# DIRECTIONAL DEPENDENCY OF CEPSTRUM ON VOCAL TRACT LENGTH

*Daisuke SAITO*[1], *Ryo MATSUURA*[1], *Satoshi ASAKAWA*[1], *Nobuaki MINEMATSU*[1], *Keikichi HIROSE*[2]

[1]Graduate School of Frontier Sciences, The University of Tokyo
[2]Graduate School of Information Science and Technology, The University of Tokyo
{dsk_saito,matsuura,asakawa,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 the $n$ dimensional cepstrum space. In speech recognition studies, vocal tract length normalization (VTLN) techniques are widely used to cancel age- and gender-differences. In VTLN, a frequency warping is often carried out and it can be implemented as a linear transformation in a cepstrum space; $\hat{\boldsymbol{c}} = \boldsymbol{A}\boldsymbol{c}$. However, the geometric properties of this transformation matrix $\boldsymbol{A}$ have not been well discussed. In this study, its properties are made clear using $n$ dimensional geometry and it is shown that the matrix rotates any cepstrum vector similarly and apparently. Experimental results using resynthesized speech demonstrate that cepstrum vectors extracted from a speaker of 180 [cm] in height and those from another speaker of 120 [cm] in height are reasonably orthogonal. This result makes clear one of the reasons why children's speech is very difficult for conventional speech recognizers to deal with adequately.

***Index Terms***— frequency warping, cepstrum, rotation, rotation matrix, vocal tract length

## 1. INTRODUCTION

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, such as children. It means that the SI systems are not really SI.

To overcome the above problem, speaker normalization has been used in many systems. Speaker 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 $\Delta$cepstrums 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. The use of $\Delta$cepstrums is also based on subtracting the cepstrum of the previous frame from that of the current one.

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, a special emphasis is put on 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 $\Delta$s and $\Delta\Delta$s.

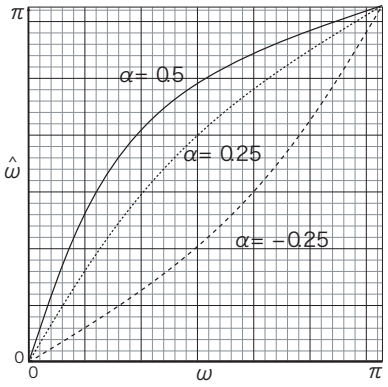


**Fig. 1**. Examples of frequency warping functions for different values of $\alpha$. $\alpha < 0$ transforms VTL longer and $\alpha > 0$ does VTL shorter.

## 2. DIFFERENCE IN VTL AND ITS EFFECTS

### 2.1. Frequency warping

The difference in VTL is often modeled by a warping function in a spectrum space. We employ a first order all-pass transform as a warping function here. The all-pass transform is described as

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

where $\alpha$ is a warping parameter and $|\alpha| < 1$; $\omega$ and $\hat{\omega}$ are frequencies before and after transformation, respectively. In case of $\alpha < 0$, formants are shifted to be lower and the VTL is transformed to be longer. $\alpha > 0$ brings about 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 [3] 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 not considered, a frequency warping can be expressed as

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

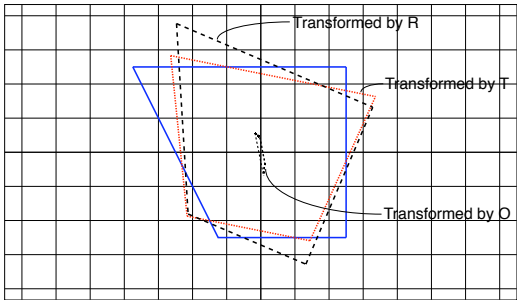


**Fig. 2**. Effects of transformations of $\boldsymbol{T}$, $\boldsymbol{R}$, and $\boldsymbol{O}$ for $\alpha = 0.2$.

where

$$\hat{\boldsymbol{c}} = (\hat{c}_1\ \hat{c}_2\ \hat{c}_3\ \hat{c}_4 \cdots)^t$$
$$\boldsymbol{A} = \begin{pmatrix} 1-\alpha^2 & 2\alpha-2\alpha^3 & \cdots & \cdots \\ -\alpha+\alpha^3 & 1-4\alpha^2+3\alpha^4 & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix} \quad (3)$$
$$\boldsymbol{c} = (c_1\ c_2\ c_3\ c_4 \cdots)^t.$$

From Pitz [4], the element $a_{ij}$ of matrix $\boldsymbol{A}$ can be written using the warping parameter $\alpha$ as

$$a_{ij} = \frac{1}{(j-1)!} \sum_{m=\max(0,j-i)}^{j} \binom{j}{m} \times \frac{(m+i-1)!}{(m+i-j)!}(-1)^{(m+i-j)}\alpha^{(2m+i-j)}, \quad (4)$$

where

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

## 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 on the first and second dimensions of the cepstrum space. Then, the discussion will be expanded into $n$ dimensions.

