\ 0 & -8\alpha^2 & \alpha & 1+32\alpha^2 & \cdots \\ \vdots & \ddots & \ddots & \ddots & \ddots \end{pmatrix} \tag{17}$$

where the diagonal elements are $1+k\alpha^2$ with $k \in \mathcal{Z}$, the elements where $|i-j|=1$ are α, those where $|i-j|=2$ are $m\alpha^2$ with $m \in \mathcal{Z}$ and the others are zero. $\mathcal{Z}$ means the set of integers. $\boldsymbol{A}_n \boldsymbol{A}_n^\top$ takes the following form.

$$\boldsymbol{A}_n \boldsymbol{A}_n^\top = \begin{pmatrix} 1+4\alpha^2 & \alpha & -4\alpha^2 & 0 & \cdots \\ \alpha & 1+10\alpha^2 & \alpha & -9\alpha^2 & \cdots \\ -4\alpha^2 & \alpha & 1+20\alpha^2 & \alpha & \cdots \\ 0 & -9\alpha^2 & \alpha & 1+34\alpha^2 & \cdots \\ \vdots & \ddots & \ddots & \ddots & \ddots \end{pmatrix} \tag{18}$$

In both the products, α^2 can be ignored using the assumption of $|\alpha| \ll 1$. Hence, both products can be regarded as a special case of a tridiagonal matrix, of which the diagonal elements are 1 and the elements where $|i-j|=1$ are α. Although $\boldsymbol{A}_n^\top \boldsymbol{A}_n$ and $\boldsymbol{A}_n \boldsymbol{A}_n^\top$ are not equal to $\boldsymbol{I}$ strictly, we can say that $\boldsymbol{A}_n$ has high orthogonality, putting it another way, matrix $\boldsymbol{A}_n$ approximately satisfies Equation 13.

We can calculate the determinant of $\boldsymbol{A}_n$ because $\boldsymbol{A}_n$ is a tridiagonal matrix [8]. The determinant can be computed recursively as

$$\det \boldsymbol{A}_n = a_{nn} \det \boldsymbol{A}_{n-1} - a_{n(n-1)} a_{(n-1)n} \det \boldsymbol{A}_{n-2}. \tag{19}$$

From Equation 15, $a_{nn}=1$ and $a_{n(n-1)}a_{(n-1)n} \sim \alpha^2$. Using also the assumption of $|\alpha| \ll 1$, we can conclude $\det \boldsymbol{A}_n \approx \det \boldsymbol{A}_{n-1} \approx \cdots \approx \det \boldsymbol{A}_1 \approx 1$ recursively.

From the discussions above, we can conclude that $\boldsymbol{A}$ in Equation 3 has a certain function of rotating any vector in an n dimensional space. However, we have to admit that the discussions above include some rough approximations and then, the rotation function which $\boldsymbol{A}$ is supposed to have has to be verified experimentally. In the experimental verification, by assuming that the vector field obtained in Figure 3 should be observed also in an n dimensional space, some new properties of $\boldsymbol{A}$ are predicted. Figure 3 shows that a vector at any point is rotated by $\boldsymbol{T}$ and, with a fixed value of α, we can say that a vector at any point will be rotated by a similar angle, where the angle is dependent only on α. In other words, dependently on α, a cepstrum vector of any phoneme or any gender will be rotated by a similar angle. Another prediction is about the rotation of Δ parameters. For simplicity, in this part, $\Delta \boldsymbol{c}$ is defined as $\boldsymbol{c}_{t+1} - \boldsymbol{c}_t$. Figure 4 shows two cepstrum vectors, $\boldsymbol{c}_t$ and $\boldsymbol{c}_{t+1}$ and its Δ vector. If the two vectors are similarly rotated, then, the Δ vector has to be rotated in the same way. It is the case with two Δ vectors and their $\Delta\Delta$ vector. Similar rotation of any cepstrum vectors means that any Δ^n vectors are also rotated similarly. As told in Section 1, differential operations play an important role in canceling some kinds of mismatch between training and testing conditions. However, we can predict that these operations are totally ineffective for cancelling distortions caused by rotation-based transformation.

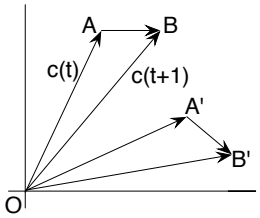


Figure 4: Rotation of two cepstrum vectors and their Δ vector.

4. Eigenvalues of transformation matrix

4.1 Diagonalization of a rotation matrix

To evaluate the rotation quantitatively, the angle calculated based on the inner product of two vectors was used in [3]. This angle will be called IP angle henceforth. However, with the IP angle only, it is difficult to characterize the rotation adequately. For example, only with the original vector and the IP angle, the direction of the warped vector cannot be determined. In this section, a new method is proposed to completely characterize the rotation by introducing multiple rotation angles, calculated through diagonalizing $\boldsymbol{A}$ based on eigenvalues.

