AFFINE INVARIANT FEATURES AND THEIR APPLICATION TO SPEECH RECOGNITION

Yu Qiao, Masayuki Suzuki, Nobuaki Minematsu

Graduate School of Engineering, The University of Tokyo, Tokyo, Japan
{qiao, suzuki, mine}@gavo.t.u-tokyo.ac.jp

ABSTRACT

This paper proposes a set of affine invariant features (AIFs) for sequence data. The proposed AIFs can be calculated directly from the sequence data, and their invariance to affine transformation is proved mathematically through algebraic calculation. We apply the AIFs to speech recognition. Since the vocal tract length (VTL) difference causes to frequency warping which can be approximated well by affine transform on cepstral features [1], the AIFs of cepstral sequence provide robust features for VTL variations. We experimentally examine the invariance of AIFs of speech signals, and apply AIFs for Japanese isolated word recognition. The experimental results show that the combination of AIFs with MFCC or MFCC+Δ can lead to higher recognition rates than MFCC or MFCC+Δ only. Especially in the mismatched experiments, the combination with AIFs can reduce the error rates about 30% when compared to MFCC or MFCC+Δ only. The AIFs are expected to have other applications than speech recognition, since their invariance is general.

Index Terms— Affine invariant feature, frequency warping, speaker normalization, speech recognition

1. INTRODUCTION

Many pattern recognition tasks face the challenge to deal with the variation exhibited by samples of the same category. In face recognition, the same face can lead to very different images due to viewpoint and illumination changes. In speech recognition, the same text can be uttered into different acoustic observations by different speakers or even by a single speaker in different conditions. Many of these variations can be modeled by transformations on input patterns. For example, the viewpoint change can correspond to different 3D projection matrices. Therefore, finding features invariant to certain groups of transforms plays an important role in many pattern recognition problems.

This paper proposes a novel set of features of sequence data, which are invariant to affine transformations. The invariance is general and we don't put any special constraints on affine transformations other than it is invertible. The affine invariant features (AIFs) proposed here capture the detailed information of local frames, and can be easily calculated. To the best of our knowledge, affine invariant features have received much more attention in image processing than in speech engineering. Flusser and Suk [2] proposed the affine-invariant moment features through algebraic calculation. Petrou and Kadyrov [3] made use of trace transform to design affine invariant features. However, unlike the AIFs proposed, most of these features are designed for 2D images, which make them unsuitable for sequence data such as speech signals.

It is well known that the difference of vocal tract length (VTL) induces frequency warping of speech signals. And the frequency warping can be approximately modeled by affine transformations of cepstral features [1]. Thus the AIFs proposed must be robust to VTL (speaker) variations. It is noted that our AIFs are different from previous speaker or VTL invariant features [4, 5, 6] which are based on a simple linear frequency warping assumption $f' = \alpha f$. Our AIFs can account for more general frequency warping functions $f' = w(f)$. The AIFs are also different from the invariant structural representation (ISR) introduced in our previous works [7, 8, 9]. ISR is based on the Bhattacharyya distance between distributions. Thus to calculate ISR, one needs to convert a cepstrum sequence to a sequence of distributions.

The reminder of this paper is organized as follows. Section 2 formulates the affine invariant features and proves their invariance. In Section 3, we experimentally examine the invariance of AIFs for speech data and test the usage of AIFs for speech recognition. Finally, the paper is concluded in Section 4.

2. AFFINE INVARIANT MEASURES

This section describes the affine invariant features (AIFs) and proves their invariance to affine transformation. Let $X = [x_1, x_2, ..., x_n]$ denote a sequence of samples, where each sample (frame) is represented by a d-dimension vector x_i. In speech processing, x_i can be a spectral feature vector. $X^{s:e} = [x_s, x_{s+1}, ..., x_e]$ represents a segment (sub-sequence) of X.