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) as $\boldsymbol{T}$, and $\boldsymbol{T}$ can be decomposed into

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

where

$$\boldsymbol{R} = \begin{pmatrix} 1-2\alpha^2 & 2\alpha(1-\frac{1}{2}\alpha^2) \\ -2\alpha(1-\frac{1}{2}\alpha^2) & 1-2\alpha^2 \end{pmatrix}, \quad (8)$$
$$\boldsymbol{O} = \begin{pmatrix} \alpha^2 & -\alpha^3 \\ -\alpha & -2\alpha^2+3\alpha^4 \end{pmatrix}. \quad (9)$$

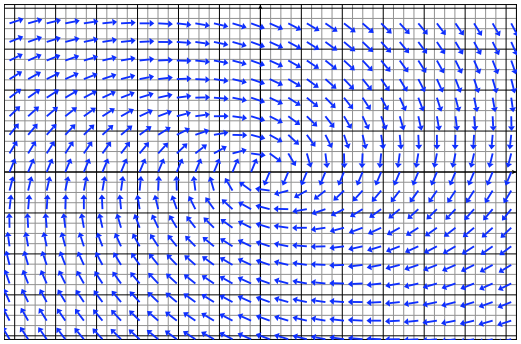

**Fig. 3**. Vector filed given by Equation (12) for $\alpha = 0.2$.

$\boldsymbol{R}$ 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 $2\theta$ arround 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 $\alpha^n$ 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 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 2 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

$$\boldsymbol{R}^t\boldsymbol{R} = \boldsymbol{R}\boldsymbol{R}^t = \boldsymbol{I} \quad (13)$$
$$\det \boldsymbol{R} = +1. \quad (14)$$

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

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

in which $\alpha^n$s with $n \geq 2$ are ignored. The elements $a_{ij}$ of $\boldsymbol{A}_n$ are

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

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

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

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

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

In both products, $\alpha^2$ 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 $\alpha$. Although $\boldsymbol{A}_n^t \boldsymbol{A}_n$ and $\boldsymbol{A}_n \boldsymbol{A}_n^t$ 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 [5]. 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 include some rough approximations and then, the rotation function which $\boldsymbol{A}$ is supposed to have has to be verified experimentally. Here, by assuming that the vector field obtained in Figure 3 should be observed also in an $n$ dimensional space, some properties of $\boldsymbol{A}$ are additionally predicted. Figure 3 shows that a vector at any point is rotated by $\boldsymbol{T}$ and, with a fixed value of $\alpha$, we can say that a vector at any point will be rotated by a similar angle, where the angle is dependent only on $\alpha$. In other words, dependently on $\alpha$, a cepstrum vector of any phoneme or any gender will be rotated by a similar angle. Another prediction is about the rotation of $\Delta$ parameters. Figure 4 shows two cepstrum vectors, $\boldsymbol{c}_t$ and $\boldsymbol{c}_{t+1}$ and its $\Delta$ vector. If the two vectors are similarly rotated, then, the $\Delta$ vector has to be rotated in the same way. It is the case with two $\Delta$ vectors and their $\Delta\Delta$ vector. Similar rotation of any cepstrum vectors means that any $\Delta^n$ 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 rotation-based transformation.

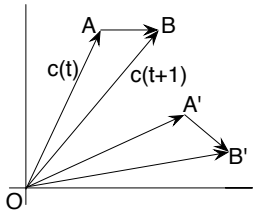


**Fig. 4**. Rotation of two cepstrum vectors and their $\Delta$ vector.


**Fig. 5**. The original speech (left) and its warped version (right).

## 4. EXPERIMENTS

### 4.1. Experimental conditions

For evaluating the properties of rotation caused by the difference of VTL, experiments using resynthesized speech samples /aiueo/ were carried out. We used speech data 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 [6]. 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 MFCC, its $\Delta$, and its $\Delta^2$). Their direction at the central position of each transient segment (/a/ to /i/, /i/ to /u/, /u/ to /e/ and /e/ to /o/) was focused on and they were calculated as a function of the estimated body height of the speaker, where the direction at the original height had 0 deg. The angle between two vectors, $a$ and $b$, was calculated as

$$\theta = \arccos \frac{a \cdot b}{|a||b|}. \tag{20}$$

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

$$\hat{\omega} = \begin{cases} \frac{1}{m}\omega & (0 \leq \omega < \frac{m}{1+m}\pi) \\ m(\omega - \pi) + \pi & (\frac{m}{1+m}\pi \leq \omega \leq \pi). \end{cases} \tag{21}$$

This was to obtain the relation explicitly 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 $\alpha$ can be approximately represented by

$$\frac{1}{m} = \frac{3}{5}\left(-1 + \frac{\pi}{\arccos \alpha}\right) + \frac{2}{5}\frac{(1+\alpha)^2}{1-\alpha^2}. \tag{22}$$

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.