Let $\boldsymbol{R}_n$ be an n-dimensional complete rotation matrix. $\boldsymbol{R}_n$ can be diagonalized with an $n \times n$ unitary matrix $\boldsymbol{U}_n$ and a diagonal matrix $\boldsymbol{D}_n$ which includes complex elements,

$$\boldsymbol{R}_n = \boldsymbol{U}_n \boldsymbol{D}_n \boldsymbol{U}_n^\dagger. \tag{20}$$

$(\cdot)^\dagger$ denotes the conjugate transpose of a matrix. For example, a two-dimensional rotation matrix $\boldsymbol{R}_2(\theta)$ whose θ is its rotation angle can be diagonalized into

$$\boldsymbol{R}_2(\theta) = \boldsymbol{U}_2 \boldsymbol{D}_2(\theta) \boldsymbol{U}_2^\dagger, \tag{21}$$

where

$$\boldsymbol{R}_2(\theta) = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \tag{22}$$

$$\boldsymbol{D}_2(\theta) = \begin{pmatrix} e^{j\theta} & 0 \\ 0 & e^{-j\theta} \end{pmatrix} \tag{23}$$

When n=2m, the eigen-equation of $\boldsymbol{R_n}$ has m sets of complex conjugate roots whose absolute value is 1. When n=2m+1, another root is obtained as 1 in addition to the m sets of roots [9]. Therefore, $\boldsymbol{D}_n$ in Equation 20 can be described as

$$\boldsymbol{D}_n(\Theta) = \begin{cases} \begin{pmatrix} \boldsymbol{D}_2(\theta_1) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \boldsymbol{D}_2(\theta_m) \end{pmatrix} & (n : \text{even}) \\ \begin{pmatrix} 1 & & \cdots & 0 \\ & \boldsymbol{D}_2(\theta_1) & & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & \boldsymbol{D}_2(\theta_m) \end{pmatrix} & (n : \text{odd}), \end{cases} \tag{24}$$

where Θ is an m dimensional vector to define $\boldsymbol{D}_n(\Theta)$;

$$\Theta = [\theta_1, \theta_2, \cdots, \theta_m]^\top. \tag{25}$$

Based on Equation 20, $\boldsymbol{R}_n$ can also be decomposed in a different way using rotation matrices containing only real numbers [9] as

$$\boldsymbol{R}_n(\Theta) = \boldsymbol{R}''_n \boldsymbol{R}'_n(\Theta) \boldsymbol{R}''^\top_n. \tag{26}$$

Here, when n=2m+1,

$$\boldsymbol{R}'_n(\Theta) = \begin{pmatrix} 1 & & \cdots & 0 \\ & \boldsymbol{R}_2(\theta_1) & & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & \boldsymbol{R}_2(\theta_m) \end{pmatrix}. \tag{27}$$

When n=2m, $\boldsymbol{R}'_n(\Theta)$ is obtained from the above matrix by removing the first column and the first row. In the below, only the case of n=2m is considered. Since $\boldsymbol{R}'_n(\Theta)$ has all $\boldsymbol{R}'_2(\theta_i)$ located in diagonal blocks, the operation of $\boldsymbol{R}'_n(\Theta)$ can be completely decomposed into m independent 2-dimensional rotations and can be completely characterized as Θ. In Equation 26, for vector X, $\boldsymbol{R}''^\top_n$ functions just as transforming the bases so that an n-dimensional rotation of vector $\boldsymbol{R}''^\top_n X$ can be done as m independent 2-dimensional rotations. After that, by $\boldsymbol{R}''_n$, the new bases are transformed back to the original ones. It is clear that all the rotational properties owe to Θ. In this study, each of the two dimensional subspaces affected by only $\boldsymbol{R}_2(\theta_i)$ is called a rotational plane. A rotational plane corresponding to $\boldsymbol{R}_2(\theta_i)$ is derived from its corresponding eigenvectors in Equation 20 or 26. Here, the eigenvectors a and b that correspond to eivenvalues of $\boldsymbol{D}_2(\theta_i)$ is selected, and $\frac{a+b}{2}$ and $\frac{a-b}{2j}$ become the bases of the rotational plane, since the corresponding elements of a and b are conjugate complex numbers with each other. For example, in the case of the rotation of the earth, an equatorial plane corresponds to the rotational plane. Since each of the rotational planes are orthogonal, the rotation caused by $\boldsymbol{R}_2(\theta_i)$ affects just to the corresponding rotational plane. Hence, the operation of $\boldsymbol{R}'_n(\Theta)$ is completely decomposed into m independent two-dimensional rotations. In the following section, $\boldsymbol{R}'_n(\Theta)$ is further discussed.

4.2 Relation between IP angle and Θ

How is an IP angle decomposed into rotation parameter vector Θ? Here, even if we consider $\boldsymbol{R}''_n$ as $\boldsymbol{I}$, generality is not lost. When n=2m, n-dimensional vector Y, transformed from vector X=$[x_1, x_2, \ldots, x_n]^\top$ by $\boldsymbol{R}'_n(\Theta)$, is described as

$$\begin{aligned} Y &= \boldsymbol{R}'_n(\Theta) X \\ &= [\boldsymbol{R}_2(\theta_1) v_1, \boldsymbol{R}_2(\theta_2) v_2, \ldots, \boldsymbol{R}_2(\theta_m) v_m]^\top, \end{aligned} \tag{28}$$

where $v_i = [x_{2i-1}, x_{2i}]^\top$ and $|Y| = |X|$. Now we can introduce the relation between IP angle θ' and rotation parameter vector Θ.

$$\begin{aligned} \cos\theta' &= \frac{X \cdot Y}{|X||Y|} \\ &= \frac{1}{|X|^2} \sum_{i=1}^{m} v_i^\top \boldsymbol{R}_2(\theta_i) v_i \\ &= \frac{1}{|X|^2} \sum_{i=1}^{m} |v_i|^2 \cos\theta_i. \end{aligned} \tag{29}$$

v_i is a vector projected onto the rotational plane corresponding to $\boldsymbol{R}_2(\theta_i)$. Equation 29 means that the cosine similarity based on the IP angle is represented as a weighted sum of the cosine similarities on each rotational plane. When n=2m+1, by considering that $\cos\theta_{2n+1}$=1 where θ_{2n+1} corresponds to the root of 1, we can find the same representation.