Consider an affine (linear) transformation on x_i,

$$x_i' = Ax_i + c, \tag{1}$$

where transformation parameter A is a $d \times d$ full rank matrix and c is a d-dimension vector. We use $X' = [x_1', x_2', ..., x_n']$ to denote the sequence after transformation. Our objective is to find feature (measure) M for sequence X and X', which is invariant to the affine transformation, in other words, $M(X) = M(X')$. But the AIF defined for the whole sequence X only contains limited information, and cannot capture the detailed local features. Therefore, here we are interested in AIFs for local frames, denoted by $M(X, i)$, where i is a frame index. One may suggest to define AIF for an individual sample (frame) x_i, such as $M(x_i)$. However, for two arbitrary cepstrum vectors x and x' which may come from different acoustic events (phonemes), we can always find an affine transformation between them, e.g., $A = I$ and $c = x' - x$. This means if we use frame level AIF , this will judge different acoustic events as the same one. Actually, it can be shown that frame level AIF $M(x_i)$ must be a constant.

For this reason, here we define AIF $M(X, i)$ at segment (sub-sequence) level. Consider segment $X^{i-k_1:i+k_2} = [x_{i-k_1}, x_{i-k_1+1}, ..., x_i, ..., x_{i+k_2}]$, which starts k_1 frames before i and ends k_2 frames after index i. Note k_1 has not to be equal to k_2. In the next, we will introduce a set of AIFs, which actually capture the difference between the segment $X^b = X^{i-k_1:i}$ before index i and the segment

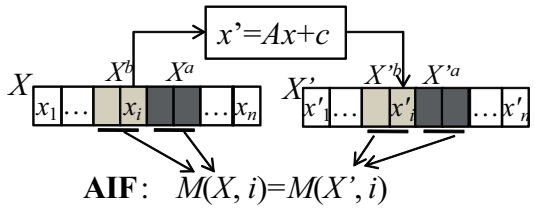


Fig. 1. Calculation of affine invariant features

$X^a = X^{i+1:i+k_2}$ after i (Fig. 1). As a preparation, we define the mean $\mu_{X^{s:e}}$ and the covariance matrix $\Sigma_{X^{s:e}}$ for subsequence $X^{s:e}$ as

$$\mu_{X^{s:e}} = \frac{1}{e-s+1}\sum_{i=s}^{e} x_i, \tag{2}$$

$$\Sigma_{X^{s:e}} = \frac{1}{e-s+1}\sum_{i=s}^{e}(x_i - \mu_{X^{s:e}})(x_i - \mu_{X^{s:e}})^T. \tag{3}$$

For simplicity, let $\mu_i^b = \mu_{X^{i-k_1:i}}$, $\Sigma_i^b = \mu_{X^{i-k_1:i}}$, $\mu_i^a = \mu_{X^{i+1:i+k_2}}$ and $\Sigma_i^a = \mu_{X^{i+1:i+k_2}}$ denote the means and covariance matrices of X^a and X^b. Then we have seven types of affine invariant features (AIFs) as follows,

$$M_{\mathrm{T}1}(X,i) = (\mu_i^b - \mu_i^a)^T(\Sigma_i^b)^{-1}(\mu_i^b - \mu_i^a), \tag{4}$$

$$M_{\mathrm{T}2}(X,i) = (\mu_i^b - \mu_i^a)^T(\Sigma_i^a)^{-1}(\mu_i^b - \mu_i^a), \tag{5}$$

$$M_{\mathrm{T}3}(X,i) = (\mu_i^b - \mu_i^a)^T(\Sigma_i^b + \Sigma_i^a)^{-1}(\mu_i^b - \mu_i^a), \tag{6}$$

$$M_{\mathrm{T}4}(X,i) = \mathrm{Trace}\left((\Sigma_i^a)^{-1}\Sigma_i^b\right), \tag{7}$$

$$M_{\mathrm{T}5}(X,i) = \mathrm{Trace}\left((\Sigma_i^b)^{-1/2}\Sigma_i^a(\Sigma_i^b)^{-1/2}\right), \tag{8}$$

$$M_{\mathrm{T}6}(X,i) = \frac{|\Sigma_i^a|}{|\Sigma_i^b|}, \tag{9}$$

$$M_{\mathrm{T}7}(X,i) = \frac{|\Sigma_i^a|}{|\Sigma_i^a + \Sigma_i^b|}. \tag{10}$$

