interspeech 2008 incorporating SST 08

Decomposition of Rotational Distortion Caused by VTL Difference Using Eigenvalues of Its Transformation Matrix

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

[1]Graduate School of Engineering, The University of Tokyo,
[2]Graduate School of Information Science and Technology, The University of Tokyo
{dsk_saito,mine,hirose}@gavo.t.u-tokyo.ac.jp

Abstract

In speech recognition studies, vocal tract length normalization (VTLN) techniques are widely used to cancel age- and gender-difference. In VTLN, the distortion is often modeled as a linear transform in a cepstrum space; $\hat{c}$=$\boldsymbol{A}c$. In our previous study, the geometrical properties of $\boldsymbol{A}$ were discussed and it was shown that the matrix can be approximated as rotation matrix. In this study, a new method of better approximating $\boldsymbol{A}$ is proposed. Using eigenvalues of $\boldsymbol{A}$, its quasi-rotational distortion is factorized into multiple rotation operations and multiple magnification operations. Using this method, the intrinsic ambiguity of the rotation angle used in our previous study is resolved. Instead, multiple rotation angles are introduced to understand better what kind of geometrical distortions $\boldsymbol{A}$ induces to cepstrum vectors. Experiments show the validity of the new method and a new speech feature is also derived by the new method.

Index Terms: 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 cancel the other kinds of information. In every speech application, a feature vector is used to characterize the focused information well. 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. Here, we're interested in the geometrical aspect of the modification in the feature space. If some tractable properties are found in the geometrical aspect of the modification, it will be possible to exclude the unfocused information appropriately so that a collection of a huge amount of data may not be needed.

Recently in [1], we proved mathematically and experimentally that the distortion caused by VTL differences can be approximated to rotate a cepstrum vector. In this previous work, the rotation angle based on the inner product of two vectors, IP angle henceforth, was adopted to evaluate the rotation quantitatively. However, we have to admit that the IP angle is a rather rough index to evaluate the rotation in an n-dimensional space. In other words, the same IP angle can show different rotation properties and, in this sense, the IP angle is ambiguous. In the current paper, to understand the geometrical properties of $\boldsymbol{A}$ better, multiple rotation angles are introduced by diagonalizing $\boldsymbol{A}$ with eigenvalues. Although the IP angle was seen speaker- and phoneme-independent in [1], in this paper, some dependencies are discussed mathematically and verified experimentally.

2. Difference in VTL and its effects

2.1. Frequency warping

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

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

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.

2.2. Linear modeling of frequency warping

Emori [2] 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

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

where

$$\boldsymbol{A} = \begin{pmatrix} 1-\alpha^2 & 2\alpha-2\alpha^3 & \cdots & \cdots \\ -\alpha+\alpha^3 & 1-4\alpha^2+3\alpha^4 & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}. \tag{3}$$

From Pitz [3], a_{ij} of $\boldsymbol{A}$ can be written using α as

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

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

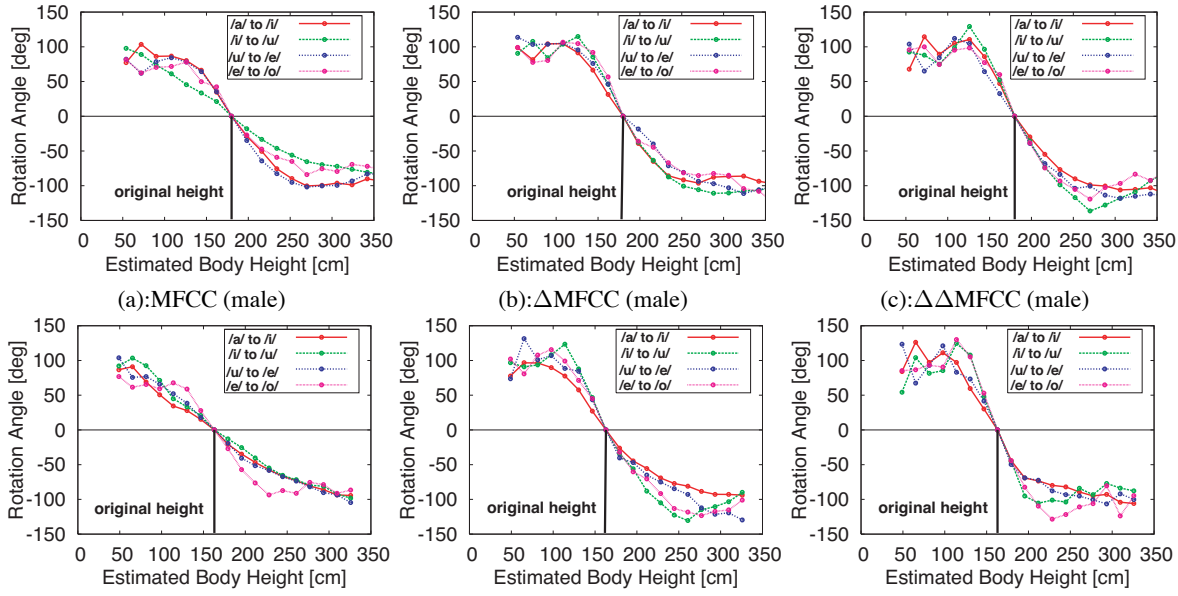
$$\binom{j}{m} = \begin{cases} {}_jC_m & (j \geq m) \\ 0 & (j < m). \end{cases} \tag{5}$$