Further, Equation 29 indicates that, even with the same Θ, the IP angle can be changed according to v_i assigned to each rotational plane. Since v_i depends on $\boldsymbol{R}''^\top_n X$, the IP angle will depend on X. Although the IP angle was observed speaker- and phoneme-independent in our previous study [3], the above discussion mathematically predicts the dependence of the IP angle on these factors. Shortly, this will be verified experimentally.

4.3 Decomposition of quasi-rotation matrix of $\boldsymbol{A}$

The previous section discusses that the geometrical property of a complete rotation matrix can be decomposed into that on each rotational plane. Strictly speaking, however, $\boldsymbol{A}$ is not a complete rotation matrix. Here, we assume that $\boldsymbol{A}$ can be approximated as a combination of rotation matrix and magnification matrix. On this assumption, $\boldsymbol{R}_2(\theta_i)$ in Equation 27 and 29 is replaced with $r_i \boldsymbol{R}_2(\theta_i)$, where r_i is a magnification parameter for i-th rotational plane. Hence, Equation

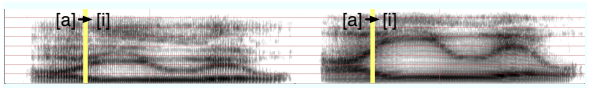


Figure 5: The original speech (left) and its warped version (right).

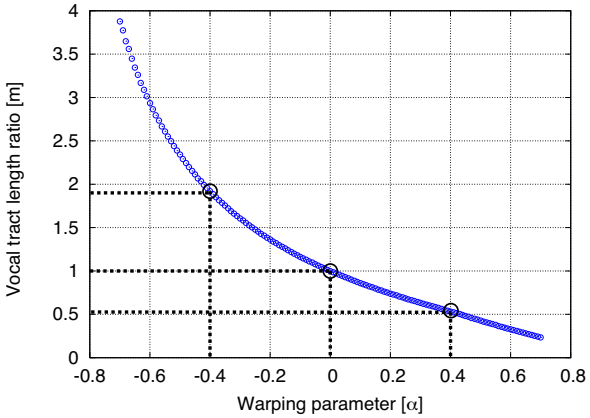


Figure 6: Relation between the warping parameter α and the vocal tract length ration m.

29 is obtained for $\boldsymbol{A}$;

$$\cos\theta' = \frac{\sum_{i=1}^{m} r_i |v_i|^2 \cos\theta_i}{\sqrt{(\sum_{i=1}^{m} |v_i|^2)\,(\sum_{i=1}^{m} r_i^2 |v_i|^2)}}. \tag{30}$$

If it can be assumed that the rotational planes of $\boldsymbol{A}$ are independent of α, meaning that $\boldsymbol{R}_n''$ is independent of α, and if it can also be assumed that only Θ and r_i are dependent on α, meaning that only $\boldsymbol{R}_n'(\Theta)$ is dependent on α, we can say the following. $\boldsymbol{A}$, characterized with θ_i and r_i, can be decomposed into m independent 2-dimensional rotations and magnifications. Then, for vector $\boldsymbol{R}_n''^{\top} X$, a new speech feature with a unit of two dimensional elements, $[|v_1|^2, |v_2|^2, ..., |v_m|^2]^{\top}$, may correspond to a certain aspect of speech. For example, by applying any 2-dimensional rotations of $\boldsymbol{R}_2(\theta_i)$, the above vector is not changed although the rotations change the VTL easily. This consideration implies that this vector may correspond to some speech features not influenced by VTL differences.

5. Experiments

5.1 Experimental conditions

We conducted three kinds of experimental evaluation for the properties of rotation and magnification caused by the difference of VTL. First, in order to prove the basic rotational properties of the matrix $\boldsymbol{A}$ quantitatively, we calculated the rotational angles based on the IP angle using original and resynthesized speech samples. Second, in order to demonstrate the effectiveness of the decomposition based on the eigenvalue analysis, we compared the IP angle and the estimated IP angle by Equation 30 using original speech samples. Finally, we conducted the analysis based on the rotational properties using real speech samples in order to evaluate the effectiveness to focus on the geometrical properties in the cepstral space.

5.1.1 Experiment 1

In the first experimental evaluations, we used speech data of /aiueo/ utterances from 2 speakers (1 male and 1 female). All the spectrum slices from each speech sample were converted to their warped versions through STRAIGHT analysis [4]. These warped versions correspond to speech samples with different VTL. Each speech sample was digitized at a sampling rate of 16 kHz, and analyzed in 25 ms length Hamming window and 5 ms frame shift.