To prove the affine invariance, we need to show $M_{\mathrm{T}k}(X,i) = M_{\mathrm{T}k}(X',i)$ $(k = 1,2,...,7)$. In the next, we only provide proofs for $k = 1$ and $k = 6$, and the others can be examined in the similar ways. Remind Eq. 1, we have $\mu_{X'^{s:e}} = A\mu_{X^{s:e}} + c$ and $\Sigma_{X'^{s:e}} = A\Sigma_{X^{s:e}}A^T$. For type-1 (Eq. 4),

$$\begin{aligned} M_{\mathrm{T}1}(X',i) &= (\mu'^b_i - \mu'^a_i)^T(\Sigma'^b_i)^{-1}(\mu'^b_i - \mu'^a_i) \\ &= (A\mu_i^b - A\mu_i^a)^T(A\Sigma_i^bA^T)^{-1}(A\mu_i^b - A\mu_i^a) \\ &= M_{\mathrm{T}1}(X,i). \end{aligned} \tag{11}$$

And for type-6 (Eq. 8),

$$\begin{aligned} M_{\mathrm{T}6}(X',i) &= \frac{|\Sigma'^a_i|}{|\Sigma'^b_i|} = \frac{|A\Sigma_i^aA^T|}{|A\Sigma_i^bA^T|} \\ &= M_{\mathrm{T}6}(X,i). \end{aligned} \tag{12}$$

The types 1-3 AIFs (Eq. 4,5,6) resemble the Mahalanobis distances, which have been widely used to measure the similarity between data sets in pattern recognition. We also notice that the type-5 AIF (Eq. 8) has a similar form to an invariant Riemannian metric for covariance matrices introduced in [10, 11]. More generally, it can be shown that the general eigenvalues of Σ_i^a and Σ_i^b are invariant to affine transformation. (The general eigenvalues λ are obtained by solving $|\Sigma_i^a - \lambda\Sigma_i^b| = 0$.)

We can also generalize the above AIFs to their weighted versions. These weighted AIFs are also invariant to affine transformations. Let $W^b = \{w_j^b\}_{j=i-k_1}^{i}$ denote a set of nonnegative weights for the samples of $X^{i-k_1:i}$. It is required that $\sum_j w_j^b = 1$ for normalization. Let W_b represent a $(k_1+1)\times(k_1+1)$ diagonal matrix with $\mathrm{Diag}(W_b) = [w_{i-k_1}^b, ..., w_i^b]^T$. The weighted mean $\mu_i^{W_b}$ and covariance matrix $\Sigma_i^{W_b}$ can be calculated by,

$$\mu_i^{W_b} = \sum_{j=i-k_1}^{i} w_j^b x_j, \tag{13}$$

$$\Sigma_i^{W_b} = \sum_{j=i-k_1}^{i} w_j^b (x_j - \mu_i^{W_b})(x_j - \mu_i^{W_b})^T. \tag{14}$$

Similarly, we can define $\mu_i^{W_a}$ and $\Sigma_i^{W_a}$ for $X^{i+1:i+k_2}$. Note that the weights W_b and W_a can have different values. Then we have the weighted versions of Eq. 4, Eq. 8, Eq. 9 as follows,

$$M_{\mathrm{T}1}^W(X,i) = (\mu_i^{W_b} - \mu_i^{W_a})^T(\Sigma_i^{W_b})^{-1}(\mu_i^{W_b} - \mu_i^{W_a}), \tag{15}$$

$$M_{\mathrm{T}5}^W(X,i) = \mathrm{Trace}((\Sigma_i^{W_b})^{-1/2}\Sigma_i^{W_a}(\Sigma_i^{W_b})^{-1/2}), \tag{16}$$

$$M_{\mathrm{T}6}^W(X,i) = \frac{|\Sigma_i^{W_a}|}{|\Sigma_i^{W_b}|}. \tag{17}$$

We can define the weighted versions of the Eq. 5, Eq. 6, Eq. 7 and Eq. 10 similarly.