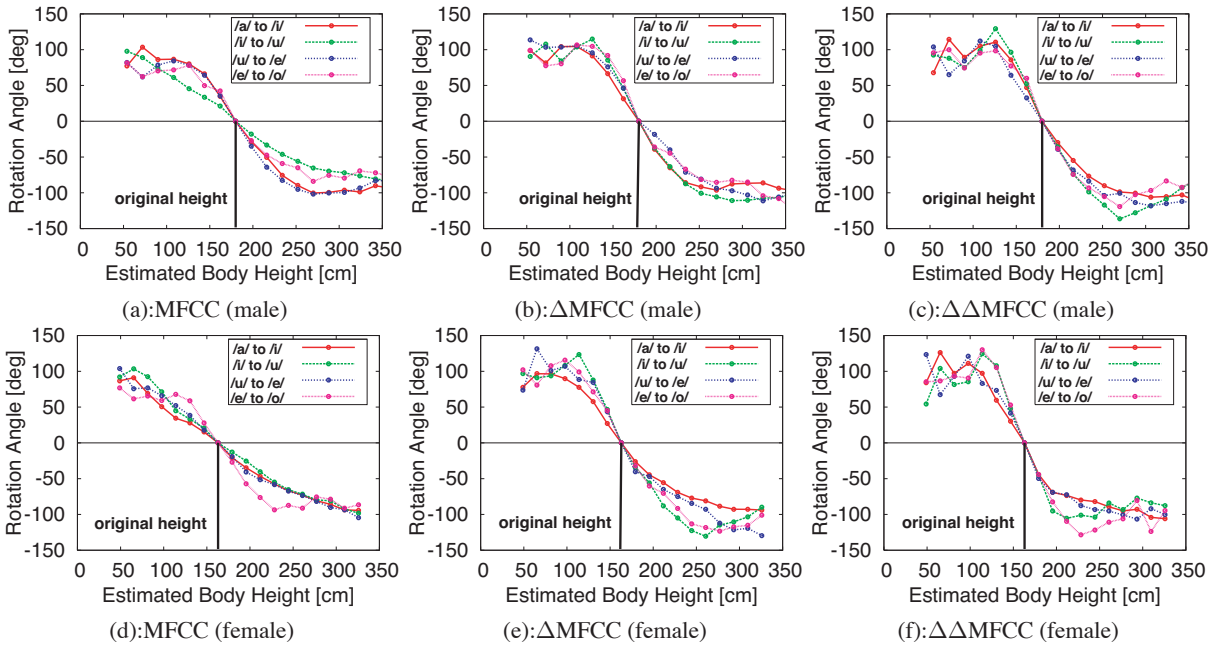


**Fig. 6**. Relation between the rotation angle and the estimated body height. (a) to (c) are from a male speaker of 180 cm in height and (d) to (f) are from a female speaker of 163 cm in height.

### 4.2. Results and discussions

Figure 6 shows the rotation angles calculated as a function 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 MFCC, its Δ, and its ΔΔ, respectively. Each graph contains the results obtained at the four transient positions in the /aiueo/ utterance. As we predicted in the previous section, the rotation is observed reasonably irrespective of gender, phoneme, and the number of differential operations. It is interesting to see in Figure 6(b), for example, that ΔMFCCs of a male speaker of 180 cm in height and those of its warped speaker to be 120 cm in height are orthogonal. We can say that the direction of cepstrum-based parameters is rotated slowly as the speaker grows up. These results imply that the directional dependency of cepstrum coefficients on VTL can be used as one of the effective features for age (VTL) estimation. Further, we consider that these results clarify quantitatively one of the reasons why conventional speech recognizers work poorly with children's speech.

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. We already proposed another framework which uses only *scalar*-based parameters which are invariant with the above two types of distortions. [7] showed that a small number of training speakers could provide the acoustic models for SI speech recognition because the proposed *scalar*-based parameters cannot see the two types of distortions at all.

## 5. 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. This fact was expected qualitatively in our previous study [8]. Further, the rotation angle is shown to be independent of phoneme, speaker, and the number of differential operations. It is also shown that two vectors in *one* category can be orthogonal if they are from speakers with very different body height. The conventional acoustic modeling framework collected these very different data to be modeled as *one* statistical model. We consider whether this strategy is reasonable enough geometrically.

## 6. 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," *ICASSP96*, vol. 1, pp. 346–348, 1996.

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

[4] 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.

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

[6] H. Kawahara et al., "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.

[7] S. Asakawa et al., "Multi-stream parameterization for structural speech recognition," *ICASSP 2008*, 2008 (to appear).

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