3. Rotation in a cepstrum space

How does $\boldsymbol{A}$ influence cepstrum vectors geometrically? In our previous work, we proved that $\boldsymbol{A}$ has a strong function of rotation [1]. A complete rotation matrix $\boldsymbol{R}$ is defined as

$$\boldsymbol{R}^t\boldsymbol{R} = \boldsymbol{R}\boldsymbol{R}^t = \boldsymbol{I} \tag{6}$$

$$\det \boldsymbol{R} = +1, \tag{7}$$



(a):MFCC (male) (b):ΔMFCC (male) (c):ΔΔMFCC (male)

(d):MFCC (female) (e):ΔMFCC (female) (f):ΔΔMFCC (female)

Figure 1: 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.

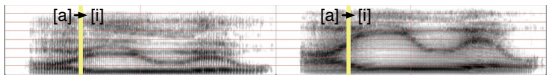


Figure 2: An original utterance and its warped version

Table 1: Conditions for acoustic analysis.

sampling	16 bit / 16 kHz
window	25 ms length and 5 ms shift
parameters	MFCC (1 to 12), its Δ, and its Δ^2

where $\boldsymbol{R}^t$ is the transpose of $\boldsymbol{R}$ and $\boldsymbol{I}$ is an n-dimensional identity matrix. In [1], we showed that $\boldsymbol{A}$ approximately satisfies these conditions and that VTL differences rotate cepstrums.

This rotational distortion in a cepstrum space was also verified experimentally in [1]. /aiueo/ utterances from 2 speakers (1 male and 1 female) were transformed into their warped versions by using STRAIGHT [4]. One example is shown in Figure 2 and the conditions for acoustic analysis are shown in Table 1. From each utterance, four frames with high spectral transition were detected, e.g., the central position of /a/ to /i/ transition. At each frame, MFCC, its Δ, and its Δ^2 vectors were calculated. For each vector, the IP angle was calculated between before and after the warping. Let a and b be vectors before and after the warping. Then, their IP angle is calculated simply as

$$\theta = \arccos \frac{a \cdot b}{|a||b|}. \tag{8}$$

Figure 1 shows the direction of the IP angle as a function of the estimated body height of the speaker, where the direction at the original height is 0 deg. For each case of MFCC, its Δ, and its Δ^2, we can say that $\boldsymbol{A}$ has a strong function of rotation

4. Eigenvalues of transformation matrix

4.1. Diagonalization of a rotation matrix

In [1], the IP angle was used to evaluate the rotation quantitatively. However, with the IP angle only, it is difficult to characterize the rotation adequately. For example, given the original vector, only the IP angle cannot determine the direction of the warped vector. 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,

$$\boldsymbol{R}_n = \boldsymbol{U}_n \boldsymbol{D}_n \boldsymbol{U}_n^{\dagger}. \tag{9}$$

For example, a two-dimensional rotation matrix $\boldsymbol{R}_2(\theta)$ whose θ is its rotation angle can be diagonalized into

$$\boldsymbol{R}_2(\theta) = \boldsymbol{U}_2 \boldsymbol{D}_2(\theta) \boldsymbol{U}_2^{\dagger}, \tag{10}$$

where

$$\boldsymbol{R}_2(\theta) = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \tag{11}$$

$$\boldsymbol{D}_2(\theta) = \begin{pmatrix} e^{j\theta} & 0 \\ 0 & e^{-j\theta} \end{pmatrix} \tag{12}$$

When $n{=}2m$, the eigen-equation of $\boldsymbol{R_n}$ has m sets of complex conjugate roots whose absolute value is 1. When $n{=}2m{+}1$, 1 becomes another root in addition to the m sets of roots [5]. Therefore, $\boldsymbol{D}_n$ in Equation 9 can be described as

$$\boldsymbol{D}_n(\Theta) = \begin{cases} \begin{pmatrix} \boldsymbol{D}_2(\theta_1) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \boldsymbol{D}_2(\theta_m) \end{pmatrix} & (n:\text{even}) \\ \begin{pmatrix} 1 & & \cdots & 0 \\ & \boldsymbol{D}_2(\theta_1) & & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & \boldsymbol{D}_2(\theta_m) \end{pmatrix} & (n:\text{odd}), \end{cases} \tag{13}$$