After transformation of Eq. 1, we have the weighted mean $\mu'^{W_b}_i$ and covariance matrix $\Sigma'^{W_b}_i$ of sub-sequences X'^b as,

$$\mu'^{W_b}_i = AX^{i-k_1:i}W_b + c, \tag{18}$$

$$\Sigma'^{W_b}_i = AX^{i-k_1:i}\left(W_b - \frac{1}{(k_1+1)^2}\mathbf{1}\right)(X^{i-k_1:i})^TA^T. \tag{19}$$

where $\mathbf{1}$ denotes a $(k_1+1)\times(k_1+1)$ matrix with all the elements as 1. Then for weighted type-1 AIF, we have,

$$\begin{aligned} M_{\mathrm{T}1}^W(X',i) =& (\mu'^{W_b}_i - \mu'^{W_a}_i)^T(\Sigma'^{W_b}_i)^{-1}(\mu'^{W_b}_i - \mu'^{W_a}_i) \\ =& (A\mu_i^bW_b - A\mu_i^aW_a)^T(A\Sigma_i^{W_b}A^T)^{-1} \\ & (A\mu_i^bW_b - A\mu_i^aW_a) \\ =& M_{\mathrm{T}1}^W(X,i). \end{aligned} \tag{20}$$

Similarly, we can examine the invariance of other weighted AIFs, $M_{\mathrm{T}k}^W(X,i) = M_{\mathrm{T}k}^W(X',i)$ $(k = 2,...,7)$.

3. AFFINE INVARIANT FEATURES FOR SPEECH RECOGNITION

One of the basic problems in automatic speech recognition (ASR) is the inter-speaker variations of acoustic features. It is well known that the speaker independent (SI) ASR systems have higher error rates than the speaker dependent (SD) ASR systems. Roughly speaking, there are two approaches to deal with the speaker variations: one is speaker adaptation, such as, MAP [12] and MLLR [13]; the other is vocal-tract length normalization (VTLN) [14, 15]. Speakers have vocal tracts with different lengths. This physical difference largely causes the acoustic variations among different speakers. The VTL

variation can be modeled by frequency warping. Based on this observation, Umesh et al. used scale transform to construct speaker invariant feature [4], and Irino and Patterson dealt with the VT size by Mellin transform [5]. More recently, Mertins and Rademacher introduced VTL invariant feature through wavelet transform [6]. All these previous methods are based on a simple linear frequency warping function: $f' = \alpha f$, where α is a constant. However, this is just a rough approximation. In this paper, we consider a more general frequency warping function $f' = w(f)$. In [1], Pitz et al. showed that the general frequency warping equals to linear (affine) transformation of cepstral features (cepstrum, mel-cepstrum or MFCC). In Section 2, we have developed a set of affine invariant features (AIFs) to affine transformation. If we calculate AIFs for cepstral feature sequence, these AIFs must be approximately invariant to frequency warping and thus provide robust features to VTL variations.

3.1. Corpus and acoustic analysis conditions

We use a subset of the Tohoku University and Panasonic isolated spoken word database (TMW) for evaluation [16]. The database consists of 212 isolated Japanese words and each word is once uttered by 30 males and 30 females. The sampling frequency was converted to 16kHz in our experiments. For each word, we calculated the cepstral features from speech signals by using 25ms Hamming windows with 10ms shift.