The analysis yielded three vectors (12-dimensional FFT-based cepstrum, its Δ, and its Δ^2). Their directions at the central position of each transient segment (/a/ to /i/, /i/ to /u/, /u/ to /e/ and /e/ to /o/) were focused on and they were calculated as functions of the estimated body height of the speaker, where the direction at the original height had 0 deg. The IP angle between two vectors, a and b, is defined as

$$\theta = \arccos \frac{a \cdot b}{|a||b|}. \tag{31}$$

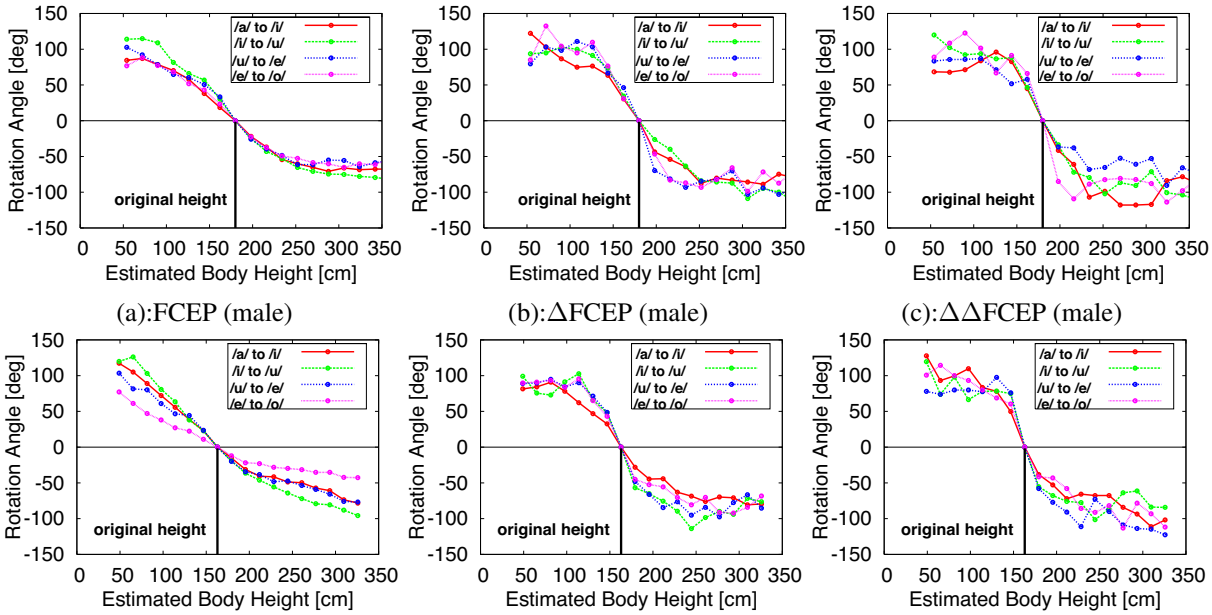
To resynthesize warped speech, we did not use Equation 1 or 3 directly but used two kinds of piece-wise linear functions. First, we adopted a piece-wise linear approximation of Equation 1;

$$\hat{\omega} = \begin{cases} \frac{1}{m}\omega & (0 \le \omega < \frac{m}{1+m}\pi) \\ m(\omega - \pi) + \pi & (\frac{m}{1+m}\pi \le \omega \le \pi). \end{cases} \tag{32}$$

This was to obtain the explicit relation between the rotation angle and the ratio of the VTL of the warped speaker to that of the original speaker. m in the above equation corresponds to the ratio of the two VTLs. Relation between m and α can be approximately represented by

$$\frac{1}{m} = \frac{3}{5}\left(-1 + \frac{\pi}{\arccos\alpha}\right) + \frac{2}{5}\frac{(1+\alpha)^2}{1-\alpha^2}. \tag{33}$$

Figure 5 shows two speech samples which are an original one and its warped version. The left is the original speech and the right is its warped version, where formant locations are clearly shifted. Figure 6 shows the relation between the warping parameter α and the vocal tract length ratio m. From Figure 6, in the case of a speaker of 180 cm in height, the warping of $\alpha = 0.4$ and $\alpha = -0.4$ correspond to a speaker of about 90 cm and 360 cm in height, respectively.



(a):FCEP (male) (b):ΔFCEP (male) (c):ΔΔFCEP (male)

(d):FCEP (female) (e):ΔFCEP (female) (f):ΔΔFCEP (female)

Figure 7: Relation between the rotation angle and the estimated body height. (a) to (c) are from a male speaker of 180 cm in height and (d) to (f) are from a female speaker of 163 cm in height. FCEP denotes FFT-based cepstrum.

Second, we used another piece-wise linear function defined as follows;

$$\hat{\omega} = \begin{cases} \frac{1}{m}\omega & (0 \le \omega < \omega_0) \\ \frac{1}{m}\omega_0 + \frac{m\pi - \omega_0}{m\pi - m\omega_0}(\omega - \omega_0) & (\omega_0 \le \omega \le \pi), \end{cases} \quad (34)$$

where

$$\omega_0 = \begin{cases} \frac{7}{8}\pi & (m \ge 1) \\ \frac{7 \cdot m}{8}\pi & (m < 1). \end{cases} \quad (35)$$