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

$$\Theta = (\theta_i | i = 1, 2, \cdots, m). \tag{14}$$

Based on Equation 9, $\boldsymbol{R}_n$ can also be decomposed in a different way using rotation matrices containing only real numbers [5] as

$$\boldsymbol{R}_n(\Theta) = \boldsymbol{R}''_n \boldsymbol{R}'_n(\Theta) \boldsymbol{R}''^t_n. \tag{15}$$

Here, when n=$2m$+1,

$$\boldsymbol{R}'_n(\Theta) = \begin{pmatrix} 1 & & \cdots & 0 \\ & \boldsymbol{R}_2(\theta_1) & & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & \boldsymbol{R}_2(\theta_m) \end{pmatrix}. \tag{16}$$

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 15, for vector X, $\boldsymbol{R}''^t_n$ functions just as transforming the bases so that an n-dimensional rotation of vector $\boldsymbol{R}''^t_n X$ can be done as m independent 2-dimensional rotations. After that, by $\boldsymbol{R}''_n$, the 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 spaces of $\boldsymbol{R}'_n(\Theta)$ is called a rotational plane. 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? As told above, even if we consider $\boldsymbol{I}$ as $\boldsymbol{R}''_n$, generality is not lost. When n=$2m$, n-dimensional vector Y, transformed from vector X=$(x_i | i$=$1, 2, \cdots, n)$ by $\boldsymbol{R}'_n(\Theta)$, is described as

$$\begin{aligned} Y &= \boldsymbol{R}'_n(\Theta) X \\ &= (\boldsymbol{R}_2(\theta_i) v_i | i = 1, 2, \cdots, m), \end{aligned} \tag{17}$$

where $v_i = (x_{2i-1}, x_{2i})^t$. Now we can introduce the relation between IP angle θ' and rotation parameter vector Θ.

$$\begin{aligned} \cos\theta' &= \frac{X \cdot Y}{|X||Y|} \\ &= \frac{1}{|X|^2} \sum_{i=1}^{m} v_i^t \boldsymbol{R}_2(\theta_i) v_i \\ &= \frac{1}{|X|^2} \sum_{i=1}^{m} |v_i|^2 \cos\theta_i. \end{aligned} \tag{18}$$

v_i is a vector projected onto the rotational plane corresponding to rotation angle θ_i. Equation 18 means that the cosine similarity based on the IP angle is represented as the 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 18 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 $\boldsymbol{R}''^t_n X$, the IP angle has to depend on X. Although the IP angle was observed speaker- and phoneme-independent in our previous study [1], the above discussion mathematically predicts the dependence of the IP angle on these factors. Shortly, this will be verified experimentally.

Table 2: Conditions for acoustic analysis.

sampling	16 bit / 16 kHz
window	25 ms length and 5 ms shift (Hamming)
parameters	Δ vector of FFT Cepstrum (1 to 12)

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 16 and 18 is replaced with $r_i \boldsymbol{R}_2(\theta_i)$, where r_i is a magnification parameter for i-th rotational plane. Hence, Equation 18 is obtained for $\boldsymbol{A}$;

$$\cos\theta' = \frac{\sum_{i=1}^{m} r_i |v_i|^2 \cos\theta_i}{\sqrt{\left(\sum_{i=1}^{m} |v_i|^2\right)\left(\sum_{i=1}^{m} r_i^2 |v_i|^2\right)}}. \tag{19}$$

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-dimeinsional rotations and magnifications. Then, for vector $\boldsymbol{R}''^t_n X$, a new speech feature with a unit of two dimensional elements, $(|v_1|^2, |v_2|^2, ..., |v_m|^2)$, may corresponds 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