We make use of MFCC sequences for calculating AIFs due to their good performances on speech recognition. The calculation of AIFs requires covariance matrices. We have executed the same experiments with spectral features other than MFCC. The results are similar, but we omit them due to space limitation. In our problem, there usually only exist a dozen of frames (10-20) for estimating covariance matrices. Remind different dimensions of MFCC are highly uncorrelated. In the following experiments, we assume to use diagonal covariance matrix without specific notification. An important parameter for calculating AIFs is length of sub-sequences X^b and X^a. In our experiments, both are set as 16, that is, $k_1 + 1 = k_2 = 16$. It was shown that most of the useful linguistic information is in modulation frequency components at the range between 1 and 16 Hz, with the dominant component at around 4 Hz [17]. This roughly corresponds to sub-sequence length 16.

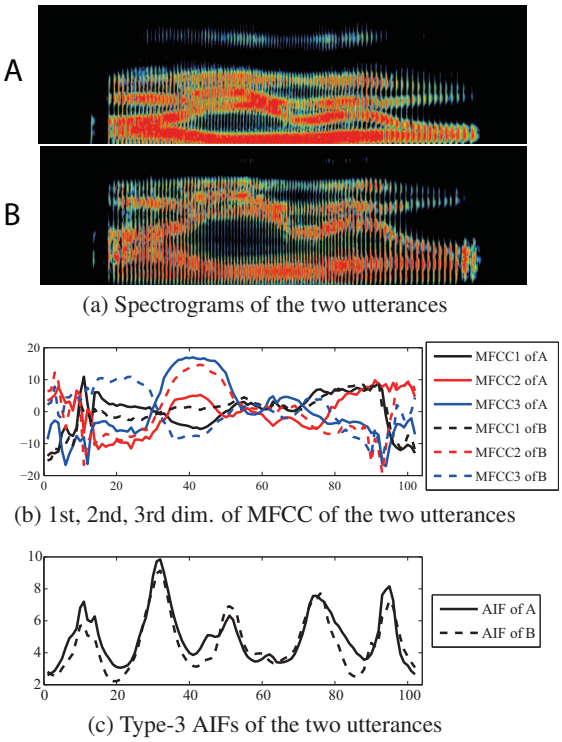
3.2. Invariance of AIFs for speech signals

This section examines the invariance of AIFs for speech signals. As example, we obtained two utterances of /aiueo/ pronounced through VTs of different length and calculated their spectrograms, MFCCs and AIFs. The results are shown in Fig. 2. One can find that the spectrograms and MFCCs of the two utterances are very different from each other. On the other hand, it can be seen that their AIFs are similar.

We conducted quantitative evaluation experiments on the invariance of AIFs by using TMW. The normalized DP matching score (NDPMS) between the two feature sequence $X = \{x_1, ..., x_n\}$ and $Y = \{y_1, ..., y_m\}$ is used as evaluation measure,

$$\text{NDPMS}(X, Y) = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n} |x_i - y_{wp(i)}|^2}}{(\sqrt{\text{Var}(X)} + \sqrt{\text{Var}(Y)})/2}, \qquad (21)$$

where $wp()$ denotes the DP warping path between X and Y, and Var() represents the variance function. For each pair of utterances with the same linguistic content but uttered by different speakers, we calculate the NDPMS between them. Five kinds of features



(a) Spectrograms of the two utterances

(b) 1st, 2nd, 3rd dim. of MFCC of the two utterances

(c) Type-3 AIFs of the two utterances

Fig. 2. Examples of spectrograms, MFCCs and AIFs of two utterances of /aiueo/.

type-1 AIF, type-3 AIF, MFCC, mel-cepstrum and FFT-cepstrum are considered for comparison. And their average NDPMSs within males, within females and between males and females are shown in Fig. 3. We can find that the AIFs show much higher invariance than MFCC, mel-cepstrum and FFT-cepstrum with respect to speaker variance. Especially, the NDPMSs of AIFs in the mismatched case (between males and females) is near to that of the un-mismatched cases (within males or within females). However, for MFCC, mel-cepstrum and FFT-cepstrum, the differences of their NDPMSs in the mismatched and un-mismatched cases are significant.