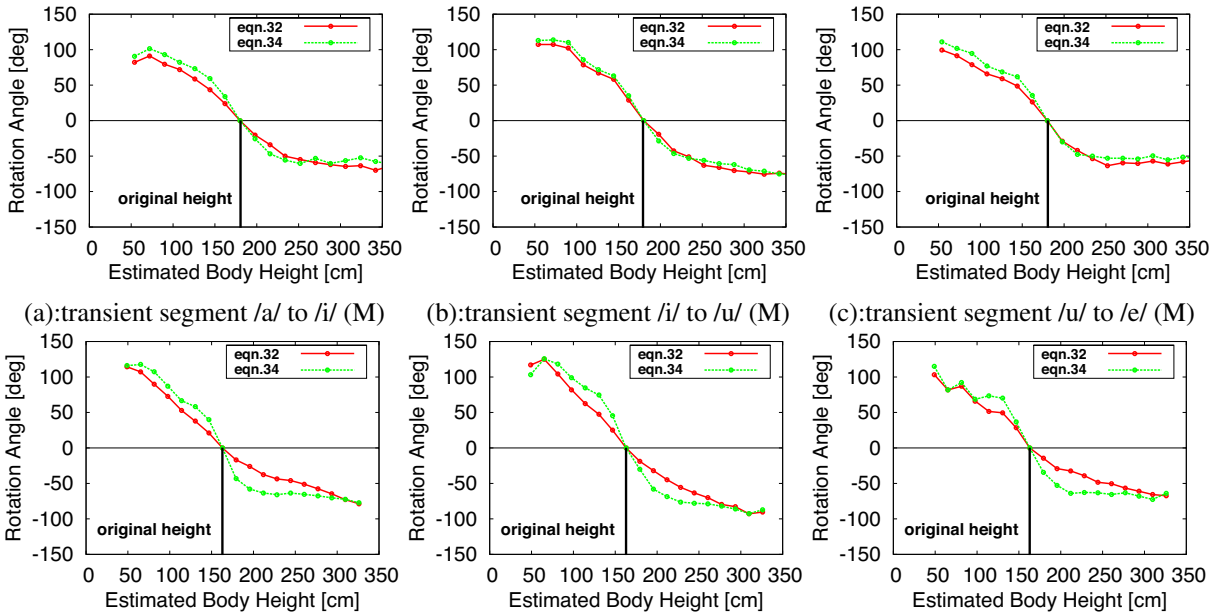
Compared with Equation 32, the first equation in Equation 34 covers wider range of ω. Hence, it is expected that Equation 34 represents the change of vocal tract length more precisely. We also compared properties of Equation 32 with those of Equation 34 with respect to rotational properties in cepstral space.

5.1.2 *Experiment 2*

In the second experiment, we compared the estimated IP angle based on Equation 30 with the IP angle based on Equation 31 . In this experiment, two issues should be investigated experimentally. The first issue is whether the rotational planes can be obtained from $\boldsymbol{A}$ irrespective of α or not. The second issue is whether $\boldsymbol{A}$ can be better approximated as modified rotation matrix, where $r_i \boldsymbol{R}_2(\theta_i)$ is used instead of $\boldsymbol{R}_2(\theta_i)$ only. For these objectives, the same utterances /aiueo/ to the first experiment, from 1 male and 1 female speakers, were acoustically analyzed. For each utterance, 3 frames were extracted and they were the central positions of spectral transition (/a/ to /i/, /i/ to /u/ and /u/ to /e/). For each frame, FFT-based Δ cepstrums were extracted and used. From $\boldsymbol{A}$, four sets of the rotational planes ($\boldsymbol{R}''_n$) were calculated for each case of α=$-0.2, -0.1, 0.1$ and 0.2. Let us define these sets as $B_{-0.2}, B_{-0.1}, B_{-0.1}$ and $B_{0.2}$, respectively. Then, for each set of the rotational planes, each of the focused Δcepstrum vectors was projected to the rotational planes and all the $|v_i|^2$s were calculated. Using these parameters and Equation 30, the IP angle was estimated as a function of the estimated body height of the speaker. Note that warped speech samples are not needed for this calculation, unlike the previous experiment. θ' is calculated only analytically. For comparison, based on Equation 31, θ' was calculated from the original utterances and their warped versions generated by applying $\boldsymbol{A}$ on the original utterances using STRAIGHT.

5.1.3 *Experiment 3*

In the third experimental evaluation, we used the connected Japanese vowel utterances from 8 speakers (4 males and 4 females), to evaluate the effectiveness of the analysis based on the rotational properties for a greater variety of speaker pairs. Each word is a concatenation of the five Japanese vowels 'a', 'i', 'u', 'e', and 'o', such as "uoiea". There are totally 120 words. The analysis based on the IP angle and the rotation angle in every rotational plane defined by $\boldsymbol{R}''_n$ was performed for 120 $\times$ 36 (=${}_8C_2 + 8$) utterance pairs. In each pair, two utterances of the same word were compared. A time alignment was performed with DTW, and rotational angles were calculated in each pair of aligned feature vectors. Note that



(a):transient segment /a/ to /i/ (M) (b):transient segment /i/ to /u/ (M) (c):transient segment /u/ to /e/ (M)

(d):transient segment /a/ to /i/ (F) (e):transient segment /i/ to /u/ (F) (f):transient segment /u/ to /e/ (F)

Figure 8: Comparison the results based on Equation 32 with those based on Equation 34. The angles between FCEPs and their warped ones are plotted. (a) to (c) are from a male speaker of 180 cm in height and (d) to (f) are from a female speaker of 163 cm in height.