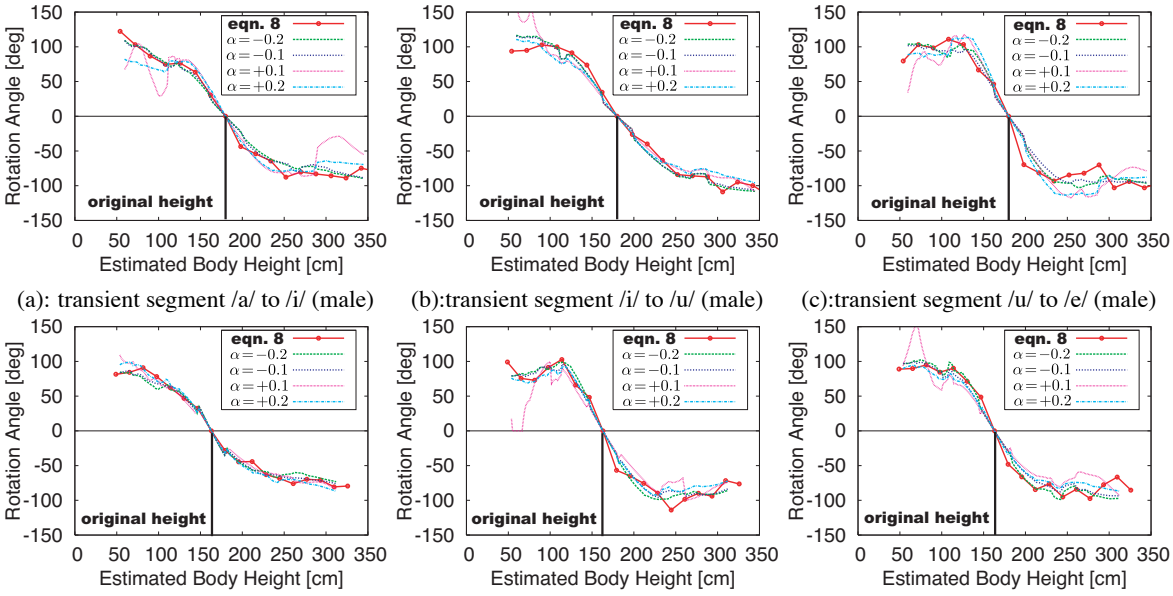
Two issues should be investigated experimentally. The first issue is whether the rotational planes can be obtained from $\boldsymbol{A}$ irrespective of α or not. The second issue is whether $\boldsymbol{A}$ can be better approximated as modified rotation matrix, where $r_i \boldsymbol{R}_2(\theta_i)$ is used instead of $\boldsymbol{R}_2(\theta_i)$ only. For these, two utterances of /aiueo/, from 1 male and 1 female speakers, were acoustically analyzed. The conditions are shown in Table 2. 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 $\boldsymbol{A}$, the rotational planes ($\boldsymbol{R}''_n$) were calculated for each case of α=$-0.2, -0.1, 0.1$ and 0.2. Then, for each set of the rotational planes, each of the Δ vectors was projected to the selected rotational planes and all the $|v_i|^2$s were calculated. Using these parameters and Equation 19, the IP angle was estimated as a function of the estimated body height of the speaker. For this calculation, unlike the previous study, warped speech samples are not needed. θ' is calculated only analytically.

For comparison, based on Equation 8, θ' was calculated from the original utterances and their warped versions generated by applying $\boldsymbol{A}$ on the original utterances using STRAIGHT.

5.2. Results and discussion

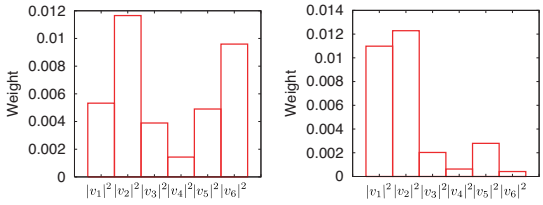
Figure 3 shows the IP angles calculated by Equation 8 and those by Equation 19 for each set of the rotational planes. The top three figures are from the male speaker and the bottom three



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

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

Figure 3: 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.



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

Figure 4: Weights $|v_i|^2$ of each rotational plane. These data are for the transient position /a/ to /i/.

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 values of α show very small differences. This indicates that the rotational planes are reasonably independent of α and the first issue is solved. In each graph from (a) to (f), we can say that the four curves drawn by Equation 19 are well fitted to the curve drawn by Equation 8. This means that modified rotation matrix is a better 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 8 curve of (d) and that of (e) show rather different patterns. Even in these cases, the curves drawn by Equation 19 follow precisely the two different Equation 8 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 our proposed method can characterize the rotational distortion caused by $\boldsymbol{A}$ even more precisely.

As for the new feature parameter discussed in Section 4.3, a small experiment was carried out. Figure 4 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 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.

6. Conclusions

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're going to carry out some experiments using real data, voices of children and adults. In the experiments, we're 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 [6].

7. References

[1] D. Saito *et al.*, "Directional dependency of cepstrum on vocal tract length," *ICASSP 2008*, pp. 4485–4488, 2008.

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

[3] 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.

[4] 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.

[5] T. Takahashi *et al.*, "Interpolation between eigenspaces using rotation in multiple dimensions," *ACCV 2008*, vol. 2, pp. 774–783, 2007.

[6] S. Asakawa *et al.*, "Multi-stream parameterization for structural speech recognition," *ICASSP 2008*, pp. 4097–4100, 2008.