3.3. AIFs for speech recognition

In this section, we study the recognition performances of AIFs on TMW corpus. The evaluation results here are limited to type-3 for its good performance and the page limitation. The training part of TMW includes utterances from 30 speakers (15 males and 15 females), and the testing part from another 30 speakers (15 males and 15 females). We also conducted mismatched experiments, that is, we used male utterances for training and female utterances for testing, and vice versa. Because the type-3 AIFs (Eq. 6) highly compress the features of all dimensions into one dimension and this can lead to loss of useful information. We do stream division and calculate AIFs for each stream. The stream division may reduce the invariance of AIFs, but on the other hand it can preserve more information for classification. The similar technique has been used in our former works on invariant structural representation [9, 8] for speech recognition. More details of stream division are discussed in [9]. We used word-HMM for acoustic modeling and classification, and made comparisons among AIFs, weighted AIFs (wAIFs) and classical fea-

Table 1. Recognition rates of AIFs and weighted AIFs

Method	MFCC	MFCC+AIF	MFCC+W-AIF	MFCC+Δ	MFCC+AIF+Δ	MFCC+W-AIF+Δ
Un-mismatched training+testing	98.35%	99.24%	99.32%	99.47%	99.51%	**99.65%**
Male training+female testing	72.71%	83.22%	83.99%	82.79%	88.35%	**88.75%**
Female training+male testing	70.59%	83.25%	80.30%	85.34%	**89.88%**	88.41%

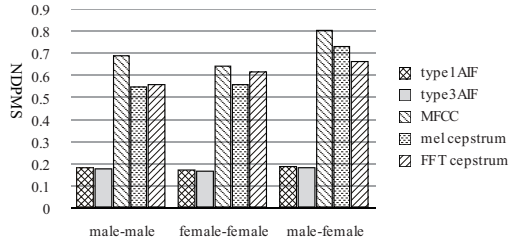


Fig. 3. NDPMS of AIFs, cepstrum, mel-cepstrum and MFCC

tures such as MFCC, MFCC+Δ. For wAIFs, the weight of the k-th sample is set as $|k-i-0.5|$. The recognition rate when using AIFs is 91.47% and the recognition rate of wAIFs is 93.90%, both of which are lower than 98.35% of MFCC. This is because although AIFs are more invariant to VTL (speaker) variation, the calculation of AIFs compresses and smoothes the original features, and this calculation can reduce certain discriminative information, which are useful for classification.

We also carried out recognition experiments on joint feature vectors, such as, MFCC+AIFs, MFCC+Δ+AIFs, MFCC+wAIFs, and MFCC+Δ+wAIFs. The experimental results are summarized in Table 1. We find that the combination of AIFs or wAIFs with MFCC and MFCC+Δ can lead to better recognition rates than MFCC and MFCC+Δ, respectively. In the mismatched experiments, the combination of AIFs with MFCC can reduce the error rates 27.3% for male training+female testing, and 29.4% for female training+male testing. Similar results are also observed when using wAIFs for combination. It can be seen that the use of weights can increase the recognition rates for the un-mismatched evaluation, however its effect to mismatched evaluation is not significant. This is partly because the weighted AIFs release the effect of smoothing, but at the same time this may decrease their robustness to noise and variations. It is noted that the combination of feature vectors and the simple weights are only a preliminary step to show the usefulness of AIFs and wAIFs. We are going to consider new layers in HMM for AIFs and other weights for wAIFs.