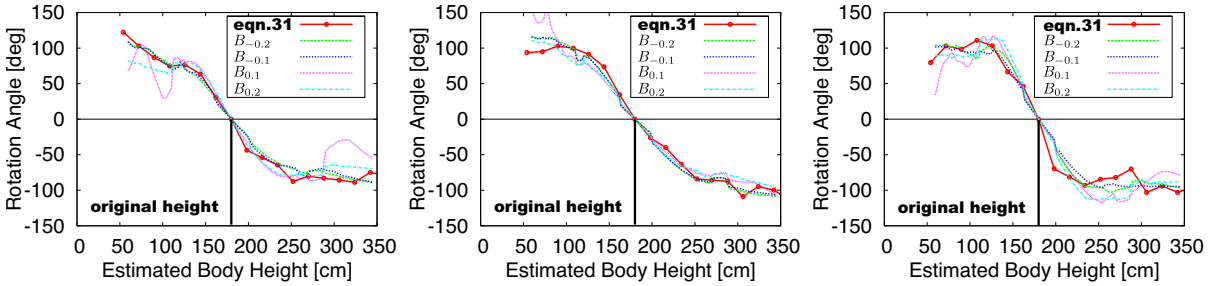
the utterance pairs from the same speaker were also included. We used the 12-dimensional FFT-based cepstrum as a feature vector. In order to calculate the rotation angle on each rotational plane, the rotational planes ($\boldsymbol{R}''_n$) were calculated for $\alpha = -0.05$ by the preliminary experiment. The indices of the rotational planes were numbered according to the absolute of the complex eigenvalues. The rotational plane that corresponds to the largest absolute was numbered 1, the plane that corresponds to the second largest one was 2, and so on. The angle on each rotational plane was calculated based on the projected vectors of the focused feature vectors. That is to say, 6 (=12/2) angles were calculated from each pair of feature vectors.

5.2 Results of experiments

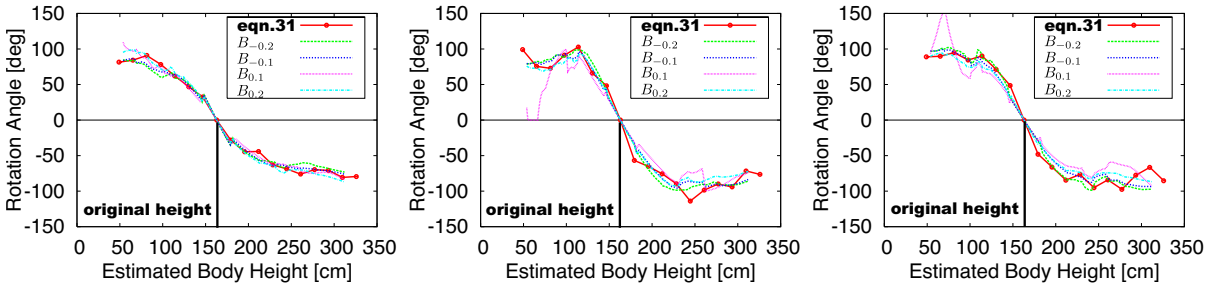
5.2.1 Results of the first experiment

Figure 7 shows the rotation angles calculated as functions of the estimated body height. The top three are from the male speaker and the bottom three are from the female speaker. The two in the left, the two in the center, and the two in the right are for FFT-based cepstrum, its Δ, and its $\Delta\Delta$, respectively. Each graph contains the results obtained at the four transient positions in the /aiueo/ utterance. Equation 32 was used for the warping function. As we predicted in the previous section, the rotation is observed reasonably irrespective of gender, phoneme, and the number of differential operations. Especially, in any condition, strong rotational properties are observed from about 120 cm to 230 cm in the case of male, from about 110 cm to 210 cm in the case of female. These range roughly correspond to $-0.16 \leq \alpha \leq 0.16$. In this range, $\boldsymbol{A}$ may be reasonably approximated to the rotation matrix. We can say that the direction of cepstrum-based parameters is rotated as the length of VTL changes. These results imply that the directional dependency of cepstrum coefficients on VTL can be useful to represent some characteristics of speakers.

Figure 8 shows the rotation angles adopting Equation 32 and Equation 34 as the frequency warping functions. Similarly to Figure 7, the top three are from the male and the bottom three are from the female. The results are for FFT-based cepstrum. Even in the case that we adopt Equation 34, the rotation is reasonably observed. According to [7], several frequency warping functions for VTLN have the similar forms when the transformation in cepstral space is adopted. The results in Figure 8 may be a proof of this opinion. In the case of the male speaker, the results from Equation 32 and those from Equation 34 are very similar. On the other hand, in the case of the female speaker, the result shows that cepstrum vectors are rotated more strongly by the frequency warping based on Equation 34 rather than that based on Equation 32. Equation 34 represents the change of vocal tract length more precisely than Equation 32. Since formant frequencies of utterances from female are higher than those from male, such a difference between Equation 32 and Equation 34 may appear in the results. Although properties of Equation 32 and



(a):transient segment /a/ to /i/ (M) (b):transient segment /i/ to /u/ (M) (c):transient segment /u/ to /e/ (M)

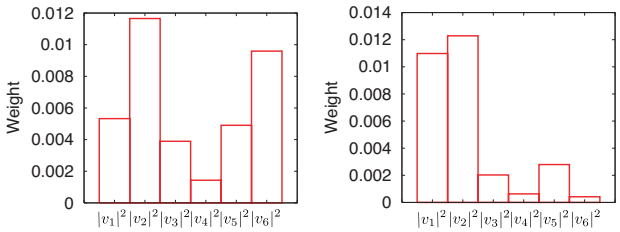


(d):transient segment /a/ to /i/ (F) (e):transient segment /i/ to /u/ (F) (f):transient segment /u/ to /e/ (F)

Figure 9: Relation between the IP angles and the estimated body height. (a) to (c) are from a male speaker of 180 cm in height and (d) to (f) are from a female speaker of 163 cm in height.

those of Equation 34 are slightly different from each other, rotational properties of VTL difference are observed.

As told in Section 1, some acoustic distortions can be effectively canceled by differential operations but the distortion examined in this paper cannot be canceled by these operations at all. If a parameter is defined as *vector* in an acoustic space, such as Δcepstrum, it will inevitably has this kind of distortion.



(a): $|v_i|^2$s for a male. (b):$|v_i|^2$s for a female.

Figure 10: Weights $|v_i|^2$ of each rotational plane. These data are for the transient position /a/ to /i/.

5.2.2 Effectiveness of the decomposition based on eigenvalues of rotation matrix

Figure 9 shows the estimated IP angles calculated by Equation 30 for each set of the rotational planes ($B_{-0.2}, B_{-0.1}, B_{0.1}$ and $B_{0.2}$). In Figure 9, IP angles calculated by Equation 31 are also plotted for comparison. The top three figures are from the male speaker and the bottom three ones are from the female speaker. The two in the left, the two in the center, and the two in the right are for the transient positions of /a/ to /i/, /i/ to /u/, and /u/ to /e/, respectively.

The four curves drawn with four different sets of rotational planes show very small differences. This indicates that the rotational planes are reasonably independent of α and the first issue is solved. In any graph, we can say that the four curves drawn by Equation 30 are well fitted to the curve drawn by Equation 31. This means that modified rotation matrix is a proper solution to approximate $\boldsymbol{A}$. The second issue is also solved here. Further, we can find some dependence of the IP angle on phonemes or speakers. For example, the Equation 31 curve of (d) and that of (e) show rather different patterns. Even in these cases, the curves drawn by Equation 30 follow precisely the two different Equation 31 curves. This means that the proposed method can capture phoneme- and speaker-dependence very well. This performance is considered due to the weights $|v_i|^2$ assigned to the individual rotational planes. Considering these results, we can conclude that the decomposition based on eigenvalues of quasi-rotational matrix can characterize the rotational distortion caused by $\boldsymbol{A}$ even more precisely.

As for these weights, a small experiment was carried out. Figure 10 shows the projection weights calculated to the individual rotational planes for a male speaker and a female speaker. Frames in /a/ to /i/ transition were used in both the cases. We can say that the weights depend heavily on speaker. A physical and phonetic meaning of this vector will be discussed experimentally in our future work.

Table 1: Analysis of the cosine similarity based on Equation 31. The numbers within parentheses indicate the number of speaker pairs.

	the mean of cosine similarity
Intra-speaker (8)	0.979
Within-gender [male] (6)	0.914
Within-gender [female] (6)	0.958
Cross-gender (16)	0.897

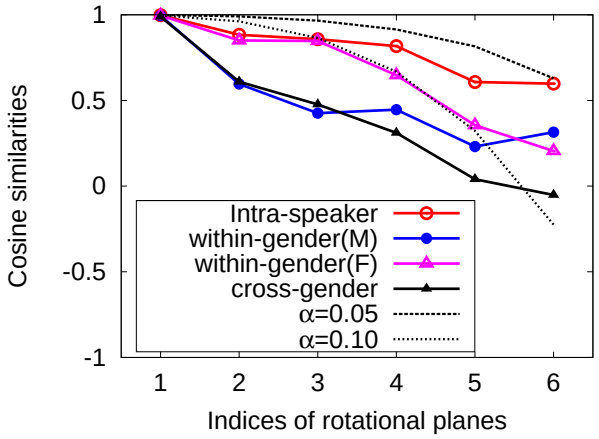


Figure 11: Result of the cosine similarities on each rotational plane. The kinds of pairs are the same to those of Table 1.

5.2.3 Results of the third experiment

Table 1 shows the mean of cosine similarity for four kinds of speaker pairs. "Intra-speaker" means a pair of utterances of the same word from the same speaker. "Within-gender" means a pair of utterances of the same word from two speakers of the same gender. "Cross-gender" means a pair of utterances of the same word from two speakers of the different gender. The cosine similarity of "Cross-gender" was smaller than that of the other kinds. This indicates that acoustic changes caused by vocal tract length difference are represented as the rotational property in a cepstral space even in the case of real speech samples.

Figure 11 shows the cosine similarities on each rotational plane. In Figure 11, as the indices of the rotational planes increase, the cosine similarities are decreasing. In other words, as the indices of the planes increase, projected vectors are rotated more strongly. The reason of this result corresponds to properties of cepstrum. Since the identity matrix is also one of the rotation matrices, the original cepstrum space also can be divided into the rotational planes. Then, the first rotational plane in Figure 11 is similar to the (c_1, c_2) subspace in cepstral space, the second one is similar to the (c_3, c_4) subspace, and so on. Cepstrum coefficients in low quefrencies correspond to envelopes of power spectrum, and those in high quefrencies correspond to fine ranges of frequencies. Hence, when the frequency axis of spectrum is warped by the frequency warping, the coefficients in low quefrencies change gradually and those in high quefrencies change rapidly. These phenomena may be reflected to the results of rotation angles in Figure 11.

Compared with the results of Table 1, the difference between "Cross-gender" and the other kinds was observed as rotational properties more clearly. Especially in the cases from the fourth to the sixth planes, we could observe the clear differences. Compared with the rotational angles from $\boldsymbol{A}$ (they were shown as dotted and dashed lines), stronger rotational differences were observed in the cases of "Cross-gender" and "Within-gender (M)". These results indicate that the decomposition based on the eigenvalue analysis of the rotational matrix can reasonably capture the difference of vocal tract length, and it can be used for representing the speaker information.