4. CONCLUSIONS

This paper introduces a set of affine invariant features (AIFs) for sequence data. The AIFs capture the relative information of sequence data and can be calculated directly. We apply AIFs to speech recognition. Because the VTL difference can be approximated by affine transformations on cepstrum features, the AIFs of cepstrum sequence yield robust features to VTL variations. We experimentally showed that AIFs have higher invariance to speaker difference than MFCC, mel-cepstrum and cepstrum for speech signals. We also found the combination of AIFs with MFCC or MFCC+Δ can lead to better recognition rates than using MFCC or MFCC+Δ only through a Japanese isolated word classification task. In the mismatched experiments, the combination of AIFs with MFCC or MFCC+Δ can reduce the error rates about 30%. The AIFs proposed have very general invariance, which is expected to have other applications. However, the too general invariance of AIFs may cause loss of useful information and affect the final performance of AIFs. Now we are developing techniques to deal with this strong invariance. We are also considering conducting experiments on a larger database and making comparison with speaker normalization and adaption techniques [14, 13]. The experimental results of this paper are preliminary. We are also going to apply AIFs for continuous speech recognition.

5. REFERENCES

[1] M. Pitz and H. Ney, "Vocal Tract Normalization Equals Linear Transformation in Cepstral Space," *IEEE Trans. on SAP*, vol. 13, no. 5, pp. 930–944, 2005.

[2] J. Flusser and T. Suk, "Pattern recognition by affine moment invariants," *Pattern Recognition*, vol. 26, no. 1, pp. 167–174, 1993.

[3] M. Petrou and A. Kadyrov, "Affine Invariant Features from the Trace Transform," *IEEE Trans. on PAMI*, pp. 30–44, 2004.

[4] S. Umesh, L. Cohen, N. Marinovic, and DJ Nelson, "Scale transform in speech analysis," *IEEE Trans. on SAP,*, vol. 7, no. 1, pp. 40–45, 1999.

[5] T. Irino and R.D. Patterson, "Segregating information about the size and shape of the vocal tract using a time-domain auditory model: The stabilised wavelet-Mellin transform," *Speech Communication*, vol. 36, no. 3-4, pp. 181–203, 2002.

[6] A. Mertins and J. Rademacher, "Vocal Tract Length Invariant Features for Automatic Speech Recognition," *Proc. ASRU*, pp. 33–37, 2005.

[7] N. Minematsu, "Mathematical Evidence of the Acoustic Universal Structure in Speech," *Proc. ICASSP,*, vol. 1, 2005.

[8] Y. Qiao, S. Asakawa, and N. Minematsu, "Random discriminant structure analysis for automatic recognition of connected vowels," *Proc. ASRU*, pp. 576–581, 2007.

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

[10] W. Forstner and B. Moonen, "A metric for covariance matrices," *Tech. Rep. of Stuttgart Univ.*, pp. 113–128, 1999.

[11] X. Pennec, P. Fillard, and N. Ayache, "A Riemannian Framework for Tensor Computing," *Intl. Journal of Computer Vision*, vol. 66, no. 1, pp. 41–66, 2006.

[12] J.L. Gauvain and C.H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixtureobservations of Markov chains," *IEEE Trans. on SAP*, vol. 2, no. 2, pp. 291–298, 1994.

[13] M.J.F. Gales and P.C. Woodland, "Mean and variance adaptation within the MLLR framework," *Computer Speech & Language*, vol. 10, no. 4, pp. 249–264, 1996.

[14] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," *Proc. ICASSP*, vol. 1, 1996.

[15] L. Lee and R. Rose, "A frequency warping approach to speaker normalization," *IEEE Trans. on SAP,*, vol. 6, no. 1, pp. 49–60, 1998.

[16] "Tohoku univ. - Matsushita isolated word database (TMW)," *http://research.nii.ac.jp/src/eng/list/detail.html#TMW*.

[17] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel, "On the relative importance of various components of the modulation spectrum for automatic speech recognition," *Speech Communication*, vol. 28, no. 1, pp. 43–55, 1999.