6. Conclusions

We have mathematically and experimentally proved that cepstrum coefficients are strongly dependent on vocal tract length difference and this dependency is represented as rotation in a cepstrum space. Although this fact was expected qualitatively in our previous study [10], this paper has proved it quantitatively. In addition, we have analyzed and factorized the rotational distortion caused by matrix $\boldsymbol{A}$ using its eigenvalues in a cepstrum space. We have proved that vocal tract *length* difference is mainly represented as rotation angles on multiple rotational planes. Further, a new speech feature, projection weights to the rotational planes, is introduced and we consider that it may be able to capture vocal tract *shape* difference. In future works, we are going to carry out some experiments using these rotational properties for speaker recognition [11]. In the paper, we are interested especially in the separation of speech information between vocal tract length and vocal tract shape. This separation can be used in many speech applications including structural speech recognition, which has been proposed recently by some of the authors of this paper [12].

References

[1] B. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J.Acoust.Soc.America*, vol. 55, pp. 1304–1312, 1974.

[2] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," *Proc. of ICASSP96*, vol. 1, pp. 346–348, 1996.

[3] D. Saito, R. Matsuura, S. Asakawa, N. Minematsu, and K. Hirose, "Directional dependency of cepstrum on vo-

cal tract length," *Proc. of ICASSP 2008*, pp. 4485–4488, 2008.

[4] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, pp. 187–207, 1999.

[5] T. Fukuda, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," *Proc. of ICASSP 92*, pp. 137–140, 1992.

[6] T. Emori and K. Shinoda, "Rapid vocal tract length normalization using maximum likelihood estimation," *Proc. of Eurospeech2001*, pp. 1649–1652, 2001.

[7] M. Pitz and H. Ney, "Vocal tract normalization equals linear transformation in cepstral space," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 5, pp. 930–944, 2005.

[8] R. Horn and C. Johnson, *Matrix Analysis*. Cambridge University Press, 1985.

[9] T. Takahashi, Lina, I. Ide, Y. Mekada, and H. Murase, "Interpolation between eigenspaces using rotation in multiple dimensions," *Proc. of ACCV 2008*, vol. 2, pp. 774–783, 2007.

[10] N. Minematsu, "Mathematical evidence of the acoustic universal structure in speech," *Proc. of ICASSP2005*, pp. 889–892, 2005.

[11] H. Tang, S. M. Chu, and T. S. Huang, "Generative model-based speaker clustering via mixture of von mises-fisher distributions," *Proc. of ICASSP2009*, pp. 4101–4104, 2009.

[12] S. Asakawa, N. Minematsu, and K. Hirose, "Multi-stream parameterization for structural speech recognition," *Proc. of ICASSP 2008*, pp. 4097–4100, 2008.

Photo

Daisuke Saito received the B.E. and M.S. degrees in 2006 and 2008, respectively, from the University of Tokyo, Tokyo, Japan. From 2008, he is a Ph.D student in the Graduate School of Engineering, the University of Tokyo. From 2010, he is a research fellow (DC2) of the Japan Society for the Promotion of Science. He is interested in various areas of speech engineering, including speech synthesis, voice conversion, acoustic analysis, speaker recognition, and speech recognition. He is a member of the Institute of Electrical and Electronics Engineers, the International Speech Communication Association, the Acoustical Society of Japan, the Institute of Electronics, Information and Communication Engineers,the Japanese Society for Artificial Intelligence, and the Institute of Image Information and Television Engineers.

Photo

Nobuaki Minematsu received the Ph.D. degree in electronic engineering in 1995 from the University of Tokyo. In 1995, he was an Assistant Researcher with the Department of Information and Computer Science, Toyohashi University of Technology, and in 2000, he was an Associate Professor with the Graduate School of Engineering, University of Tokyo. Since 2009, he has been an Associate Professor with the Graduate School of Information Science and Technology, University of Tokyo. From 2002 to 2003, he was a visiting researcher at Kungl Tekniska Högskolan (KTH), Sweden. He has a wide interest in speech from science to engineering, including phonetics, phonology, speech perception, speech analysis, speech recognition, speech synthesis, and speech applications. Dr. Minematsu is a member of ISCA, IPA, the Computer Assisted Language Instruction Consortium, the Institute of Electronics, Information and Communication Engineering, the Acoustical Society of Japan, the Information Processing Society of Japan, the Japanese Society for Artificial Intelligence, and the Phonetic Society of Japan. He received 2007 paper award from Research Institute of Signal Processing.

Photo

Keikichi Hirose received the B.E. degree in electrical engineering in 1972, and the Ph.D. degree in electronic engineering in 1977, respectively, from the University of Tokyo, Tokyo, Japan. In 1977, he joined the University of Tokyo as a Lecturer in the Department of Electrical Engineering, and, in 1994, became a Professor in the Dept. of Electronic Engineering. From 1996, he was a Professor at the Graduate School of Engineering, Department of Information and Communication Engineering, University of Tokyo. On April 1, 1999, he moved to the University's Graduate School of Frontier Sciences (Department of Frontier Informatics), and again moved to Graduate School of Information Science and Technology

(Department of Information and Communication Engineering) on October 1, 2004. From March 1987 until January 1988, he was a Visiting Scientist of the Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, U.S.A. His research interests cover widely spoken language information processing. He led a project "Realization of advanced spoken language information processing from prosodic feature," Scientific Research on Priority Areas, Grant in Aid on Scientific Research, Ministry of Education, Culture, Sports, Science and Technology, Japanese Government. He is a member of the Institute of Electrical and Electronics Engineers, the Acoustical Society of America, the International Speech Communication Association, the Institute of Electronics, Information and Communication Engineers (Fellow), the Acoustical Society of Japan, and other professional organizations. He received 2007 paper award from Research Institute of Signal Processing.